log-RRIM: Yield Prediction via Local-to-global Reaction Representation Learning and Interaction Modeling

Xiao Hu[1], Ziqi Chen[1], Bo Peng[1], Daniel Adu-Ampratwum[2], Xia Ning[1,2,3,4] ✉

[1]Computer Science and Engineering, The Ohio State University, Columbus, OH 43210. [2]Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, The Ohio State University, Columbus, Ohio 43210. [3]Biomedical Informatics, The Ohio State University, Columbus, OH 43210. [4]Translational Data Analytics Institute, The Ohio State University, Columbus, OH, 43210. ✉ning.104@osu.edu

Accurate prediction of chemical reaction yields is crucial for optimizing organic synthesis, potentially reducing time and resources spent on experimentation. With the rise of artificial intelligence (AI), there is growing interest in leveraging AI-based methods to accelerate yield predictions without conducting in vitro experiments. We present log-RRIM, an innovative graph transformer-based framework designed for predicting chemical reaction yields. Our approach implements a unique local-to-global reaction representation learning strategy. This approach initially captures detailed molecule-level information and then models and aggregates intermolecular interactions, ensuring that the impact of varying-sizes molecular fragments on yield is accurately accounted for. Another key feature of log-RRIM is its integration of a cross-attention mechanism that focuses on the interplay between reagents and reaction centers. This design reflects a fundamental principle in chemical reactions: the crucial role of reagents in influencing bond-breaking and formation processes, which ultimately affect reaction yields. log-RRIM outperforms existing methods in our experiments, especially for medium to high-yielding reactions, proving its reliability as a predictor. Its advanced modeling of reactant-reagent interactions and sensitivity to small molecular fragments make it a valuable tool for reaction planning and optimization in chemical synthesis. The data and codes of log-RRIM are accessible through https://github.com/ninglab/Yield_log_RRIM.

Chemical yield prediction is crucial for optimizing organic synthesis, offering chemists an efficient tool to identify high-yielding reactions while reducing time and resource expenditure.[1] Traditionally, chemists have relied on expertise and systematic experimentation to optimize reactions.[2] While foundational, these methods can become resource-intensive when scaling up.[3] Consequently, there is an increasing interest in developing artificial intelligence (AI)-based methods.[4–10] These AI-based methods allow chemists to accelerate precise yield prediction without doing in vitro experiments, potentially enhancing the efficiency of organic synthesis optimization. Despite the importance of the task, AI-based computational methods have received comparatively little attention in yield prediction compared to other chemistry-related tasks (e.g. forward prediction,[11,12] retrosynthesis[13,14]). We aim to bridge the gap and introduce novel and effective AI methods for yield prediction.

Early AI-based methods focused on identifying effective chemical knowledge-based reaction descriptors[15,16] and employing traditional machine learning models[17,18] over such descriptors for chemical yield prediction. However, these methods often produce unsatisfactory results, suggesting the limitation of the chemical knowledge-based descriptors, as well as the companion traditional machine learning models. The advent of language models[19,20] has enabled sequence-based approaches for chemistry-related tasks.[4–8] These models are typically pre-trained on large molecular datasets[21] using SMILES[22] representations and then fine-tuned on specific datasets for yield prediction with the entire reaction's SMILES string as input. However, this pre-training and fine-tuning framework may not be optimal for chemistry-specific tasks like yield prediction,[4,6] as it lacks features that account for unique characteristics of yield prediction, such as explicit modeling of reactant-reagent interactions. Moreover, these models, using the entire reaction as input, tend to overlook the contributions of small yet influential molecular fragments,[23] as their attention mechanisms may not be sensitive enough to focus on these critical elements. Additionally, building such pre-trained foundation models is resource-intensive. In contrast to the sequence-based models, graph neural networks (GNNs) have recently been employed to represent molecules and reactions as graphs, learning molecular structural information for yield prediction.[9,10] This approach allows for a more intuitive representation of molecular structure compared to sequence-based models. However, most GNN-based methods lack effective modeling of molecular interactions. This limitation is particularly significant in yield prediction, as the interactions between reactants and reagents, like catalysts, can substantially impact reaction outcomes.[24,25]

To address these challenges, we introduce log-RRIM: a graph transformer-based local-to-global reaction representation learning and interaction modeling for yield prediction. log-RRIM employs a local-to-global graph transformer-based reaction representation learning process, which first learns representations at the molecule level for each component individually and then models their interactions. This information is then aggregated, ensuring a more balanced attention mechanism that considers molecules of all sizes, preventing small fragments from being overlooked in the whole reaction for yield prediction. Additionally, log-RRIM incorporates a cross-attention mechanism between the reagents and reaction center atoms to simulate a principle of chemical reactions: reagents have a huge impact on the bond-breaking and formation of the reaction, thus affecting the yield changes. This design more effectively captures the interactions between molecules (reactants and reagents), thereby improving the prediction accuracy.

Performance evaluation on the commonly investigated datasets[6,26,27] demonstrates log-RRIM's superior prediction accuracy, particularly for medium to high-yielding reactions. This suggests its potential for enhancing reaction yield optimization accuracy in practical synthetic chemistry. Our analyses further reveal log-RRIM's effectiveness in capturing complex molecular (reactant-reagent) interactions and accurately assessing small molecular fragments' contributions to yield. These capabilities highlight log-RRIM's potential for optimizing synthetic routes through informed modifications of reactants and reagents, providing chemists with a sophisticated instrument for reaction design and optimization.

Related Work

Reaction yield prediction has evolved primarily through three types of approaches, each addressing the challenges of representing complex molecular structures and modeling their interactions in different ways. The approaches started with traditional machine learning models based on chemical knowledge-based descriptors. Next, sequence-based models were developed, representing each molecule as a SMILES string. These models are typically pre-trained on large molecule datasets to learn general molecule representations and then fine-tuned specifically for yield prediction tasks. Most recently, graph-based models have emerged as a powerful tool for learning molecular structures, treating molecules as graphs, and aggregating molecular information for prediction.

Traditional Machine Learning Models

Early approaches to yield prediction utilized traditional machine learning models, such as random forest (RF)[18] and support vector machine (SVM),[17] to predict yields. These models relied on chemical knowledge-based descriptors to depict the molecule properties, which include density functional theory calculations,[15,16] one-hot encoding ,[28] and fingerprint features .[29] These methods were primarily evaluated on reaction datasets containing a single reaction class.[26,30] However, they often demonstrated unsatisfactory performance.[15,16,28,29] This highlighted two main limitations. First, the modeling ability of traditional machine learning methods is insufficient for this complex problem. Second, relying solely on predefined chemical descriptors for constructing reaction representations is inadequate. The suboptimal results obtained from these methods suggest that more sophisticated and effective approaches are needed to capture the complex information between molecular structures and reaction yields.

Sequence-based Models

Transformer-based models have recently gained prominence in chemical tasks.[4–8] These models are typically pre-trained on large molecular datasets represented by SMILES strings, learning general molecular representations. They are then fine-tuned on specific datasets containing yield information for the prediction. During fine-tuning, the models learn to process the SMILES string of the entire reaction as input, enabling them to capture relationships between all reaction components. For example, Schwaller *et al.* introduced YieldBERT,[4] which employs the SMILES string of a whole reaction as input to a BERT-based yield predictor.[20] This BERT-based yield predictor is obtained from fine-tuning a yield regression head layer on a reaction encoder.[31] Similarly, Lu and Zhang developed T5Chem,[6] utilizing the Text-to-Text Transfer Transformer (T5) model.[19] T5Chem, pre-trained on the PubChem dataset,[21] is designed for multiple reaction prediction tasks (e.g., product prediction, retrosynthesis) and employs a fine-tuned regression head for yield prediction purposes. The sophisticated sequence modeling techniques enable these methods to learn more informative reaction representation than handcrafted chemical knowledge-based descriptors by capturing contextual information embedded in the SMILES string of the entire reactions. Consequently, they demonstrate commendable prediction performance on datasets containing a single reaction class.

However, the efficacy diminishes when testing on datasets with a wide variety of reaction types and diverse substances, such as the US Patent database (USPTO).[32] Additionally, treating the whole reaction as input makes it challenging for the sequence-based models to distinguish the effects of different components in a reaction, as reactants and reagents have distinct impacts on yield. Also, small modifications in the molecules, even those involving only a few fragments (atoms, functional groups, or small-size molecules), can significantly affect reaction outcomes.[23] When sequence-based models treat the entire reactions as inputs, they tend to overlook the contributions of those small yet influential fragments. This occurs because the attention mechanisms used in these sequence-based models may not be sufficiently sensitive to those critical fragments, potentially leading to inaccurate predictions.

To address these challenges, we propose to apply a local-to-global learning process to ensure equal attention is allocated to molecules of varying sizes. The local-to-global learning process treats each reactant, reagent, and product separately before interacting and aggregating their information, intuitively depicting the role of different components in the reaction. This prevents the model from ignoring the impact of small fragments. Our experiment and analysis demonstrate the effectiveness of our modeling design.

Graph-based Models

Recent advancements have established graph neural networks (GNNs) as powerful tools for analyzing molecules and predicting reaction yields.[9,10,33–36] These approaches represent chemical structures as graphs, using GNNs to learn structural information and typically employing multilayer perceptrons (MLPs) to predict yields after aggregating molecular information into vector representations. Saebi *et al.* developed YieldGNN,[9] which uses Weisfeiler-Lehman networks (WLNs)[37] to aggregate atom and bond features over their neighborhood and finally obtain the high-order structural information. These learned structural features and the selected chemical knowledge-based reaction descriptors are then combined to predict the reaction yield through a linear layer. Their results highlight the importance of learned molecular structural

features over the chemical descriptors. Yarish *et al.* introduced RD-MPNN,[36] which first uses directed message passing networks (D-MPNN)[33] to generate atom and bond embeddings from reactant and product graphs. Then, it creates the chemical transformation encoding according to the atom and bond mapping between the reactants and the products, which is combined with pre-computed molecular descriptors to predict the yield. Li *et al.* proposed SEMG-MIGNN,[10] which similarly employs a GNN to update atom features and obtain molecule representations. Then, it applies an attention mechanism based on all involved components to model the molecular interplays and derive the reaction representation for prediction.

While these graph-based methods demonstrate satisfactory performance on datasets of a single reaction class, they have not been extensively tested on more challenging datasets like USPTO. Furthermore, these approaches exhibit certain limitations in molecular interaction design. RD-MPNN and YieldGNN lack explicit modeling of interactions among reactants and reagents, while SEMG-MIGNN's design may not effectively capture the full complexity of molecular interactions.

To address these limitations and better enable the model to learn the interactions between reactants and reagents, we propose to explicitly characterize the function of reagents on the reaction center. This approach uses a cross-attention mechanism[38] to capture the complex interplay between different reaction components (reactants and reagents) more effectively, potentially leading to improved yield predictions. Our experiments and analysis demonstrate that this design improves the effectiveness of molecular interaction modeling.

Materials

Datasets

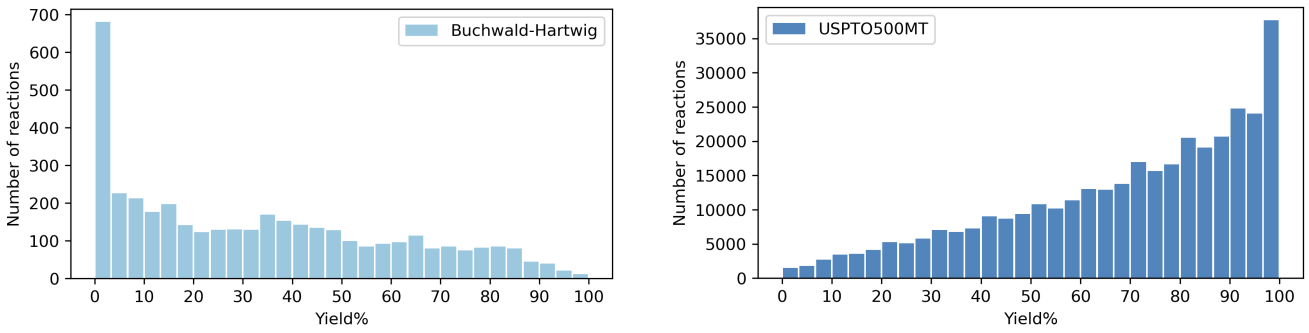
USPTO500MT Dataset

USPTO500MT is derived from USTPO-TPL[31] by the authors of T5Chem.[6] USTPO-TPL comprises 445,000 reactions, with yield reported, partitioned into 1,000 strongly imbalanced reaction types. USPTO500MT is obtained by extracting the top 500 most frequently occurring reaction types from USPTO-TPL. It consists of 116,360 reactions for training, 12,937 reactions for validation, and 14,238 reactions for testing purposes. The reactants, reagents, and products are encoded as SMILES strings. The yield distribution is summarized in Figure 1b and the entire dataset is skewed towards high-yielding reactions. Within the USPTO500MT dataset, approximately 95.5% of the reactions (129,437) are unique. Additionally, about 3.7% of the products (4,949) are documented with two distinct synthesized processes. Only a small fraction (0.1%) of products are synthesized through over five different processes. Moreover, the number and the function of reagents are varying among each reaction. These showcase the diversity and complexity of the reactions within the dataset.

Buchwald–Hartwig Amination Reaction Dataset

The Buchwald-Hartwig dataset, constructed by Ahneman *et al.* ,[26] has become a benchmark for assessing the performance of yield prediction models. This dataset comprises 3,955 palladium-catalyzed C-N cross-coupling reactions, with yields obtained through high-throughput experimentation (HTE). The dataset encodes information on reactants, reagents, and products as SMILES strings. It includes 15 distinct aryl halides paired with a single amine as reactants. These reactant pairs undergo experimentation with 3 different bases, 4 Buchwald ligands, and 22 isoxazole additives, resulting in 5 different products. The yield distribution, illustrated in Figure 1a, reveals a notable skew due to a substantial proportion of non-yielding reactions.

In comparison to broader datasets such as USPTO500MT, the Buchwald-Hartwig dataset is limited to a single reaction type and features a constrained set of reaction components. Moreover, reagent information is consistently organized, with each reaction entry containing ligand, base, and solvent information in a consistent order. While this structured format may facilitate easier predictive model learning, it potentially misrepresents real-world scenarios where chemical data is often comprehensive and less organized. This underscores the limitation in this dataset's ability to reflect the complexity and variability of practical chemical information, despite its value as a benchmark for yield prediction models.



(a) Buchwald–Hartwig reactions yield distribution

(b) USPTO500MT reactions yield distribution

Fig. 1 | Overview reactions yield distributions of the two datasets

Training data generation

Basic atom features

We follow Maziarka *et al.*[39] and employ the open-source RDKit toolkit to extract the basic chemical features for atoms in molecules represented by SMILES strings. The basic atom features utilized in log-RRIM are delineated in Table 1. These features describe the basic chemical properties and environment, serving as the input of log-RRIM$_b$.

Table 1 | Basic atom features used in log-RRIM

Indices	Description
0-11	Atom type of B, N, C, O, F, P, S, CL, BR, I, Dummy, Other (One-hot encoded)
12-17	Number of connected heavy atoms of 0, 1, 2, 3, 4, 5 (One-hot encoded)
18-22	Number of connected hydrogen of 0, 1, 2, 3, 4 (One-hot encoded)
23-25	Formal charge of -1, 0, 1 (One-hot encoded)
26	If the atom is in a ring (Binary)
27	If it is aromatic (Binary)

Learned atom representations from pre-trained models

To investigate the impact of atom features chosen on log-RRIM, we employ two approaches: one using the basic atom features directly, and another using learned atom representations derived from a pre-trained model MAT by Maziarka *et al.*.[39] We name the log-RRIM trained on basic atom features as log-RRIM$_b$, and the version trained on learned atom representations as log-RRIM$_l$. The pre-trained model MAT takes the basic atom features as input and utilizes node-level self-supervised learning [40] on a subset of 2 million molecules from the Zinc15 dataset [41] for molecule representation learning. These learned atom representations are then input for log-RRIM$_l$, potentially capturing more complex atomic relations and information. The hyperparameters of the pre-trained model are delineated in Table A8 and remain consistent across all experiments.

Reaction center identification

Identifying reaction centers is crucial for log-RRIM as it allows us to pinpoint the specific atoms involved in the chemical transformation. We follow GraphRetro's[42] approach to identify these reaction center atoms by comparing the changed bonds between the mapped reactant and product molecules. In log-RRIM, we model the interactions between these reaction centers and reagents, which enables us to more effectively capture the key information (reagents have an impact on bond-breaking and formation) that influences the reaction yield, potentially improving the accuracy of predictions.

Experimental setting

For the USPTO500MT dataset, we adopt the training, validation, and testing split used by T5Chem. We adhere to the data-splitting protocol for the Buchwald-Hartwig dataset as YieldGNN, using 10-fold 70/30 random train/test splits. We further allocate 10% of the training data for validation. After determining the optimal hyperparameters using three data splits, we apply the model across all ten data splits and compare its performance against other baselines. In addition, we exclude reactions that cannot be processed by the reaction center identification method. The reaction center identification process ensures that all reactions in the dataset have well-defined reaction centers and identifiable mechanistic pathways, which is critical for accurate modeling of reaction mechanisms and yield predictions. While this process does not filter out any reactions in the Buchwald–Hartwig dataset, it results in 78,201 reactions filtered out for training, 8,716 for validation, and 9,497 for testing in USPTO500MT. This curation enhances the overall integrity of USPTO500MT, allowing for more precise reactions to be considered. All performance comparisons are conducted on these curated datasets to maintain consistency in our evaluations.

Model evaluation

We use mean absolute error (MAE) and root mean squared error (RMSE) for evaluation purposes. Their calculations are given by the following equations:

$$\text{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}, \tag{1}$$

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

where $\hat{y}_i$ is the predicted yield, y_i is the ground-truth yield, and N is the number of samples. The smaller the MAE and RMSE are, the more accurate the yield predictor model is. Previous methods[6,26] use the coefficient of determination (R^2) to evaluate the goodness of fit of the regression model, which is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \tag{3}$$

where $\bar{y}$ is the mean of N ground-truth yields and a larger value of R^2 implies a better goodness of fit of the models. However, R^2 is not an ideal metric to evaluate the accuracy and relationship, as it has several limitations.[43] One significant issue is that R^2 can be heavily influenced by outliers, potentially giving a distorted view of the model's overall fit. This

sensitivity means that a few extreme error predictions can lead to a very low R^2, even if the majority of predictions are accurate. Therefore, it is challenging to draw definitive conclusions from R^2, especially when it is low. While we still present the results in R^2 in line with the literature, the evaluation is primarily via MAE and RMSE.

Experiment results

Performance on the USPTO500MT dataset

Overall performance

Table 2 presents the performance comparison of log-RRIM$_b$, log-RRIM$_l$, and baseline methods YieldBERT and T5Chem on the USPTO500MT dataset. log-RRIM$_l$ demonstrates the best performance in terms of MAE and RMSE, achieving

Table 2 | Model performance comparison on USPTO500MT

Method	MAE	RMSE	R^2
YieldBERT	0.191	0.245	0.090
T5Chem	0.190	0.249	**0.212**
log-RRIM$_b$	**0.181**	**0.228**	0.122
log-RRIM$_l$	**0.179**	**0.226**	0.144

The best performance is highlighted in bold.

the lowest MAE of 0.179 and RMSE of 0.226. These results represent statistically significant improvements of 5.8% on MAE over the previous best-performing method T5Chem. The statistical significance of this improvement is underscored by a p-value of 5e-12 at a significance level of 5%, obtained from a paired t-test comparing the Absolute Errors (AE) of log-RRIM$_l$ and T5Chem (Unless otherwise specified, the p-values mentioned in the following paper are all derived from this paired t-test).

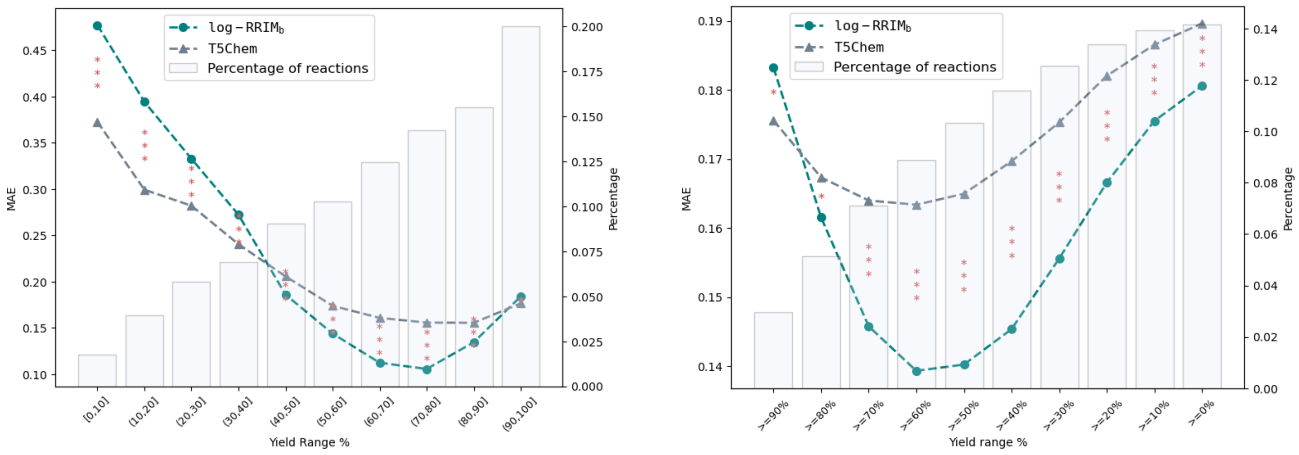
log-RRIM$_b$, which utilizes the basic atom features in contrast to log-RRIM$_l$ utilizing the learned atom representations, achieved comparable results to log-RRIM$_l$ with an MAE of 0.181. log-RRIM$_b$ is still significantly better than T5Chem (p-value = 1e-8). We attribute the superior performance of log-RRIM to its effective framework design, specifically engineered to model and learn fundamental factors influencing reaction yield. The local-to-global learning scheme employed by log-RRIM allows for equal attention to all molecules of varying sizes before modeling their interactions, preventing the oversight of the contributions from small yet influential fragments (e.g., atoms, functional groups, or small molecules). This approach contrasts with sequence-based models like T5Chem and YieldBERT, which treat the entire reaction as input, where the attention mechanisms may not be sufficiently sensitive to critical fragments. Furthermore, log-RRIM's molecular interaction design explicitly models the function of reagents on reaction centers, more closely mimicking the synthetic reaction principle: reagents like catalysts have a huge impact on bond-breaking and formation. This targeted design is more effective than T5Chem and YieldBERT's interaction modeling, which indiscriminately applies global attention to all atoms. It is also worth noting that log-RRIM is pre-training-free, whereas T5Chem and YieldBERT are based on foundation models pre-trained on extensive molecule datasets (e.g. 97 million molecules from PubChem[21]). log-RRIM's superior performance suggests that pre-training may not be necessary if the training dataset is sufficiently large (e.g., 78K for USPTO500MT) when the reactions are modeled in a targeted and explicit way. By incorporating more effective designs, log-RRIM achieves better performance while saving huge resources required for pre-training.

log-RRIM$_b$ and log-RRIM$_l$ exhibit nearly identical performance, with MAE values of 0.181 and 0.179, respectively. The former employs basic atom features, while the latter utilizes atom representations derived from the pre-trained MAT model.[39] The incorporation of learned representations does not obtain a substantial improvement in yield prediction accuracy over basic features. This outcome suggests that the atom representations acquired through the MAT model, which was originally developed for general molecule representation learning[39], lack the specificity required for reaction-oriented tasks. Although basic atom features only provide elementary information about molecular properties, our findings underscore that the key to enhancing yield prediction accuracy lies in more sophisticated and effective modeling of intermolecular interactions.

While other graph-based yield prediction methods[9,10,36] exist, they are primarily designed for datasets with fixed reagent structures, such as the Buchwald-Hartwig dataset, which includes very specific reagent information (additive, base, solvent, and ligand).[26] However, these methods do not apply to the USPTO500MT dataset used in this study due to its varying number of reagents across reactions and lack of standardized reagent information. However, the USPTO500MT dataset more closely resembles real-world scenarios where reaction compositions are not strictly structured. In this context, log-RRIM, T5Chem, and YieldBERT demonstrate greater potential for practical applications compared to the graph-based methods just mentioned. log-RRIM's superior performance among those methods, as demonstrated in the previous results, combined with its flexibility in handling diverse inputs, positions it as a promising approach for accurate yield prediction in practical usage.

Performance comparison over different yield ranges

To gain deeper insights into the performance differences between log-RRIM$_b$ and T5Chem, we conducted a detailed analysis of predictions across various yield ranges. Figure 2 visualizes these comparisons, with stacked asterisks indicating the level of statistical significance of the performance difference across yield ranges (see Table A1 for exact values). Figure 2a shows that log-RRIM$_b$ outperforms T5Chem in predicting yields within the 40% to 100% with t-test p-values all less than 0.05, indicating statistical significance at the 5% level. This pattern suggests that log-RRIM$_b$ is a more reliable predictor for medium to high-yielding reactions, a crucial advantage in practical synthesis scenarios.[44,45] Also, Figure 2b suggests



(a) Performance on reactions within each yield range

(b) Performance on reactions within cumulative yield range

Fig. 2 | Performance comparison of log-RRIM$_b$ and T5Chem across yield ranges on the USPTO500MT testing set. Left y-axis: MAE of predicted yields. Right y-axis: percentage of reactions in the testing set for each yield range. 5% significance level: * for p-values < 0.05, ** for p-values < 0.005, *** for p-values < 0.0005.

the overall prediction performance of log-RRIM$_b$ is significantly better. This improved overall accuracy is particularly valuable in the context of exploring new reactions, where precise yield data may not be available for reference. In such scenarios, log-RRIM$_b$'s overall more reliable predictions can offer more accurate guidance for reaction planning and optimization. However, for reaction yields below 40%, log-RRIM$_b$ exhibits inferior performance than T5Chem. We attributed this to T5Chem's leveraging of foundation models pre-trained on extensive molecule datasets, compensating for a potential shortage of training samples encountered by log-RRIM$_b$ on reactions with yields below 40% (18.1% of the training set). Nevertheless, log-RRIM$_b$ remains the preferred choice for chemists seeking reliable yield predictions, particularly for medium to high-yielding reactions or when no preliminary reaction yield data can be referred to. This reliability can significantly aid chemists in experimental planning, reducing the number of optimization iterations and minimizing resource consumption.

Effectiveness in reactant-reagent interactions modeling

To assess the model's capacity to capture the influence of molecular interactions on yield, specifically how reactants and reagents affect each other in the context of a reaction, we conducted two analyses on the testing set of USPTO500MT. First, we identified 76 reaction pairs (152 reactions) with identical reactants but different reagents and yields. This setup allowed us to evaluate how our method is sensitive to the effects of reagents on yields. In this context, "interactions" refer to how the introduction of different reagents influences the reaction outcome with the same reactants. log-RRIM$_b$ achieved a prediction MAE of 0.145, outperforming T5Chem's 0.182. Furthermore, log-RRIM$_b$ correctly predicted the yield difference (how much the yield increases or decreases) in 62% (47 out of 76) of reaction pairs, compared to T5Chem's 38%. This suggests that log-RRIM$_b$ is more sensitive to reagent changes and their effects on yield. Case 1 in Figure 4 illustrates this: in two identical aryl nitration reactions, adding ether as a solvent increases the ground-truth yield from 42.0% to 57.7%. log-RRIM$_b$ correctly predicts this upward trend, while T5Chem does not. This shows log-RRIM$_b$'s ability to capture how the addition of a solvent (ether) interacts with the existing reactants to influence the yield.

Secondly, we examined 3,698 reactions grouped into 619 sets, each containing two or more reactions with identical reagents but different reactants. This analysis aimed to evaluate the models' ability to predict yields when the same reagents interact with various reactants. Here, "interactions" refer to how the same set of reagents behaves differently with varying reactants. log-RRIM$_b$ exhibited more accurate predictions in 58% of sets (357 out of 619), with a lower MAE of 0.147 compared to T5Chem's 0.222. Case 2 in Figure 4 demonstrates log-RRIM$_b$'s consistently more accurate predictions when the same reagents (carbon disulfide and bromine) interact with two different reactants. This indicates log-RRIM$_b$'s enhanced capability to learn and model specific reagent functions across different reaction contexts, capturing how the same reagents behave differently with varying reactants.

Overall, These analyses suggest that log-RRIM$_b$ is more sensitive to changes in reactant-reagent combinations, indicating better modeling of their interactions. This enhanced capability makes log-RRIM$_b$ a potential aid for chemists in selecting and optimizing reactants or reagents during synthesis planning. We attribute this superiority to log-RRIM$_b$'s explicit modeling of reagent function to reaction centersd. This approach, implemented through a cross-attention mechanism, aligns with fundamental reaction principles. It allows log-RRIM$_b$ to directly model how reagents influence the reaction center, providing a more nuanced understanding of the reaction process. An ablation study on the removal of explicit reagent function modeling, provided in Table A5, further supports this design choice. As a result, log-RRIM$_b$ demonstrates an enhanced ability to capture and interpret complex reactant-reagent interactions, leading to more accurate yield predictions across diverse reaction component combinations.

Sensitivity to small fragments modifications

To evaluate the models' ability to capture the influence of involved small fragments on reaction yields, we conducted a comparative analysis of their performance on similar reactions with small differences only on a few small fragments in

reactants or reagents. Given the absence of a standardized method for quantifying reaction similarity, we propose a novel similarity metric $Sim(\mathcal{X}_i, \mathcal{X}_j)$ between reactions $\mathcal{X}_i$ and $\mathcal{X}_j$, defined as the average of reactant and reagent similarities:

$$Sim(\mathcal{X}_i, \mathcal{X}_j) = \frac{1}{2}\left[s\left(\mathcal{R}_i, \mathcal{R}_j\right) + s\left(\mathcal{A}_i, \mathcal{A}_j\right)\right] \tag{4}$$

where $\mathcal{X} : \mathcal{R} \xrightarrow{\mathcal{A}} \mathcal{P}$ refers to the reaction, $\mathcal{R}$ and $\mathcal{A}$ are the concatenation of all reactants and reagents in the reaction, respectively. $s(\cdot, \cdot)$ is the Tanimoto coefficient between the two chemical structures of Morgan fingerprint.[46]

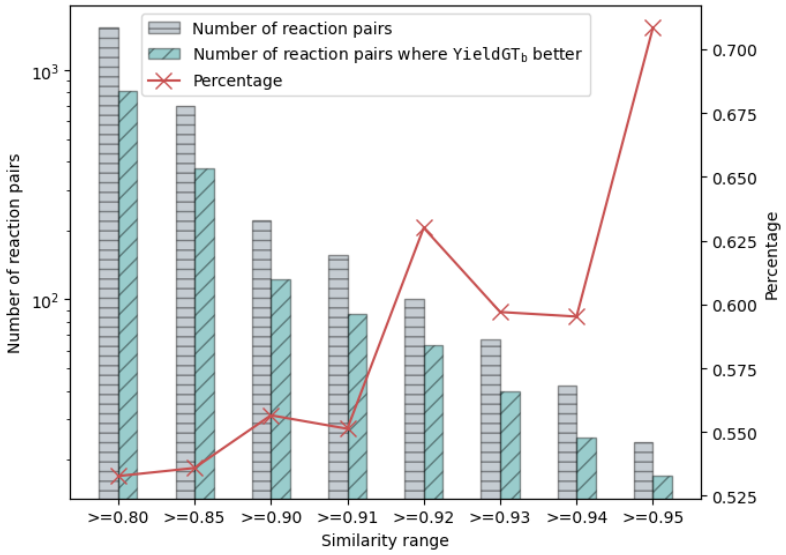


Fig. 3 | Model performance on reaction pairs categorized by similarity. The left y-axis displays the number of reaction pairs on a logarithmic scale. Grey bars indicate the number of reaction pairs within each similarity range. Green bars represent the number of reaction pairs where log-RRIM$_b$ predicts more accurately than T5Chem. The right y-axis shows the percentage of reaction pairs with more accurate predictions by log-RRIM$_b$ relative to the total number of reactions in each similarity range, as depicted by the red line.

We evaluated reaction pairs across a range of similarity thresholds (0.8-0.95), comparing the performance of log-RRIM$_b$ and T5Chem in predicting yield differences between the two reactions in the pair. The results are illustrated in Figure 3. Specifically, for reaction pairs with $Sim \geq 0.80$ (1526 pairs, 3052 reactions), log-RRIM$_b$ outperformed T5Chem on 53% (813/1526) pairs, with overall MAEs of 0.158 and 0.159 respectively. This advantage becomes more pronounced as the reaction similarity increases. On 221 pairs with $Sim \geq 0.9$, log-RRIM$_b$ surpassed T5Chem on 56% (123/221), with MAEs of 0.162 and 0.165 respectively. The trend culminated with highly similar reaction pairs ($Sim \geq 0.95$, 24 pairs), where log-RRIM$_b$ demonstrated marked superiority, outperforming T5Chem on 71% (17/24), with MAEs of 0.150 and 0.170 respectively. These results reveal a clear trend: log-RRIM$_b$'s accuracy in capturing yield differences improves as reaction similarity increases. This indicates that log-RRIM$_b$ exhibits enhanced sensitivity to subtle component changes that impact reaction yields, particularly for highly similar reactions.

The capability is also demonstrated in several cases. In Figure 4 case 3, the two reactions differ only in their *ortho*-substitution (methoxy vs fluoro group), resulting in a yield decrease from 68.2% to 48.9%. log-RRIM correctly predicts this change, while T5Chem incorrectly predicts the opposite trend. Similarly, case 4 in Figure 4 presents two alkylations of hydroxyquinoline with different alkylating agents. The ground-truth yield changes minimally in this situation, which log-RRIM$_b$ correctly predicts, whereas T5Chem makes an erroneous prediction. These results indicate that the log-RRIM$_b$ excels in predicting yield changes triggered by those small modifications in atoms, functional groups, or small molecules in reactants or reagents. This capability is essential for optimizing reactions in complex chemical systems, where small adjustments to reactants and reagents can significantly impact yields. log-RRIM$_b$'s precision in predicting the effects of these subtle changes enhances its utility for guiding synthetic strategies and fine-tuning reactions. By offering reliable forecasts for small modifications, log-RRIM$_b$ can potentially streamline the optimization process, reducing the number of experimental iterations required and saving time and resources in research and industrial settings.

This capability stems from log-RRIM$_b$'s unique local-to-global learning strategy. By first analyzing each molecule separately and then modeling their interactions, the model ensures equal consideration of all molecules, regardless of their size. This approach differs from sequence-based models like T5Chem, which process the entire reaction SMILES string simultaneously. Such models may overlook crucial smaller fragments that significantly impact the overall yield, as the global attention mechanisms might not be sufficiently sensitive to these critical molecular fragments.

Overall, the performance differences presented in these analyses underscore our belief that model frameworks should be carefully designed based on specific task characteristics rather than solely relying on foundation models. While they have shown great promise in many areas, a basic fine-tuning strategy may not always be optimal for specialized tasks like reaction yield prediction. Such an approach lacks task-specific module designs that capture the intricate characteristics of chemical reactions, potentially limiting the performance.

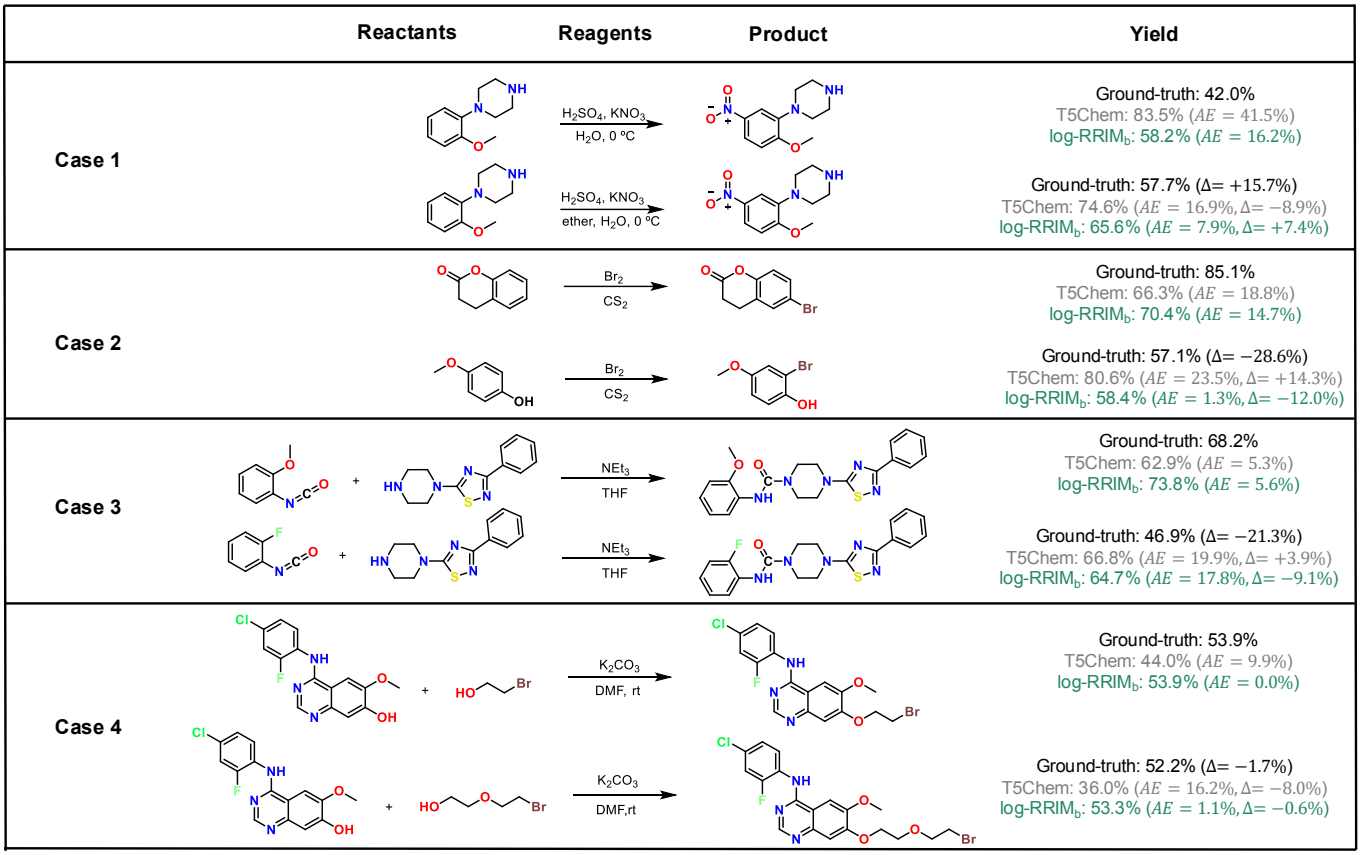


Fig. 4 | Cases analysis on the USPTO500MT dataset. Each reaction is reported with reactants, reagents, products, and the ground-truth and predicted yields by T5Chem and log-RRIM$_b$. *AE* in parentheses represents the Absolute Error between the predicted and ground-truth yields. Δ in parentheses represents the change of the ground-truth and predicted yields in the second reaction to the corresponding value in the first reaction.

Performance on the external dataset CJHIF

To assess our model's performance on external datasets, we conducted an evaluation using a subset of the CJHIF dataset.[27] This approach involves using models trained on USPTO500MT and testing them on a subset of the CJHIF dataset, which comprises 3,219,165 reactions sourced from high-impact factor journals. Our assessment involved 1,000 zero-yielding chemical reactions randomly selected from the initial 50,000 reactions in the CJHIF dataset. We specifically chose reactions with reported non-zero yields because CJHIF treats unreported yields as zeros, and we aimed to evaluate our model on reactions with confirmed, measurable outcomes. Importantly, these 1,000 reactions are not included in the training or testing data of USPTO500MT, thus providing an independent testing set for assessing our model's performance on external reactions.

Overall, log-RRIM$_b$ achieved an MAE of 0.149, representing a 16.8% improvement over T5Chem's MAE of 0.179. The results of analyzing performance across yield ranges are illustrated in Figure 5. log-RRIM$_b$ significantly outperformed T5Chem for reactions with yields between 60% to 100% (confidence level 95%, more details are provided in Table A2). This superior performance aligns closely with our observations from the USPTO500MT dataset, particularly in log-RRIM$_b$'s enhanced accuracy for medium to high-yielding reactions, which suggests that log-RRIM$_b$'s improved predictive power for high-yielding reactions is a generalizable feature, not limited to a specific dataset. We attribute this generalizability to log-RRIM$_b$'s molecular interaction design which uses the cross-attention mechanism to effectively model the function of reagents in relation to the reaction center. This allows log-RRIM$_b$ to learn fundamental principles about how reagents impact bond-breaking and formation, which are key factors affecting reaction yield. The extensive data in USPTO500MT training data enables log-RRIM$_b$ to learn such principles to achieve better test performance on external datasets.

To further validate that log-RRIM has effectively learned key factors influencing reaction yield, we visualized the contribution (weight) of each atom when log-RRIM$_b$ aggregates atom embeddings and constructs the molecule representation. Three exemplar reactions are shown in Figure 6. In reaction **A**, a sulfonylation reaction, the sulfur-bearing sulfonyl chloride group on the *p*-toluenesulfonyl chloride and the free hydroxyl (OH) group on the alcohol are the two reacting centers. The oxygen (O) acts as the nucleophile that displaces the chlorine (Cl) atom, and these atoms influence the yield of the reaction. Reaction **B** is an imine reduction reaction of the compound N-(4-methoxyphenyl)-1-phenylethylamine. The polar C=N bond between the Nitrogen (N) and Carbon (C) is the reactive site, and these two atoms influence the yield of the reaction, which results in the single C-N bond in the corresponding amine. In reaction **C**, the two atoms that ultimately influence the yield are Sulfur (S) of benzene sulfonyl chloride and Nitrogen (N) of the indole, producing the final compound. Combined with the weights highlighted by the colormap in Figure 6, we found that the atoms mentioned above that have a greater impact on yield are given higher weights by log-RRIM$_b$, and these atoms are also the atoms in the reaction center. This finding aligns with the fundamental chemical principle that reaction center atoms play a crucial role in the bond-breaking and bond-forming steps in the transition state, thereby exerting substantial influence on the

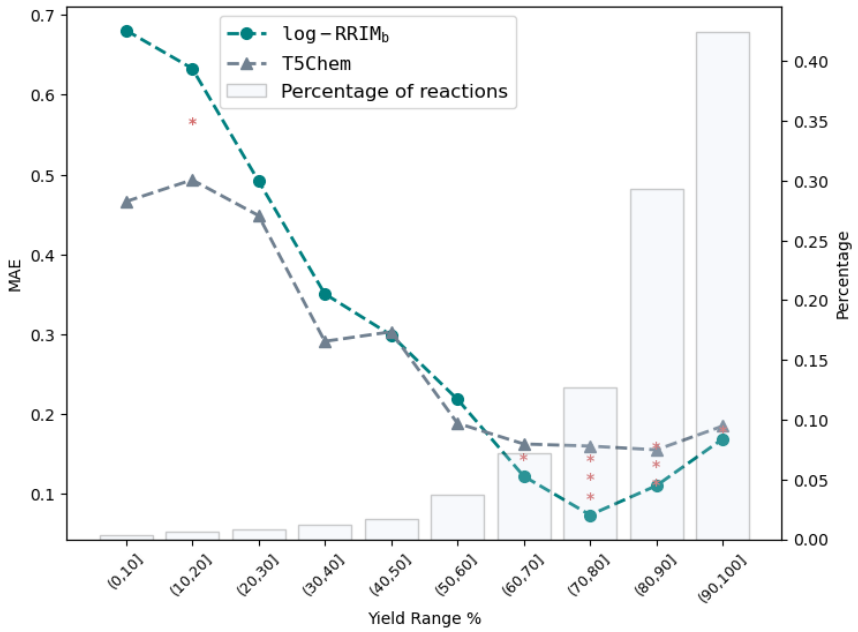


Fig. 5 | log-RRIM$_b$ and T5Chem performance comparison over each yield range on a subset of CJHIF. Left y-axis: MAE of predicted yields. Right y-axis: percentage of reactions in the testing set for each yield range. 5% significance level: * for p-values < 0.05, ** for p-values < 0.005, *** for p-values < 0.0005.

yield. The ability of log-RRIM$_b$ to prioritize these critical atoms in learning the molecule representation is essential for building more accurate models for predicting reaction yield, and our method demonstrates particular effectiveness in this regard.

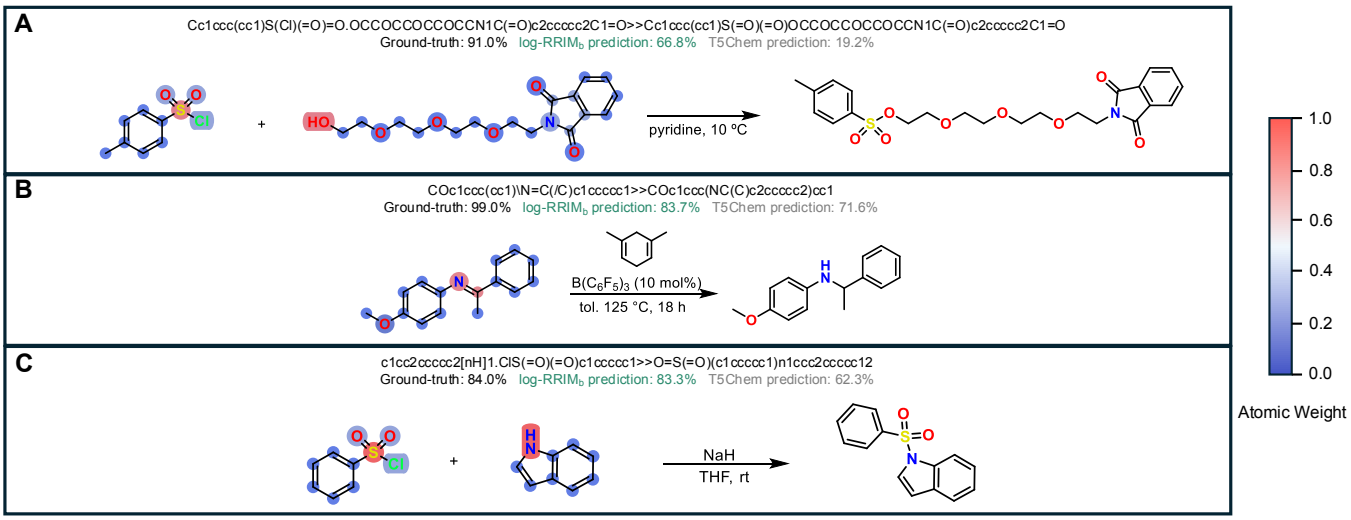


Fig. 6 | Visualizations of atom contribution in learning molecule representation. The contribution is quantified by the color and the three exemplar reactions are selected from the CJHIF dataset.

Performance on the Buchwald-Hartwig dataset

On the Buchwald-Hartwig dataset, we conducted a performance comparison among pre-training-free models (YieldGNN, SEMG-MIGNN, and RD-MPNN), using 10-fold cross-validation,[47] and reported the testing results averaged over the 10 folds. Table 3 reports the mean and standard deviation (in parentheses) of MAE, RMSE, and R^2. Table A4 provides a detailed comparison of testing MAE values for different models on each fold. Our method, log-RRIM$_b$, outperforms other pre-training-free (also graph-based) methods across all evaluation metrics (MAE 0.0348, RMSE 0.0544, and R^2 0.953). Notably, it achieves a 14.7% improvement in MAE over the best-performing baseline, YieldGNN. We attribute log-RRIM$_b$'s superior performance to its more effective molecular interaction design, explicitly modeling the reagents' function to the reaction center. By incorporating this design, log-RRIM$_b$ captures crucial chemical insights that other methods may overlook, leading to more accurate predictions. Compared to the second-best baseline method, SEMG-MIGNN, log-RRIM$_b$ improves the MAE by 17.9%. To put this improvement in context, it's worth recalling that SEMG-MIGNN focuses on building more informative atom features (digitalized steric and electronic information). In contrast, log-RRIM$_b$ emphasizes learning the characteristics of the reaction itself and molecular interactions. The performance difference between these approaches suggests that for yield prediction tasks, the latter strategy may be more effective. In summary, log-RRIM$_b$

outperforms other pre-training-free and graph-based models substantially. These results demonstrate the importance of focusing on reaction characteristics and molecular interactions in yield prediction tasks.

As shown in Table 4, when compared to the pre-training-based methods, which are also sequence-based models (T5Chem and YieldBERT), log-RRIM$_l$ also shows competitive performance by achieving the MAE of 0.0347, which has a 16.4% improvement over YieldBERT but inferior to T5Chem by 11.6%. These results indicate our method is comparable to the best-performing sequence-based model T5Chem while using only 2% of the pre-training dataset size compared to T5Chem (2M vs 97M). To further validate this comparison, we conducted statistical analysis on the predicted values of log-RRIM$_l$ and T5Chem for each yield range on the first data split (with detailed p-values shown in Table A3). The analysis shows that, for all ranges except the 10%-20%, 40%-50%, and 70%-80% ranges (a total of 27.6% of the test reactions), the differences between log-RRIM$_l$ and T5Chem are not statistically significant at the 95% confidence interval. This indicates that the performance of log-RRIM$_l$ and T5Chem is largely comparable across most yield ranges. Note that the Buchwald-Hartwig dataset involves only a single reaction type with limited components. On this specific reaction type, log-RRIM could underperform T5Chem. However, real-world chemical synthesis generally involves reactions of multiple types. Thus, methods that could accurately predict yields of various types are highly demanded. As shown on the USPTO500MT and CJHIF datasets, which contain numerous reaction types, our method demonstrates superior performance. On these more diverse and complex datasets, log-RRIM$_b$ outperforms T5Chem. These results demonstrate the potential superior utility of log-RRIM$_b$ over T5Chem in real-world chemical synthesis applications.

We also note that the performance of log-RRIM$_b$ and log-RRIM$_l$ are nearly identical (MAE 0.0348 vs 0.0347) on the Buchwald-Hartwig dataset. This observation aligns with the results we obtained on the USPTO500MT dataset. These consistent findings across different datasets suggest that effective molecular interaction modeling may play a more crucial role than using pre-trained models to generate informative atom representations in yield prediction tasks.

Table 3 | Pre-training-free models performance comparison on the Buchwald-Hartwig dataset

Method	Metrics		
	MAE	**RMSE**	**R^2**
RD-MPNN	0.0746(0.005)	0.1040(0.007)	0.854(0.018)
SEMG-MIGNN	0.0424(0.001)	0.0605(0.002)	0.951(0.004)
YieldGNN	0.0408(0.002)	0.0575(0.002)	**0.956(0.003)**
log-RRIM$_b$	**0.0348(0.002)**	**0.0544(0.004)**	0.953(0.009)

Each value is the mean and standard deviation (in parentheses), averaging 10 folds. The best performance is highlighted in bold.

Table 4 | Pre-training-based models performance comparison on the Buchwald-Hartwig dataset

Method	Metrics		
	MAE	**RMSE**	**R^2**
YieldBERT	0.0415(0.001)	0.0641(0.005)	0.945(0.008)
T5Chem	**0.0311(0.001)**	**0.0482(0.002)**	**0.971(0.002)**
log-RRIM$_l$	0.0347(0.001)	0.0528(0.003)	0.957(0.006)

Each value is the mean and standard deviation (in parentheses), averaging 10 folds. The best performance is highlighted in bold.

Discussion

In conclusion, in this paper, we present log-RRIM, a novel graph-transformer-based reaction representation learning framework for yield prediction. log-RRIM leverages a local-to-global representation learning process and incorporates a cross-attention mechanism to model reagent-reaction center interactions, facilitating improved capture of small fragment contributions and interactions between reactant and reagent molecules. This approach allows log-RRIM to tap into crucial aspects of chemical knowledge, particularly the importance of reagent effects and reaction center dynamics in determining reaction outcomes. Without reliance on pre-training tasks, log-RRIM demonstrates superior accuracy and effectiveness compared to other graph-based methods and state-of-the-art sequence-based approaches, particularly for medium to high-yielding reactions. Our analyses further show log-RRIM's advanced modeling of reactant-reagent interactions and sensitivity to small molecular fragments, making it a valuable asset for reaction planning and optimization in chemical synthesis.

The log-RRIM framework requires that predicted reactions consist of three parts (reactant, reagent, and product) and that reaction center atoms be correctly identifiable. While this may limit its practical applications in some scenarios, it enables log-RRIM to more effectively model the crucial intermolecular dynamics that significantly influence reaction outcomes. This approach underscores the importance of incorporating chemical-specific information into model architecture design, rather than directly adapting general-purpose foundation models for chemical tasks like yield prediction.

While log-RRIM makes significant strides in leveraging chemical knowledge, particularly in modeling reagent-reaction center interactions, there remains a vast body of chemical expertise that could potentially be incorporated to further enhance the performance. For instance, research has elucidated detailed mechanisms for different reaction types, like transition states,[48,49] which are not yet explicitly incorporated into our model. Furthermore, chemists have developed a deep understanding of the relative reactivity of different functional groups under various conditions,[50,51] which represents another rich source of knowledge that could be integrated into the model. Incorporating such additional aspects of chemical knowledge presents both a challenge and an opportunity for future research. It could potentially enhance the

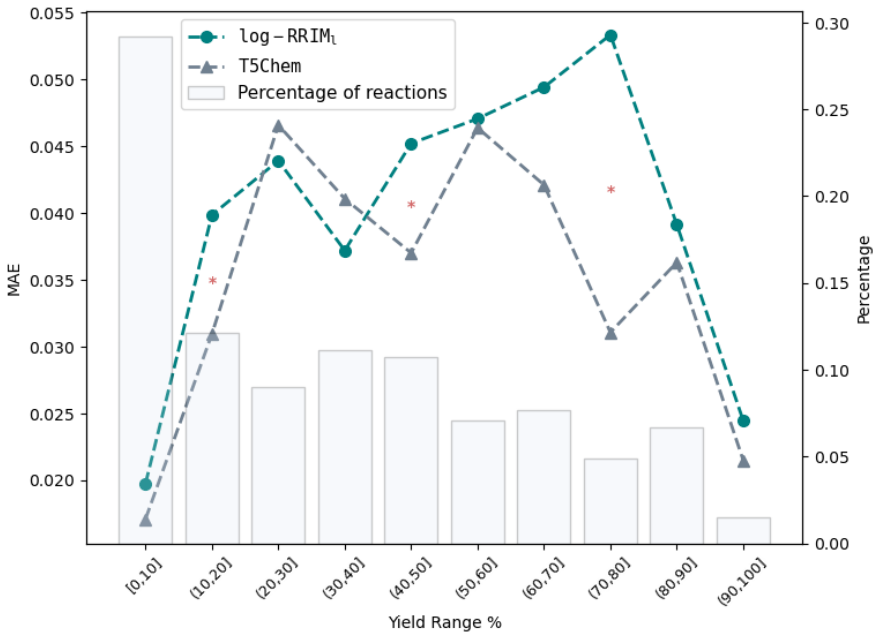


Fig. 7 | Performance comparison of log-RRIM$_l$ and T5Chem across yield ranges on the first data split of the Buchwald-Hartwig dataset. Left y-axis: MAE of predicted yields. Right y-axis: percentage of reactions in the testing set for each yield range. 5% significance level: * for p-values < 0.05, ** for p-values < 0.005, *** for p-values < 0.0005.

model's predictive power, improve its generalization to diverse reaction types, and provide more interpretable insights into the factors driving yield predictions. Another promising direction for future research is the exploration of multi-task learning approaches, where the model could be trained simultaneously on yield prediction, reaction condition optimization, retrosynthesis planning, etc. This could lead to a more comprehensive understanding of chemical reactivity and potentially improve performance across all tasks.

log-RRIM represents a significant step forward in reaction yield prediction by leveraging graph-based representations and modeling reagent-reaction center interactions, and there is still room for further integration of chemical knowledge and enhancement of the model's capabilities. By continuing to merge data-driven techniques with established chemical principles, it is crucial to develop more robust, versatile, and reliable models for computational chemistry.

Data Availability

The data used in this manuscript is made publicly available at https://github.com/ninglab/Yield_log_RRIM.

Code Availability

The code for log-RRIM is made publicly at https://github.com/ninglab/Yield_log_RRIM.

Acknowledgements

This project was made possible, in part, by support from the National Science Foundation grant no. IIS-2133650 (X.N.) and the National Library of Medicine grant no. 1R01LM014385-01 (X.N., D.A.). Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

Method

Our method, log-RRIM, is a novel local-to-global graph-transformer-based reaction representation learning and molecular interaction modeling for yield prediction. It employs a local-to-global learning process for reaction representation learning, beginning with molecule (reactants, reagents, and product) representation learning. It subsequently models the molecule interactions (between reactants and reagents) and ultimately represents the entire reaction. log-RRIM further uses the reaction representation to predict yield.

Specifically, log-RRIM consists of the following three modules: **(1)** Molecule Representation Learning (MRL) module: which uses graph transformers[39] with multi-head self-attention layers to encode molecular structural information into atom embeddings, and then aggregate atom embeddings into molecule embeddings through Atomic Integration (AI). **(2)** Molecule Interaction (MIT) module: which learns the interactions between reactants and reagents through the cross-attention mechanism, resulting in interaction-aware embeddings for reaction centers. **(3)** Reaction Information Aggregation (RIA) module: which employs Molecular Integration (MI) to derive a comprehensive reaction representation from all involved molecules and their interaction representations. Finally, this reaction representation is utilized to predict the yield. An overview of log-RRIM is depicted in Figure 8.

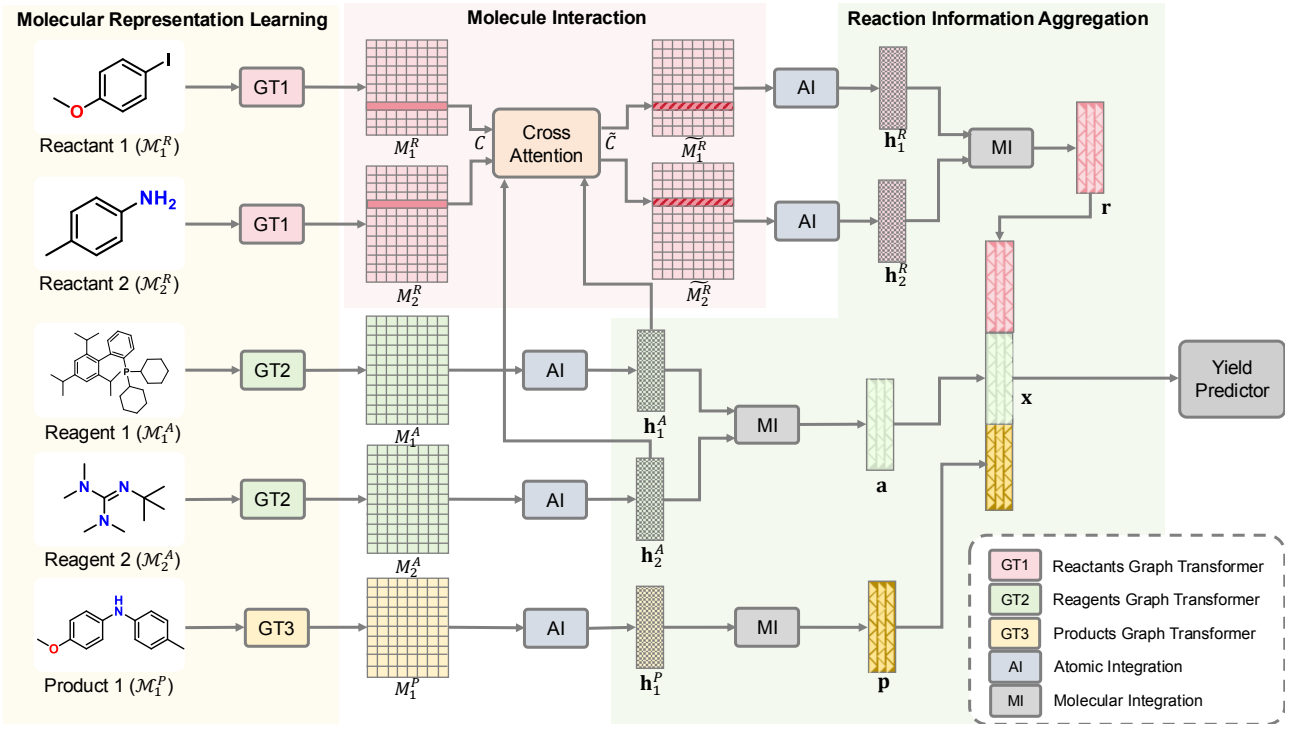


Fig. 8 | Pipeline of log-RRIM

Notations

In a reaction $\mathcal{X}$, each reactant, reagent, and product is a molecule. We view each molecule $\mathcal{M}$ as a graph, with basic node (atom) features $I \in \mathbb{R}^{n\times s}$, graph adjacent matrix $J \in \{0,1\}^{n\times n}$, and inter-atomic distance matrix $D \in \mathbb{R}^{n\times n}$, where n is the number of atoms in the molecule and can be different for each molecule, s is the dimension of basic atom features. The reaction $\mathcal{X}$ is represented as $(\mathcal{R}, \mathcal{A}, \mathcal{P}, y)$, where $\mathcal{R} = \{\mathcal{M}_1^R, \ldots, \mathcal{M}_{n_r}^R\}$, $\mathcal{A} = \{\mathcal{M}_1^A, \ldots, \mathcal{M}_{n_a}^A\}$, and $\mathcal{P} = \left\{\mathcal{M}_1^P, \ldots, \mathcal{M}_{n_p}^P\right\}$ are the set of n_r reactants, n_a reagents and n_p products in the reaction, and y is the reaction yield. n_r, n_a, and n_p can be different for each reaction. Notably, we denote the reaction center atom embeddings in reactants as $C \in \mathbb{R}^{|C|\times d}$, where $|C|$ refers to the number of reaction center atoms.

In MRL module, the atom embeddings of each molecule after the $l \in [1, n_l]$-th self-attention layer is denoted as $M(l) \in \mathbb{R}^{n\times d}$, where n_l is the number of self-attention layers used in MRL and d is the model dimension.

$\mathbf{h}_m \in \left\{\mathbf{h}_1^R, \ldots, \mathbf{h}_{n_r}^R, \mathbf{h}_1^A, \ldots, \mathbf{h}_{n_a}^A, \mathbf{h}_1^P, \ldots, \mathbf{h}_{n_p}^P\right\}$ is a d-dimension vector and denoted as the representation of each molecule in reactants, reagents, and products. The reactant, reagent, and product representations are named as $\mathbf{r} \in \mathbb{R}^d$, $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{p} \in \mathbb{R}^d$. The representation of the whole reaction is $\mathbf{x}$ and the predicted yield is denoted by $\hat{y}$. We summarize the key notations in Table 5. We use uppercase letters to denote matrices, lowercase bold letters to denote row vectors, and lower-case non-bold letters to represent scalars.

Molecule Representation Learning (MRL)

Given a reaction consisting of molecules represented in graphs, we first employ the Molecule Attention Transformer (MAT) [39] to learn the molecule representations. MAT recursively propagates information between atoms to learn the

Table 5 | Key Notations

Notation	Description
$\mathcal{X}$	Reaction
$\mathbf{x}$	Reaction representation
y, $\hat{y}$	Ground truth reaction yield and the predicted yield
$\mathcal{R}$, $\mathcal{A}$, $\mathcal{P}$	Reactant, reagent and product
$\mathbf{r}, \mathbf{a}, \mathbf{p}$	Reactant, reagent, and product representation
n_r, n_a, n_p	Number of molecules in the reactant, reagent, and product
$\mathcal{M}$	Molecule
n	Number of atoms of the molecule
M	Atom embeddings of the molecule
$\mathbf{h}_m$	Molecule representation

structural information of molecules via multi-head molecule self-attention layers as follows:

$$\begin{aligned} Q_j^l = M(l-1)E_j^l, K_j^l = M(l-1)F_j^l, V_j^l = M(l-1)G_j^l, \\ \text{HEAD}_j^l = \left(\lambda_a \rho\left(\frac{Q_j^l \left(K_j^l\right)^T}{\sqrt{d}}\right) + \lambda_d \rho(D) + \lambda_g J\right) V_j^l, \\ H^l = \left[\text{HEAD}_1^l, \text{HEAD}_2^l, \ldots, \text{HEAD}_{n_h}^l\right] O^l, \end{aligned} \tag{5}$$

where $M(l-1)$ is the atom embeddings from the $(l-1)$-th molecule self-attention layer and $M(0) = I$ is the input of first layer ; Q_j^l, K_j^l, and V_j^l are the query, key, and value matrix derived from $M(l-1)$ with learnable parameters E_j^l, F_j^l and $G_j^l \in \mathbb{R}^{d \times \frac{d}{n_h}}$; $\lambda_a, \lambda_d, \lambda_g$ are the scalars to balance the importance of the self-attention, distance matrix, and adjacency matrix, and ρ is the softmax function. Each molecule attention layer has n_h heads and HEAD_j^l is the output from the j-th attention head; $O^l \in \mathbb{R}^{d \times d}$ is a learnable matrix to integrate the attention heads. After each molecule self-attention layer, MAT includes a feed-forward layer to introduce non-linearity which is a fully connected network (FCN) described below:

$$M(l) = \sigma\left(H^l W^l + B^l\right), \tag{6}$$

where $W^l \in \mathbb{R}^{d \times d}$ and $B^l \in \mathbb{R}^{n \times d}$ are learnable parameters, $\sigma(\cdot)$ is ReLU[52] activation function. After n_l molecule self-attention layers, the molecule's structural information is encoded into the atom embeddings $M(n_l)$. When no ambiguity arises, for simplicity, we eliminate n_l for $M(n_l)$ and only use M to represent atom embeddings of molecule $\mathcal{M}$ as the output of the last molecule self-attention layer.

Compared to the original Transformer,[38] MAT integrates the interactions among atoms, the geometric information in the molecule, and the topology of the molecule to better learn expressive atom embeddings, and captures the structural information of the molecule. Given the atom embeddings M learned from MAT, we utilize the Atomic Integration(AI) module to aggregate atom embeddings and generate the molecule representation $\mathbf{h}_m$. Particularly, AI uses a gating mechanism to capture the importance of different atoms in the aggregation as follows:

$$\begin{aligned} \boldsymbol{\alpha} &= M\mathbf{w}_1, \\ \mathbf{h}_m &= \sum_{k=1}^{n} [M]_k \times [\boldsymbol{\alpha}]_k, \end{aligned} \tag{7}$$

where $\mathbf{w}_1 \in R^d$ is a learnable vector and $\boldsymbol{\alpha}$ is the vector where each element represents the contribution of each atom embedding to the molecule representation.

Additionally, in the MRL module, AI is only performed on reagents and products to get their molecule representations and is omitted for reactants. This is because the reaction center atom embeddings in the reactants will undergo further updates in the Molecule Interaction(MIT) module. The reactant molecule representations will be obtained through AI afterward.

Molecule Interaction (MIT)

Reagents, such as catalysts, significantly impact reaction yield by promoting or inhibiting bond breaking and formation. We explicitly model their function to the reaction center atoms to better capture the interaction between reactants and reagents. Specifically, given the reaction center atom embeddings $C \in \mathbb{R}^{|C| \times d}$ in reactant molecules, and the reagent molecule representations $\mathbf{h}_i^A \in \left\{\mathbf{h}_1^A, \ldots, \mathbf{h}_{n_a}^A\right\}$, we update the reaction center atom embeddings by applying a multi-head

cross-attention mechanism, described as follows:

$$
\begin{aligned}
&Q_j = CW_j^Q, K_j = H^A W_j^K, V_j = H^A W_j^V, \\
&\mathrm{HEAD}_j = \rho\left(\frac{Q_j\left(K_j\right)^T}{\sqrt{d}}\right) V_j, \\
&H = \left[\mathrm{HEAD}_1, \mathrm{HEAD}_2, \ldots, \mathrm{HEAD}_{n_h}\right] O,
\end{aligned}
\tag{8}
$$

where Q_j is the linear projection of reaction center atom embeddings C; K_j, and V_j are the linear projection of the concatenated molecule representations of reagents $H^A = [\mathbf{h}_1^A, \ldots, \mathbf{h}_{n_a}^A] \in \mathbb{R}^{n_a \times d}$. A cross attention layer has n_h attention heads and HEAD_j is the output from j-th attention head, $O \in \mathbb{R}^{d \times d}$ is a learnable parameter to integrate the attention heads. The updated reaction center atom embeddings $\tilde{C}$ are obtained by passing H to an FCN:

$$
\tilde{C} = \sigma\left(HW^c + B^c\right), \tag{9}
$$

where $W^c \in \mathbb{R}^{d \times d}$ and $B^c \in \mathbb{R}^{|C| \times d}$ are learnable parameters. After updating the reaction center atom embeddings in the reactants, we use AI to derive the reactant molecule representations $\mathbf{h}_i^R \in \left\{\mathbf{h}_1^R, \ldots, \mathbf{h}_{n_r}^R\right\}$.

MIT uses a cross-attention layer to transform and integrate reagent information into the reaction center atoms, enabling the model to consider relationships between various reaction components. This makes log-RRIM learn a more chemically meaningful reaction representation by emphasizing reaction centers and reagent interactions. We further show and analyze the benefits that MIT brings to log-RRIM in Table A5, demonstrating its contribution to the overall performance of our model.

Reaction Information Aggregation (RIA)

After the derivation of representations for all the molecules involved in reactants, reagents, and products, we introduce RIA to aggregate all the molecular information. This module explicitly describes the interaction of the involved molecules in the reaction and their contribution to yield.

Specifically, given the reactant molecule representations $\mathbf{h}_i^R \in \left\{\mathbf{h}_1^R, \ldots, \mathbf{h}_{n_r}^R\right\}$, reagent molecule representations $\mathbf{h}_i^A \in \left\{\mathbf{h}_1^A, \ldots, \mathbf{h}_{n_a}^A\right\}$, and the product molecule representations $\mathbf{h}_i^P \in \left\{\mathbf{h}_1^P, \ldots, \mathbf{h}_{n_p}^P\right\}$, we first apply MI to respectively derive three representations $\mathbf{r}$, $\mathbf{a}$, and $\mathbf{p}$ for reactant, reagent, and product. MI uses a gating mechanism to aggregate the information from involved molecules. Taking reactant molecules as an example, this process can be described as follows:

$$
\begin{aligned}
&\beta_i = \langle \mathbf{h}_i^R, \mathbf{w}_2 \rangle \\
&\mathbf{r} = \sum_{i=1}^{n_r} \mathbf{h}_i^R \times \beta_i,
\end{aligned}
\tag{10}
$$

where $\mathbf{w}_2 \in R^d$ is a learnable vector to map each molecule representation to its weight β_i. This step allows log-RRIM to capture the collective properties within each group of molecules, providing a more compact and informative representation for subsequent processing. These three representations are then concatenated to form a comprehensive representation of the entire reaction $\mathbf{x} = [\mathbf{r}, \mathbf{a}, \mathbf{p}] \in R^{3d}$. $\mathbf{x}$ then serves as the input for the yield predictor.

RIA processes reactant, reagent, and product molecules separately and aggregates information hierarchically to achieve a nuanced representation of the reaction. This design allows log-RRIM to capture each component's unique role and contribution to the reaction process, leading to a nuanced overall representation.

Yield Predictor

Provided with the comprehensive reaction representation $\mathbf{x}$, we stack two FCNs to predict the yield $\hat{y}$. The process is described below:

$$
\hat{y} = (\mathbf{x}) = f\left(\sigma\left(\mathbf{x}W_3 + \mathbf{b}_1\right)\mathbf{w}_4 + b_2\right), \tag{11}
$$

where $W_3 \in \mathbb{R}^{3d \times d}$, $\mathbf{w}_4 \in \mathbb{R}^d$, $\mathbf{b}_1 \in \mathbb{R}^d$ and $b_2 \in \mathbb{R}^1$ are learnable parameters, and $f(\cdot)$ is a sigmoid function to control the predicted yield within the range $[0\%, 100\%]$.

Model training and hyperparameters optimization

During training, the mean absolute error (MAE) loss is optimized using adaptive moment estimation (Adam).[53]

$$
\mathrm{MAE} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N} \tag{12}
$$

The initial learning rate is treated as a hyperparameter. Additionally, we utilize the validation set to schedule the learning rate decay patience and decay factor required in *lr_scheduler.ReduceLROnPlateau* provided by PyTorch.[54] All the searched hyperparameters and their respective search ranges are summarized in Table A6 and Table A7, respectively.

Appendix

Exact paired t-test p-values between log-RRIM and T5Chem on each dataset

The tables below include all the MAE differences and p-values of the paired t-tests between log-RRIM$_b$ and T5Chem.

Table A1 | log-RRIM$_b$ and T5Chem performance comparison across different yield ranges on USPTO500MT

Yield range	**[0%,10%]**	**[10%,20%]**	**[20%,30%]**	**[30%,40%]**	**[40%,50%]**	**[50%,60%]**	**[60%,70%]**	**[70%,80%]**	**[80%,90%]**	**[90%,100%]**
MAE difference (log-RRIM$_b$- T5Chem)	0.104	0.095	0.051	0.032	-0.020	-0.030	-0.048	-0.050	-0.021	0.008
T-test p-values	1e-10	2e-17	2e-1	1e-6	5e-5	3e-12	7e-40	3e-40	6e-8	3e-2
Cumulative yield range	**[90%, 100%]**	**[80%, 100%]**	**[70%, 100%]**	**[60%, 100%]**	**[50%, 100%]**	**[40%, 100%]**	**[30%, 100%]**	**[20%, 100%]**	**[10%, 100%]**	**[0%, 100%]**
MAE difference (log-RRIM$_b$- T5Chem)	0.008	-0.006	-0.018	-0.024	-0.025	-0.024	-0.020	-0.015	-0.011	-0.009
T-test p-values	3e-2	3e-2	4e-17	3e-37	2e-46	3e-50	4e-35	1e-22	4e-12	1e-8

Table A2 | log-RRIM$_b$ and T5Chem performance comparison across different yield range on 1,000 sampled non-zero-yielding reactions from CJHIF

Yield range	**(0%,10%]**	**(10%,20%]**	**(20%,30%]**	**(30%,40%]**	**(40%,50%]**	**(50%,60%]**	**(60%,70%]**	**(70%,80%]**	**(80%,90%]**	**(90%,100%]**
MAE difference (log-RRIM$_b$- T5Chem)	0.215	0.140	0.043	0.059	-0.004	0.030	-0.040	-0.086	-0.045	-0.016
T-test p-values	2e-1	5e-2	4e-1	3e-1	9e-1	2e-1	8e-3	3e-14	4e-7	4e-2

Table A3 | log-RRIM$_l$ and T5Chem performance comparison across different yield ranges on the first data split of Buchwald-Hartwig dataset

Yield range	**[0%,10%]**	**(10%,20%]**	**(20%,30%]**	**(30%,40%]**	**(40%,50%]**	**(50%,60%]**	**(60%,70%]**	**(70%,80%]**	**(80%,90%]**	**(90%,100%]**
MAE difference (log-RRIM$_l$- T5Chem)	0.0027	0.0089	-0.0027	-0.0039	0.0082	0.0007	0.0073	0.0223	0.0028	0.0030
T-test p-values	6e-2	1e-2	5e-1	3e-1	3e-2	9e-1	2e-1	2e-2	5e-1	6e-1

Performance on each Buchwald-Hartwig data split

In Table A4, we present the performance (MAE on the testing set) of each model across every data split.

Impact of explicit reagents function modeling

We conducted an ablation study to investigate the importance of explicitly modeling reagent effects on reaction yield by removing the second module Molecular Interaction (MIT) from our proposed log-RRIM$_b$ framework. In this modified version, atom embeddings of the reactant molecules are directly passed to the Atomic Interaction (AI) module to derive reactant molecule representations without using the cross-attention mechanism to update the reaction center atom embeddings. We compared the performance of this ablated model with the original log-RRIM$_b$ on the first data split of the Buchwald-Hartwig dataset. The results are summarized in Table A5. Results show that removing the MIT module hugely decreased prediction performance, with the MAE of log-RRIM$_b$ increasing by 45.0%. This substantial drop in accuracy underscores the critical role of explicitly modeling reagent effects in yield prediction tasks. The observed performance degradation can be attributed to the MIT module's ability to capture fundamental characteristics of chemical reactions, particularly the influence of substances such as catalysts on bond breaking and formation at reaction centers. By incorporating this knowledge, log-RRIM enhances its capacity to construct more informative molecular and reaction representations, ultimately making more accurate reaction yield predictions.

Hyperparameters

Table A6 and Table A7 summarizes the searched hyperparameters and their ranges on the USPTO500MT and Buchwald-Hartwig datasets. The selected values for log-RRIM$_b$ are highlighted by underlining, while those for log-RRIM$_l$ are indicated in bold. Table A8 summarizes the hyperparameters used in the pre-trained MAT model.

Table A4 | Model performance over each data split on Buchwald-Hartwig

Method	Data split										
	1	2	3	4	5	6	7	8	9	10	mean(std)
YieldBERT	0.0424	0.0425	0.0408	0.0410	0.0417	0.0411	0.0401	0.0424	0.0403	0.0432	0.0416(0.001)
T5Chem	**0.0323**	**0.0311**	**0.0311**	**0.0314**	**0.0303**	**0.0297**	**0.0315**	**0.0332**	**0.0298**	**0.0309**	**0.0311(0.001)**
RD-MPNN	0.0758	0.0698	0.0694	0.0758	0.0802	0.0726	0.0727	0.0866	0.0697	0.0734	0.0746(0.005)
SEMG-MIGNN	0.0440	0.0418	0.0397	0.0432	0.0439	0.0418	0.0433	0.0426	0.0410	0.0424	0.0424(0.001)
YieldGNN	0.0423	0.0415	0.0391	0.0397	0.0410	0.0418	0.0386	0.0394	0.0406	0.0439	0.0408(0.002)
log-RRIM$_b$	0.0338	0.0339	0.0331	0.0364	0.0360	0.0328	0.0344	0.0379	0.0344	0.0352	0.0348(0.002)
log-RRIM$_l$	0.0363	0.0342	0.0326	0.0371	0.0346	0.0340	0.0338	0.0364	0.0338	0.0343	0.0347(0.001)

Each value is the model's MAE on the corresponding testing set, the best performance is highlighted in bold. The mean and standard deviation (in parentheses) are obtained by averaging across 10 data splits.

Table A5 | log-RRIM$_b$ without MIT performance on the first data split of the Buchwald-Hartwig dataset

Method	Metrics		
	MAE	**RMSE**	**R^2**
log-RRIM$_b$ without MIT	0.0490	0.0651	0.931
log-RRIM$_b$	**0.0338**	**0.0530**	**0.957**

Table A6 | Hyperparameters and their searched ranges on the USPTO500MT dataset

Name	Description	Range
eb_Nlayer	The number of pre-trained self-attention layers to initialize atom features	0, **8**
Nlayer	The number of self-attention layers	4, **5**, 6
Nheads	The number of attention heads	**16**
hs	Model dimension	128, **256**, 512, 1024
e	Epoch	**35**
bs	Batch size	**32**
init_lr	Initial learning rate	**1e-5**, 3e-5, 1e-4
lr_decay_step	Learning rate decay patience	5, 10, **15**
lr_decay_factor	Learning rate decay factor	0.85, 0.90, **0.95**
dp	Dropout	**0**, 0.1, 0.2
gnorm	Gradient norm clipping threshold	0.5, 1, 5, **None**
wd	Weight Decay	**0**, 1e-6, 1e-5

Table A7 | Hyperparameters and their searched ranges on the Buchwald-Hartwig dataset

Name	Description	Range
eb_Nlayer	The number of pre-trained self-attention layers to initialize atom features	0, **8**
Nlayer	The number of self-attention layers	1, 2, 3, 4, **5**, 6, 7, 8
Nheads	The number of attention heads	**16**
hs	Model dimension	128, **256**, 512, 1024
e	Epoch	**300**
bs	Batch size	**32**
init_lr	Initial learning rate	1e-5, **3e-4**, 1e-4, 3e-4
lr_decay_step	Learning rate decay patience	10, 15, **20**
lr_decay_factor	Learning rate decay factor	0.85, **0.90**, 0.95
dp	Dropout	**0**, 0.1, 0.2
gnorm	Gradient norm clipping threshold	0.5, 1, 5, **None**
wd	Weight Decay	0, 1e-6, **1e-5**

Table A8 | Hyperparameters used in the pre-trained MAT model

Name	Description	Value
eb_Nlayer	The number of pre-trained MAT self-attention layers	8
hs_pretrain	Pre-trained MAT model dimension	1024
Nheads_pretrain	Pre-trained MAT attention heads number	16
Npff	The number of dense layers in the position-wise feed-forward block	1
k	Distance matrix kernel	'EXP'
dp_pretrain	Dropout	0
wd_pretrain	Weight decay	0
$\lambda_a, \lambda_d, \lambda_g$	The scalars weighting the naive self-attention, distance, and adjacency matrices	0.33

References

[1] Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).

[2] Reizman, B. J. & Jensen, K. F. An automated continuous-flow platform for the estimation of multistep reaction kinetics. *Organic Process Research & Development* **16**, 1770–1782 (2012).

[3] Sigman, M. S., Harper, K. C., Bess, E. N. & Milo, A. The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Accounts of chemical research* **49**, 1292–1301 (2016).

[4] Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2**, 015016 (2021).

[5] Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital discovery* **1**, 91–97 (2022).

[6] Lu, J. & Zhang, Y. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling* **62**, 1376–1387 (2022).

[7] Chen, X. *et al.* Sequence-based peptide identification, generation, and property prediction with deep learning: a review. *Molecular Systems Design & Engineering* **6**, 406–428 (2021).

[8] Li, J. & Jiang, X. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing* **2021**, 7181815 (2021).

[9] Saebi, M. *et al.* On the use of real-world datasets for reaction yield prediction. *Chemical science* **14**, 4997–5005 (2023).

[10] Li, S.-W., Xu, L.-C., Zhang, C., Zhang, S.-Q. & Hong, X. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nature Communications* **14**, 3569 (2023).

[11] Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5**, 1572–1583 (2019).

[12] Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **10**, 370–377 (2019).

[13] Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science* **3**, 1103–1113 (2017).

[14] Chen, Z., Ayinde, O. R., Fuchs, J. R., Sun, H. & Ning, X. G 2 retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry* **6**, 102 (2023).

[15] Kariofillis, S. K. *et al.* Using data science to guide aryl bromide substrate scope analysis in a ni/photoredox-catalyzed cross-coupling with acetals as alcohol-derived radical sources. *Journal of the American Chemical Society* **144**, 1045–1055 (2022).

[16] Yada, A. *et al.* Machine learning approach for prediction of reaction yield with simulated catalyst parameters. *Chemistry Letters* **47**, 284–287 (2018).

[17] Cortes, C. & Vapnik, V. Support-vector networks. *Machine learning* **20**, 273–297 (1995).

[18] Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).

[19] Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**, 1–67 (2020).

[20] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[21] Kim, S. *et al.* Pubchem 2019 update: improved access to chemical data. *Nucleic acids research* **47**, D1102–D1109 (2019).

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

[23] Neel, A. J., Milo, A., Sigman, M. S. & Toste, F. D. Enantiodivergent fluorination of allylic alcohols: data set design reveals structural interplay between achiral directing group and chiral anion. *Journal of the American Chemical Society* **138**, 3863–3875 (2016).

[24] Carlson, R. & Carlson, J. E. *Design and optimization in organic synthesis* (Elsevier, 2005).

[25] Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of chemical research* **51**, 1281–1289 (2018).

[26] Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in c–n cross-coupling using machine learning. *Science* **360**, 186–190 (2018).

[27] Jiang, S. *et al.* When smiles smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access* **9**, 85071–85083 (2021).

[28] Chuang, K. V. & Keiser, M. J. Comment on "predicting reaction performance in c–n cross-coupling using machine learning". *Science* **362**, eaat8603 (2018).

[29] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379–1390 (2020).

[30] Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018).

[31] Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nature machine intelligence* **3**, 144–152 (2021).

[32] Lowe, D. Chemical reactions from us patents (1976-sep2016) (2017). URL "https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873".

[33] Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **59**, 3370–3388 (2019).

[34] Jo, J., Kwak, B., Choi, H.-S. & Yoon, S. The message passing neural networks for chemical property prediction on smiles. *Methods* **179**, 65–72 (2020).

[35] Tang, M., Li, B. & Chen, H. Application of message passing neural networks for molecular property prediction. *Current Opinion in Structural Biology* **81**, 102616 (2023).

[36] Yarish, D. *et al.* Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. *Journal of Computational Chemistry* **44**, 76–92 (2023).

[37] Lei, T., Jin, W., Barzilay, R. & Jaakkola, T. Deriving neural architectures from sequence and graph kernels. In *International Conference on Machine Learning*, 2024–2033 (PMLR, 2017).

[38] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
[39] Maziarka, Ł. *et al.* Molecule attention transformer. *arXiv preprint arXiv:2002.08264* (2020).
[40] Hu, W. *et al.* Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265* (2019).
[41] Sterling, T. & Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling* **55**, 2324–2337 (2015).
[42] Somnath, V. R., Bunne, C., Coley, C. W., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. In *Thirty-Fifth Conference on Neural Information Processing Systems* (2021). URL https://openreview.net/forum?id=SnONpXZ_uQ_.
[43] Balaji, N. N. A., Beaulieu, C. L., Bogner, J. & Ning, X. Traumatic brain injury rehabilitation outcome prediction using machine learning methods. *Archives of Rehabilitation Research and Clinical Translation* **5**, 100295 (2023).
[44] Ma, Y. *et al.* Are we making much progress? revisiting chemical reaction yield prediction from an imbalanced regression perspective. In *Companion Proceedings of the ACM on Web Conference 2024*, 790–793 (2024).
[45] Kawasaki, H., Kihara, N. & Takata, T. High yielding and practical synthesis of rotaxanes by acylative end-capping catalyzed by tributylphosphine. *Chemistry Letters* **28**, 1015–1016 (1999).
[46] Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation* **5**, 107–113 (1965).
[47] Geisser, S. The predictive sample reuse method with applications. *Journal of the American statistical Association* **70**, 320–328 (1975).
[48] Carey, F. A. & Sundberg, R. J. *Advanced organic chemistry: part A: structure and mechanisms* (Springer Science & Business Media, 2007).
[49] Anslyn, E. *Modern Physical Organic Chemistry*, vol. 227 (University Science Books, 2006).
[50] Clayden, J., Greeves, N. & Warren, S. *Organic chemistry* (Oxford University Press, USA, 2012).
[51] Smith, M. B. *March's advanced organic chemistry: reactions, mechanisms, and structure* (John Wiley & Sons, 2020).
[52] Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807–814 (2010).
[53] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[54] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).