Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Exploring Molecular Pretraining and Mechanism-Aware Modeling for Reaction Yield Prediction

Yongkyu Lee[ab], Won Bo Lee[b], Lauren Ye Seol Lee[a]*

[a] Department of Chemical Engineering, University College London, Torrington Place, London, WC1E 7JE, United Kingdom Torrington Place, London, WC1E
[b] Seoul National University, Department of Chemical and Biological Engineering, Gwanak-ro 1, Gwanak-gu, Seoul 08826, Republic of Korea
* Corresponding Author: lauren.lee@ucl.ac.uk.

ABSTRACT

Accurately predicting chemical reaction yields can accelerate reaction optimization by prioritizing promising conditions; however, the complexity of multicomponent reactions and the limited availability of high-quality datasets remain significant challenges. While machine learning has achieved substantial progress in molecular property prediction, reaction-level modeling requires representations that capture three-dimensional structures, intercomponent interactions, and mechanistically critical features. In this study, we investigate the effective extension of molecular pretraining to reaction yield prediction from two complementary perspectives. First, we apply a reaction-aware architectural design based on molecular representation models pretrained via partial denoising to multicomponent reactions. Each reaction component is encoded as a three-dimensional stereoisomer embedding and concatenated into a reaction-level representation, with multi-head attention modeling intercomponent dependencies. We show that this approach achieves robust performance and generalizes well to reactions involving previously unseen components. Second, we design an architecture that extracts intermolecular relational embeddings through mechanistically meaningful pre-supervised learning using electron source-sink annotations derived from reaction mechanisms, followed by an active fine-tuning process. This mechanism-aware extension leverages large-scale reaction data to provide informative inductive bias and improves performance in data-constrained settings. We further confirm that this approach outperforms models that simply combine individual molecular embeddings, as it directly captures chemically meaningful information relevant to reaction outcomes. These results suggest promising directions for developing machine-learning models for reaction yield prediction that are more closely aligned with underlying chemical mechanisms.

Keywords: Deep learning, Reaction yield, Transfer learning

1. INTRODUCTION

Chemical reactions play a central role in diverse fields, including the synthesis of new compounds, process optimization, and materials development. Accurate prediction of chemical reaction outcomes can significantly reduce experimental costs and enable more efficient optimization strategies. However, reaction outcome prediction remains highly challenging due to the complex interplay of numerous factors, such as reaction conditions, catalysts, and the structural properties of reactants. Consequently, predicting reaction outcomes a priori, without performing physical experiments, remains an open and difficult problem.

Meanwhile, recent advances in deep learning have demonstrated strong performance in single-molecule property prediction tasks [1, 2], and the rapid growth of large-scale chemical reaction databases [3-5] has highlighted the potential of data-driven approaches to advance reaction modeling. Inspired by natural language processing, the BERT architecture has been applied to reaction yield prediction through large-scale pretraining on reaction corpora such as the Pistachio dataset[4]. The

resulting YieldBERT [6] model has shown superior performance compared to existing data-driven methods and has been widely adopted as a benchmark in subsequent studies [7]. However, as molecules are represented as linearized SMILES strings, key structural characteristics—such as explicit atomic connectivity and three-dimensional conformational information—are not directly encoded, which may limit the ability of such models to fully capture the chemical features governing reaction.

At the single-molecule level, numerous graph-based approaches have been developed to better capture intrinsic molecular characteristics [2, 8]. By representing molecules as graphs, these methods naturally encode atomic connectivity and, in some cases, incorporate three-dimensional coordinates into molecular embeddings. Pretraining strategies based on denoising techniques [2] have demonstrated the ability to learn physically meaningful representations beyond simple masking-based objectives. At the same time, reaction modeling inherently involves multiple interacting molecular species, and existing studies often address this challenge by aggregating or concatenating embeddings from individual molecules; however, it remains unclear whether such straightforward combinations are sufficient to fully capture complex inter-molecular interactions.

Accordingly, in this work we investigate reaction yield prediction using approaches specifically designed to address two essential aspects: intrinsic molecular representations that capture graph structure and three-dimensional conformational information, and data representations and architectures that explicitly model intermolecular interactions. First, we adapt and extend a denoising-based molecular property prediction model to reaction-level prediction, focusing on capturing intrinsic structural characteristics of individual molecules. This approach demonstrates stable generalization in scenarios involving unseen products. Second, from a complementary perspective, we incorporate mechanistically relevant supervision by leveraging electron source-sink annotations derived from reaction mechanisms. Rather than treating interactions implicitly, the model is pretrained to identify atomic sites that play decisive roles in electron flow, and this mechanistic information is subsequently integrated into reaction-level embeddings during fine-tuning. Taken together, these results suggest that explicitly reflecting fundamental mechanistic aspects of chemical reactions, rather than relying solely on molecular aggregation or architectural complexity, provides a promising direction for developing more chemically informed machine-learning models for reaction yield prediction.

2. METHODS

2.1 Structure-Aware Molecular Representation for Reaction Yield Prediction

2.1.1 Overview

To incorporate three-dimensional molecular conformations and chemically meaningful representations for reaction, without performing expensive geometry optimization, we build upon the fractional denoising molecular pretraining framework, denoted as Frad, introduced by Ni *et al* [2]. Frad is a denoising-based pretraining approach based on TorchMD-Net [9] that learns structure-aware molecular representations by reconstructing molecular conformation perturbed from their equilibrium geometries. A key feature of Frad is a chemically aware noise (CAN) designed to capture physically relevant perturbations, including rotations and vibrational modes. This enables the model to encode conformational variability beyond small harmonic distortion and has been shown to produce molecular embeddings that improve molecular property prediction and remain robust to the quality of input geometry, including those generated using inexpensive classical methods. In this work, the pretrained Frad encoder is extended from single molecule property prediction to reaction-level modeling and reaction yield prediction on benchmark reaction datasets. Each reaction is represented as a set of molecular components, and their embeddings are aggregated to obtain a reaction-level representation, while treating the relationship between reaction components implicitly.

2.1.2 Input Construction and Model Architecture

Reactions are represented using three-dimensional structure-based encoding produced by the pretrained Frad molecular encoder, in which each molecular component is treated as an independent molecular graph. To accommodate reactions with a variable number of components, a dummy molecular graph with atomic number zero and fixed coordinates at (0, 0, 0) is introduced to represent missing species. This allows all reactions to be expressed in a consistent input format. Each component-specific graph is then processed independently by the pretrained encoder to generate a fixed-dimensional molecular embeddings. To adapt the Frad designed only for single-molecule for reaction-level modeling, the component embeddings are concatenated in a fixed order and pass through a multi-head attention module. This attention mechanism serves as a learnable aggregation function that adaptatively capture correlations among reaction constituents. The resulting interaction-aware representation is subsequently passed through two multi-layer perceptron (MLP) layers to produce a scalar prediction of the reaction yield (see Figure 1).

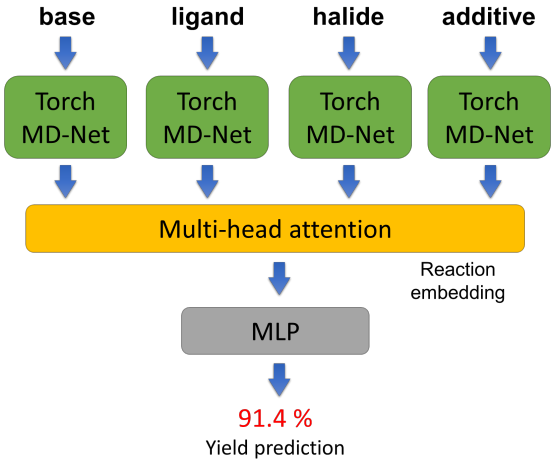

Figure 1. Reaction yield prediction using structure-aware molecular representations (SA-YIELD).

2.1.3 Datasets and Evaluation Protocol

Model performance for reaction yield prediction in this section is evaluated on the Buchwald-Hartwig coupling dataset [3], which was generated through high-throughput experimentation (HTE). Each reaction is defined by its constituent molecular components and represented using SMILES strings. In the HTE study, reaction conditions, such as concentration and temperature, are fixed across the dataset and are therefore not included as model inputs. The reactions involve four molecular components: aryl halide, ligand, base, and additive.

Reactions with zero yield frequently do not produce an isolable product, and including the product information as model input may introduce information leakage. To avoid this, each reaction instance is defined using only the four reactant-side components—aryl halide, ligand, base, and additive—while excluding the product from the input representation. This formulation ensures that yield prediction is based solely on information available prior to reaction execution and aligns with realistic experimental design constraints.

Model performance is evaluated under different generalization settings: random and component-wise unseen splits. For random split setting, reactions are randomly divided into training, validation, and test sets with a 6:2:2 ratio. To assess extrapolation to unseen molecular components, component-wise hold-out splits are applied. For bases and ligands, which consist of three and four distinct species, respectively, splits are generated by holding out one component at a time for testing, while using the remaining two bases and three ligands for training. For aryl halides and additives, which exhibit greater chemical diversity, unseen splits are generated by randomly selecting one quarter of the component combinations as the test set. Performance metrics reported for unseen component prediction are obtained by averaging results across all corresponding splits.

2.2 Interaction-Aware Pretraining Approach via Supergraph Representation

2.2.1 Overview

Many reaction prediction models form reaction representations by combining features derived from individual components, leaving cross-component relationship to be inferred implicitly from yield supervision. While the structure-aware formulation in Section 2.1 enriches molecular embeddings with conformation information, it similarly does not incorporate explicit mechanism derived relation across reaction components. In many cases, reaction outcomes depend on how specific atomic sites across different molecules participate in bond formation and electron transfer. Even for the same molecular species, the atoms involved in bond formation or braking can vary depending on the co-reactants, which limits what can be inferred from representations learned for each molecule in isolation.

To address this limitation, we introduce a mechanism-derived reaction modelling framework that incorporate explicit relational structure and supervision. First, reactions are represented using a pairwise supergraph formulation in which molecular components are decomposed into all combinatorial pairs. Second, electron source-sink annotations derived from reaction mechanisms are used as a pretraining objective to introduce mechanism-inspired relational bias into the learned representation. The pretrained model is then fine-tuned on a smaller reaction yield dataset, during which the predicted source-sink probabilities are incorporated to adaptively modulate how atomic information is aggregated into supergraph embedding, rather than relying on uniform pooling.

2.2.2 Supergraph Representation for interaction awareness

To encode intermolecular relationship explicitly, reactions are represented using a hierarchical, relation-centric supergraph formulation rather than as an aggregation of individual molecular embeddings. The model architecture is adapted from the HiMol [11] hierarchical graph framework, which augments atom-level graph with higher-order (e.g., fragment- or molecular-level) structural nodes to enable information exchange across multiple scales via deep message-passing neural network (D-MPNN)[12].

In the work, the hierarchical representation is extended from single molecules to reactions by constructing pairwise molecular supergraph. For a reaction involving multiple molecular components, all combinatorial molecular pairs are enumerated, and each pair is encoded as an independent supergraph. Within each supergraph, atom-level nodes capture local chemical structure, while higher-level nodes aggregate information at molecular-

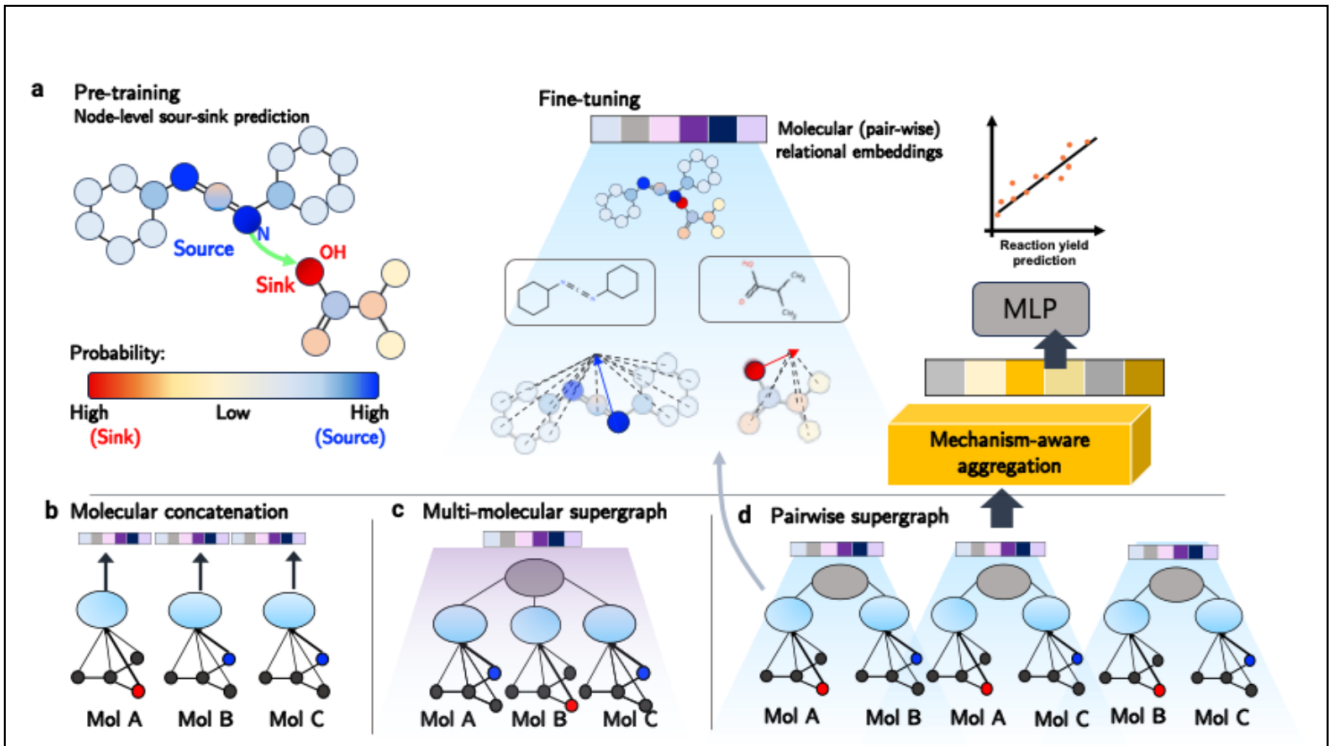

Figure 2. Mechanism-aware relational modeling for reaction yield prediction. (a) Source-sink pretraining and fine-tuning: predicted atom-level source-sink probabilities are used to weight atomic contributions during pooling when constructing pairwise relational embeddings for yield prediction. (b) Molecular concatenation baseline. (c) Multi-molecular supergraph. (d) Pairwise (combinatorial) supergraph formulation with mechanism-aware aggregation.

and reaction-levels. Message passing is then performed across all hierarchical levels of nodes, enabling information flow among atom-, molecular-, and reaction-level representations.

This pairwise construction is motivated by the observation that, even in multicomponent reactions, mechanistically relevant relationships are more naturally and effectively represented as pairwise interactions between specific components rather than as uniformly distributed multimolecular correlations. Key mechanistic process, such as electron flow and bond rearrangement, typically arises from interactions between two reacting species at a time, rather than from simultaneous many-body effects. Representing reactions using a pairwise supergraphs therefore provides a closer abstraction of the underlying reaction mechanism and aligns direclty with the pretraining objective, which is designed to capture electron source-sink relationships defined between interacting molecular pairs.

For reaction-level, embeddings extracted from all molecular pairs are integrated via a masked attention pooling. This aggregation yields a fixed-dimensional reaction representation and allows different molecular pairs to be weighted adaptively. A schematic overview of the pairwise supergraph representation and its integration into a reaction-level embedding is shown in Figure 2.

2.2.3 Source–Sink–Aware Pretraining and Active Finetuning for Reaction Yield Prediction

Reaction yield datasets are often limited in size, which constrains the ability of supervised models to learn complex intermolecular relationships reliably from yield labels alone. To mitigate this limitation, a transfer learning strategy is employed in which the pairwise supergraph encoder is first trained on a source-sink prediction task to introduce chemically motivated inductive bias, and is subsequently transferred to downstream reaction yield prediction.

Pretraining is performed using the Mech-USPTO-31k dataset [10], which provides reaction mechanisms annotated in an arrow-pushing format and include atomic- or bond-level labels identifying electron sources and sinks based on electron pair transfer. These annotations provide atom-level supervision that depends on the joint context of reacting molecular pairs and are therefore well aligned with the pairwise supergraph representation. Source-sink prediction is formulated as a node-level classification tasks on pairwise supergraphs, encouraging the model to assign higher likelihood to atomic sites associated with electron transfer in a reaction context.

Following pretraining, the model is fine-tuned on a liquid-phase amide coupling dataset reported by Zhang *et al* [13]. This dataset comprises reaction between

amines and carboxylic acids conducted in solvents in the presence of base and and coupling agent. During fine-tuning, pretrained parameters are used as initialization, and the source–sink prediction head, trained during pretraining as a classification objective with cross-entropy loss, is directly reused to generate softmax-normalized atom-level probabilities. These probabilities are employed as node-wise weighting factors during supergraph embedding aggregation, biasing the pooled representation toward atomic regions inferred to be more mechanistically relevant to electron transfer. The resulting weighted pairwise embeddings are then aggregated to form the final reaction-level representation for yield prediction. It is worth noting that the full distribution of source-sink probabilities is retained during pooling rather than selecting a single reactive atom, which improves robustness when reactive sites are distributed across multiple atoms or are uncertain at the atom level.

This pretraining-fine-tuning strategy allows mechanism-derived information learned from large-scale reaction data to influence downstream yield prediction indirectly, without requiring mechanistic annotations in the yield datasets. Figure 2 summarize our workflow.

2.3 Implementation Details

All experiments for the structure-aware reaction models (Section 2.1) were conducted on an NVIDIA A100 GPU, with batch size of 32 and the learning rate 5×10^{-5}. Each molecular component embedding had dimension 256, matching the dimensionality of the pretrained Frad encoder. Unless otherwise stated, all remaining hyperparameters, including the optimizer, scheduler, and activation functions, were kept identical to those used in the original Frad implementation [2]. All Frad encoders were initialized from checkpoints pretrained on the PCQM4Mv2 dataset. Reaction-level representations used a four-head multi-head self-attention module, with component embedding used as query, key, and value. Attention outputs were concatenated before being passed to the subsequent fully connected layers. Each task was trained independently and required approximately 12 hours to complete.

Experiments for the pairwise supergraph model with source-sink pretraining (Section 2.2) were conducted on a single NVIDIA RTX 4090 GPU. For source-sink pretraining on Mech-USPTO-31k, a random 0.85:0.15 train-validation split was used with AdamW optimizer (learning rate $1x10^{-4}$, batch size 256). Source-sink prediction was formulated as a graph-wise node selection task, where each atom in a given (super)graph is treated as a candidate class and the number of classes thus varies with graph size. Two parallel heads (source and sink) are trained with cross-entropy loss (with non-reactive nodes masked), and predictions are obtained by taking the argmax over atoms for each head. For yield prediction, the pretrained model was fine-tuned on the amide coupling dataset[13] using a random 0.85:0.15 split, learning rate 1×10^{-6} and batch size 16. Reactions were represented as a hierarchical graph with atom-, molecule-, and molecule-pair-level nodes, using a five-layer D-MPNN backbone, with embeddings projected to dimension 256. Pairwise embeddings were aggregated using masked four-head attention pooling. During fine-tuning, predicted source-sink probabilities were used for probability-weighted pooling. Dropout rates of 0.1 (encoder) and 0.5 (yield head) were applied.

3. RESULTS

3.1 Reaction Yield Prediction with Structure-Aware Molecular Representations

This section evaluates whether incorporating three-dimensional, structure-aware molecular representation improve reaction yield prediction, with particular emphasis on generalization to reactions involving previously unseen components. The proposed structure-aware reaction model (SA-YIELD), described in Section 2.1, is evaluated on Buchwald-Hartwig coupling reaction dataset.

First, we note that SA-YIELD achieves a MAE of 4.18 under a random split setting. This performance is comparable to that reported for Yield-BERT [5], which achieves an average MAE of approximately 4.0 on the same dataset. Considering that Yield-BERT relies on representations pretrained on reaction SMILES from the USPTO patent corpus (approximately 3.7 million reactions), these results indicate that incorporating structure-aware molecular representations at the reaction level can achieve competitive yield prediction accuracy without requiring geometries derived from crystal structures or large-scale reaction pretraining.

Next, we compare the SA-YIELD to representative baseline models, sequence-based Yield-BERT [5] and structure-aware CRYSTAL-YIELD7. All performance of components splits are summarized in Table 1. In this setting, prediction accuracy decreases substantially for all models, as it assesses extrapolation to reactions containing held-out components. For example, the MAE is approximately 4 in random split, whereas it increases into a substantially higher range for reactions involving unseen components, reaching their largest values for unseen aryl halides up to 24.

This behavior reflects the fact that aryl halides span a broader chemical space and play a central role in determining reactivity in Buchwald–Hartwig couplings. As primary electrophilic partners, they directly determine the identity of the reacting electrophile and play a central role in bond formation. Withholding such components therefore induces a pronounced shift in the underlying yield distribution, which limits extrapolation based solely on patterns learned from the remaining reaction

components. This effect is substantially more pronounced than for other reaction partners, such as ligands, and exposes limitations in models that encode reactants in isolation, without explicitly accounting for how reactivity depends on interacting reaction partners.

Despite these challenges, SA-YIELD consistently achieve the lowest MAE across all unseen-component categories. The average MAE is reduced from 24.4 for Yield-BERT and 15.65 for CRYSTAL-YIELD to 13.5, with the largest improvement observed for aryl halides—the most challenging unseen component, where SA–YIELD achieves an MAE of 18.8. Comparisons between models with (SA-YIELD) and without pretraining (SA-YIELD-RAW) indicate that these gains primarily arise from reaction-level architectural design choices rather than from denoising-based molecular pretraining alone.

While the Frad framework effectively captures three-dimensional molecular structure, molecular-level pretraining alone provides limited benefit for extrapolative prediction. This suggests that accurately modeling reaction outcomes requires representations that go beyond intrinsic molecular features. This motivates a shift toward reaction-level architectures and supervision strategies that explicitly reflect reaction mechanisms, which we explore in the following section.

Table 1: Reaction yield prediction performance on the Buchwald-Hartwig coupling dataset under component-wise unseen splits. Reported values are mean absolute error (MAE) in yield percentage points (mean ± s. e. across splits where applicable)

Model	Base	Ligand	Halide	Additive	Average
Yield-BERT	24.3 ± 1.6	24.3 ± 1.4	24.7 ± 2.1	24.1 ± 0.7	24.4
CRYSTAL-YIELD	13.4 ± 0.3	11.7 ± 2.2	21.3 ± 3.3	16.2 ± 0.4	15.65
SA-YIELD	11.8 ± 1.6	11.3 ± 2.6	18.8 ± 0.8	12.1 ± 1.3	13.5
SA-YIELD-RAW	11.5 ± 1.5	11.7 ± 3.0	19.5 ± 1.4	12.1 ± 1.3	13.7

3.2 Reaction Yield Prediction from explicit relation embedding

3.2.1 Source-Sink Pretraining Performance

To assess the effectiveness of mechanism-aware pretraining, the pretrained model was evaluated on the electron source-sink identification task defined by the Mech-USPTO-31k dataset. The evaluation follows the first-step source-sink labeling scheme described in Section2.2.3. Consistent with the dataset formulation, source–sink prediction is treated as a graph-wise node selection problem, in which each atomic node within a pairwise molecular supergraph is assigned a relative score, and the source and sink atoms are identified by selecting the highest-scoring nodes for each role. These scores are normalized via a softmax operation during training with cross-entropy loss, allowing the model to learn a single source and a single sink atom per supergraph.

As summarized in Table 2, our model demonstrates high accuracy in both source and sink predictions. On the test set comprising 4, 705 reactions, 4, 288 cases (91.2%) ere correctly predicted for both source and sink components. Partial predictions were observed in 264 cases with only the source correctly identified and in 137 cases with only the sink correctly identified, indicating greater difficulty in sink-site prediction. Notably, only 16 reactions (0.3%) were misclassified for both sites, highlighting the robustness of the model in distinguishing reactive and non-reactive atomic positions. This reliability allows the predicted source–sink signals to be used for active fine-tuning. Rather than treating source-sink prediction as a standalone objective, the present model incorporates source-sink supervision within a pairwise molecular supergraph architecture that explicitly encodes interactions between reacting species. In this framework, source-sink prediction is not an endpoint, but a means of biasing learned representations toward interaction-dependent atomic features.

Table 2: Prediction Outcomes on the Test Set

Outcome	Counts	%
Both correct	4288	91.1
Source only	264	5.6
Sink only	137	2.9
None	16	0.3

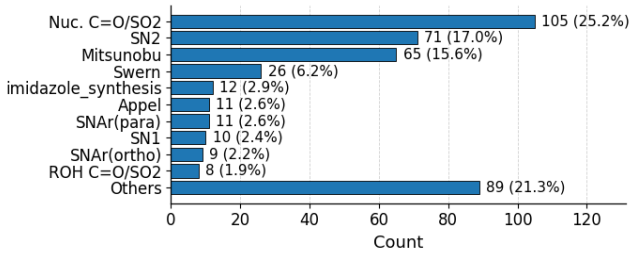

Figure 3. Reaction-Class Distribution of Failed Predictions

To examine the limitations of the pre-training approach, we analyzed the reaction classes associated with prediction errors, including both complete and partial mispredictions (Figure 3). Prediction errors were frequently observed in nucleophilic attacks on (thio)carbonyl or sulfonyl groups, S_N2 reactions, and Mitsunobu-type transformations, accounting for 105 (25.2%), 71 (17.0%), and 65 cases (15.6%), respectively. As discussed in Ref 10, S_N2 reactions are inherently challenging to annotate using discrete electron-flow representations, as bond formation and cleavage occur concertedly and multiple mechanistic interpretations can be chemically

plausible. In nucleophilic attack reactions involving (thio)carbonyl or sulfonyl groups, misassignments frequently involved intramolecular sites rather than external nucleophiles, reflecting a fundamental limitation of the molecular representation, as the 2D graph encoding does not capture steric accessibility or approach geometry. For Mitsunobu reactions, detailed inspection revealed that a substantial fraction of errors originated from the symmetric structure of diethyl azodicarboxylate, in which two nitrogen atoms are chemically equivalent, and the model frequently selected the alternative symmetric site that was treated as incorrect under the current labeling scheme despite being mechanistically equivalent. These observations indicate that many Mitsunobu-related mispredictions arise from annotation artifacts rather than genuine mechanistic failures, suggesting that the effective prediction accuracy of our model is likely higher than the nominal value reported here.

Despite these limitations, the pretrained model exhibits informative probabilistic behavior. Even when the highest predicted source or sink probability does not coincide with the ground-truth, the model frequently assigns relatively high probabilities to other chemically plausible atomic sites. Because the downstream framework does not rely on discrete source-sink assignments but instead incorporate the full probability distribution during the pooling, such partially correct predictions can still contribute useful interaction-relevant information.

3.2.2 Effect of Mechanism-Aware Pretraining and Active Finetuning

Reaction yield prediction performance on the amide coupling dataset is summarized in Table 3, comparing different reaction embedding strategies with and without mechanism-aware pretraining. Three embedding formulations are evaluated. Molecular concatenation aggregates independently encoded molecular embeddings without explicit intermolecular message passing and serves as a non-relational baseline. Multi-molecular supergraph encodes all reaction components within a single graph, enabling message passing across all atoms simultaneously. Pairwise supergraph, as described in Section 2.2.2, constructs relational embeddings by encoding molecular components as combinatorial molecular pairs. The first two settings are included as ablation baselines to assess the effect of explicitly modeling pairwise intermolecular relationships.

When trained without source-sink pretraining, the model based on molecular concatenation achieve slightly better performance than the pairwise supergraph. In this setting, reaction yield labels alone provide limited information about which molecular pairs and atomic regions are responsible for determining the reaction outcome. As a result, the model is unable to reliably identify and exploit pair-specific interaction patterns from the amide coupling dataset, which contains on the order of 500 reactions. which limits the effectiveness of relational representations and can lead to degraded generalization performance.

In contrast, when the pairwise supergraph encoder is initialized using source-sink pretraining and actively leveraging the mechanistic information, a substantial improvement in prediction accuracy is observed. This improvement reflects the ability of the large-scale pretraining to bias the model toward interaction-dependent atomic features that are relevant to reaction outcomes. Because source-sink annotations encode electron-transfer roles that arise from interactions between specific reacting species, the resulting inductive bias is naturally aligned with pairwise relational representations and cannot be readily captured by models that operate solely on independently encoded molecules.

Finally, comparison between relational formulations shows that the pairwise (combinatorial) supergraph outperforms the multi-molecular supergraph when combined with mechanism-aware pretraining. This behavior is consistent with the structure of the source-sink supervision, which is defined at the level of interacting molecular pairs rather than across all reaction components simultaneously. Encoding all molecules within a single supergraph can dilute pair-specific information, whereas the pairwise construction preserves interaction-specific contexts prior to aggregation. In summary, these results confirm that the proposed framework effectively pretrains mechanistically meaningful interaction information and selectively deploys it during fine-tuning, providing a principled and data-efficient strategy for yield prediction.

Table 3: Effect of embedding strategy and source-sink pretraining on amide coupling yield prediction. Reported values are MAE on the held-out test set

Embedding strategy	No pre-training	With pre-training
Molecular concatenation	0.17	-
Multi-molecular supergraph	-	0.16
Pairwise supergraph	0.19	0.14

4. CONCLUSION

In this study, two complementary approaches for reaction yield prediction were developed and evaluated, incorporating three-dimensional molecular structure and mechanism-derived relational information. First, a denoising-based molecular representation model pretrained on three-dimensional conformational data was extended from single-molecule property prediction to reaction-level modeling. This structure-aware formulation improved robustness under component-wise extrapolation, although molecular-level pretraining alone provided limited gains when transferred directly to reaction yield

prediction.

Second, a mechanism-aware framework was introduced that represents reactions using a pairwise supergraph formulation and leverages electron source-sink annotations for pretraining. By biasing representations toward interaction-dependent atomic regions associated with electron transfer, this approach improved yield prediction in data-constrained settings and outperformed direct training without such pretraining. These results indicate that reaction-level learning benefits from aligning representation granularity and pretraining tasks with mechanistically motivated constructs.

The findings suggest that yield prediction models benefit from moving beyond aggregation of isolated molecular features and instead incorporating interaction-specific, mechanism-derived information. While the present study used two-dimensional graph representations for mechanistic pretraining and was limited by available dataset scale, it establishes a basis for integrating three-dimensional structure and richer mechanistic annotations in future work.

ACKNOWLEDGEMENTS

This research was supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Ministry of Trade, Industry & Energy (MOTIE), Korea Government (RS-2024-00436106, Human Resource Development Program for Industrial Innovation).

AUTHOR IDENTIFIERS

Author ORCIDs:
Lee Y: 0009-0007-5662-0494
Lee WB: 0000-0001-7801-083X
Lee L: 0000-0003-1847-7838

REFERENCES

1. Jiang J, Ke L, Chen L, Dou B, Zhu Y, Liu J, Zhang B, Zhou T, Wei G. Transformer technology in molecular science. WIREs Comput Mol Sci 14: (2024). https://doi.org/10.1002/wcms.1725
2. Ni Y, Feng S, Hong X, Sun Y, Ma WY, Ma ZM, Ye Q, Lan Y. Pre-training with fractional denoising to enhance molecular property prediction. Nat Mach Intell 6:1169-1178 (2024). https://doi.org/10.1038/s42256-024-00900-z
3. Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C–N cross-coupling using machine learning. Science 360:186-190 (2018). https://doi.org/10.1126/science.aar5169
4. Pistachio (Nextmove Software). http://www.nextmovesoftware.com/pistachio.html.
5. Lowe DM. Extraction of chemical structures and reactions from the literature. In: University of Cambridge, editor.; 2012.
6. Schwaller P, Vaucher AC, Laino T, Reymond JL. Prediction of chemical reaction yields using deep learning. Mach. Learn.: Sci. Technol. 2:015016 (2021). https://doi.org/10.1088/2632-2153/abc81d
7. King-Smith E. Transfer learning for a foundational chemistry model. Chem. Sci. 15:5143-5151 (2024). https://doi.org/10.1039/d3sc04928k
8. Fang X, Liu L, Lei J, He D, Zhang S, Zhou J, Wang F, Wu H, Wang H. Geometry-enhanced molecular representation learning for property prediction. Nat Mach Intell 4:127-134 (2022). https://doi.org/10.1038/s42256-021-00438-4
9. Singh J, Singh C. Learning invariant representations for equivariant neural networks using orthogonal moments. 2022 International Joint Conference on Neural Networks (IJCNN) :1-8 (2022). https://doi.org/10.1109/ijcnn55064.2022.9892412
10. Chen S, Babazade R, Kim T, Han S, Jung Y. A large-scale reaction dataset of mechanistic pathways of organic reactions. Sci Data 11: (2024). https://doi.org/10.1038/s41597-024-03709-y
11. Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction. Commun Chem 6: (2023). https://doi.org/10.1038/s42004-023-00825-5
12. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, Palmer A, Settels V, Jaakkola T, Jensen K, Barzilay R. Analyzing learned molecular representations for property prediction. J. Chem. Inf. Model. 59:3370-3388 (2019). https://doi.org/10.1021/acs.jcim.9b00237
13. Zhang C, Lin Q, Yang C, Kong Y, Yu Z, Liao K. Intermediate knowledge enhanced the performance of the amide coupling yield prediction model. Chem. Sci. 16:11809-11822 (2025). https://doi.org/10.1039/d5sc03364k

© 2026 by the authors. Licensed to PSEcommunity.org and PSE Press. This is an open access article under the creative commons CC-BY-SA licensing terms. Credit must be given to creator and adaptations must be shared under the same terms. See https://creativecommons.org/licenses/by-sa/4.0/