Reaction Prediction via Interaction Modeling of Symmetric Difference Shingle Sets

Runhan Shi, Letian Chen, Gufeng Yu, Yang Yang*
School of Computer Science, Shanghai Jiao Tong University
{han.run.jiangming, clt2001, jm5820zz}@sjtu.edu.cn, yangyang@cs.sjtu.edu.cn

Abstract

Chemical reaction prediction remains a fundamental challenge in organic chemistry, where existing machine learning models face two critical limitations: sensitivity to input permutations (molecule/atom orderings) and inadequate modeling of substructural interactions governing reactivity. These shortcomings lead to inconsistent predictions and poor generalization to real-world scenarios. To address these challenges, we propose ReaDISH, a novel reaction prediction model that learns permutation-invariant representations while incorporating interaction-aware features. It introduces two innovations: (1) symmetric difference shingle encoding, which computes molecular shingle differences to capture reaction-specific structural changes while eliminating order sensitivity; and (2) geometry-structure interaction attention, a mechanism that models intra- and inter-molecular interactions at the shingle level. Extensive experiments demonstrate that ReaDISH improves reaction prediction performance across diverse benchmarks. It shows enhanced robustness with an average improvement of 8.76% on R^2 under permutation perturbations.[1]

1 Introduction

Accurate modeling of chemical reactions is a fundamental problem in organic chemistry, as it provides critical insights into reaction mechanisms, predicts reaction outcomes, and guides experimental design [1–4]. Reaction representation learning is central to tasks such as reaction yield prediction [5], enantioselectivity prediction [6], conversion rate estimation [7], and reaction type classification [8]. These tasks have gained considerable attention with the rise of machine learning (ML). Nevertheless, representing reactions for ML is challenging due to the complexity of reaction spaces and the multitude of factors influencing chemical experiments like substrates, catalysts, and reaction conditions [9–11]. While many reactions may appear theoretically feasible [12], in practice, successfully executing them requires a deeper understanding of how these factors interact. Even slight variations in any of these elements can significantly influence the outcome [13, 14], making the reaction prediction problem a complex and nuanced challenge.

Despite advances in ML models, they still face two major limitations that hinder their broader applicability, as shown in Figure 1. First, many models, especially those based on sequential representations like SMILES [15], fail to account for the inherent permutation invariance of chemical reactions [16, 17]. They tend to produce inconsistent predictions when changing the ordering of input molecules (e.g., swapping reactants and reagents) or the ordering of atoms (e.g., alternative SMILES). Such sensitivity to input ordering undermines reliability and generalizability. While data augmentation techniques can partially alleviate this issue, training on all possible permutations is

*Corresponding author. This work was supported by the National Natural Science Foundation of China (No. 62272300).

[1]The code is available at https://github.com/Meteor-han/ReaDISH.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

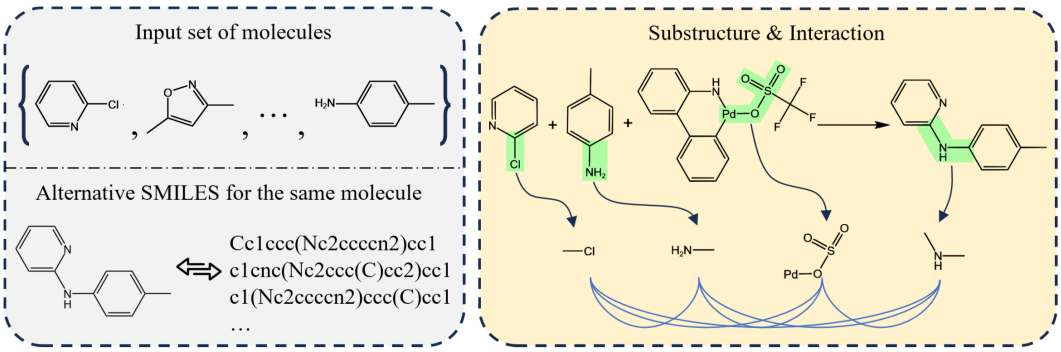


Figure 1: **Challenges for reaction representation learning**. (a) Permutation perturbations by inter-molecular order variation (top) and intra-molecular SMILES token randomization (bottom). (b) Key substructures that determine the outcomes of reactions and their inherent interactions.

computationally inefficient and impractical at scale [18, 19]. Achieving permutation invariance is essential for robust and unbiased reaction modeling across diverse chemical scenarios.

Second, current works often fail to capture the structural interactions that drive chemical reactivity in reaction modeling. Many approaches overlook critical substructures, particularly those that change from reactants to products (e.g., reaction centers [20]), and lack explicit modeling of interactions between such substructures. Although atom-level interactions have been widely explored in molecular representation learning [21–23], capturing higher-order interactions among chemically meaningful substructures across molecules remains limited in the reaction prediction problem. This lack of interaction awareness leads to suboptimal predictions and diminishes the model's ability to capture underlying reaction mechanisms [24]. As a result of these two limitations, many models perform poorly on out-of-sample reactions that better reflect real-world chemical diversity [25–27].

To address these challenges, we propose the ReaDISH model for **Rea**ction prediction via interaction modeling of symmetric **DI**fference **SH**ingle sets. First, we construct chemically meaningful substructure sets named *molecular shingle* [28–30] sets via the circular topology for each molecule to capture key structural components. Then we compute and encode the *symmetric difference* of shingle sets between reactants and products, which is naturally robustly permutation-invariant. Second, we add intra- and inter-molecular interaction pair representation on shingle-level attention based on geometric distance, structural distance, and chemical connectivity of shingles to improve interaction awareness. We conduct numerical experiments on various reaction representation learning tasks and benchmarks, including **reaction yield prediction**, **enantioselectivity prediction**, **conversion rate estimation**, and **reaction type classification**. Experimental results show that compared with baseline models, our proposed ReaDISH achieves SOTA on prediction accuracy and uncertainty estimation in most scenarios. It increases R^2 by an average of 8.76% under out-of-sample splits when performing permutation perturbations, presenting a better generalization capability. We present related work in Appendix A. Our key contributions are summarized as follows:

- We introduce a new way to present reaction structures based on the symmetric difference of molecular shingles between reactants and products, which is inherently permutation-invariant and highlights critical substructures that drive reaction outcomes.
- We design an interaction-aware attention mechanism that integrates geometric distance, structural distance, and chemical connectivity between shingles. This pair representation enables the model to capture both intra- and inter-molecular interactions.
- Extensive results across a wide range of tasks demonstrate that our proposed ReaDISH outperforms baseline models, achieving higher prediction accuracy and lower prediction uncertainty, especially under out-of-sample conditions.

2 Background

2.1 Problem definition

We denote the dataset of chemical reactions and their associated labels as $\mathcal{D} = \{(\mathcal{R}_i, y_i)\}_{i=1}^{N}$, where each reaction $\mathcal{R}_i$ consists of a set of participating molecules and $y_i \in \mathcal{Y}$ is the corresponding property label. Depending on the task, y_i may represent (1) a real-valued scalar for regression tasks, such as reaction yield or energy barrier, i.e., $y_i \in \mathbb{R}$, or (2) a categorical label for classification tasks, such as reaction type or success/failure, i.e., $y_i \in \mathbb{N}$.

Formally, each reaction is denoted by $\mathcal{R}_i = \left\{\mathcal{M}_i^{\mathrm{r}}; \mathcal{M}_i^{\mathrm{p}}\right\} = \left\{M_i^{\mathrm{r}_1}, \ldots, M_i^{\mathrm{r}_m}; M_i^{\mathrm{p}_1}, \ldots, M_i^{\mathrm{p}_n}\right\}$, where $\mathcal{M}_i^{\mathrm{r}}$ ($\mathcal{M}_i^{\mathrm{p}}$) denotes the set of reactant (product) molecules and m (n) is the number of reactant (product) molecules. For simplicity and consistency, we refer to molecules other than the products collectively as reactants (including catalysts, solvents, etc.). Each molecule M_i^j is represented as the 3D conformer $C_i^j = \{(a_k, \mathbf{x}_k)\}_{k=1}^{N_{\mathrm{a}}}$, where a_k is the atomic type, $\mathbf{x}_k \in \mathbb{R}^3$ is the spatial coordinate, and N_{a} is the number of atoms. The goal of reaction representation learning is to design a mapping function $f_\phi : \mathcal{R} \to \mathcal{Y}$, parameterized by ϕ, such that the predicted label $\hat{y}_i = f_\phi(\mathcal{R}_i)$ approximates the true label y_i. During training, the model minimizes an objective function, such as root mean squared error (RMSE) for regression or cross-entropy loss for classification.

2.2 Molecular shingles

In cheminformatics, *molecular shingles* refer to structured fragments designed to capture local connectivity patterns in molecules. These shingles are typically defined as sets of atoms and their connectivity within a defined neighborhood, which are also utilized in classical fingerprints like ECFP [31], effectively representing the structural information of the molecule.

Formally, let a molecule M be represented by a conformer $C = (V, E, \mathbf{X})$ with atoms V, bonds E, and coordinates $\mathbf{X}$. We define an r-sized shingle of an atom $v \in V$ as a connected subgraph induced by the center atom and its neighboring atoms with radius r, along with the corresponding bonds and their 3D positions. Let $\mathcal{N}_r(v)$ denote the set of neighboring atoms to atom v with radius r. The corresponding shingle is then given by

$$S^{(r)}(v) = C\left[\{v\} \cup \mathcal{N}_r(v)\right], \tag{1}$$

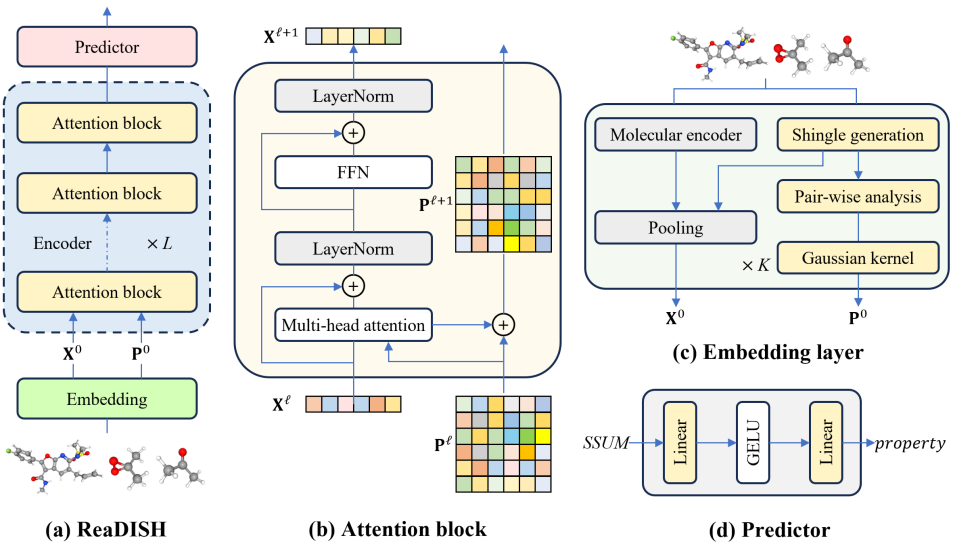
where $C[\cdot]$ denotes the sub-conformer of C restricted to the specified subset of atoms $\{v\} \cup \mathcal{N}_r(v)$. The set of all r-sized shingles in M is denoted as $\mathcal{S}_M^{(r)}$, and its representation is invariant to the arrangement of atoms.

3 Method

3.1 Model architecture

We propose ReaDISH, a novel framework for reaction property prediction that learns permutation-invariant and interaction-aware representations of chemical reactions. The overall architecture is illustrated in Figure 2. It comprises three main components: an embedding layer that processes molecular inputs, an encoder that captures geometric and structural features, and a lightweight predictor that outputs reaction properties.

Given a reaction $\mathcal{R} = \left\{C_i^{\mathrm{r}_1}, \ldots, C_i^{\mathrm{r}_m}; C_i^{\mathrm{p}_1}, \ldots, C_i^{\mathrm{p}_n}\right\}$ consisting of reactant and product molecules in 3D conformer format, ReaDISH first encodes each molecule C_i^j using a 3D molecular encoder f_{mol} to generate atom-level representations $\mathbf{X}^{\mathrm{a}} \in \mathbb{R}^{N_{\mathrm{a}} \times F}$, where F is the embedding dimension. A shingle-generation algorithm then extracts molecular shingles and computes the *symmetric difference* between reactant and product shingle sets to capture reaction-specific transformations. Next, ReaDISH aggregates the atom-level representations within each shingle through a pooling operation, producing initial shingle-level embeddings $\mathbf{X}^0 \in \mathbb{R}^{N_{\mathrm{s}} \times F}$, where N_{s} is the number of shingles. We compute three types of pairwise relations to model interactions between shingles: geometric distances, structural distances, and chemical connectivity. These are encoded as pairwise matrices $\mathbf{P} = \{\mathbf{P}_{\mathrm{g}}, \mathbf{P}_{\mathrm{e}}, \mathbf{P}_{\mathrm{s}}\} \in \mathbb{R}^{3 \times N_{\mathrm{s}} \times N_{\mathrm{s}}}$. To incorporate these relationships into the attention mechanism, we apply K Gaussian kernels to each pairwise matrix to produce the initial pair representation $\mathbf{P}^0$. The ReaDISH encoder



(a) ReaDISH **(b) Attention block** **(c) Embedding layer** **(d) Predictor**

Figure 2: **Overall architecture of ReaDISH** (a). It consists of an embedding layer (c), an encoder incorporating L attention blocks (b), where the extended self-attention module with the Gaussian kernel is depicted in Figure 4(a), and a lightweight predictor (d) for predicting reaction properties.

then processes the shingle-level representations using L transformer-style [32] attention blocks. Each block contains a multi-head attention (MHA) layer that integrates the pairwise attention biases $\mathbf{P}^{\ell}$ at each layer ℓ. Finally, ReaDISH uses a special `[SSUM]` token to summarize the reaction representation and passes it to a multilayer perceptron (MLP) head to predict reaction property $\hat{y}$.

3.2 Embedding layer for shingles

Molecular encoder. The 3D molecular encoder f_{mol} transforms each molecule $C = \{(a_k, \mathbf{x}_k)\}_{k=1}^{N_{\text{a}}}$, comprising atom types a_k and 3D coordinates $\mathbf{x}_k$, into atom-level feature representations $\mathbf{X}^{\text{a}} \in \mathbb{R}^{N_{\text{a}} \times F}$ followed by layer normalization:

$$\mathbf{X}^{\text{a}} = \text{LayerNorm}\left(f_{\text{mol}}(C)\right). \tag{2}$$

Symmetric difference shingle set generation. In reaction modeling, comparing shingles across reactants and products allows for the precise identification of structural transformations [33]. For each reaction, we treat molecules except products as reactants. Shingles within a given radius are generated for reactants and products, respectively. The generated shingles are expressed in SMILES strings to perform the symmetric difference operation, as shown in Figure 3. This operation yields the set of shingles that are exclusive to either reactants or products, thereby identifying structural changes associated with the reaction mechanism. The complete algorithm and illustration for computing the symmetric difference shingle set are provided in Appendix C. Formally, for a reaction $\mathcal{R} = \{\mathcal{M}^{\text{r}}; \mathcal{M}^{\text{p}}\}$ and a radius r_{max}, we define the symmetric difference shingle set $\mathcal{S}_{\mathcal{R}}$ as

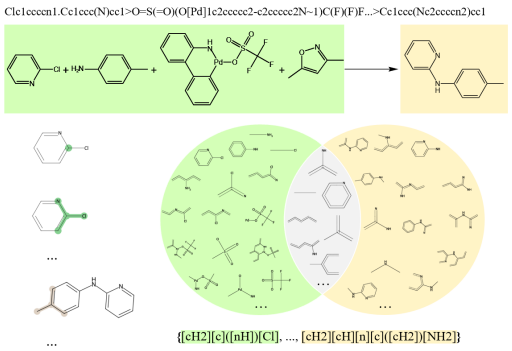


Figure 3: **Shingles generation**. We remove the intersection part (in gray), and keep the remaining shingles for reactants (in green) and products (in yellow).

$$\mathcal{S}_{\mathcal{R}} = \mathcal{S}_{\text{react}} \triangle \mathcal{S}_{\text{prod}} = \left(\bigcup_{i=1}^{r_{\text{max}}} \bigcup_{j=1}^{m} \mathcal{S}^{(i)}_{M^{\text{r}}_j} \right) \triangle \left(\bigcup_{i=1}^{r_{\text{max}}} \bigcup_{k=1}^{n} \mathcal{S}^{(i)}_{M^{\text{p}}_k} \right), \tag{3}$$

where $\mathcal{S}_{\text{react}}$ ($\mathcal{S}_{\text{prod}}$) is the shingle set for reactants (products), and $\triangle$ denotes the symmetric difference operator $\mathcal{A} \triangle \mathcal{B} = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A})$. This set-based strategy is invariant to the ordering of molecules.

Shingle pooling. To construct shingle-level representations from atom-level embeddings, we apply a shingle pooling operation over the set of symmetric difference shingles. Each shingle $S \in \mathcal{S}_{\mathcal{R}}$ consists of a subset of atoms, and its representation is computed by averaging the embeddings:

$$\mathbf{h}_S = \text{Pooling}\left(S, \mathbf{X}^{\text{a}}\right) = \frac{1}{\mid \mathcal{A}(S) \mid} \sum_{i \in \mathcal{A}(S)} \mathbf{x}_i^{\text{a}}, \tag{4}$$

where $\mathcal{A}(S) \subseteq \{1, \ldots, N_{\text{a}}\}$ denotes the indices of atoms in shingle S and $\mathbf{x}_i^{\text{a}}$ is the embedding of the i-th atom. This operation aggregates localized chemical information within each shingle, enabling the model to learn chemically meaningful shingle-level representations. The resulting set of initial shingle-level representations is denoted as

$$\mathbf{X}^0 = [\, \mathbf{h}_{S_i} \,]_{i=1}^{N_{\text{s}}} \in \mathbb{R}^{N_{\text{s}} \times F}. \tag{5}$$

These pooled embeddings serve as the input to the subsequent encoder, which models interactions between shingles to capture the overall reaction context.

Pairwise interaction. We introduce an interaction framework that explicitly captures pairwise dependencies between molecular shingles. This formulation allows ReaDISH to model both intra- and inter-molecular interactions, leveraging geometric and structural features to represent chemical transformations. As illustrated in the right panel of Figure 4(b), we define three pairwise metrics between shingles S_i and S_j as

$$\begin{aligned} d_{ij}^{\text{g}} &= \begin{cases} \|\mathbf{c}_i - \mathbf{c}_j\|_2, & \text{if } S_i, S_j \text{ belong to the same molecule,} \\ 0, & \text{otherwise;} \end{cases} \\ d_{ij}^{\text{e}} &= \begin{cases} 1, & \text{if } S_i, S_j \text{ belong to the same molecule,} \\ 0, & \text{otherwise;} \end{cases} \\ d_{ij}^{\text{s}} &= 1 - \text{sim}(S_i, S_j), \end{aligned} \tag{6}$$

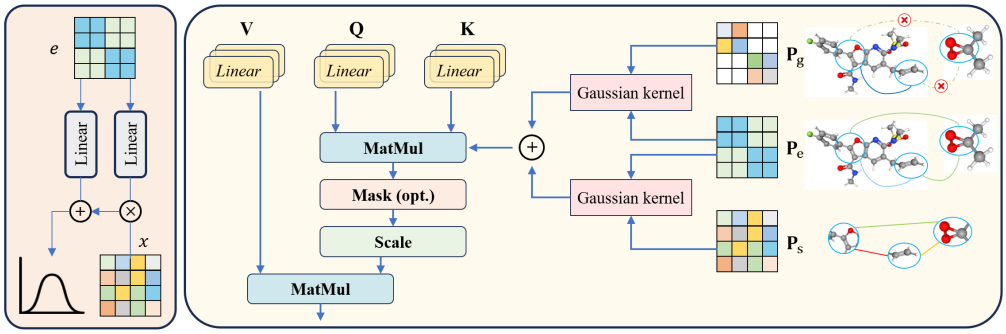
where $\mathbf{c}_i$ and $\mathbf{c}_j$ denote the geometric centers of shingles S_i and S_j, respectively, and $\text{sim}(\cdot, \cdot)$ is the Tanimoto similarity [34] of their Morgan fingerprints [35]. These metrics yield a structured pairwise representation $\mathbf{P} = \{\mathbf{P}_{\text{g}}, \mathbf{P}_{\text{e}}, \mathbf{P}_{\text{s}}\} = \left[\, d_{ij}^{\text{g}}, d_{ij}^{\text{e}}, d_{ij}^{\text{s}} \,\right]_{i,j=1}^{N_{\text{s}}} \in \mathbb{R}^{3 \times N_{\text{s}} \times N_{\text{s}}}$, where each component captures a specific interaction modality across all N shingles. The geometric distance d_{ij}^{g} encodes spatial relationships within the same molecule, analogous to atom-level distance modeling. The binary chemical connectivity d_{ij}^{e} indicates whether two shingles are part of the same molecule. The structural distance d_{ij}^{s} emphasizes functional dissimilarity, which is essential for modeling reactive interactions across different molecules. These pair interactions enable a finer-grained understanding of reaction mechanisms and facilitate more accurate prediction of reaction outcomes.

Gaussian kernel. After computing the pairwise interactions, we integrate these geometry-aware and structure-aware signals into the attention mechanism to better capture chemically meaningful context. To this end, we adopt the Gaussian Kernel with Pair Type (GKPT) [36], a technique effectively applied in molecular representation learning [21]. GKPT applies an affine transformation to pairwise distances based on the interaction type, followed by a classical Gaussian kernel, as illustrated in Figure 4(a). Formally, the GKPT is defined as

$$\text{GKPT}\left((x, e), \boldsymbol{\mu}, \boldsymbol{\sigma}\right) = \mathcal{G}\left(\mathbf{E}_1(e) \cdot x + \mathbf{E}_2(e), \boldsymbol{\mu}, \boldsymbol{\sigma}\right), \tag{7}$$

where $\mathcal{G}\left(x', \boldsymbol{\mu}, \boldsymbol{\sigma}\right) = \frac{1}{\sqrt{2\pi}\boldsymbol{\sigma}} \exp\left(-\frac{1}{2}\left(\frac{x' - \boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)^2\right)$ is the standard Gaussian kernel. Here, x is the input distance, e is the pair type index, $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N_{\text{e}} \times K}$ are learnable embedding layers, N_{e} is the number of pair types, and $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^K$ are learnable parameters for K Gaussian kernels. We apply separate GKPT modules to process geometric distance (d_{ij}^{g}) and structural distance (d_{ij}^{s}) along with chemical connectivity (d_{ij}^{e}) as shown in Figure 4(b). Each output is projected into the attention space and combined to form the pairwise bias term as

$$p_{ij} = \text{GKPT}_{\text{g}}\left((d_{ij}^{\text{g}}, d_{ij}^{\text{e}}), \boldsymbol{\mu}, \boldsymbol{\sigma}\right) \mathbf{w}_{\text{g}} + \text{GKPT}_{\text{s}}\left((d_{ij}^{\text{s}}, d_{ij}^{\text{e}}), \boldsymbol{\mu}, \boldsymbol{\sigma}\right) \mathbf{w}_{\text{s}}, \tag{8}$$



(a) Gaussian kernel (b) Geometry- and structure-enhanced attention

Figure 4: **Interaction-aware attention**. (a) Gaussian kernel with learned pair type transformations. (b) Self-attention enhanced by geometric and structural interactions. Each pairwise representation incorporates one intra-molecular relationship (geometric distance) and two inter-molecular relationships (structural distance and chemical connectivity).

where $\mathbf{w}_{\mathrm{g}}, \mathbf{w}_{\mathrm{s}} \in \mathbb{R}^K$ are learnable projection vectors. The full interaction-aware pairwise bias matrix used in attention is denoted as

$$\mathbf{P}^0 = [\,p_{ij}\,]_{i,j=1}^{N_{\mathrm{s}}} \in \mathbb{R}^{N_{\mathrm{s}} \times N_{\mathrm{s}}}. \tag{9}$$

This pairwise representation encodes chemically relevant interactions and informs the subsequent attention computation, enabling the model to learn complex spatial and structural dependencies in chemical reactions.

3.3 Enhanced attention block

To incorporate geometric and structural information, we introduce an attention mechanism enhanced by pair representation. Specifically, we augment the standard self-attention with a learnable pairwise bias derived from the previously defined pairwise representations. At each layer ℓ, we update the pairwise bias p_{ij}^{ℓ} between shingle S_i and S_j using the query-key interaction:

$$p_{ij}^{\ell+1} = \frac{\mathbf{Q}_i^{\ell} \left(\mathbf{K}_j^{\ell}\right)^{\top}}{\sqrt{d}} + p_{ij}^{\ell}, \tag{10}$$

where $\mathbf{Q}_i^{\ell}$ and $\mathbf{K}_j^{\ell}$ are the query and key vectors of shingles i and j at layer ℓ, respectively, and d is the hidden dimension. The attention score between shingles i and j is then computed by incorporating this updated bias into the scaled dot-product attention:

$$\text{Attention}\left(\mathbf{Q}_i^{\ell}, \mathbf{K}_j^{\ell}, \mathbf{V}_j^{\ell}\right) = \text{softmax}\left(\frac{\mathbf{Q}_i^{\ell} \left(\mathbf{K}_j^{\ell}\right)^{\top}}{\sqrt{d}} + p_{ij}^{\ell-1}\right)\mathbf{V}_j^{\ell}, \tag{11}$$

where $\mathbf{V}_j^{\ell}$ is the value vector of shingle j. This formulation allows the attention mechanism to be guided by chemically meaningful geometric and structural priors. Additionally, we prepend a learnable `[SSUM]` token to the shingle sequence $\mathbf{X}^{\ell}$ at each layer, which serves as a global summary token for downstream reaction property prediction.

3.4 Predictor

ReaDISH utilizes a lightweight predictor, as depicted in Figure 2(d), to predict reaction properties. It consists of two linear layers with GELU activation [37]. This simple yet effective architecture maps the final `[SSUM]` token representation, which encodes the global context of the reaction, to the target prediction space.

3.5 Pre-training via pseudo-reaction-type classification

To enhance the generalization capability of ReaDISH, we introduce a pseudo-reaction-type classification task as the pre-training objective on a large-scale dataset. Specifically, we assign each reaction

a pseudo-label by clustering its structural representation at different granularities. This multi-scale classification setting encourages the encoder to capture both coarse-grained and fine-grained structural semantics. We generate K_{t} pseudo-labels for each reaction. Formally, the pre-training loss is defined as

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K_{\mathrm{t}}}\mathcal{L}\left(\mathbf{W}_k(f_\phi(\mathbf{X}_n^0, \mathbf{P}_n^0)),\ y_n^{(k)}\right), \tag{12}$$

where f_ϕ denotes the ReaDISH encoder, $\mathcal{L}$ is the cross-entropy loss, $\mathbf{W}_k \in \mathbb{R}^{F\times N_k}$ with target dimension N_k for $k \in \{1,\ldots,K_{\mathrm{t}}\}$ are parameters of fully connected layers, respectively, and $y_n^{(k)} \in \mathbb{N}$ are pseudo-labels for the n-th sample. More information can be found in Appendix D.

4 Experiments

4.1 Settings

Pre-training datasets. We collect 3.7M chemical reactions for pre-training based on the United States Patent and Trademark Office (USPTO) dataset [38] and the Chemical Journals with High Impact Factor (CJHIF) dataset [39]. We employ the DRFP [33] method with the K-means algorithm to compute pseudo-labels. More information can be found in Appendix B.

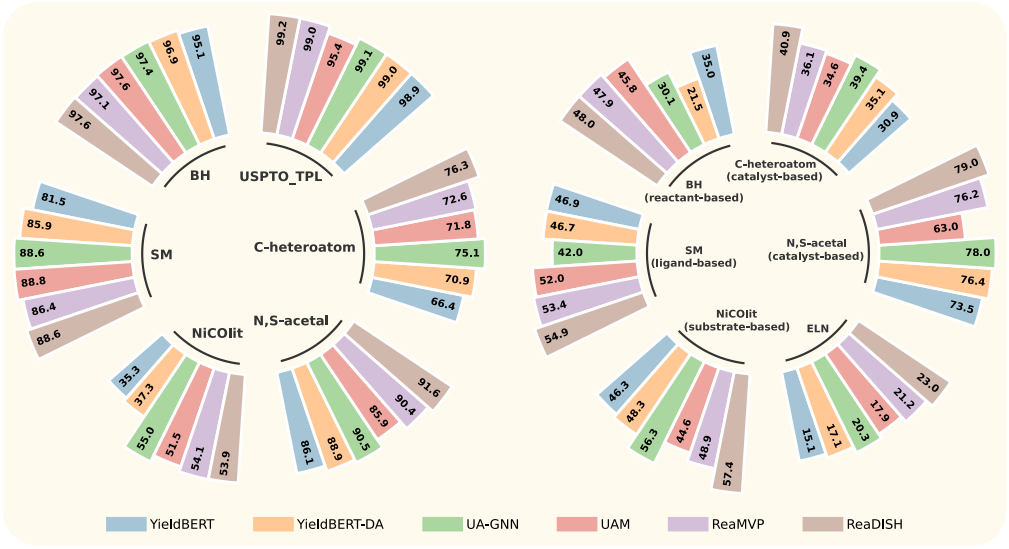
Downstream datasets. To comprehensively evaluate the effectiveness of ReaDISH, we use seven datasets across a wide range of chemical tasks, including: (1) yield prediction, the Buchwald-Hartwig (BH) dataset [13], the Suzuki-Miyaura (SM) dataset [14], the real-world electronic laboratory notebook (ELN) dataset [40], and the Ni-catalyzed C-O bond activation (NiCOlit) dataset [41]; (2) enantioselectivitiy prediction, the asymmetric N,S-acetal formation (N,S-acetal) dataset [42]; (3) conversion rate estimation, the C-heteroatom-coupling reactions (C-heteroatom) dataset [43]; and (4) reaction type classification, the USPTO_TPL dataset [8]. We consider both random and more challenging out-of-sample splits, which better reflect practical deployment scenarios. More information can be found in Appendix B.

Baselines and metrics. We use five representative models as baselines in our experiments: YieldBERT [5], YieldBERT-DA [44], UA-GNN [45], UAM [46], and ReaMVP [47], which we introduce in related work in Appendix A. We additionally compare our method with a cross-attention baseline that directly models reactant-product interactions as provided in Appendix E. Given the importance of spatial structural patterns in molecules, we adopt the SE(3)-invariant [48] Uni-Mol [21] as our 3D molecular encoder f_{mol} for its excellent performance. For regression tasks, we report mean absolute error (MAE $\downarrow$), root mean squared error (RMSE $\downarrow$), and coefficient of determination (R^2 $\uparrow$). For classification tasks, we measure accuracy (ACC $\uparrow$), Matthews correlation coefficient (MCC $\uparrow$), and confusion entropy (CEN $\downarrow$). More information can be found in Appendix F.

4.2 Main results

ReaDISH achieves SOTA or competitive performance across diverse benchmark datasets. As shown in Figure 5(a), ReaDISH achieves top or competitive results on six datasets under random splits. It ranks first on the BH, N,S-acetal, and C-heteroatom datasets for R^2, and the USPTO_TPL dataset for accuracy. On the SM and NiCOlit datasets, ReaDISH consistently matches or outperforms existing models. Figure 5(b) presents results on more challenging out-of-sample splits, where the test distribution differs from the training one. ReaDISH consistently outperforms all baseline models across six benchmark datasets. It shows a greater advantage under out-of-sample splits, which better reflects real-world scenarios and offers greater practical relevance than random splits, demonstrating ReaDISH's strong robustness across diverse reaction types and chemical spaces.

On the particularly challenging ELN dataset, however, all methods show limited performance with R^2 below 0.3. This underscores the difficulty in generalizing to complex, real-world reaction data. The poor results can be attributed to two main factors: (1) the ELN dataset includes a broader range of substrates, ligands, and solvents than other datasets, with significantly greater structural variability, making it harder for models to capture meaningful patterns [26]; and (2) limited fine-tuning data, which constrains the model's ability to adapt to the diverse chemical contexts. Performance regarding training size is presented in Appendix G. Full results are available in Appendix H.



(a) Random splits **(b) Out-of-sample splits**

Figure 5: **Performance comparison**. (a) Results under random splits for six datasets and (b) results under out-of-sample splits for six datasets, where we report accuracy (%) for the USPTO_TPL dataset and R^2 (%) for other datasets.

Table 1: Impact of permutation-invariant modeling on prediction uncertainty scores in out-of-sample splits. ^{a} denotes methods that perform permutation data augmentation, and ^{b} denotes permutation-invariant methods. Bold entries highlight the best performance.

Method	BH (reactant-based)			SM (ligand-based)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
YieldBERT	17.65 ± 1.13	23.96 ± 1.17	0.342 ± 0.082	14.98 ± 0.38	20.05 ± 0.39	0.447 ± 0.029
YieldBERT-DA^a	18.41 ± 0.48	26.30 ± 0.33	0.215 ± 0.020	15.78 ± 0.24	19.64 ± 0.26	0.467 ± 0.014
UA-GNN^b	16.95 ± 0.21	24.82 ± 0.63	0.301 ± 0.035	15.59 ± 0.36	20.49 ± 0.39	0.420 ± 0.022
UAM	17.34 ± 0.63	22.61 ± 1.24	0.421 ± 0.072	15.84 ± 0.53	19.61 ± 0.53	0.503 ± 0.042
ReaMVP	17.62 ± 1.04	22.52 ± 1.47	0.432 ± 0.068	13.91 ± 0.29	18.64 ± 0.40	0.516 ± 0.026
ReaDISH^b	$\mathbf{16.29 \pm 0.30}$	$\mathbf{21.76 \pm 0.80}$	$\mathbf{0.480 \pm 0.032}$	$\mathbf{14.22 \pm 0.33}$	$\mathbf{18.05 \pm 0.37}$	$\mathbf{0.549 \pm 0.019}$

Permutation-invariant modeling enhances prediction robustness. A key strength of ReaDISH is its permutation-invariant design, which ensures consistent predictions regardless of the ordering of input molecules or SMILES tokens. To test whether this property enhances robustness in out-of-sample settings, we measure the standard deviation of model predictions across five runs, each using different random molecule orderings and SMILES permutations, the same as YieldBERT-DA [44]. As shown in Table 1, ReaDISH consistently shows the lowest prediction variance. It increases R^2 by 11.11% and 6.40% on the BH and SM out-of-sample splits (average 8.76%), respectively. In contrast, models without permutation invariance show greater variability and reduced performance under input perturbations. We show an example in Figure 6.

These findings highlight the practical benefits of permutation invariance in enhancing model reliability. By modeling interactions between symmetric difference shingles in an order-independent manner, ReaDISH offers more stable and reliable predictions under out-of-sample conditions.

4.3 Ablation study

To evaluate the contributions of individual components within ReaDISH, we conduct a series of ablation studies concerning pair representation, symmetric difference, pre-training strategies, and radius of shingles. Table 2 presents the results for the BH and SM datasets under out-of-sample splits.

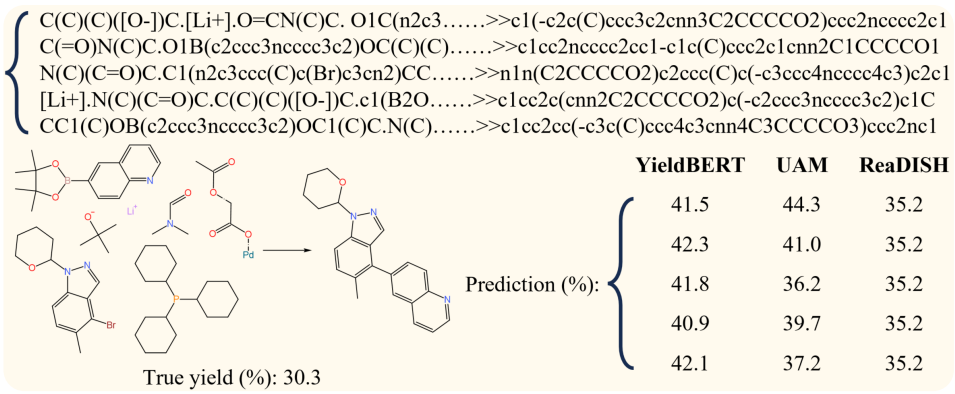


Figure 6: **Permutation influence on predictions**. We sample one reaction from the SM dataset under the out-of-sample split and perform five permutations on the molecule and SMILES token orderings with consistent conformers. While YieldBERT and UAM produce varying predictions with standard deviations of 0.49 and 2.88, respectively, ReaDISH consistently yields an invariant result with the least error.

Table 2: Ablation study results on key components of ReaDISH, evaluated under out-of-sample splits for the BH and SM datasets. Bold entries highlight the best performance.

Method	BH (reactant-based split)			SM (ligand-based split)		
	MAE	RMSE	R^2	MAE	RMSE	R^2
w/o P^0	16.82 ± 0.34	22.24 ± 0.87	0.430 ± 0.038	14.81 ± 0.29	19.24 ± 0.40	0.506 ± 0.026
w/o P_g	16.55 ± 0.44	22.10 ± 0.91	0.457 ± 0.040	14.47 ± 0.40	18.36 ± 0.42	0.535 ± 0.028
w/o P_s	16.41 ± 0.36	21.87 ± 0.78	0.464 ± 0.034	14.30 ± 0.31	18.30 ± 0.41	0.530 ± 0.023
w/o $\triangle$	16.52 ± 0.45	21.95 ± 0.98	0.465 ± 0.039	14.35 ± 0.37	18.21 ± 0.41	0.538 ± 0.017
w/o pre-training	16.40 ± 0.35	21.80 ± 0.85	0.472 ± 0.035	14.26 ± 0.32	18.10 ± 0.43	0.540 ± 0.020
ReaDISH (r=2)	16.50 ± 0.15	22.11 ± 0.25	0.465 ± 0.014	14.41 ± 0.33	18.57 ± 0.44	0.523 ± 0.228
ReaDISH (r=4)	17.91 ± 0.40	23.90 ± 0.01	0.421 ± 0.006	14.88 ± 0.56	18.89 ± 0.82	0.505 ± 0.043
ReaDISH (r=3)	$\mathbf{16.29 \pm 0.30}$	$\mathbf{21.76 \pm 0.80}$	$\mathbf{0.480 \pm 0.032}$	$\mathbf{14.22 \pm 0.33}$	$\mathbf{18.05 \pm 0.37}$	$\mathbf{0.549 \pm 0.019}$

(i) Impact of pair representation. A central feature of ReaDISH is its use of pair representation to capture interactions between molecular shingles. We evaluate three ablated variants: (1) w/o $\mathbf{P}^0$, which entirely removes pairwise modeling from self-attention, (2) w/o $\mathbf{P}_g$, which excludes geometric distance information, and (3) w/o $\mathbf{P}_s$, which omits structural distance features. The results show that removing the full pair representation causes a clear performance drop. Excluding either geometric or structural features also leads to consistent degradation in performance.

(ii) Efficacy of symmetric difference shingles. To test the effectiveness of the symmetric difference encoding strategy, we compare against a variant (w/o $\triangle$) that processes all shingles from both reactants and products without filtering duplicates. This modification leads to a noticeable increase in error, suggesting that without explicitly focusing on molecular transformations, the model is exposed to noise from irrelevant or redundant molecular features, which hampers learning.

(iii) Contribution of pre-training. We evaluate a version trained from scratch (w/o pre-training) to measure the effectiveness of the pseudo-reaction-type pre-training task. Removing this pre-training leads to reduced performance across the board. Though the performance gap is modest compared to other ablations, it highlights that even a simple pre-training task helps the model converge to better representations of the reaction space.

(iv) Impact of shingle radius. We assess how varying the maximum shingle generation radius affects downstream performance. Using a radius of 4 yields the poorest performance, likely because it produces excessive shingles, introducing redundancy and higher complexity. A radius of 2 performs slightly worse than radius 3, indicating that radius 3 offers a good balance between coverage and efficiency.

5 Conclusion

In this work, we present ReaDISH, a novel method for chemical reaction prediction that leverages the symmetric difference of molecular shingles to model chemical interactions. Our method incorporates geometric and structural information through a shingle-level attention mechanism that captures both intra- and inter-molecular interactions. Through comprehensive experiments across multiple reaction prediction tasks, we demonstrate that ReaDISH consistently outperforms baseline models and exhibits improved robustness to input permutations, particularly under out-of-sample scenarios. ReaDISH offers a flexible and effective framework for reaction modeling and paves the way for more interpretable and generalizable machine learning models in computational chemistry. We discuss limitations and impact statements in Appendix I and J, respectively. Future directions include extending ReaDISH to more complex tasks, such as retrosynthesis planning and the incorporation of reaction conditions to enable more accurate predictions.

References

[1] C. W. Coley, R. Barzilay, T. S. Jaakkola, and et al. Prediction of organic reaction outcomes using machine learning. *ACS Central Science*, 3(5):434–443, 2017.

[2] Ian W. Davies. The digitization of organic synthesis. *Nature*, 570:175–181, 2019.

[3] Markus Meuwly. Machine learning for chemical reactions. *Chemical reviews*, 2021.

[4] W. L. Williams, L. Zeng, T. Gensch, and et al. The evolution of data-driven modeling in organic chemistry. *ACS Central Science*, 7(10):1622–1637, 2021.

[5] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1):015016, 2021.

[6] J. P. Reid and M. S. Sigman. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature*, 571 (7765):343–348, 2019.

[7] P. L. Kang and Z. P. Liu. Reaction prediction via atomistic simulation: from quantum mechanics to machine learning. *iScience*, 24(1), 2021.

[8] Philippe Schwaller, Daniel Probst, Alain C. Vaucher, Vishnu H. Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.*, 3:144–152, 2021.

[9] Sina Stocker, Gábor Csányi, Karsten Reuter, and Johannes T Margraf. Machine learning in chemical reaction space. *Nature communications*, 11(1):5505, 2020.

[10] Philippe Schwaller, Alain C Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, and Teodoro Laino. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1604, 2022.

[11] Yifei Yang, Runhan Shi, Zuchao Li, Shu Jiang, Bao-Liang Lu, Qibin Zhao, Yang Yang, and Hai Zhao. Batgpt-chem: A foundation large model for chemical engineering. *Research*, 8:0827, 2025. doi: 10.34133/research.0827.

[12] Xin Hong, Qi Yang, Kuangbiao Liao, Jianfeng Pei, Mao Chen, Fanyang Mo, Hua Lu, Wenbin Zhang, Haisen Zhou, Jiaxiao Chen, Lebin Su, Shuoqing Zhang, Siyuan Liu, Xu Huang, Yizhou Sun, Yuxiang Wang, Zexi Zhang, Zhunzhun Yu, Sanzhong Luo, Xuefeng Fu, and Shuli You. Ai for organic and polymer synthesis. *Science China Chemistry*, pages 1–36, 2024. ISSN 1674-7291. doi: 10.1007/s11426-024-2072-4.

[13] Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018.

[14] Damith Perera, Joseph W Tucker, Shalini Brahmbhatt, Christopher J Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W Sach. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434, 2018.

[15] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[16] Z. Tu and C. W. Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of Chemical Information and Modeling*, 62(15):3503–3513, 2022.

[17] M. Tavakoli, A. Shmakov, F. Ceccarelli, and et al. Rxn hypergraph: A hypergraph attention model for chemical reaction representation. *arXiv preprint arXiv:2201.01196*, 2022.

[18] E. J. Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*, 2017.

[19] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019. doi: 10.1021/acscentsci.9b00576.

[20] X. Wang, C. Y. Hsieh, X. Yin, and et al. Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research*, 6:0231, 2023.

[21] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] Yoni Choukroun and Lior Wolf. Geometric transformer for end-to-end molecule properties prediction. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2895–2901. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/401. Main Track.

[23] Y. Wang, S. Li, T. Wang, and et al. Geometric transformer with interatomic positional encoding. *Advances in Neural Information Processing Systems*, 36:55981–55994, 2023.

[24] G. N. Simm, A. C. Vaucher, and M. Reiher. Exploration of reaction pathways and chemical transformation networks. *The Journal of Physical Chemistry A*, 123(2):385–399, 2018.

[25] Martin Fitzner, Georg Wuitschik, Raffael J Koller, Jean-Michel Adam, and Torsten Schindler. Machine learning C–N couplings: Obstacles for a general-purpose reaction yield prediction. *ACS Omega*, 8:3017 – 3025, 2023.

[26] Wiktor Beker, Rafał Roszak, Agnieszka Wołos, Nicholas H. Angello, Vandana Rathore, Martin D. Burke, and Bartosz A. Grzybowski. Machine learning may sometimes simply capture literature popularity trends: A case study of heterocyclic suzuki–miyaura coupling. *Journal of the American Chemical Society*, 144 (11):4819–4827, 2022. doi: 10.1021/jacs.1c12005. PMID: 35258973.

[27] J. Bradshaw, A. Zhang, B. Mahjour, and et al. Challenging reaction prediction models to generalize to novel chemistry. *ACS Central Science*, 11(4):539–549, 2025.

[28] D. Probst and J. L. Reymond. A probabilistic molecular fingerprint for big data settings. *Journal of Cheminformatics*, 10:1–12, 2018.

[29] A. Capecchi, D. Probst, and J. L. Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12:1–15, 2020.

[30] M. Orsi and J. L. Reymond. One chiral fingerprint to find them all. *Journal of Cheminformatics*, 16(1):53, 2024.

[31] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[33] Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital Discovery*, 2022.

[34] D. Bajusz, A. Rácz, and K. Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:1–13, 2015.

[35] H. L. Morgan. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

[36] M. Shuaibi, A. Kolluru, A. Das, and et al. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021.

[37] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[38] Daniel Mark Lowe. Extraction of chemical structures and reactions from the literature, 2012. PhD thesis, University of Cambridge.

[39] Shu Jiang, Zhuosheng Zhang, Hai Zhao, Jiangtong Li, Yang Yang, Bao-Liang Lu, and Ning Xia. When smiles smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access*, 9:85071–85083, 2021.

[40] Mandana Saebi, Bozhao Nan, John E. Herr, Jessica Wahlers, Zhichun Guo, Andrzej M. Zurański, Thierry Kogej, Per-Ola Norrby, Abigail G. Doyle, Nitesh V. Chawla, and Olaf Wiest. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.*, 14:4997–5005, 2023. doi: 10.1039/D2SC06041H.

[41] Jules Schleinitz, Maxime Langevin, Yanis Smail, Benjamin Wehnert, Laurence Grimaud, and Rodolphe Vuilleumier. Machine learning yield prediction from nicolit, a small-size literature data set of nickel catalyzed c–o couplings. *Journal of the American Chemical Society*, 144(32):14722–14730, 2022.

[42] Andrew F. Zahrt, Jeremy J. Henle, Brennan T. Rose, Yang Wang, William T. Darrow, and Scott E. Denmark. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science*, 363, 2019.

[43] Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin P. Bateman, Louis-Charles Campeau, Jonathan E Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe G. Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim A Cernak, and Spencer D. Dreher. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347:49–53, 2015.

[44] Philippe Schwaller, Alain C. Vaucher, Teodoro Laino, and Jean-Louis Reymond. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *Machine Learning for Molecules Workshop @ NeurIPS 2020*, 11 2020. doi: 10.26434/chemrxiv.13286741.

[45] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, and Seokho Kang. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *Journal of Cheminformatics*, 14(1):1–10, 2022.

[46] Jiayuan Chen, Kehan Guo, Zhen Liu, Olexandr Isayev, and Xiangliang Zhang. Uncertainty-aware yield prediction with multimodal molecular features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8274–8282. AAAI, 2024.

[47] Runhan Shi, Gufeng Yu, Xiao Huo, et al. Prediction of chemical reaction yields with large-scale multi-view pre-training. *Journal of Cheminformatics*, 16(1):22, 2024.

[48] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

[49] Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. A structure-based platform for predicting chemical reactivity. *Chem*, 2019.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[51] Jieyu Lu and Yingkai Zhang. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*, 2022.

[52] C. Raffel, N. Shazeer, A. Roberts, and et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[53] Shuan Chen and Yousung Jung. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4:772–780, 2022.

[54] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST@EMNLP*, 2014.

[55] R. Shi, G. Yu, L. Chen, and et al. Yieldfcp: Enhancing reaction yield prediction via fine-grained cross-modal pre-training. *Artificial Intelligence Chemistry*, 3(1):100085, 2025.

[56] X. Q. Lewell, D. B. Judd, S. P. Watson, and M. M. Hann. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences*, 38:511–522, 1998.

[57] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, and M. Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3:1503–1507, 2008.

[58] Z. Ji, R. Shi, J. Lu, and et al. Relmole: Molecular representation learning based on two-level graph similarities. *Journal of Chemical Information and Modeling*, 62(22):5361–5372, 2022.

[59] Greg Landrum. Rdkit: Open-source cheminformatics, 2024. `https://zenodo.org/records/12782092`.

[60] Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of chemical information and modeling*, 55(12):2562–2574, 2015.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[63] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are clearly presented in the abstract and introduction, and are further elaborated and substantiated throughout the subsequent sections of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Appendix I.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Refer to Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive descriptions for each step of the proposed method, along with all necessary settings and training details to facilitate experiment replication. Additionally, we commit to making the code publicly available by the time of the camera-ready submission.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide comprehensive descriptions for each step of the proposed method, along with all necessary settings and training details to facilitate experiment replication. The source data and code will be made public upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to Section 4.1 and Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and variance of the experiment results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Refer to Appendix J.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of all assets used in the paper are properly credited and cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

A Related work

Reaction performance prediction. Chemical reactions involve complex transformations among multiple molecular entities, such as reactants, catalysts, and reagents. Accurate prediction of reaction outcomes requires expressive and chemically informed representations. Early approaches rely heavily on handcrafted features as molecular fingerprint features derived from domain knowledge, to encode atomic and physicochemical properties [13, 14, 49, 40, 41]. Recent advances in deep learning for reaction prediction can be grouped into three main categories: sequence-based, graph-based, and conformer-based models. Sequence-based models represent chemical reactions using linear notations such as SMILES and apply neural sequence architectures to learn patterns in tokenized strings. Notable examples include YieldBERT [5] and YieldBERT-DA [44], which leverage BERT-based [50] encoders; MolFormer [19], which adopts a transformer [32] architecture; and T5Chem [51], which builds on the T5 language model [52]. Graph-based methods encode molecular structures as graphs and extract meaningful representations using graph neural networks (GNNs). Rxn Hypergraph [17] employs hypergraph attention to capture molecular interactions. UA-GNN [45] sums reactant embeddings via GNNs and concatenates them with product representations. UAM [46] combines multi-view inputs, including graphs, SMILES, and molecular fingerprints, and aggregates their embeddings. LocalTransform [53] focuses on local structural changes by encoding atom-level differences between reactants and products. Conformer-based models incorporate 3D atomic coordinates to learn geometry-aware representations, which can better capture stereoelectronic factors influencing reactivity. ReaMVP [47] combines Bi-GRU networks [54] with multi-view pre-training on conformers and SMILES. YieldFCP [55] introduces a cross-modal projector to align conformer and SMILES representations for yield prediction. Despite their respective advantages, these methods often overlook the rich intra- and inter-molecular interactions that influence reaction outcomes. In this work, we address this gap by introducing a geometry- and structure-enhanced modeling approach that explicitly incorporates such interactions into the representation learning process.

Molecular substructure learning. Substructures (functional groups, motifs, and fragments) play a pivotal role in determining molecular properties and, by extension, reaction outcomes. Several fragment-based approaches have been proposed to decompose molecules into chemically meaningful parts. RECAP [56] introduces predefined cleavage rules to generate fragments at drug-like bond types, while BRICS [57] focuses on retrosynthetically relevant splits. ReLMole [58] further generalizes this process using graph-based heuristics to automatically extract relevant substructures. In parallel, fingerprinting methods such as ECFP [31] iteratively encode local atomic environments, capturing circular subgraphs of increasing radii. These subgraphs, referred to as molecular shingles [28–30], serve as a compact and effective representation of chemical structures. DRFP [33] extracts shingles and applies a symmetric difference operation on reactant and product shingle sets to obtain reaction fingerprints. Unlike DRFP, our method directly models the interaction-aware symmetric difference via a transformer, allowing it to capture fine-grained transformations at the shingle level.

B Datasets statistics

We use two datasets for pre-training and seven datasets for downstream evaluation, as summarized in Table 3. We remove duplicate records and invalid reactions for pre-training by RDKit [59]. To assess the generalizability of our approach, we consider both random and out-of-sample splits. In the out-of-sample split, the test set contains reactions involving molecules that do not appear in the training set. This setup mimics real-world scenarios, where unseen molecular structures are encountered during reaction property prediction. To strike a balance between accuracy in spatial coordinates and computational efficiency, we utilize the ETKDG algorithm [60] to generate up to 100 conformers and sample one for each molecule.

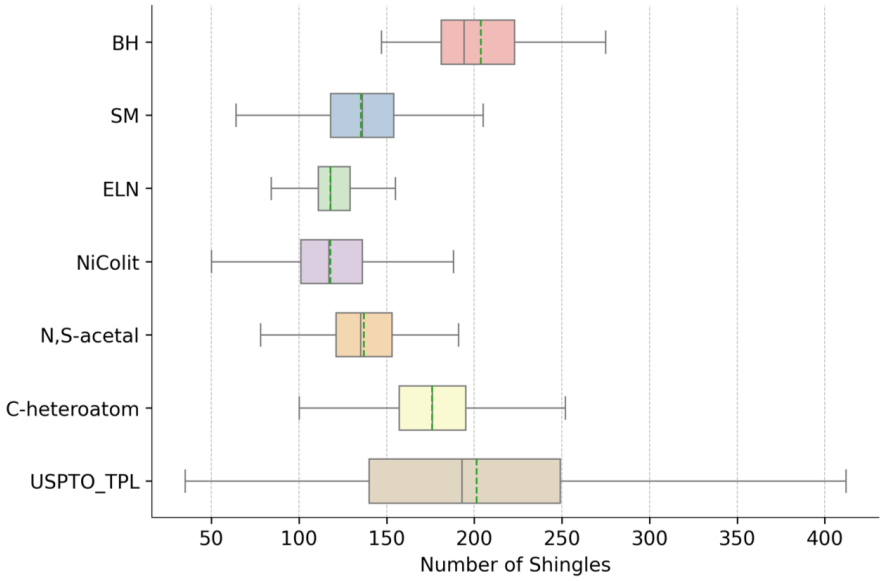
- **The United States Patent and Trademark Office (USPTO) dataset** [38] was collected from 1976 to September 2016, containing over 1.8 million chemical reactions stored in the form of SMILES arbitrary target specification (SMARTS).

Table 3: The statistics of pre-training datasets (first row) and evaluation datasets (remaining rows).

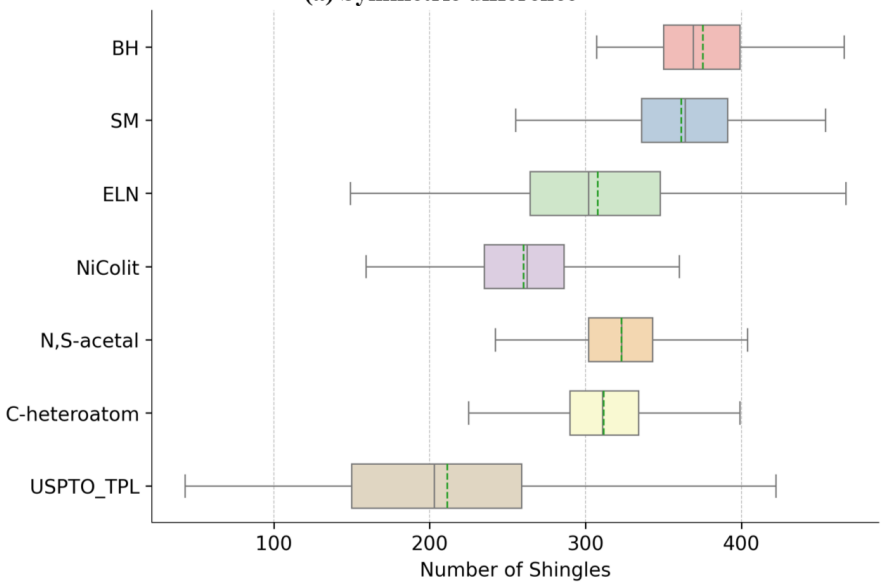
Dataset	No. reactions	Split type	Out-of-sample	No. training	No. test
USPTO [38] & CJHIF [39]	3,728,503	stratified	✗	3,542,077	186,426
BH [13]	3,955	random	✗	2,768	1,187
	3,955	reactant-based	✓	2,372	1,583
SM [14]	5,760	random	✗	4,032	1,728
	5,760	ligand-based	✓	4,320	1,440
NiCOlit [41]	1,406	random	✗	1,124	282
	1,406	substrate-based	✓	1,012	394
ELN [40]	750	random	✓	525	225
N,S-acetal [42]	1,075	random	✗	600	475
	688	catalyst-based	✓	384	304
C-heteroatom [43]	1,536	random	✗	1,075	461
	1,536	catalyst-based	✓	1,152	384
USPTO_TPL [8]	445,115	random	✗	400,604	44,511

- **The Chemical Journals with High Impact Factor (CJHIF) dataset** [39] included over 3.2 million chemical reactions in the form of SMARTS extracted from chemistry journals by Chemical.AI.
- **The Buchwald-Hartwig (BH) dataset** [13] contained 3,955 reactions from high-throughput experiments (HTEs) with 1,536-well plates on the class of Pd-catalyzed Buchwald-Hartwig C-N cross-coupling reactions. We choose the pyridyl reactants as the pivot to construct the out-of-sample split condition, where training includes nine pyridyl reactants and testing uses three non-pyridyl reactants, as shown in Figure 8(a).
- **The Suzuki-Miyaura (SM) dataset** [14] was constructed from high-throughput experiments on the class of Suzuki-Miyaura cross-coupling reactions, resulting in measured yields for a total of 5,760 reactions. We choose a set of ligands as the pivot to construct the out-of-sample split condition, where nine ligands (including "None") are used for training and three for testing, as shown in Figure 8(b).
- **The Ni-catalyzed C-O bond activation (NiCOlit) dataset** [41] was extracted from organic reaction publications to form C-C and C-N bonds, containing 1,406 reactions. We choose the OPiv substrates as the pivot to construct the out-of-sample split condition, where 247 OPiv substrates are used for training and 42 non-OPiv substrates for testing, as shown in Figure 8(c).
- **The real-world electronic laboratory notebook (ELN) dataset** [40] was created for Buchwald-Hartwig reactions from electronic laboratory notebooks, including 750 reactions. The structural diversity of the real-world ELN dataset is much higher than that of the HTE datasets. The random split also simulates an out-of-sample scenario.
- **The asymmetric N,S-acetal formation using CPA catalysts (N,S-acetal) dataset** [42] included combinatorial variations of CPA catalysts, N-acyl imines, and thiols, resulting in a total of 1,075 reactions. We choose a set of catalysts (test-cat) as the pivot to construct the out-of-sample split condition, including 24 catalysts for training and 19 for testing, as shown in Figure 8(d).
- **The nanomole-scale reactivity evaluation of C-heteroatom-coupling reactions (C-heteroatom) dataset** [43] was performed by an automated high-throughput screening on a nanomole scale, yielding 1,536 reactions. We choose a set of catalysts (test-cat) as the pivot to construct the out-of-sample split condition, where the split uses 12 catalysts for training and 4 for testing, as shown in Figure 8(e).
- **USPTO_TPL dataset** [8] labels were generated by extracting the 1,000 most common templates from the USPTO dataset, containing 445,115 reactions.

Figure 7 presents distributions of the number of symmetric difference and union shingles per reaction for each dataset.



(a) Symmetric difference



(b) Union

Figure 7: **Distributions of the number of shingles per reaction**. (a) The symmetric difference of shingles removes many duplicate shingles in the intersection. (b) The union of shingles from reactants and products.

C Shingle generation

In practice, we also treat ring structures within molecules as shingles, as they capture rich and informative substructures. We set the max radius to 3. To balance expressiveness and computational efficiency, we consider up to 280 shingles per reaction and 100 shingles per molecule, allowing each unique shingle to appear up to 10 times. The per-reaction bound should be adapted to the downstream dataset at hand. The per-molecule bound prevents rare cases where one molecule generates an excessive number of shingles and dominates the representation. The generation of the symmetric difference shingle set is depicted in Algorithm 1.

Note that for the USPTO_TPL dataset, we generate shingles solely from reactants. This choice is motivated by the observation that the symmetric difference of substructures between reactants and

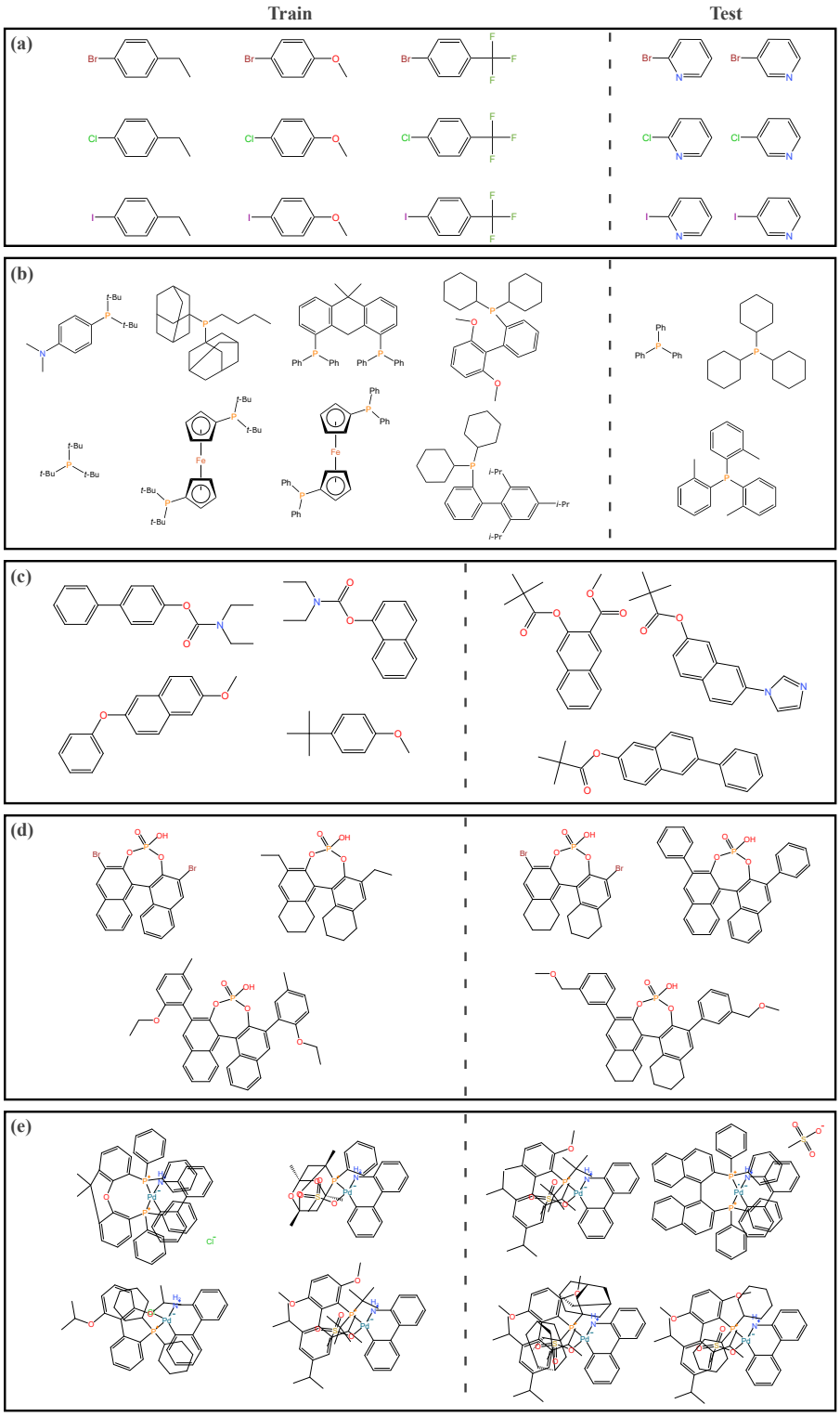


Figure 8: **Example molecules used in out-of-sample splits**. (a) Reactant-based for the BH dataset. (b) Ligand-based for the SM dataset. (c) Substrate-based for the NiCOlit dataset. (d) Catalyst-based for the N,S-acetal dataset. (e) Catalyst-based for the C-heteroatom dataset.

products is often minimal. In practice, we can apply the symmetric difference, reactants-only, and union strategies, and select the one that yields the best performance.

Algorithm 1 Pipeline to extract symmetric difference shingles

Input: the reaction $\mathcal{R} = \{\mathcal{M}^{\mathrm{r}}; \mathcal{M}^{\mathrm{p}}\}$ divided into reactants and products, and the maximum radius $r_{\max}$ for shingles
Output: symmetric difference shingle set $S_{\mathcal{R}}$ between reactants and products

$S_{\text{react}} \leftarrow \{\}$ ▷ initialize reactants shingle set
$S_{\text{prod}} \leftarrow \{\}$ ▷ initialize products shingle set
for m in $\mathcal{M}^{\mathrm{r}}$ **do** ▷ Enumerate molecules and atoms
  for v in m **do**
    for $r \leftarrow 1$ to $r_{\max}$ **do**
      calculate and add shingles $S^{(r)}(v)$ to S_{react} ▷ Generate shingles for reactants
    end for
  end for
end for
for m in $\mathcal{M}^{\mathrm{p}}$ **do** ▷ Enumerate molecules and atoms
  for v in m **do**
    for $r \leftarrow 1$ to $r_{\max}$ **do**
      calculate and add shingles $S^{(r)}(v)$ to S_{prod} ▷ Generate shingles for products
    end for
  end for
end for
$S_{\mathcal{R}} = (S_{\text{react}} \setminus S_{\text{prod}}) \cup (S_{\text{prod}} \setminus S_{\text{react}})$ ▷ Compute symmetric difference
return $S_{\mathcal{R}}$

D Pre-training settings

We introduce three pseudo-reaction-type classification tasks as pre-training objectives. The underlying intuition is that similar chemical structures tend to exhibit consistent semantic behavior across various reactions. Specifically, we employ the DRFP [33] with default parameters, which is a 1024-length one-hot descriptor as the reaction fingerprint. These reaction fingerprint sequences can be used for clustering, where shorter distances between fingerprints indicate a higher likelihood of belonging to the same cluster. We apply K means clustering by scikit-learn [61] with different values of K (100, 1,000, 4,000, see Figure 9 concerning selection of K) to cluster reactions to obtain clusters with different granularities from coarse-grained to fine-grained. Subsequently, we use three classification heads with output sizes of 100, 1,000, and 4,000, respectively, to predict the pseudo labels associated with each cluster. This approach leverages the structural consistency of chemical reactions, enabling the model to learn more robust and transferable representations.

E Comparison with standard cross-attention architecture

We implement a cross-attention baseline that explicitly models interactions between reactants and products. In this architecture, product embeddings serve as queries, while reactant embeddings are used as keys and values within a standard multi-head cross-attention module [32]. The model consists of 4 cross-attention layers, 64 attention heads, and a hidden dimension of 512, and it shares the same molecular encoders as ReaDISH.

We evaluate this architecture for six datasets to compare it with ReaDISH. As shown in Table 4, the cross-attention baseline consistently underperforms our proposed model. This result suggests that simple cross-attention treats inter-molecular interactions uniformly, lacking chemical or geometric guidance. In contrast, ReaDISH explicitly incorporates both inter- and intra-molecular relationships through structural distances, edge types, and symmetric-difference shingles, resulting in more chemically grounded representations.

Table 4: Comparison between ReaDISH and the cross-attention model for six datasets.

Model	BH	SM	NiCOlit	N,S-acetal	C-heteroatom	ELN (OOS)
Cross-attention	0.946 ± 0.005	0.858 ± 0.009	0.431 ± 0.007	0.844 ± 0.005	0.705 ± 0.076	0.156 ± 0.006
ReaDISH	0.976 ± 0.001	0.886 ± 0.008	0.539 ± 0.024	0.916 ± 0.007	0.763 ± 0.026	0.230 ± 0.047

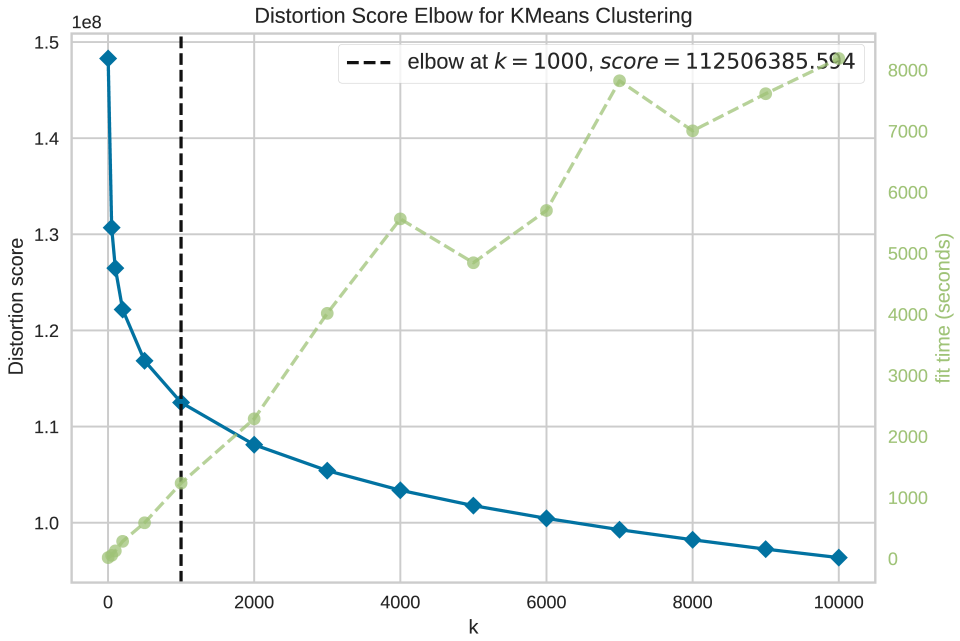


Figure 9: **The elbow for the K means algorithm**. We use the distortion score to find the elbow and choose $K = 100$, 1,000, 4,000 for reaction clustering.

F Implementation details

We extract the results of the baseline models from the existing literature as much as possible, and otherwise reproduce them based on publicly available code. Note that UAM applies contrastive pre-training on the whole dataset in downstream tasks, which may introduce data leakage. Hence, we randomly initialize the model weights to fine-tune the model during reproduction. All methods are tested on (1) the same ten random splits and (2) the same out-of-sample split across five random runs to ensure fair comparisons, with the average results reported.

Table 5 and 6 present hyper-parameters used in ReaDISH during pre-training and fine-tuning, respectively. We use Pytorch [62] with the Adam [63] optimizer and the cosine learning rate decay strategy for training. All experiments are executed on 4 NVIDIA RTX3090 GPUs. Pre-training the model takes about 8 hours, whereas fine-tuning on downstream datasets requires up to 1 hour. Generating shingles for the pre-training dataset takes around 7 hours using 50 CPU threads, while processing a downstream dataset completes in under 4 minutes. ReaDISH contains about 16.3M trainable parameters.

Table 5: Parameters during pre-training.

Parameters	Value	Parameters	Value
Number of encoder layers	4	Number of encoder attention heads	64
Encoder FFN embedding dimension	2048	Encoder embedding dimension	512
Batch size	64	Initial learning rate	5e-5
Max epochs	3	Minimum learning rate	5e-6
Warmup steps	2000	Warmup learning rate	1e-6
Dropout rate	0.1	Number of Gaussian kernels	128

G Dependence of the prediction performance on the size of the training data

We scale training size on BH and SM benchmarks to measure performance changes, using subsets of 5%, 10%, 20%, 30%, 50%, and 70% while testing on the full 30% test split. Results in Table 7 and

Table 6: Search space of parameters during fine-tuning.

Parameters	Search space
Max epochs	150, 200
Batch size	64, 128
Initial learning rate	5e-3, 1e-3, 5e-4
Minimum learning rate	5e-4, 1e-4, 5e-5

Figure 10 show that predictive accuracy (R^2) steadily improves without saturation, suggesting further gains with larger datasets. But the rate of increase slows down.

Table 7: Performance regarding training size on the BH and SM datasets.

Dataset	5%	10%	20%	30%	50%	70%
BH	0.713 ± 0.017	0.808 ± 0.017	0.874 ± 0.016	0.919 ± 0.006	0.955 ± 0.002	0.976 ± 0.001
SM	0.508 ± 0.018	0.649 ± 0.008	0.751 ± 0.010	0.810 ± 0.001	0.851 ± 0.010	0.886 ± 0.008

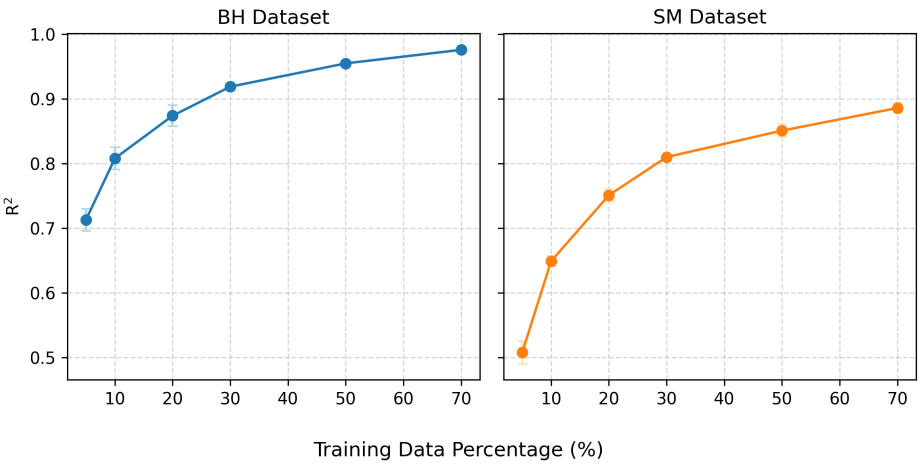


Figure 10: **The learning curve regarding training size** for the BH (left) and SM (right) datasets.

H Additional experimental results

The full experimental results are detailed in Table 8 and Table 9.

I Limitations

While ReaDISH effectively models chemical reactions through the symmetric difference of molecular shingles, one limitation is its computational overhead. In cases where all substructures are considered, the number of shingles can exceed the number of atoms, leading to increased memory consumption and longer training times. We mitigate this challenge by setting an upper bound on the number of shingles per reaction, balancing expressiveness with efficiency. Future work could explore adaptive pruning strategies to further reduce resource demands without compromising predictive performance.

J Impact statements

Advances in chemical reaction prediction hold the potential to accelerate scientific discovery in areas such as drug development, materials science, and sustainable chemistry. However, enhanced

Table 8: Results under random splits. Bold entries highlight the best performance.

Method	BH			SM			NiCOlit		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
YieldBERT	3.99 ± 0.15	6.01 ± 0.27	0.951 ± 0.005	8.13 ± 0.34	12.07 ± 0.46	0.815 ± 0.013	19.96 ± 1.95	27.23 ± 1.38	0.353 ± 0.070
YieldBERT-DA	3.09 ± 0.12	4.80 ± 0.26	0.969 ± 0.004	6.60 ± 0.27	10.52 ± 0.48	0.859 ± 0.012	19.53 ± 1.82	26.53 ± 1.28	0.373 ± 0.060
UA-GNN	2.92 ± 0.06	4.43 ± 0.09	0.974 ± 0.001	6.12 ± 0.22	9.47 ± 0.46	0.886 ± 0.010	16.19 ± 0.26	22.23 ± 0.43	$\mathbf{0.550 \pm 0.017}$
UAM	$\mathbf{2.89 \pm 0.06}$	$\mathbf{4.36 \pm 0.10}$	$\mathbf{0.976 \pm 0.001}$	$\mathbf{6.04 \pm 0.18}$	$\mathbf{9.23 \pm 0.40}$	$\mathbf{0.888 \pm 0.009}$	16.51 ± 0.87	22.91 ± 1.12	0.515 ± 0.053
ReaMVP	3.11 ± 0.07	4.63 ± 0.14	0.971 ± 0.002	6.59 ± 0.20	10.37 ± 0.42	0.864 ± 0.010	16.21 ± 0.87	22.61 ± 1.12	0.541 ± 0.053
ReaDISH	2.98 ± 0.05	$\mathbf{4.36 \pm 0.09}$	$\mathbf{0.976 \pm 0.001}$	6.09 ± 0.18	9.55 ± 0.39	0.886 ± 0.008	$\mathbf{15.36 \pm 0.28}$	$\mathbf{21.89 \pm 0.57}$	0.539 ± 0.024
Method	N,S-acetal			C-heteroatom			USPTO_TPL		
	MAE	RMSE	R^2	MAE	RMSE	R^2	ACC	MCC	CEN
YieldBERT	0.16 ± 0.01	0.23 ± 0.02	0.861 ± 0.015	1.23 ± 0.14	2.61 ± 0.11	0.664 ± 0.057	0.989	0.989	0.006
YieldBERT-DA	0.16 ± 0.01	0.22 ± 0.01	0.889 ± 0.017	1.03 ± 0.12	2.58 ± 0.06	0.709 ± 0.043	0.990	0.990	0.006
UA-GNN	0.15 ± 0.00	0.21 ± 0.01	0.905 ± 0.007	0.90 ± 0.08	2.30 ± 0.17	0.751 ± 0.038	0.991	0.991	**0.005**
UAM	0.19 ± 0.01	0.26 ± 0.02	0.859 ± 0.012	1.04 ± 0.11	2.65 ± 0.10	0.718 ± 0.047	0.954	0.954	0.250
ReaMVP	0.14 ± 0.01	0.21 ± 0.01	0.904 ± 0.012	0.90 ± 0.09	2.57 ± 0.11	0.726 ± 0.043	0.990	0.990	0.006
ReaDISH	$\mathbf{0.14 \pm 0.00}$	$\mathbf{0.20 \pm 0.01}$	$\mathbf{0.916 \pm 0.007}$	$\mathbf{0.87 \pm 0.04}$	$\mathbf{1.94 \pm 0.10}$	$\mathbf{0.763 \pm 0.026}$	**0.992**	**0.992**	**0.005**

Table 9: Results under out-of-sample splits. Bold entries highlight the best performance.

Method	BH (reactant-based)			SM (ligand-based)			NiCOlit (substrate-based)		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
YieldBERT	17.44 ± 1.01	23.90 ± 1.13	0.350 ± 0.060	14.85 ± 0.36	19.59 ± 0.39	0.469 ± 0.021	15.53	21.91	0.463
YieldBERT-DA	18.41 ± 0.48	26.30 ± 0.33	0.215 ± 0.020	15.78 ± 0.24	19.64 ± 0.26	0.467 ± 0.014	15.05	**21.21**	0.483
UA-GNN	16.95 ± 0.21	24.82 ± 0.63	0.301 ± 0.035	15.59 ± 0.36	20.48 ± 0.39	0.420 ± 0.022	**14.82**	22.19	0.563
UAM	16.99 ± 0.55	21.78 ± 1.13	0.458 ± 0.051	14.46 ± 0.37	18.63 ± 0.41	0.520 ± 0.029	17.64	22.99	0.446
ReaMVP	17.17 ± 0.83	$\mathbf{21.40 \pm 1.15}$	0.479 ± 0.056	$\mathbf{13.90 \pm 0.29}$	18.36 ± 0.35	0.534 ± 0.018	15.01	21.43	0.489
ReaDISH	$\mathbf{16.29 \pm 0.30}$	21.76 ± 0.80	$\mathbf{0.480 \pm 0.032}$	14.22 ± 0.33	$\mathbf{18.05 \pm 0.37}$	$\mathbf{0.549 \pm 0.019}$	15.09	21.91	**0.574**
Method	ELN			N,S-acetal (catalyst-based)			C-heteroatom (catalyst-based)		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
YieldBERT	21.98 ± 2.27	27.67 ± 2.14	0.151 ± 0.102	0.258	0.361	0.735	2.11	3.70	0.309
YieldBERT-DA	21.58 ± 2.19	26.97 ± 1.98	0.171 ± 0.112	0.239	0.340	0.764	1.96	3.58	0.351
UA-GNN	20.64 ± 1.13	26.50 ± 1.03	0.203 ± 0.054	0.239	0.329	0.780	1.77	3.46	0.394
UAM	22.10 ± 1.35	26.65 ± 1.78	0.179 ± 0.050	0.305	0.429	0.630	2.01	3.62	0.346
ReaMVP	20.69 ± 1.33	26.36 ± 1.29	0.212 ± 0.057	0.234	0.342	0.762	2.04	3.54	0.361
ReaDISH	20.82 ± 1.01	$\mathbf{25.99 \pm 1.12}$	$\mathbf{0.230 \pm 0.047}$	**0.226**	**0.327**	**0.790**	2.02	**3.42**	**0.409**

predictive models could be misused to facilitate the design of harmful substances, including toxic chemicals or controlled compounds. This raises ethical and security concerns regarding dual-use applications. Additionally, widespread adoption of automated chemical design tools may shift traditional roles in chemical research, with potential implications for education and workforce development.