JOC The Journal of Organic Chemistry

pubs.acs.org/joc Article

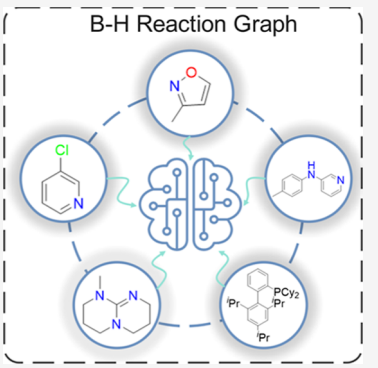
Designing Buchwald−Hartwig Reaction Graph for Yield Prediction

Weiren Zhao,[§] Shen Wang,[§] and Yang Li*

Cite This: J. Org. Chem. 2025, 90, 12975−12983 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The Buchwald−Hartwig (B−H) reaction graph, a novel graph for deep learning models, is designed to simulate the interactions among multiple chemical components in the B−H reaction by representing each reactant as an individual node within a custom-designed reaction graph, thereby capturing both single-molecule and intermolecular relationship features. Trained on a high-throughput B−H reaction data set, B−H Reaction Graph Neural Network (BH-RGNN) achieves near-state-of-the-art performance with an R^2 score of 0.971 while maintaining low computational costs. Using a perturbation-based analysis, the model reveals significant insights into the relationship between bases and reaction yields, identifying key characteristics of bases influencing yields, which provides valuable guidance for base selection in the B−H reaction. This work demonstrates the effectiveness of tailored graph representations in advancing GNN applications for chemical reaction modeling, offering a promising tool for predicting reaction yield and identifying key reaction factors.



Downloaded via KYOTO UNIV on April 14, 2026 at 07:09:18 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

1. INTRODUCTION

In recent years, machine learning methods based on graph neural networks (GNNs) have emerged as powerful tools in the field of chemical reaction research.[1] Their ability to effectively represent molecular structures through graph-based modeling has led to widespread applications in materials science, drug discovery, and synthetic and process chemistry.[2−6] In these approaches, molecules are typically encoded as undirected or directed graphs, where atoms and bonds are represented as nodes and edges, respectively. This representation allows GNNs to capture complex intermolecular relationships and has significantly advanced predictive modeling in various chemical domains.

Probst et al. proposed a GNN framework based on set representation learning, which significantly improved model performance by aggregating the outputs of the encoding layer into a global set representation.[7] However, despite the success of existing GNN-based models, most studies predominantly focus on single-molecule learning through the construction of individual molecular graphs for each reactant or product.[2,8,9] Although such representations effectively capture local structural features, they often overlook intermolecular interactions that are crucial in determining the overall reactivity and outcome of chemical transformations. This limitation highlights the need for a more comprehensive and holistic representation of chemical reactions—one that explicitly incorporates interactions among all participating species.

To address this limitation, Hong et al. proposed a strategy in which one-dimensional feature vectors, obtained through graph convolution and pooling operations, are duplicated to create two identical vectors. These vectors are then subjected to matrix multiplication, resulting in a two-dimensional representation that captures the intercomponent interaction information in chemical reactions.[10] While this approach effectively captures certain aspects of molecular interactions, it tends to produce high-dimensional feature spaces and lacks the flexibility needed to analyze the contributions of specific component pairs. Additionally, it does not provide clear interpretability regarding which intermolecular interactions are considered most significant by the model.

In this work, we propose a novel reaction graph representation specifically designed for the analysis of Buchwald−Hartwig (B−H) reaction, which is anticipated to expand to pharmaceuticals, materials science, and catalysis. As depicted in Figure 1, unlike conventional approaches that model each reactant separately, our method treats all participating components as independent nodes within a unified graph structure, which we define as a reaction graph. This framework enables GNN models to directly learn from both single-molecule and intermolecular interaction features, thereby capturing the full complexity of chemical transformations at the systems level. The various graph construction strategies were evaluated on Doyle et al.'s data set.[11] By exploring various graph construction strategies and evaluating their influence on model performance, we demonstrated that the proposed method achieves strong predictive accuracy across multiple GNN architectures. More importantly, the trained model offers interpretable insights into key reaction features, including the discovery of significant base-related characteristics and their correlation with reaction yields.

Received: June 8, 2025
Revised: September 1, 2025
Accepted: September 5, 2025
Published: September 10, 2025

ACS Publications © 2025 American Chemical Society

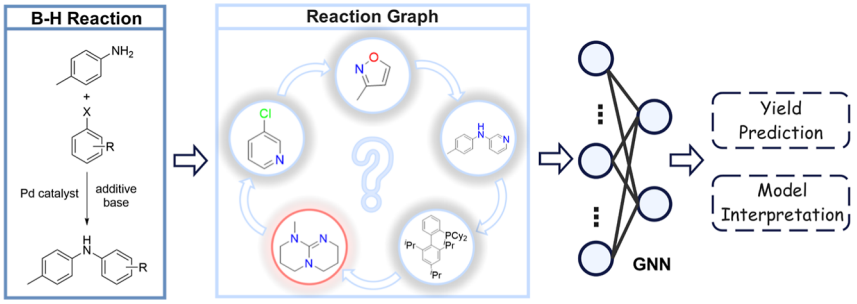


Figure 1. Schematic illustration of this work.

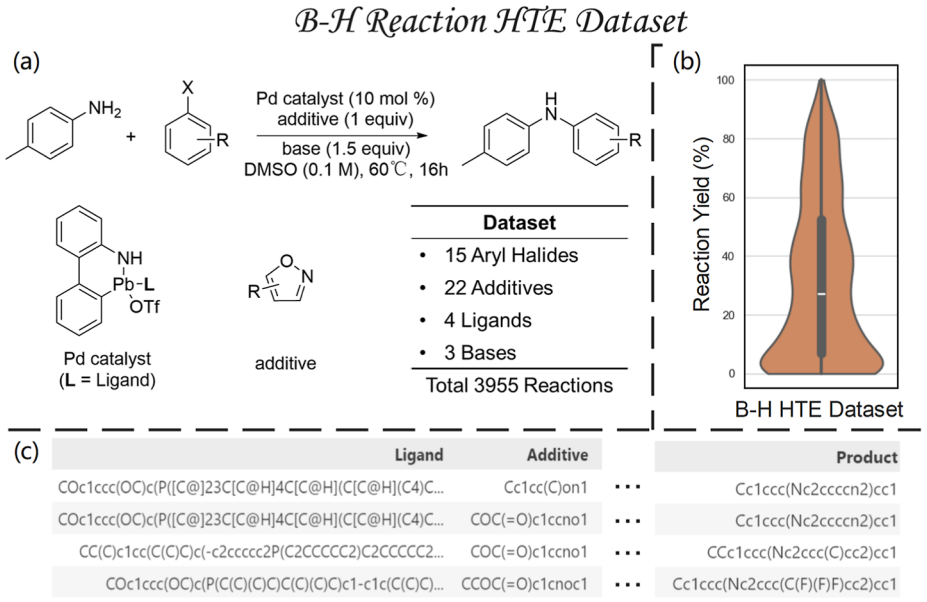


Figure 2. Introduction of the B−H reaction HTE data set. (a) B−H reaction in the data set. (b) Yield distribution. (c) Data structure.

2. RESULTS AND DISCUSSION

2.1. B−H Reaction Data Set. For our study, we utilized Doyle et al.'s data set, a data set of Pd-catalyzed C−N cross-coupling reactions between 4-methylaniline and aryl halides, illustrated in Figure 2a, as the benchmark for designing and training neural network models.[11−13] This high-quality data set comprises 3955 B−H reaction yield data points from a high-throughput experimentation (HTE), provides comprehensive statistical information that maps the full synthetic space defined by the studied reaction components, and has been widely adopted as a benchmark in machine learning studies.[14−18]

The yield distribution of the B−H reaction HTE data set is presented in Figure 2b. Analysis of the yield distribution reveals that most reactions exhibit relatively low yields. Specifically, the most common yield range is between 0% and 20%, followed by the 20% to 80% range, with very few reactions achieving yields above 90% (more details in the Supporting Information, Section S1.1). Overall, the data set shows an average yield of 32.37% and a median yield of 27.25%. This imbalanced distribution poses significant challenges for model training, since more low-yield cases make the trained model biased toward low-yields, thus underestimating the reaction yield when predicting.

Additionally, the data structure of the data set is displayed in Figure 2c. Compounds are represented using the simplified molecular input line entry system (SMILES).[19] The reactions primarily involve varying aryl halides while keeping 4-methylaniline as the other reactant constant, and the other components have a consistent skeleton with different substituents. Doyle et al. utilized SMILES as the input for the density functional theory (DFT) calculations of all compounds, and fed the obtained quantum chemical parameters into the machine learning model for yield prediction. This method captures some aspects of quantum chemical properties; the chemical interactions are neglected.

2.2. B−H Reaction Graph. Given the detailed characteristics of the B−H reaction data set, it becomes evident that traditional methods may struggle to fully capture the complex interactions among different reaction components. To address these challenges, we treated the five key components of the B−H reaction—substrate, product, additive, base, and catalyst ligand—as distinct nodes within a unified graph framework. Different connection patterns among these nodes were utilized to reflect the complex intermolecular interactions involved in the reaction system. This approach requires each component to be processed independently, with relevant physicochemical features extracted and assigned to their respective nodes.

Specifically, the Mordred package was employed to compute molecular descriptors from the SMILES strings of each reaction component.[20] Mordred is widely recognized for its scientifically sound definitions, which are derived from established QSAR (Quantitative Structure−Activity Relationships) literature, thereby ensuring the reliability of the generated features. These descriptors are primarily based on fundamental atomic properties such as atomic number, atomic mass, and valence

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983

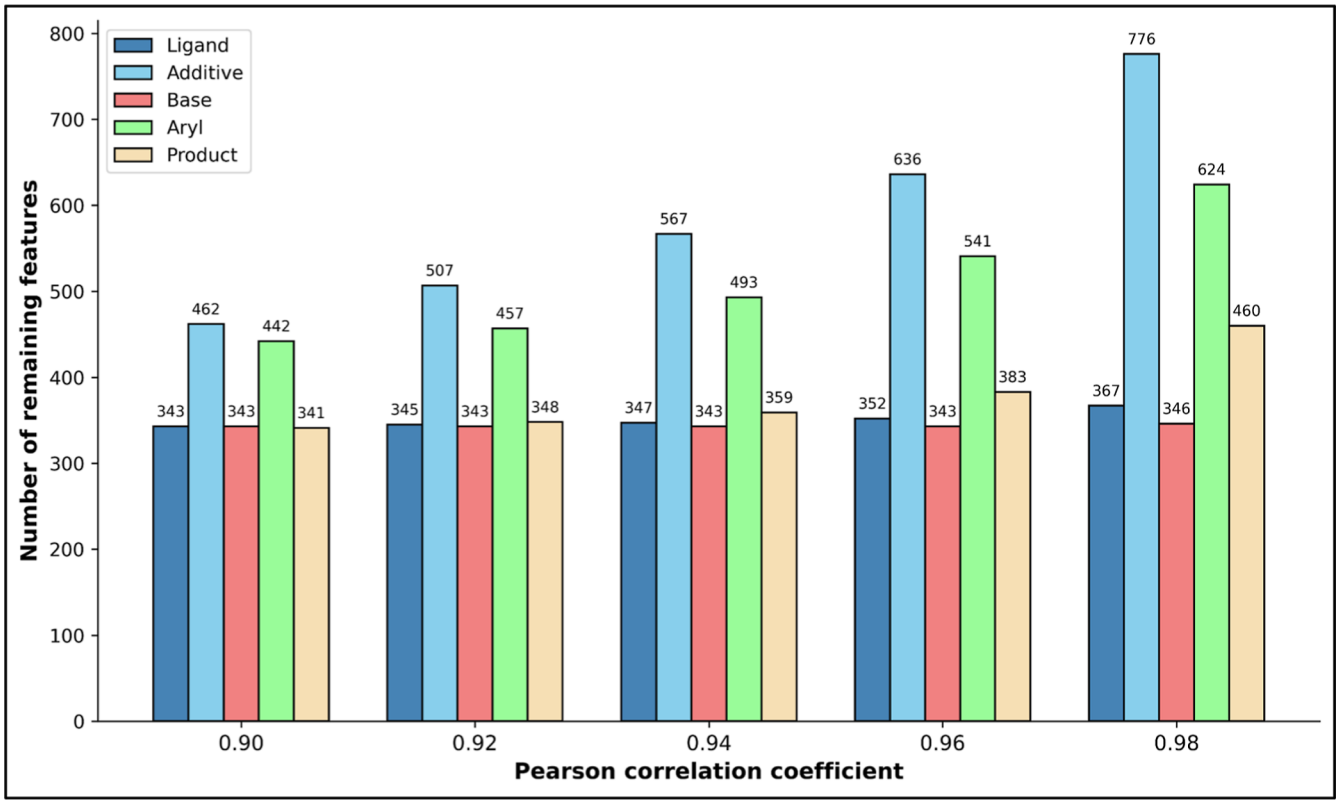


Figure 3. Preliminary feature screening based on Pearson correlation coefficients.

electron count, enabling efficient computation of a large number of physicochemical parameters in a short time.

After computing the feature sets for all five reaction components, data preprocessing was performed to remove invalid or noncomputable values. For instance, certain descriptors related to cyclic structures could not be calculated for molecules lacking such features, resulting in NAN (not a number) entries (more details in the Supporting Information, Section S1.3). Following the filtering of physicochemical meaningless (redundant parameters with the same value for all and NAN), the final feature dimensions for each component were as follows: 1335 descriptors for aryl halides (substrate), 1486 for the product, 1335 for the additive, 1430 for the base, and 1489 for the catalyst ligand. Since graph neural networks require consistent feature vector lengths across all nodes, it was necessary to standardize the dimensionality of these feature sets. This was achieved through a feature selection strategy rather than padding, to optimize computational efficiency.[21] A common subset of descriptors was selected for each component to ensure uniform feature vector lengths while preserving meaningful chemical information.

To obtain optimally matched feature vectors of equal length across all reaction components, we processed the feature sets of each component accordingly. Figure 3 shows that after removing features with Pearson correlation coefficients above 0.90, the number of remaining features for the five components in the B−H reaction becomes more balanced. This facilitates the subsequent selection of feature vectors with a common length. Therefore, a Pearson correlation coefficient threshold of 0.90 was selected for final feature screening (more details in the Supporting Information, Section S1.4 and 1.5).

The observed variation in the number of low-correlation features appears to be associated with the chemical diversity of each component. Among the five components, additives are represented by the most diverse set, comprising 22 distinct species, which likely contributes to their richer feature space and higher number of weakly correlated descriptors. In contrast, only three types of bases are present in the data set, resulting in a higher proportion of highly correlated features. This observation suggests that capturing meaningful information becomes increasingly challenging when the underlying chemical space is limited.

Based on the above analysis and considering the trade-off between computational efficiency and model accuracy, we selected 100 low-correlation features from each component for representation. To achieve this, we first calculated the pairwise Pearson correlation coefficients among all descriptors within each component's feature set and removed one of the two features in any pair where the correlation exceeded 0.90. When choosing which feature to retain from a highly correlated pair, a random elimination strategy is adopted. This approach not only simplifies the process but also introduces randomness that enhances model robustness. Following this procedure, the feature vector lengths for the five components were reduced to between 340 and 470. These filtered feature sets were then used as input to train Random Forest (RF) models, with B−H reaction yields serving as the regression targets. RF was selected due to its proven effectiveness in previous studies on the B−H data set and its favorable balance between performance and computational cost.[11,14] The well-trained RF models were utilized to rank the importance of individual features for each component. Based on the feature importance ranking, the top 100 most influential descriptors were selected from each component's feature set to form the final feature vectors used in constructing the B−H reaction graph. This strategy ensures both the relevance and consistency of the selected features while

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983

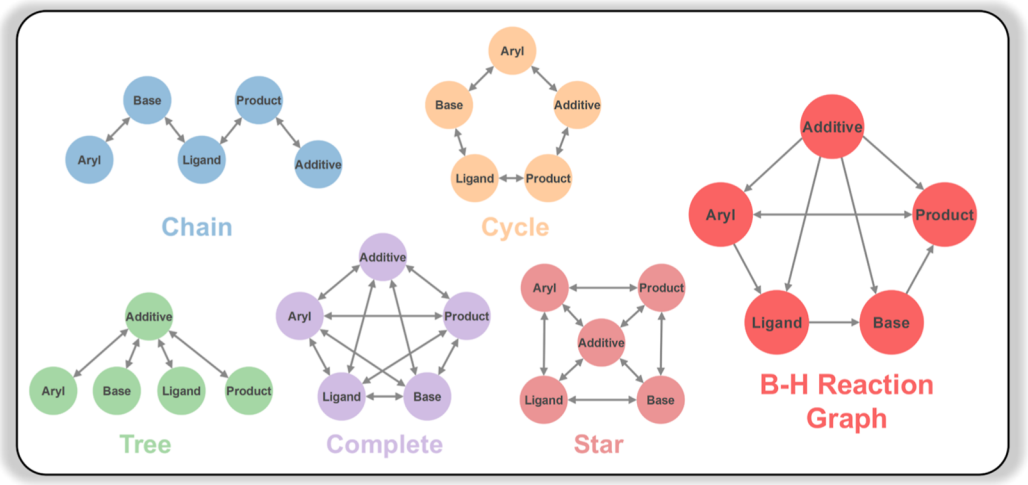


Figure 4. Common topological reaction graph (left) and the defined B−H reaction graph (right).

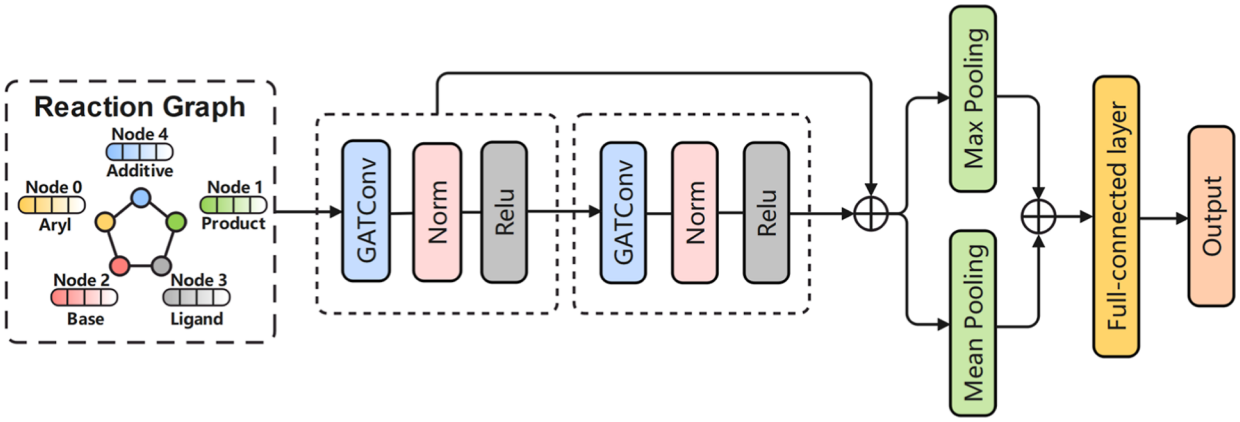


Figure 5. Framework of the reaction graph neural network (RGNN).

significantly improving model training efficiency and predictive accuracy. By reducing redundancy without sacrificing critical chemical information, the proposed feature selection method effectively balances model complexity and performance, laying a solid foundation for subsequent model development and mechanistic interpretation. Notably, the feature set for each component remained unchanged throughout the subsequent experiments and analyses.

After processing the node features, we explored the graph topologies for constructing B−H reaction graphs, which serve as the basis for model training and evaluation.[22] As shown in the left panel of Figure 4, these topologies include star-shaped, chain-like, cycle, tree-like, and complete configurations. In each topology, nodes represent individual reaction components (substrate, product, additive, base, and catalyst ligand), and undirected edges are used to simulate their interactions. Wherein, the cycle topology allows for potential exploration of cyclic or feedback mechanisms within the reaction system; the completed topology enables all components to interact directly, offering a comprehensive view of intercomponent effects; the tree-like structure provides a hierarchical perspective, useful for understanding branching pathways in complex reaction networks. The star-shaped topology assumes one central component influencing all others, while the chain-like configuration reflects sequential interactions among components. The study of diverse connection patterns provides complementary perspectives for the model to learn from.

It is worth noting that, apart from the completed topology, the other four topologies—cycle, chain, star, and tree—are significantly varied in the placement order of nodes. Given the vast number of possible permutations among these graph topologies, a random assignment strategy was adopted in node labeling to balance computational cost and generalization. This approach not only reduces the complexity of the experiments but also enhances the robustness of the model by introducing stochasticity. Also, a custom-designed topology based on chemical knowledge was proposed. This novel architecture, denoted as the B−H reaction graph (right panel of Figure 4), subsequently served as the foundation for further comparative analysis.

2.3. Reaction Graph Neural Network. To effectively process the B−H reaction graph, we designed the reaction graph neural network (RGNN) based on our previously proposed Graph Attention Network Unit (GATU) for yield prediction.[18] As illustrated in Figure 5, the feature vectors representing each reaction component are first assigned to their corresponding nodes in the graph.

The model employs two graph convolution operations, which leverages self-attention mechanisms to enhance the model's ability to learn from key features across the graph. To preserve information from both convolution operations, a residual

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983

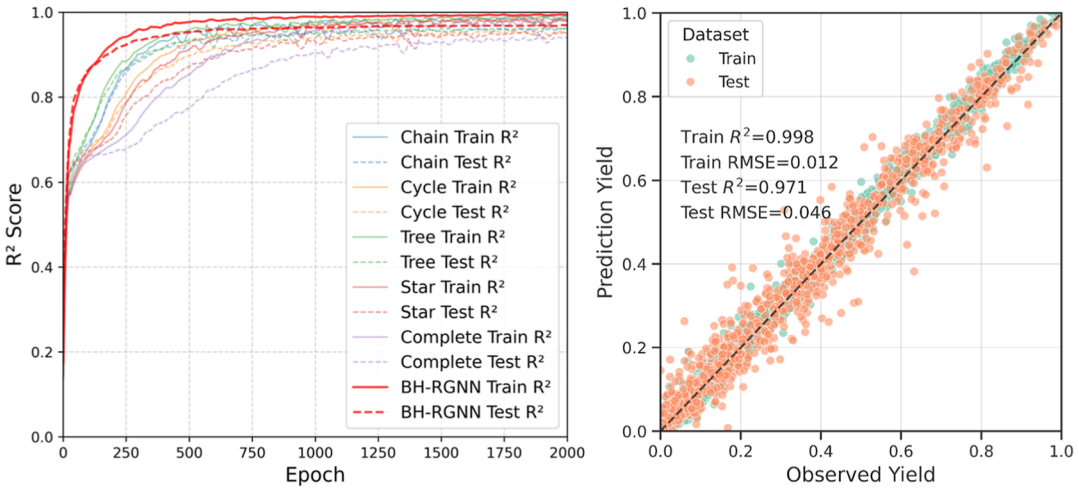


Figure 6. Performance of the model under different topological reaction graphs (left), and performance of the B−H reaction graph model (right).

Table 1. Performance of the BH-RGNN Model with Different Topological Reaction Graph Data

topological graph	train evaluation metrics			test evaluation metrics		
	R^2	RMSE	MAE	R^2	RMSE	MAE
chain	0.983 ± 0.004	0.036 ± 0.004	0.024 ± 0.001	0.960 ± 0.006	0.054 ± 0.003	0.037 ± 0.001
cycle	0.986 ± 0.002	0.032 ± 0.001	0.027 ± 0.001	0.957 ± 0.007	0.056 ± 0.004	0.038 ± 0.002
tree	0.990 ± 0.002	0.027 ± 0.003	0.020 ± 0.002	0.965 ± 0.007	0.051 ± 0.004	0.036 ± 0.002
star	0.984 ± 0.004	0.034 ± 0.004	0.023 ± 0.002	0.955 ± 0.007	0.057 ± 0.003	0.040 ± 0.002
complete	0.980 ± 0.003	0.039 ± 0.003	0.027 ± 0.002	0.943 ± 0.006	0.065 ± 0.002	0.044 ± 0.001
B−H reaction graph	**0.998 ± 0.001**	**0.012 ± 0.001**	**0.009 ± 0.001**	**0.971 ± 0.006**	**0.046 ± 0.004**	**0.031 ± 0.002**

connection is applied. This combined vector is then processed through max-pooling and min-pooling layers, followed by concatenation to retain more of the original feature information. The resulting one-dimensional vector is subsequently passed through a fully connected layer for dimensionality reduction, completing the final regression task. To mitigate overfitting, a Dropout layer is incorporated into the framework, randomly deactivating a portion of neurons during training. This strategy not only introduces stochasticity but also improves model robustness by encouraging the learning of more generalized feature representations. Through this architecture, the model is capable of capturing complex relationships among different reaction components while maintaining strong generalization performance and stability. It is worth noting that other topological reaction graphs also utilize this framework for comparison.

2.4. Model Training and Evaluation. The B−H data set was randomly split into train and test sets, with 70% of the data used for training and the remaining 30% for testing. This split facilitates comparison with prior work and effectively validates the superiority of our approach.[11] Five times random splits between the train and test sets were conducted for model performance evaluation to ensure result reliability. Wherein, the coefficient of determination (R^2), root-mean-square error (RMSE), and mean square error (MAE) are utilized as evaluation metrics. Hyperparameter optimization was performed using grid search (details in the Supporting Information, Section S2.1).

Figure 6 presents the performance of models trained on different topologies. The model using a custom-designed B−H reaction graph (shown on the left of Figure 6, red curve) demonstrated faster, smoother convergence and achieved the highest final R^2 value (More details in the Supporting Information, Section S2.3).

Table 1 provides detailed model performance metrics for comparison, showing that all five topologies yield R^2 values exceeding 0.94, indicating the effectiveness of the BH-RGNN model. Among the five common topologies, the tree-like structure achieved the best performance with an R^2 value of 0.965, while the fully connected topology, despite having the most bidirectional edges, exhibited the lowest R^2 of 0.943. This intriguing finding suggests that higher connectivity and interaction frequency between nodes do not necessarily correlate with better model performance. Specifically, simpler structures like the tree-like or chain-like topologies, which have only four bidirectional edges, outperformed more complex networks, achieving R^2 values of 0.965 and 0.960, respectively. This phenomenon implies that simpler network architectures may be more effective in handling data sets with outliers, as overly complex networks can propagate misleading information, negatively impacting prediction accuracy.[22] Models of the complete type, which have the most bidirectional edges, exhibit the lowest R^2 of 0.943. In contrast, tree- and chain-type models, with fewer bidirectional edges, achieve higher R^2 values of 0.965 and 0.960, respectively. And experimental results show that the model using this streamlined B−H reaction graph achieved superior performance, with an R^2 value of 0.971 on the test set, outperforming all other configurations. These findings demonstrate that our custom-designed B−H reaction graph with only one bidirectional edge connecting the substrate and product nodes reduces unnecessary complexity, making the model more accurately capture intercomponent interactions and limit the spread of erroneous information, thereby enhancing predictive performance.

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983

Table 2. Comparison of BH-RGNN with Other Models on the B−H Reaction HTE Dataset

		test evaluation metrics		
model	input	R^2	RMSE	MAE
RF	DRFP[13]	0.928 ± 0.002	0.073 ± 0.001	0.049 ± 0.001
XGBoost	DRFP[13]	0.946 ± 0.005	0.063 ± 0.003	0.042 ± 0.001
GAT[26]	graph	0.911 ± 0.036	0.080 ± 0.018	0.056 ± 0.011
yield-BERT[15]	SMILES	0.951 ± 0.005	0.058 ± 0.004	0.054 ± 0.003
yield-GNN[9]	graph, physicochemical parameters	0.961 ± 0.005		0.040 ± 0.002
multimodal BERT[27]	SMILES, physicochemical parameters	0.959 ± 0.005	0.055 ± 0.003	
MMHRP-GCL[18]	SMILES, graph	0.968 ± 0.002	0.049 ± 0.002	0.034 ± 0.001
uncertainty-aware GNN[8]	graph	0.974 ± 0.001	0.044 ± 0.001	0.029 ± 0.001
BH-RGNN (this work)	B−H reaction graph	0.971 ± 0.006	0.046 ± 0.001	0.031 ± 0.001

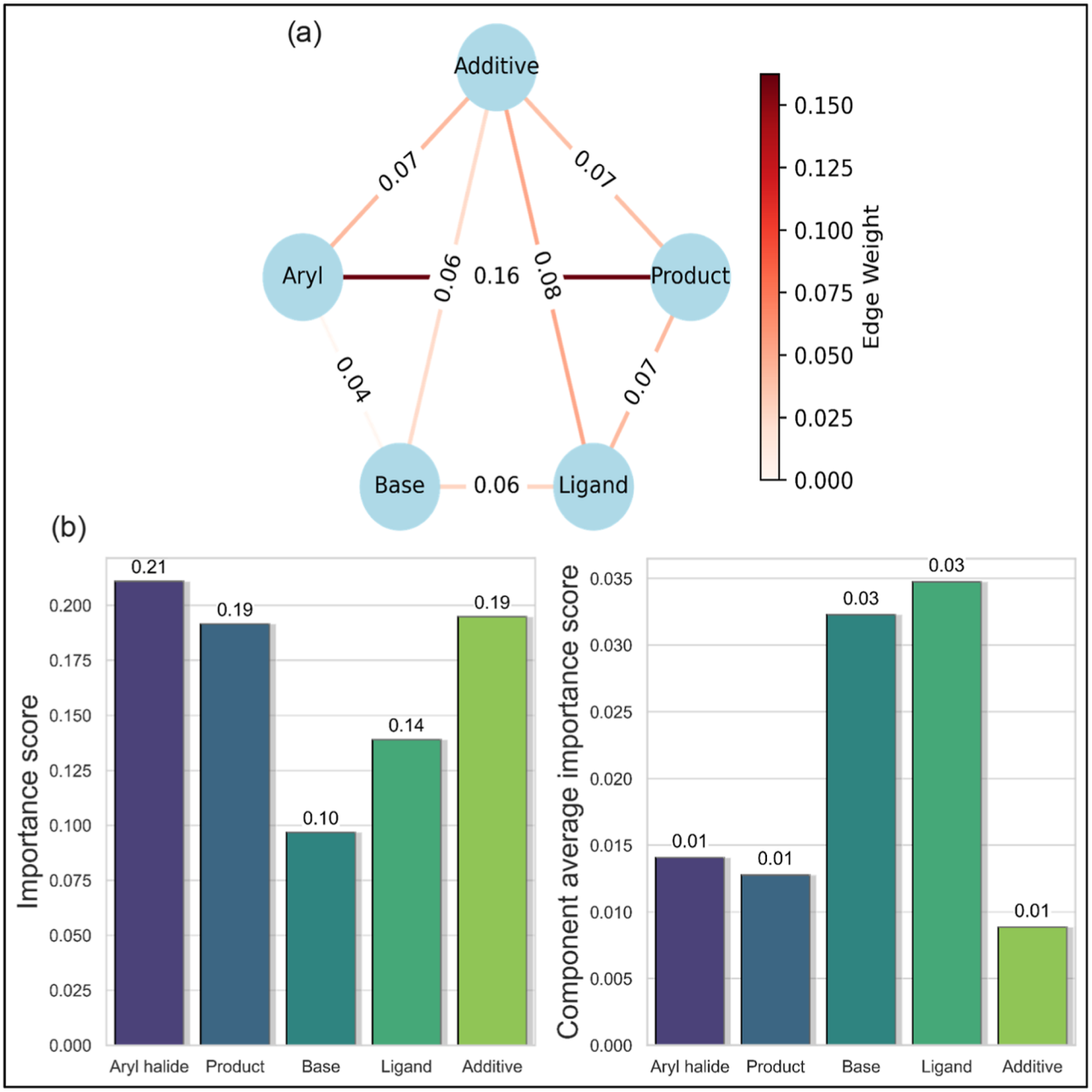


Figure 7. Model-based interpretation of the Buchwald−Hartwig reaction graph. (a) Importance of edges in the reaction graph. (b) Node importance (left) and mean importance per sample across nodes (right).

We investigated the impact of randomly permuting the node assignments corresponding to reaction components under the same topology, as shown in Figure S19 in the Supporting Information. Table S11 shows that such node permutations have

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983

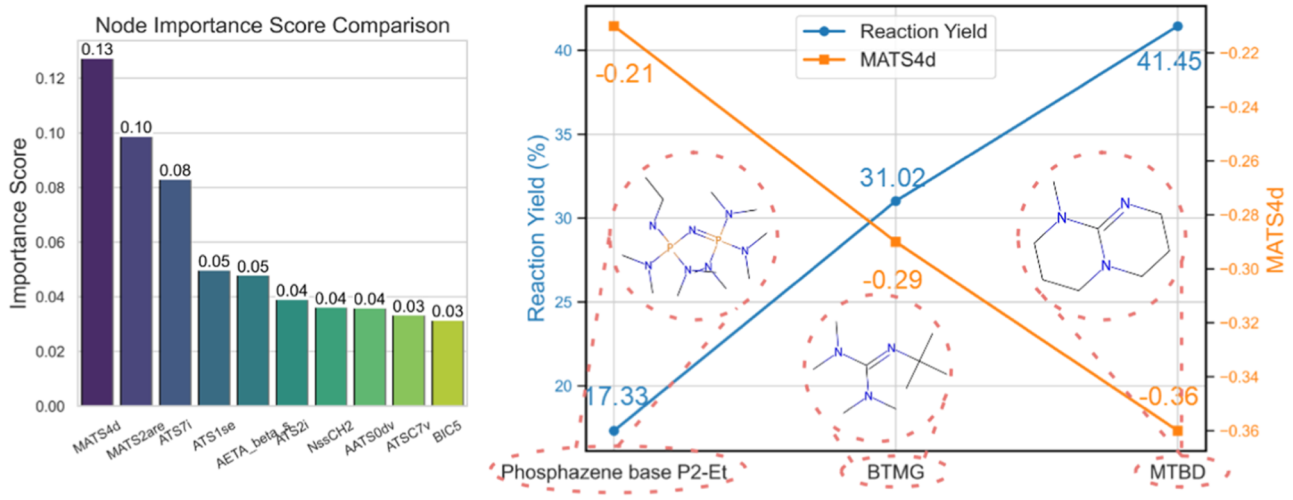


Figure 8. Model-based interpretation of the bases in the B−H reaction: node importance (left) and MATS4d value (right).

minimal effect on model performance, with R^2 ranging from 0.965 to 0.971. Therefore, when designing reaction graphs for optimal performance, the choice of topological connectivity should be the primary consideration, followed by the specific assignment of components to nodes.

We also tested the BH-RGNN using molecular fingerprints as input features. Five common 100-dimensional molecular fingerprints were utilized: Atom Pairs, Avalon, Morgan, RDkit, and Topological Torsions.[23−25] Table S12 shows that, except for RDkit fingerprint, using molecular fingerprints slightly improved model performance, with R^2 ranging from 0.973 to 0.975. However, molecular fingerprints provide limited interpretability for deriving chemical insights. Therefore, in this study, we focus on the more interpretable Mordred physicochemical descriptors.

To further validate the effectiveness of the proposed method, the BH-RGNN model trained on the B−H reaction graph was compared with existing state-of-the-art models. As shown in Table 2, the results demonstrate the superior performance of BH-RGNN based on the custom-designed B−H reaction graph. BH-RGNN achieves a near-state-of-the-art performance with an R^2 of 0.971, and RMSE and MAE values of 0.046 and 0.031, respectively, on the test set, using only non-DFT-derived physicochemical descriptors within the proposed custom topology and a single graph modality. It should be noted that while our model achieves high predictive performance on the B−H HTE data set, prior studies have demonstrated that such strong performance may be partly attributed to the relatively narrow and highly optimized chemical space inherent in this data set. For instance, the limited diversity of reaction outcomes—featuring only a few core C−N coupled products—can contribute to enhanced model accuracy. As shown in previous work, predictive power often declines significantly when models trained on such specialized data sets are applied to more diverse, real-world reaction data.[9] This outperforms models that rely on potentially computationally expensive DFT-based features or complex multimodal architectures, demonstrating both the validity of our approach and the potential of well-designed graph representations combined with appropriate model design. Although Kwon et al. developed an uncertainty-aware graph neural network and reached an R^2 of 0.974 under the same model evaluation method, their model used overly simplistic atomic and bond features, which made it difficult to gain a deeper chemical interpretation than BH-RGNN.[8]

2.5. Model Interpretation and Analysis. To further investigate the interpretability of the BH-RGNN model, we employed perturbation-based approaches.[28,29] Specifically, we set the portions of the target data to zero, and the modified inputs were fed into the model to evaluate the resulting predictions. The importance score of the masked portion was then quantified by computing RMSE between the predicted and true values. A higher change in RMSE indicates a more critical role of the masked component in the model's prediction.

As shown in Figure 7, edge masking was first applied to the custom-designed B−H reaction graph. The experimental results reveal that the bidirectional edge between the substrate (aryl halide) and the product exhibits the highest importance, with a value of 0.16. In contrast, the importance of other edges is generally below 0.1. Further analysis along the pathway from the substrate through the base and catalyst ligand to the product shows that the three unidirectional edges have importance scores of 0.04, 0.06, and 0.07, respectively. Their combined importance reaches 0.17, nearly matching that of the substrate−product bidirectional edge. Additionally, the edge importance between the additive and other components ranges from 0.06 to 0.08, showing relatively uniform contributions. These findings suggest that the model can effectively identify the relative importance of different edges. However, this study does not address the question of how to determine the optimal topological configuration for model performance. As such, we do not delve into defining or searching for the ideal B−H reaction graph. Future work exploring the precise identification of the optimal graph structure, in combination with the techniques proposed here, could offer a deeper mechanistic understanding of data derived from unknown reaction systems.

Node masking analysis was performed on the five different reaction components to evaluate their relative importance within the model. The results indicate that the substrate is the most critical component, followed by the additive, while the base exhibits the lowest importance. Notably, the diversity of additives (up to 22 types) in the B−H data set may significantly influence the importance analysis. Despite the low number of base types (only 3), their cumulative importance score of 0.10 corresponds to an average importance of approximately 0.03 per sample—almost the highest among all components—indicating

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983

a non-negligible contribution (Figure 7b). This suggests that even with limited diversity, bases meaningfully influence the reaction yield.

The BH-RGNN model was used to assess the importance of individual features within each node through feature masking. We conducted masking analysis on the features of the five components: aryl halides, additives, catalyst ligands, products, and bases. For each component, the top ten most important features were ranked. The experimental results show that for aryl halides, additives, catalyst ligands, and products, the importance scores of the top ten features are relatively close, differing by approximately 0.01 (more details in Section S3.1 in the Supporting Information). Moreover, the positions of the most important features can vary across repeated experiments, presenting challenges for deeper analysis. However, the model provides stable importance rankings specifically for bases, suggesting consistent predictive insights for this component.

As shown in Figure 8 (left), among the features of bases involved in the B−H reaction, MATS4d was identified as the most influential descriptor, with a significantly higher importance score compared to other features.[30,31] MATS4d is defined as the Moran autocorrelation coefficient of lag 4, weighted by the number of sigma (valence) electrons on each atom. This descriptor reflects the spatial pattern of valence electron density across the molecular structure, particularly capturing electronic correlations between atom pairs separated by four bonds.

We conducted a more in-depth investigation into the role of bases in the B−H reaction by analyzing the MATS4d value. In the data set, three bases were employed: Phosphazene base P2-Et, BTMG, and MTBD, which participated in 1320, 1318, and 1317 reactions, respectively. To explore the relationship between the MATS4d values of these bases and the corresponding reaction yields, statistical analysis was performed on reactions involving each base. As illustrated in Figure 8 (right), a clear trend emerges: when sufficient data is available, lower MATS4d values generally correlate with higher median reaction yields across the HTE reaction scope. Specifically, as the MATS4d value decreases from −0.21 to −0.36, the median yield increases from 17.33% to 41.45%. In this context, atoms with higher valence electron counts—such as carbon and nitrogen—contribute more significantly to the descriptor value than hydrogen. Structural analysis of the bases revealed that those with fewer alkyl groups and reduced molecular complexity tend to exhibit lower MATS4d values, consistent with a lower overall valence electron count and less extended electronic networks. The observed correlation between lower MATS4d values and higher reaction yields suggests that bases with simpler, less electron-dense architectures are associated with improved reaction efficiency.

3. CONCLUSIONS

In this study, we introduced the concept of a reaction graph for representing chemical reactions. By representing individual reaction components as graph nodes, we constructed a custom-defined reaction graph to serve as input data for the B−H reaction, enabling BH-RGNN to achieve near-state-of-the-art performance on Doyle et al.'s data set with an R^2 of 0.971. Furthermore, the model variants based on five commonly used graph topologies all outperformed the upper performance limit of the conventional RF model. Among these, the Complete topology yielded the lowest R^2 of 0.943, whereas the Tree topology achieved the highest R^2 of 0.965. Notably, the model was trained using only 100 non-DFT-derived physicochemical descriptors per node, demonstrating that high-accuracy prediction of chemical reaction outcomes can be achieved at a significantly reduced computational cost.

To gain deeper insights into how BH-RGNN interprets the degree of the B−H reaction yield, we conducted perturbation-based analyses to evaluate the model's predictive behavior. The results show that the model can effectively identify and differentiate the importance of various edges in the reaction graph. Furthermore, through feature importance analysis across all reaction components, we uncovered a meaningful trend: when sufficient data is available, bases with simpler, less electron-dense architectures tend to have lower MATS4d values, which are generally correlated with higher median reaction yields across the HTE reaction scope.

This finding provides valuable theoretical guidance for optimizing B−H reaction conditions, particularly in the rational selection of appropriate bases. It highlights the potential of combining well-designed graph representations with interpretable machine learning models to not only improve predictive accuracy but also provide chemical insights.

■ ASSOCIATED CONTENT

Data Availability Statement

The data underlying this study are available in the published article, in its Supporting Information, and openly available in GitHub at https://github.com/ZWR0/B-H-Reaction-Graph-Neural-Network.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.joc.5c01400.

Data processing, model evaluation, and interpretability analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

Authors

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

Shen Wang − *State Key Laboratory of Fine Chemicals, School of Chemical Engineering, Ocean and Life Sciences, Dalian University of Technology, Panjin 124221, P. R. China; Leicester International Institute, Dalian University of Technology, Panjin 124221, P. R. China;* orcid.org/0009-0008-4174-4301

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.joc.5c01400

Author Contributions

[§]W.Z. and S.W. contributed equally.

Funding

This research was supported by the National Natural Science Foundation of China (21903010) and 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) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **2009**, *20* (1), 61−80.
(2) Han, J.; Kwon, Y.; Choi, Y. S.; Kang, S. Improving chemical reaction yield prediction using pre-trained graph neural networks. *J. Cheminf.* **2024**, *16*, 25.
(3) Du, B. X.; Long, Y.; Li, X.; Wu, M.; Shi, J. Y. CMMS-GCL: cross-modality metabolic stability prediction with graph contrastive learning. *Bioinformatics* **2023**, *39* (8), btad503.
(4) An, H.; Liu, X.; Cai, W.; Shao, X. AttenGpKa: a universal predictor of solvation acidity using graph neural network and molecular topology. *J. Chem. Inf. Model.* **2024**, *64* (14), 5480−5491.
(5) Wang, S.; Yue, H.; Yuan, X. Accelerating Polymer Discovery with Uncertainty-Guided PGCNN: Explainable AI for Predicting Properties and Mechanistic Insights. *J. Chem. Inf. Model.* **2024**, *64* (14), 5500−5509.
(6) Lu, X.; Xie, L.; Xu, L.; Mao, R.; Xu, X.; Chang, S. Multimodal fused deep learning for drug property prediction: Integrating chemical language and molecular graph. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 1666−1679.
(7) Boulougouri, M.; Vandergheynst, P.; Probst, D. Molecular set representation learning. *Nat. Mach. Intell.* **2024**, *6* (7), 754−763.
(8) Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, S. Uncertainty-Aware Prediction of Chemical Reaction Yields with Graph Neural Networks. *J. Cheminf.* **2022**, *14* (1), 2.
(9) Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the Use of Real-World Datasets for Reaction Yield Prediction. *Chem. Sci.* **2023**, *14* (19), 4997−5005.
(10) 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* (1), 3569.
(11) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C−N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186−190.
(12) Chuang, K. V.; Keiser, M. J. Comment on "Predicting reaction performance in C−N cross-coupling using machine learning". *Science* **2018**, *362* (6416), No. eaat8603.
(13) Estrada, J. G.; Ahneman, D. T.; Sheridan, R. P.; Dreher, S. D.; Doyle, A. G. Response to Comment on "Predicting reaction performance in C−N cross-coupling using machine learning". *Science* **2018**, *362* (6416), No. eaat8763.
(14) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **2020**, *6* (6), 1379−1390.
(15) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. L. Prediction of chemical reaction yields using deep learning. *Mach. Learn.: Sci. Technol.* **2021**, *2* (1), 015016.
(16) Probst, D.; Schwaller, P.; Reymond, J. L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery* **2022**, *1* (2), 91−97.
(17) Zhao, W. R.; Li, Y. Predicting the Yield of Pd-Catalyzed Buchwald−Hartwig Amination Using Machine Learning with Extended Molecular Fingerprints and Selected Physical Parameters. *ChemistrySelect* **2024**, *9* (33), No. e202402529.
(18) Wang, S.; Zhao, W. R.; Liu, Y.; Li, Y. Multi-modal Homogeneous Chemical Reaction Performance Prediction with Graph and Chemical Language Information. *Chin. J. Chem.* **2025**, *43* (11), 1230−1238.
(19) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31−36.
(20) Moriwaki, H.; Tian, Y. S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminf.* **2018**, *10*, 4.
(21) Lopez-del Rio, A.; Martin, M.; Perera-Lluna, A.; Saidi, R. Effect of sequence padding on the performance of deep learning models in archaeal protein functional prediction. *Sci. Rep.* **2020**, *10* (1), 14634.
(22) Yu, M.; Wang, S.; Zhang, G.; Mao, J.; Yin, C.; Liu, Q.; Wang, Y. Netsafe: Exploring the topological safety of multi-agent networks. *arXiv* **2024**, arXiv:2410.15686.
(23) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25* (2), 64−73.
(24) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(25) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27* (2), 82−85.
(26) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2018**, arXiv:1710.10903.
(27) Baraka, S.; Kerdawy, A. M. E. Multimodal Transformer-Based Model for Buchwald-Hartwig and Suzuki-Miyaura Reaction Yield Prediction. *arXiv* **2022**, arXiv:2204.14062.
(28) Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. *arXiv* **2019**, arXiv:1903.03894.
(29) Wu, Z.; Wang, J.; Du, H.; Jiang, D.; Kang, Y.; Li, D.; Pan, P.; Deng, Y.; Cao, D.; Hsieh, C. Y.; et al. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat. Commun.* **2023**, *14* (1), 2585.
(30) Huang, Q.; Song, W.; Wang, L. Quantitative relationship between the physicochemical characteristics as well as genotoxicity of organic pollutants and molecular autocorrelation topological descriptors. *Chemosphere* **1997**, *35* (12), 2849−2855.
(31) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randi, M. Molecular connectivity I: Relationship to nonspecific local anesthesia. *J. Pharm. Sci.* **1975**, *64* (12), 1971−1974.

https://doi.org/10.1021/acs.joc.5c01400
J. Org. Chem. 2025, 90, 12975−12983