A unified pre-trained deep learning framework for cross-task reaction performance prediction and synthesis planning

Li-Cheng Xu,[1*] Miao-Jiong Tang,[1,3] Junyi An,[1] Fenglei Cao,[1*] Yuan Qi[1,2*]

[1]Shanghai Academy of Artificial Intelligence for Science, Shanghai 200232, China

[2]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai 201203, China

[3]Center of Chemistry for Frontier Technologies, Department of Chemistry, Zhejiang University, Hangzhou, 310027, China

*Emails: xulicheng@sais.com.cn; caofenglei@sais.com.cn; qiyuan@fudan.edu.cn

Abstract: Artificial intelligence has transformed the field of precise organic synthesis. Data-driven methods, including machine learning and deep learning, have shown great promise in predicting reaction performance and synthesis planning. However, the inherent methodological divergence between numerical regression-driven reaction performance prediction and sequence generation-based synthesis planning creates formidable challenges in constructing a unified deep learning architecture. Here we present RXNGraphormer, a framework to jointly address these tasks through a unified pre-training approach. By synergizing graph neural networks for intramolecular pattern recognition with Transformer-based models for intermolecular interaction modeling, and training on 13 million reactions via a carefully designed strategy, RXNGraphormer achieves state-of-the-art performance across eight benchmark datasets for reactivity/selectivity prediction and forward-/retro-synthesis planning, as well as three external realistic datasets for reactivity and selectivity prediction. Notably, the model generates chemically meaningful embeddings that: (1) spontaneously cluster reactions by type without explicit supervision, and (2) reveal structure-performance relationships through post-hoc interpretation. This work bridges the critical gap between performance prediction and synthesis planning tasks in chemical AI, offering a versatile tool for accurate reaction prediction and synthesis design.

Introduction

Prediction of reaction reactivity[1], selectivity[2], retrosynthetic analysis[3], and the complementary task of reaction product prediction[4] constitute four fundamental pillars of precise chemical synthesis. Establishing robust structure-performance relationship (SPR) is essential for accurate reactivity and selectivity predictions.[2,5] Retrosynthesis identifies viable precursors for a target compound, while reaction outcome prediction pinpoints the most likely products from specified reactants. Both of these tasks frequently employ formalized reaction templates derived from subgraph pattern matching in expert systems such as LHASA[3] and Chematica[6]. Template-based methods leverage well-established reaction patterns to achieve reliable predictions when matching templates are available, though they may face challenges with truly novel transformations outside existing template libraries.

Recently, the widespread adoption of data-driven approaches such as machine learning (ML) and deep learning (DL) in chemical reaction prediction has sparked a significant revolution in the precise prediction of organic reactions[1,7,8]. By utilizing correlations observed in extensive laboratory synthesis data, ML models robustly predict pivotal factors like reactivity[9] and selectivity[10,11] across diverse transformations, thereby underscoring their powerful role in enhancing synthetic performance. Doyle and colleagues[12] employed quantum chemical descriptors in tandem with random forest algorithms, whereas Glorius and co-workers[13] utilized molecular fingerprints, both achieving highly accurate predictions. Beyond descriptor-based strategies, end-to-end DL models have also proven effective in predicting reaction performances. Schwaller and Reymond's YieldBERT[14] (Fig. 1a) utilizes a pre-trained encoder to derive molecular representations from SMILES strings, enabling yield predictions via a regression layer. Additionally, Chen and Liao's team developed GraphRXN[15], which employs graph neural network (GNN) to embed each molecule within a reaction, subsequently integrating these embeddings to predict reaction performance (Fig. 1b). In parallel with advancing reaction SPR modeling, DL methodologies have profoundly influenced computer-aided synthesis in both retrosynthesis and forward synthesis planning[16] (Fig. 1c), as showcased by pioneering work from Segler[17], Jensen[18], Coley[19], and others[20–27]. A range of model architectures, from template-based to template-free, has thus emerged, each pushing the boundaries of computer-aided synthesis planning (CASP) in diverse ways[16].

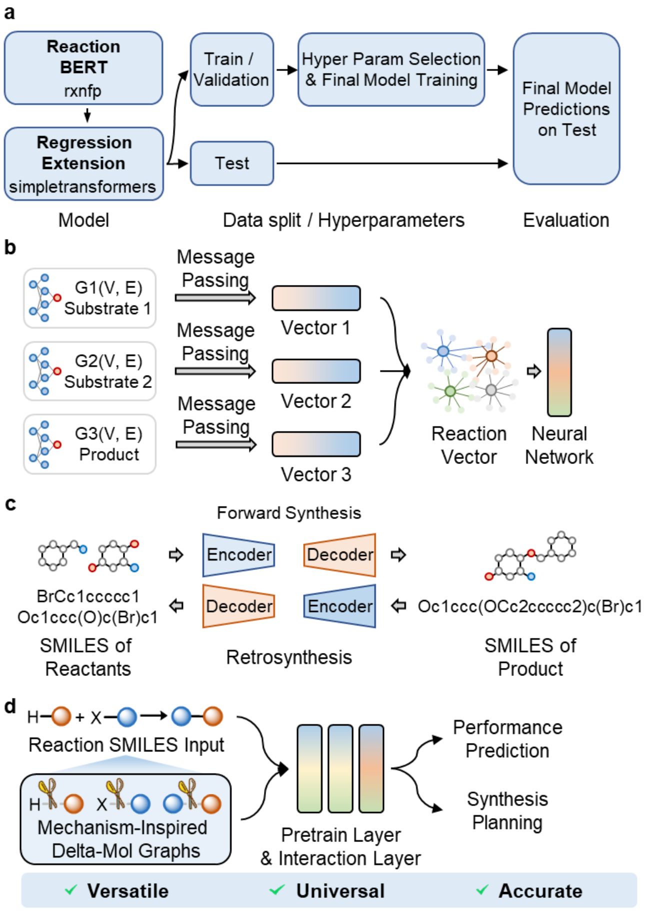


Fig. 1: Deep learning model for reaction prediction. a, The architecture of YieldBERT. **b,** The architecture of GraphRXN. **c,** The architecture of Graph2SMILES. **d,** The schematic diagram of the RXNGraphormer model architecture, which facilitates predictions of reaction performance and synthesis planning under a unified pre-training framework, incorporates delta–mol graphs inspired by reaction mechanisms specifically for the task of predicting reaction performance.

Despite the proven effectiveness of DL technologies in reaction performance prediction and synthesis planning, tackling these tasks simultaneously presents a substantial challenge due to their differing computational requirements—namely, numerical regression versus sequence generation. Although Zhang achieved cross-task prediction through plain-text representations[28], existing graph-based approaches remain fragmented—several pioneering studies have independently utilized molecular graphs and GNNs for reaction performance prediction[15,29,30] or synthesis planning[19,24,27]. We surmise that a unified framework, integrating a molecular graph encoder and an intermolecular interaction encoder, equipped with both a regression layer and a sequence decoder, can effectively deliver precise predictions of reaction performance, including reactivity and selectivity, as well as comprehensive synthesis planning, encompassing both retrosynthesis and forward synthesis, within one single system.

In this work, we introduce RXNGraphormer (Fig. 1d), a unified reaction prediction framework that integrates several innovations to precisely address multiple tasks. Our method employs a specialized reaction-encoding scheme, using a GNN[31–33] to extract molecular embeddings, which are then processed by a Transformer[34] to capture interactions among reaction components. Pre-trained on 13 million real and fictitious reactions, this encoder robustly handles diverse molecular and reaction data. We append a regression layer and a sequence decoder, enabling the prediction of reaction performance and the generation of SMILES for reactants and products. To enhance the performance of regression tasks, we generate “delta–mol” graphs by interpolating bond differences between reactants and products, thereby creating an intermediate molecular representation that incorporates richer mechanistic information about the reaction process without additional quantum calculations. RXNGraphormer delivers state-of-the-art (SOTA) prediction performance in reactivity[12,35], enantioselectivity[10], and regioselectivity[36] across four highly ordered benchmark datasets and three external literature-derived datasets, while maintaining excellent performance in both single-step retrosynthesis and forward synthesis on USPTO-50k[37], USPTO-full[38], USPTO-STEREO[39], and USPTO-480k[40]. Beyond its impressive predictive performance, the model’s pre-trained encoder can distinguish different reaction types without explicit training. Furthermore, fine-tuning for reaction performance results in embeddings that inherently encapsulate SPRs. This comprehensive design equips RXNGraphormer to tackle complex challenges in chemical reaction prediction with both accuracy and efficiency.

Results

Design of RXNGraphormer

We believe that a robust pre-trained encoder can significantly enhance cross-task prediction capabilities in chemical reaction modeling. To this end, we assembled a large-scale chemical reaction dataset of over 13 million entries and devised an innovative pre-training strategy. Specifically, we combined multiple open-source datasets and curated more than one million reaction records from the World Intellectual Property Organization (WIPO) database. For detailed information on the chemical reaction dataset, please refer to Table S1 in the Supplementary Information. After verifying SMILES correctness, canonicalizing the strings, and eliminating duplicates, we obtained approximately 6.8 million high-quality reaction records. We then applied a fragment exchange algorithm to over 4 million real product molecules derived from these records, generating an equal number of fictitious chemical reactions (comprising over 5 million algorithm-generated product molecules), thereby expanding the total dataset to over 13 million entries (Fig. 2a). More details of the fragment exchange algorithm and the processing pipeline for potential false negative data are provided in the Methods section. Our pre-training task is designed to enable the model to discriminate between real and fictitious reactions, learning the conservation of molecular frameworks and identification of reaction sites during organic transformations, thereby enhancing its representational capacity of molecular structures and reaction features.

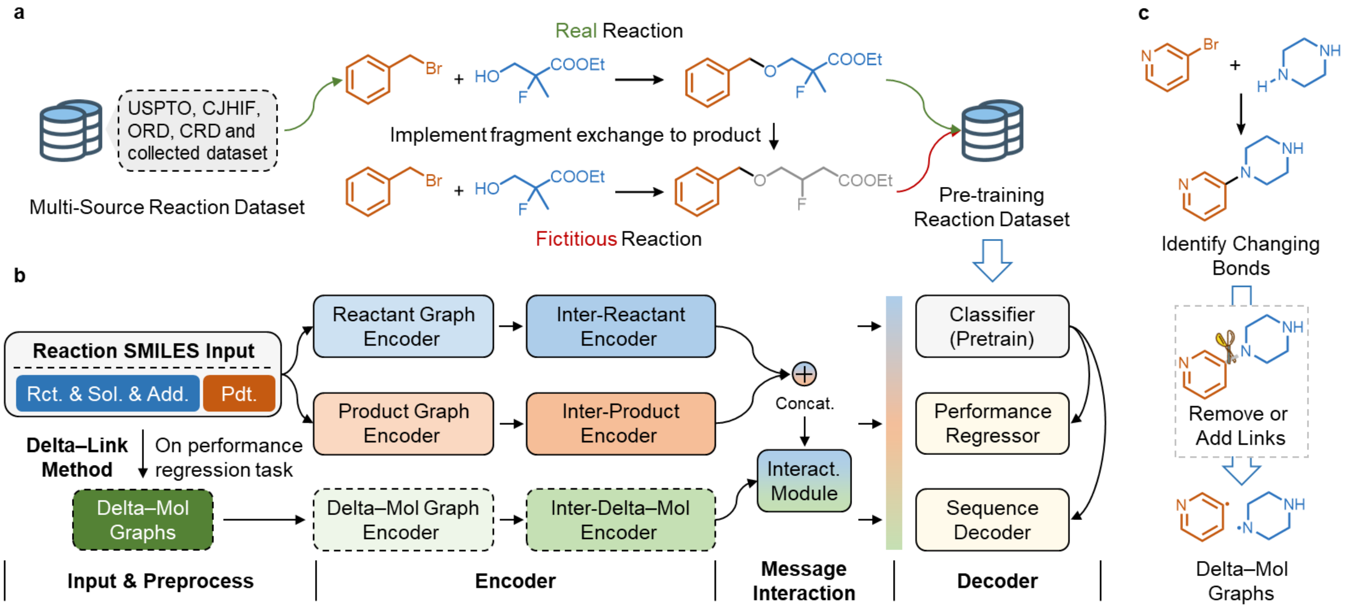


Fig. 2: Overview of the RXNGraphormer architecture. a, Dataset. 6.8 million real reactions plus an equal number of algorithm-generated fictitious reactions. **b,** Model. Processes reaction SMILES inputs and delta-mol graphs through intra- and intermolecular encoders, which are subsequently combined and fed into task-specific modules (classification, regression, and sequence decoding). **c,** Delta–link method: Identifies bond changes between reactants and products to construct reaction-specific delta-mol graphs.

Building on this pre-training strategy, we developed RXNGraphormer, a new model architecture that supports classification, regression, and sequence generation tasks (Fig. 2b). RXNGraphormer employs GNNs, adapted from Mole-BERT[31], to encode the molecules in a reaction and capture their intrinsic properties, then generating embeddings analogous to "tokens embeddings" in natural language processing (NLP). These embeddings are then processed by interaction encoders (adapted Transformers) to model intermolecular interactions. For classification pre-training, reactants and products are encoded separately through the GNNs and Transformers, and their respective embeddings—now enriched with interaction information—are concatenated to form a unified representation of the entire reaction. This representation drives a binary classification layer in the pre-training phase, enabling the model to distinguish real from fictitious reactions. Moreover, the architecture is designed to handle reactions with an arbitrary number of participants, and we canonicalized SMILES strings by RDKit to mitigate confounding effects arising from different molecular orders.

To enhance the performance of our model in downstream tasks predicting reactivity and selectivity, we extend beyond merely encoding reactants and products with a pre-trained model by additionally incorporating “delta–mol” graphs that capture information about bond changes during the reaction process. Utilizing our developed “delta-link method”, we compare the changes in chemical bonds between reactants and products, interpolate these changes, and generate intermediate “frames” of the reaction, termed delta–mol graphs (Fig. 2c). These delta–mol graphs are then processed by a graph encoder and an interaction encoder, thereby incorporating mechanistic insights into bondbreaking and bond-forming events. The resulting delta–mol graphs embedding is combined with the reactant and product embeddings, and these fused encodings are passed through an interaction module—consisting of fully connected layers—to give a comprehensive representation of the reaction (Fig. 2b). Finally, this representation is subsequently fed into a regression layer to predict reaction performance. For retrosynthesis predictions, only the product-side graph and interaction encoders are employed to generate an embedding of the product, which is then decoded via Transformer decoders into a SMILES sequence representing the reactants (Fig. 2b). Conversely, forward synthesis predictions use the reactant-side embedding to produce the product SMILES sequence.

Reaction type discrimination using pre-trained embeddings

Schwaller and colleagues introduced an innovative pre-training task to distinguish between various types of reactions[41]. Drawing inspiration from their work, we sought to determine whether our pre-trained reaction encoder might similarly discriminate among different reaction types, given that chemists routinely classify reactions based on bond changes, a principle fundamentally aligned with our pre-training objective. It is important to emphasize that our pre-training procedure was primarily designed to differentiate real reactions from fictitious ones, without incorporating any information about specific reaction types.

To examine whether the reaction embeddings produced by our pre-trained encoder inherently capture distinctions among diverse classes of chemical reactions, we eschewed training a dedicated model for reaction-type classification and instead compared pairwise distances between these embeddings. We validated our approach on the USPTO-50k dataset reported by Schneider et al.[42], which comprises 50 reaction types—each with 1,000 entries. We extracted embeddings from the penultimate layer of our pre-trained classification model and then computed Euclidean distances between reactions belonging to different types. By taking the mean distance across these comparisons, we obtained an average distance matrix between any two reaction types, culminating in the heatmap visualization shown in Fig. 3a. Further information on the 50 reaction types and the distance calculations is provided in the Supplementary Information.

In Fig. 3a, cooler (blue) hues in the heatmap indicate a shorter distance between two reaction types in the model's latent space, whereas warmer (red) hues reveal greater distances. Notably, the diagonal from the top-left to the bottom-right represents intra-type distances, which appear in shades of blue due to their close proximity. Black-outlined regions correspond to reactions in similar major categories (e.g., C–C bond formation), showing clear correlations. Moreover, Fig. 3b illustrates how the pre-trained model's reaction embeddings distinguish similarities and differences among reactions. For instance, the upper-left region of the sub-heatmap, termed "reaction type group 1", denotes a cluster of closely related reaction types, whereas the lower-right corner forms "reaction type group 2". Closer inspection of these eight reaction types confirmed that group 1 involves reactions yielding amides and sulfonamides, while group 2 predominantly features ester and sulfonic ester formations. This clustering pattern reflects the differentiated reactivity profiles: carbonyl and sulfonyl groups exhibit similar electrophilicity, causing amide/sulfonamide formations (group 1) and ester/sulfonic ester formations (group 2) to cluster separately based on their nucleophilic partners (amines versus alcohols). The comparable electronic properties of these functional groups maintain relatively small inter-group distances in the embedding space.

Significantly, the subjective and uneven granularity of USPTO-50k's reaction classification results in non-uniform blueness along the diagonal. The unusually warm region at the diagonal's lower-right corner (functional group addition) is particularly striking. This major category comprises five subtypes (Fig. 3c): bromination (A), chlorination

(B), Wohl-Ziegler bromination (C), nitration (D), and methylation (E). Methylation reactions can be primarily categorized into C–N, C–O, and C–C bond formations. The marked differences between these bond types—especially between C–C and others—create intrinsic variations within this class from a bond transformation perspective, making methylation distinct from other reaction types. Bromination subtypes include benzylic (i.e., Wohl-Ziegler), α-carbonyl, and aromatic bromination. While all involve C–Br bond formation, chemical environment differences still cause intra-class variations. Wohl-Ziegler bromination, a specific type of allylic or benzylic bromination, exhibits minimal internal variation. Nitration reactions exhibit moderate variations due to reaction sites (aromatic rings vs. carbonyl α-positions). These analyses confirms that our pre-training strategy enables the model to learn bond transformation patterns and capture implicit electronic effects in reactions.

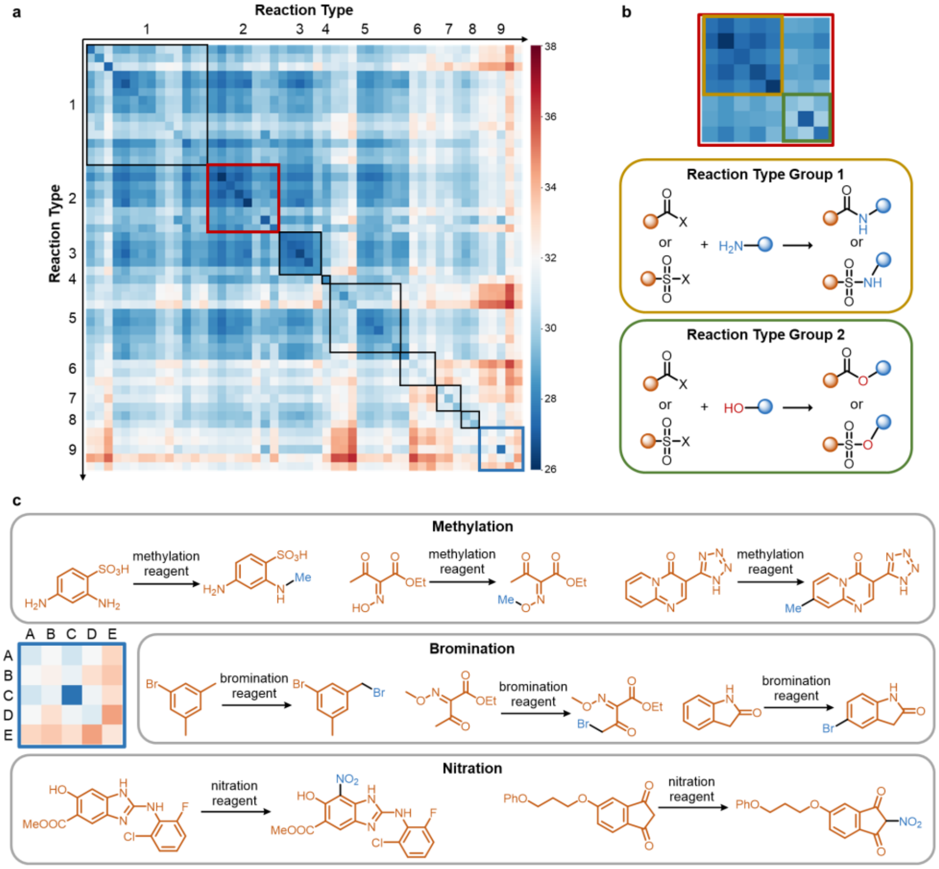


Fig. 3: Distance analysis of reaction embeddings from USPTO-50k dataset. a, Heatmap of 50,000 chemical reaction embeddings categorized into 50 reaction types across 9 major categories. The details of nine major reaction categories: 1, heteroatom alkylation and arylation; 2, acylation and related processes; 3, C–C bond formation; 4, protection; 5, deprotection; 6, reduction; 7, oxidation; 8, functional group interconversion; 9. FGA. **b,** Detailed view of two reaction type groups within the "acylation and related processes" category. **c,** Three representative functional group addition (FGA) subtypes with examples. The details of 5 reaction types in FGA: A, bromination; B, chlorination; C, Wohl-Ziegler bromination; D, nitration; E, methylation.

Regression performance of RXNGraphormer

We next evaluated the predictive capabilities of RXNGraphormer across multiple aspects of reaction performance—including reactivity, regioselectivity, and enantioselectivity—using a diverse set of high-quality and widely recognized datasets. To assess yield predictions, we employed the Buchwald–Hartwig reaction dataset (dataset 1 in Fig. 4a) reported by Doyle et al.[12], as well as the Suzuki–Miyaura reaction dataset (dataset 2 in Fig. 4a) documented by Perera et al.[35] Additionally, the radical C–H functionalization dataset (dataset 3 in Fig. 4a) from Hong et al.[36] and the asymmetric thiol addition dataset (dataset 4 in Fig. 4a) from Denmark et al.[10] were used to evaluate the model's capacity for predicting regioselectivity and enantioselectivity, respectively. Except for the C–H functionalization dataset, which was derived from high-accuracy DFT calculations, all other datasets originated from experimental results.

These datasets were randomly split into training and testing sets based on ratios used in the original publications or by previously tested models. We performed this random division and conducted ten prediction trials for each dataset. Fig. 4b illustrates a representative regression outcome on test set for each dataset. Overall, RXNGraphormer achieved outstanding predictive power across all four datasets. Remarkably, for yield prediction (datasets 1 and 2), it reached R^2 scores of 0.971 and 0.876, with corresponding mean absolute error (MAE) of 2.98 and 6.30. On the computationally derived regioselectivity dataset (dataset 3), it delivered an R^2 score of 0.992 and an MAE of 0.266 kcal/mol. The lower noise in this DFT-generated dataset likely contributes to the model's superior accuracy relative to the other three datasets. Moreover, in enantioselectivity prediction (dataset 4), RXNGraphormer attained an R^2 score of 0.915 and an MAE of 0.134 kcal/mol.

We compared RXNGraphormer against several representative SOTA methods that do not require additional DFT calculations. These include the Multiple Molecular Fingerprint combined with Random Forest (MFF+RF) approach by Glorius[13], the text-based reaction representation method YieldBERT from Schwaller et al.[14], the GraphRXN model by Chen and Liao et al.[15], and ReaMVP by Yang et al.[30], which incorporates both 2D and 3D molecular graphs for yield prediction. As shown in the panels 1 and 2 of Fig. 4c, RXNGraphormer significantly outperformed the other models on datasets 2, 3, and 4, encompassing reactivity, regioselectivity, and stereoselectivity, achieving MAE values of 6.37, 0.270 kcal/mol, and 0.136 kcal/mol, respectively. On the Buchwald-Hartwig reaction yield prediction dataset (dataset 1), RXNGraphormer and ReaMVP showed comparable performance, with MAEs of 3.02 and 3.11, respectively. Notably, ReaMVP had been specifically optimized for predicting reaction yields during its pre-training. Although ReaMVP excelled at yield prediction, its performance was markedly weaker on the two selectivity prediction datasets compared to other methods.

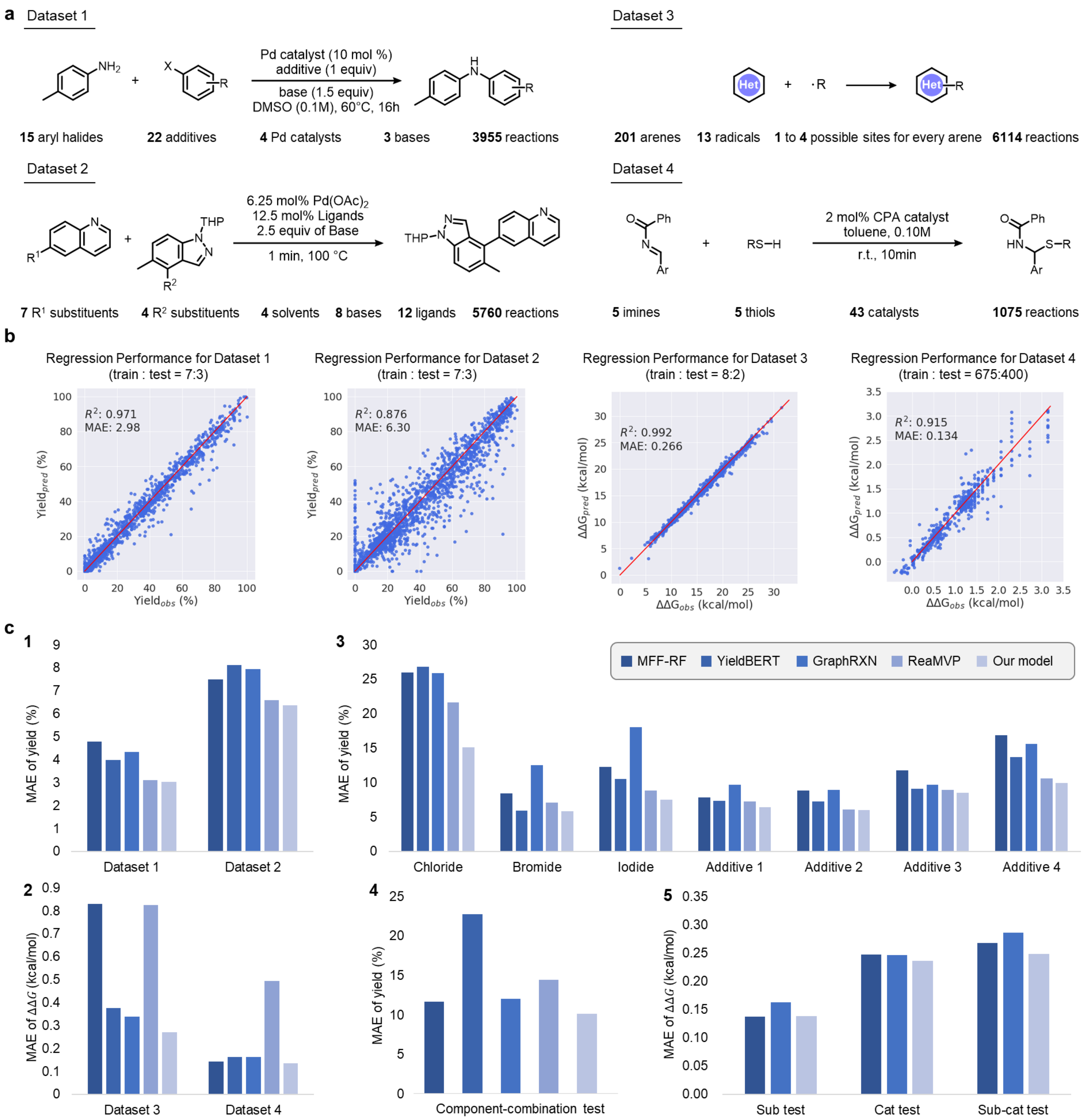


Fig. 4: Regression performance evaluation of RXNGraphormer across four benchmark datasets. **a,** Composition, structural diversity, and data volume of benchmark datasets 1 to 4. **b,** Representative scatter plots of predicted versus experimental values for RXNGraphormer on test sets (datasets 1–4). **c,** Performance comparison with other SOTA methods. Panel 1: MAE for reactivity prediction across two datasets. Panel 2: MAE for selectivity prediction across two datasets. Panel 3: MAE for single-component OOS prediction on dataset 1. Panel 4: MAE for component-combination test set of dataset 1. Panel 5: MAE for OOS prediction on substrate, catalyst, and their combinations in dataset 4.

To thoroughly evaluate generalizability, we conducted more challenging out-of-sample (OOS) tests on datasets 1 and 4. For dataset 1, we constructed three aryl halide-based OOS test sets (chloride, bromide, iodide) and four additive-based test sets. Additionally, we created a component-combination test set by selecting unseen compounds across all four dimensions (see Supplementary Information for details). For dataset 4, we followed Denmark's[10] original protocol to generate substrate (sub), catalyst (cat), and combined (sub-cat) test sets. As shown in the panel 3 of Fig. 4c, RXNGraphormer outperformed other methods across all OOS test sets of dataset 1. It maintained superiority even on the more challenging component-combination test set (panel 4 of Fig. 4c), with MAE of 10.12. Given that YieldBERT and ReaMVP are specifically optimized for yield prediction tasks, we only compared with MFF-RF and GraphRXN for dataset 4. Panel 5 of Fig. 4c shows RXNGraphormer achieved comparable accuracy to MFF-RF on sub test set (MAE 0.138 vs 0.137 kcal/mol), while demonstrating better performance on cat and sub-cat test sets. Model comparison metrics for random-split and OOS test sets are detailed in Supplementary Information. These results robustly demonstrate RXNGraphormer's capability in predicting novel reaction combinations and compounds, highlighting its advantages in SPR modeling.

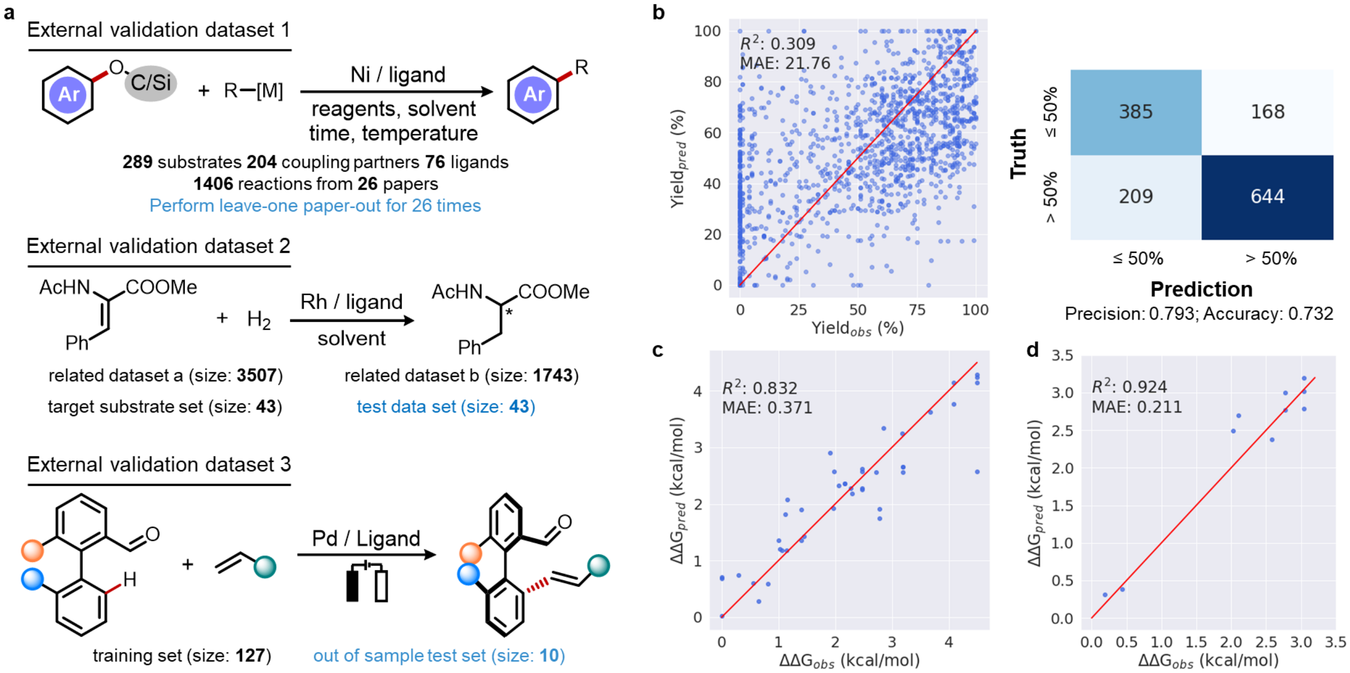


Fig. 5: External validation of RXNGraphormer on literature-derived datasets. a, Details of three literature-derived datasets; **b,** Predictive performance on external validation dataset 1 (regression and classification); **c,** Regression performance on external validation dataset 2; **d,** Regression performance on external validation dataset 3.

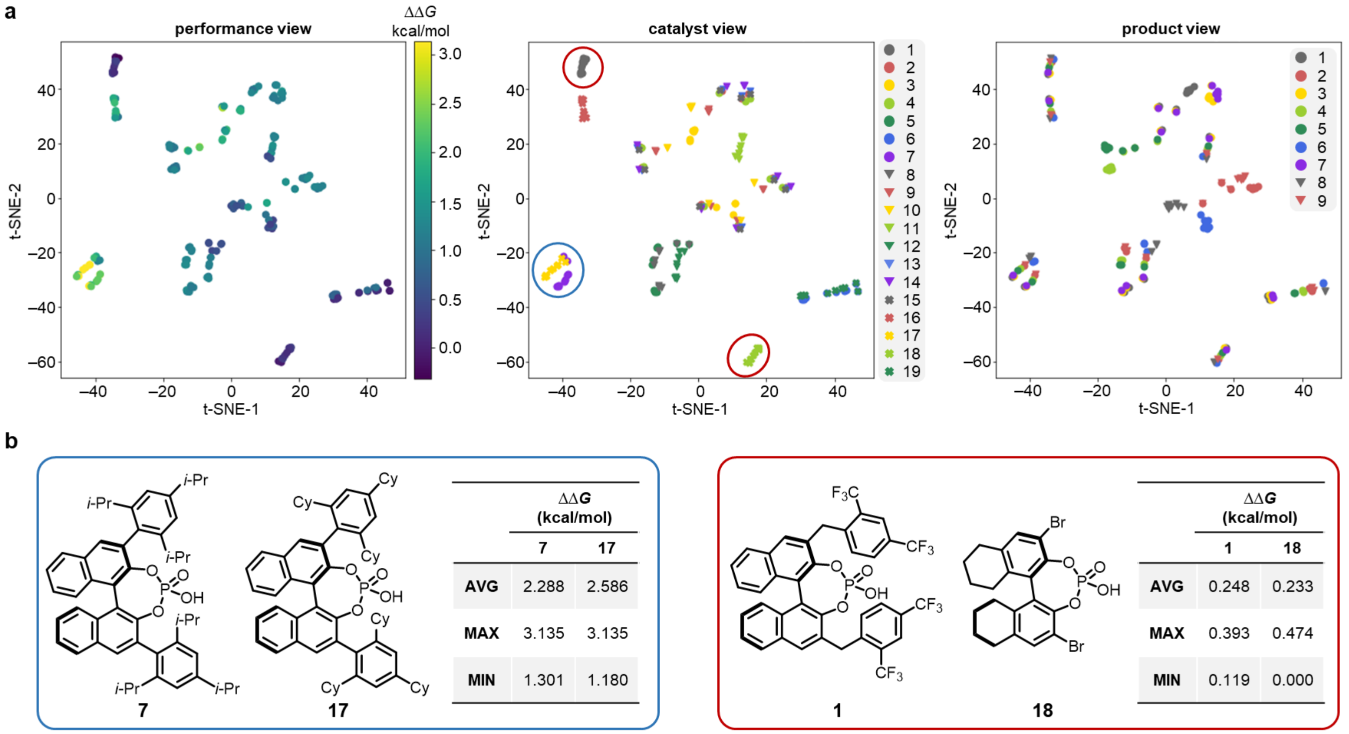
To validate RXNGraphormer's performance on more realistic scenarios, we tested it on three literature-derived datasets (Fig. 5a): nickel-catalyzed C–O coupling dataset (NiCOlit) reported by Schleinitz et al.[43], asymmetric hydrogenation of olefins dataset (AHO) by Hong et al.[44], and pallada-electrocatalyzed C–H activation dataset by Hong and Ackermann et al[45]. Using the DOI hold-out approach same as original study for NiCOlit, aggregated results from 26 validations yielded MAE of 21.76 and R^2 of 0.309 for direct yield prediction (surpassing the original 28.8 and -0.01), with binary classification precision and accuracy reaching 0.793 and 0.732 for high-yield reactions (>50%, Fig. 5b). These results demonstrate that our model can still provide effective guidance for novel chemical space exploration despite substantial data noise. On AHO dataset, following the original partitioning method that divided 86 reactions with methyl (*Z*)-2-acetamido-3-phenylacrylate as substrate into target substrate and test sets, then supplemented with thousands of related data, achieved enantioselectivity prediction with R^2 of 0.832 and MAE of 0.371 kcal/mol (Fig. 5c), comparable to the original hierarchical learning results (0.852 and 0.387 kcal/mol). Evaluation on the pallada-electrocatalysis dataset demonstrated superior performance with R^2 of 0.924 and MAE of 0.211 kcal/mol for 10 OOS test samples (Fig. 5d), exceeding the originally reported values (0.910 and 0.260 kcal/mol). These results demonstrate RXNGraphormer's excellent performance not only on highly ordered datasets but also in practical application scenarios.

Synthesis planning performance of RXNGraphormer

To assess the performance of RXNGraphormer in single-step synthesis prediction, which serves as the fundamental building block for multi-step synthesis planning, we conducted a comparative analysis with multiple SOTA models, each evaluated independently on four reaction datasets derived from the USPTO (see Extended Data Table 1). For consistency with other studies, these datasets were randomly partitioned into training, validation, and testing sets using the same ratios as employed by the competing models. Top-*n* accuracy metrics for these SOTA models were directly cited from the original reports. For retrosynthesis tasks, we excluded methods that employ template-based strategies, atom mapping, or SMILES augmentation, since these techniques can boost predictive performance independently of the underlying model architecture[19], complicating a direct comparison. In our study on the USPTO-50k dataset, we benchmarked our method against six other representative models including DMP fusion[27], Tied Transformer[23], GET[25], SCROP[22], AutoSynRoute[21], and Graph2SMILES[19]. Although Graph2SMILES[19] achieved higher top-1 accuracy at 52.9% compared to our 51.0%, RXNGraphormer excelled beyond the other contenders in top-3, top-5, and top-10 accuracies, achieving 69.0%, 74.2%, and 79.2% respectively. On the larger and more noise-prone USPTO-full dataset, RXNGraphormer surpassed three representative baselines (DMP fusion, Transformer baseline[27], and Graph2SMILES) across all four top-*n* accuracy metrics, setting new SOTA results. Without any additional performance-boost techniques, it improved the previous SOTA top-1 accuracy from 45.7% to 47.4% and raised the top-10 accuracy from 67.9% to 71.6%. In forward-synthesis prediction tasks evaluated against Molecular Transformer[20], MEGAN[24], Chemformer[26], and Graph2SMILES on USPTO-480k, RXNGraphormer achieved 90.6% top-1 accuracy—slightly below the 91.3% reported for Chemformer—while demonstrating superior performance in top-3, top-5, and top-10 accuracy. In our investigation utilizing the USPTO-STEREO dataset, which encompasses complex stereochemical information, thereby posing a greater prediction challenge, our model achieved SOTA results when compared with Molecular Transformer and Graph2SMILES. It excelled in all four top-*n* accuracy metrics, recording accuracies of 78.2%, 85.1%, 86.5%, and 87.8%. Model training parameters and details are provided in the Supplementary Information. These findings demonstrate that, thanks to our unified reaction-prediction framework and innovative pre-training–fine-tuning strategy, RXNGraphormer excels not only at regression-based reaction performance prediction but also, in nearly all cases, surpasses the top synthesis planning models for both forward- and retro-synthesis predictions.

Emergence of structure–performance insights through fine-tuning

Beyond improving overall performance, we sought to determine whether fine-tuning RXNGraphormer on a downstream task would yield deeper chemical insights than the pre-trained model alone. To this end, we first generated reaction embeddings for the randomly split test sets of regression benchmark datasets 1–4 using both the pre-trained model and its fine-tuned counterpart—trained on reaction performance prediction—and then visualized these embeddings via t-SNE[46] in two-dimensional plane (Extended Data Fig. 1). Notably, in the pre-trained model's embedding for dataset 2 (Extended Data Fig. 1b, left), higher-reactivity data points predominantly cluster in the central region (highlighted by the red circle), while lower-reactivity points tend to distribute around this central cluster. Despite this broad partition, clear linear transitions between high- and low-yield clusters were lacking; in other words, some “reactivity cliffs”[47] remained that could not be resolved by a simple linear boundary. By contrast, the fine-tuned model (Extended Data Fig. 1b, right) exhibited a strikingly linear progression of reactivity along the t-SNE-1 axis, effectively eliminating these reactivity cliffs. This characteristic linear progression of reactivity and selectivity along specific axes was similarly observed across all three other fine-tuned models (Extended Data Fig. 1a, 1c, and 1d, right panels), demonstrating the generalizability of this phenomenon.



	ΔΔG (kcal/mol)	
	7	17
AVG	2.288	2.586
MAX	3.135	3.135
MIN	1.301	1.180

	ΔΔG (kcal/mol)	
	1	18
AVG	0.248	0.233
MAX	0.393	0.474
MIN	0.119	0.000

Fig. 6: SPRs in sub-cat test set of benchmark dataset 4 revealed by t-SNE analysis. a, Performance, catalyst, and product view projections; **b,** Representative catalysts for high and low selectivity.

To complement our analysis of SPRs, we investigated more concrete correlations between molecular structures and reaction performance. Using RXNGraphormer fine-tuned on benchmark dataset 4, we generated reaction embeddings and corresponding t-SNE projections (Fig. 6a) for the sub-cat OOS test set, which was explicitly excluded from model training. Fig. 6a displays the same data distribution through three complementary perspectives: performance view, catalyst view, and product view. The performance view in Fig. 6a reveals observable spatial clustering of enantioselectivity, with high-selectivity data points concentrated in specific regions while medium-low selectivity points aggregate in other areas. Catalyst view analysis demonstrates that this performance clustering primarily originates from catalyst variations. In the catalyst view subplot, the blue-circled region contains data points for catalysts **7** and **17**, corresponding to reactions with higher enantioselectivity (mean ΔΔG > 2.2 kcal/mol, Fig. 6b left).

Conversely, the red-circled region includes reactions catalyzed by compounds **1** and **18**, showing lower selectivity (mean $\Delta\Delta G < 0.25$ kcal/mol, Fig. 6b right). Product view analysis shows no significant clustering patterns correlated with product types. These results indicate that our fine-tuned model successfully captures meaningful (albeit implicit) SPRs between catalyst structures and enantioselectivity, consistent with established chemical principles for this reaction system.

These observations not only help explain the model's remarkable accuracy in reaction performance prediction but also suggests that, following fine-tuning on a performance-focused objective, RXNGraphormer's reaction embeddings implicitly encode the SPRs relevant to reactivity and selectivity across these datasets. In other words, after fine-tuning, the learned reaction embeddings transcend the mere structural depiction of individual components and instead capture a cohesive performance landscape for the overall chemical process.

Discussion

In conclusion, this study introduces RXNGraphormer, a unified deep-learning framework for predicting reaction reactivity, selectivity, retrosynthesis pathways, and forward product formation. A pre-training dataset of more than 13 million reactions—comprising both curated real reactions and synthetically generated fictitious reactions via a custom-designed fragment-exchange algorithm—was used to enable the model to distinguish real from fictitious reactions. This large-scale pre-training dataset imbued RXNGraphormer with enriched molecular and reaction representations, which were subsequently fine-tuned for several downstream tasks. Specifically, the fine-tuned molecular and reaction encoders were applied to reaction performance prediction, assisted by the delta-link method that captures key mechanistic intermediates in delta–mol graphs, and to retrosynthesis and forward-synthesis prediction using the pre-trained molecular encoder alone.

The effectiveness of this unified deep-learning framework was rigorously validated across multiple reaction datasets, showcasing robust performance. On the Buchwald–Hartwig, Suzuki–Miyaura, radical C–H functionalization, and asymmetric thiol addition datasets, RXNGraphormer delivered superior predictive accuracy for both reactivity and selectivity, outperforming other models across both randomly partitioned test sets and OOS sets containing unseen compounds. Furthermore, the model's predictive capability under more realistic scenarios was validated on three literature-derived external validation datasets. Across four USPTO-derived datasets, RXNGraphormer displayed outstanding capabilities for retrosynthesis and forward product prediction. Most notably, on USPTO-full—comprising nearly one million reactions—the model established an overwhelming advantage over existing methods, cementing its status as a high-accuracy solution for large-scale reaction prediction.

Beyond its remarkable predictive accuracy, our novel pre-training strategy and downstream reaction-performance fine-tuning method respectively endow the model with the ability to distinguish various reaction types and encode SPRs. Since our pre-training task enables the model to learn chemical bond transformation patterns—which in most cases align with chemists' reaction classification criteria—the pre-trained model naturally clusters diverse reaction classes into distinct regions of latent space. Fine-tuning on the asymmetric thiol addition dataset for enantioselectivity prediction embeds implicit SPRs between selectivity and chiral catalysts. Comprehensive analysis of pre-trained and fine-tuned models converts RXNGraphormer from a "black box" into a more interpretable "gray box". RXNGraphormer bridges the critical gap between performance prediction and synthesis planning by establishing a unified framework, which facilitates data-driven synthetic transformation design.

Methods

The details of pre-train dataset

Data collection

The dataset employed for model pre-training includes multiple open-source chemical reaction datasets, such as USPTO, the Open Reaction Database, the Chemical Reaction Database, and CJHIF dataset et al. Additionally, this dataset incorporates over one million chemical reaction records collected from the WIPO database. A comprehensive list of these open-source datasets, complete with specific access links, is provided in Supplementary Table 1.

Fictitious reaction dataset generation

The generation of fictitious molecules structurally similar to real products was guided by the invariant risk minimization (IRM) theory. This approach encourages the model to distinguish between molecular regions that should remain invariant and those that should undergo bond changes in chemical reactions. For generating negative samples meeting these requirements in pre-trained classification tasks, RDKit is used to simulate fragment exchange reactions on the products of documented chemical reactions, creating synthetic datasets. The detailed process and Python implementation are depicted in Supplementary Figure 1. When multiple potential fictitious products arise from a single molecule, one is randomly selected to ensure parity between the numbers of fictitious and real reactions.

Filtering process for false negative fictitious product molecules

A two-stage filtration protocol was implemented to eliminate false negative fictitious product molecules:

(1) When generating fictitious reactions from real ones, if the exchanged adjacent atoms and their attached fragments are identical, the fragment exchange algorithm would produce molecules identical to the original real ones. Such fictitious molecules matching real products are filtered out during the generation phase.

(2) Given the vast data volume and occasional absence of reaction conditions, applying fragment exchange algorithm to reactions with weak selectivity would yield a minimal proportion (<0.1%) of fictitious products that might appear as real products in other records. These false negative instances are systematically removed during post-generation processing.

RXNGraphormer architecture

The molecular graph in RXNGraphormer incorporates eight atomic features: atom type, explicit atom degree, formal charge, chiral tag, aromaticity, total valence, hydrogen count, and CIP code. It also includes five bond features: bond type, bond direction, stereo information, presence in a ring, and conjugation status, as depicted in Supplementary Figure 6a.

The RXNGraphormer encoder, illustrated in Supplementary Figure 6b, begins by converting input SMILES strings into molecular graphs using RDKit, computing atomic and bond features. These features are transformed into node and edge embeddings via embedding layers. Multiple graph convolutional blocks, each containing a GCN convolutional layer and a batch normalization layer, process these embeddings to produce molecular embeddings for each molecule in the reaction. For classification and regression tasks, these embeddings are further processed through a global attention module and aligned by padding, followed by Transformer blocks that capture intermolecular interactions, resulting in interacted reactants embeddings. For sequence generation tasks, modified Transformer-XL modules are utilized to enhance interactions between molecules, with layer normalization finalizing the interacted reactants embeddings. Dependent on the task, different modules within RXNGraphormer are used, as shown in Supplementary Figures 7 to 10.

The details of delta-link method

The delta-link method uses "mixed SMILES of reactant and other reagents" and "product SMILES" as inputs. Initially, it distinguishes reactant SMILES from other reagent SMILES to identify the actual reactants. The subsequent step involves analyzing the chemical bond changes between reactant and product SMILES to detect bonds formed and broken during the reaction. The third step simulates these bond changes to generate intermediate "frames" of the reaction, termed delta-mol graphs. The final step involves streamlining the delta-mol graphs by deduplicating SMILES and omitting hydrogen atoms, resulting in a simplified graph that captures detailed reaction dynamics. This process is depicted in Supplementary Figure 11.

The details of reaction distance calculation

To evaluate the model's capability to distinguish similarities and differences across reaction types, we conducted a case study using the USPTO-50k dataset. Following pre-training, RXNGraphormer was employed to generate reaction embeddings, extracted from the penultimate layer of the model. The average Euclidean distance between pairs of reaction classes was computed using the following formula:

$$\frac{1}{N \times M}\sum_{i=1}^{N}\sum_{j=1}^{M}\sqrt{\sum_{k=1}^{d}(RE_{ik}^{1} - RE_{jk}^{2})^2}$$

Here, the variables RE^1 and RE^2 represent the reaction embeddings for two different reaction classes, which are matrices of dimensions $N \times d$ and $M \times d$, respectively. For the USPTO-50k dataset, $N = M = 1000$, representing the number of reactions per class, and $d = 768$, corresponding to the dimensionality of the reaction embeddings generated by RXNGraphormer.

The details of reaction embedding visualization using t-SNE

Reaction embeddings were extracted from the penultimate layer of RXNGraphormer under two conditions: (1) models that were pre-trained only, without fine-tuning for reaction performance prediction tasks, and (2) models fine-tuned on specific datasets, including the Buchwald–Hartwig reaction dataset, Suzuki–Miyaura reaction dataset, radical C–H functionalization dataset, and asymmetric thiol addition dataset. These embeddings were projected onto a two-dimensional plane using the t-SNE algorithm implemented in scikit-learn.

Data availability

The preprocessed dataset comprising 13 million chemical reactions used for model pre-training is accessible through https://github.com/licheng-xu-echo/RXNGraphormer. Specific datasets, including the Buchwald–Hartwig reaction dataset, Suzuki–Miyaura reaction dataset, the C–H functionalization dataset, the asymmetric thiol addition dataset, the USPTO-derived series, and three external validation datasets are available via https://github.com/licheng-xu-echo/RXNGraphormer.

Code availability

Code for data preprocessing, fictitious product molecule generation, delta–mol graph generation, model training, performance evaluation, as well as scripts for analyzing pre-trained and fine-tuned models, are available via https://github.com/licheng-xu-echo/RXNGraphormer.

Acknowledgements

Generous support by the National Natural Science Foundation of China (82394432 and 92249302, Y. Qi). The computations in this research were performed using the CFFF platform of Fudan University. The Data Platform Department of SAIS is acknowledged for their contributions in the open-source reaction dataset collection process.

Author contributions

L.-C.X., F.C., and Y.Q. conceived the initial idea for the project. F.C. and Y.Q. supervised the project. L.-C.X designed the RXNGraphormer model framework and the pre-training strategy, wrote the codes, trained the models, and analyzed the results. M.-J.T. contributed to implement the algorithm for fictitious reaction generation. J.A. contributed to design the model encoder. All authors took part in discussions. L.-C.X. wrote the manuscript with input from all the authors.

Competing interests

The authors declare no competing interests.

Reference

1. Żurański, A. M., Martinez Alvarado, J. I., Shields, B. J. & Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc. Chem. Res.* **54**, 1856–1865 (2021).

2. Zahrt, A. F., Athavale, S. V. & Denmark, S. E. Quantitative Structure–Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **120**, 1620–1689 (2020).

3. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. *Science* **228**, 408–418 (1985).

4. Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **34**, 247 (2005).

5. Cheong, P. H.-Y., Legault, C. Y., Um, J. M., Çelebi-Ölçüm, N. & Houk, K. N. Quantum Mechanical Investigations of Organocatalysis: Mechanisms, Reactivities, and Selectivities. *Chem. Rev.* **111**, 5042–5137 (2011).

6. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).

7. Crawford, J. M., Kingston, C., Toste, F. D. & Sigman, M. S. Data Science Meets Physical Organic Chemistry. *Acc. Chem. Res.* **54**, 3136–3148 (2021).

8. Rinehart, N. I., Zahrt, A. F., Henle, J. J. & Denmark, S. E. Dreams, False Starts, Dead Ends, and Redemption: A Chronicle of the Evolution of a Chemoinformatic Workflow for the Optimization of Enantioselective Catalysts. *Acc. Chem. Res.* **54**, 2041–2054 (2021).

9. Rinehart, N. I. *et al.* A machine-learning tool to predict substrate-adaptive conditions for Pd-catalyzed C–N couplings. *Science* **381**, 965–972 (2023).

10. Zahrt, A. F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).

11. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).

12. 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).

13. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **6**, 1379–1390 (2020).

14. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**, 015016 (2021).

15. Li, B. *et al.* A deep learning framework for accurate reaction prediction and its application on high-throughput experimentation data. *J. Cheminformatics* **15**, 72 (2023).

16. Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).

17. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).

18. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **3**, 434–443 (2017).

19. Tu, Z. & Coley, C. W. Permutation Invariant Graph-to-Sequence Model for Template-Free Retrosynthesis and Reaction Prediction. *J. Chem. Inf. Model.* **62**, 3503–3513 (2022).

20. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).

21. Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **11**, 3355–3364 (2020).

22. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).

23. Kim, E., Lee, D., Kwon, Y., Park, M. S. & Choi, Y.-S. Valid, Plausible, and Diverse Retrosynthesis Using Tied Two-Way Transformers with Latent Variables. *J. Chem. Inf. Model.* **61**, 123–133 (2021).

24. Sacha, M. *et al.* Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. *J. Chem. Inf. Model.* **61**, 3273–3284 (2021).

25. Mao, K. *et al.* Molecular graph enhanced transformer for retrosynthesis prediction. *Neurocomputing* **457**, 193–202 (2021).

26. Irwin, R., Dimitriadis, S., He, J. & Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **3**, 015022 (2022).

27. Zhu, J. *et al.* Dual-view Molecular Pre-training. in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 3615–3627 (ACM, Long Beach CA USA, 2023). doi:10.1145/3580305.3599317.

28. Lu, J. & Zhang, Y. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *J. Chem. Inf. Model.* **62**, 1376–1387 (2022).

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

30. Shi, R., Yu, G., Huo, X. & Yang, Y. Prediction of chemical reaction yields with large-scale multi-view pre-training. *J. Cheminformatics* **16**, 22 (2024).

31. Xia, J. *et al.* Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023* (2023).

32. Ying, C. *et al.* Do Transformers Really Perform Badly for Graph Representation? in *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 28877–28888 (Curran Associates, Inc., 2021).

33. Shi, Y. *et al.* Benchmarking Graphormer on Large-Scale Molecular Modeling Datasets. Preprint at https://doi.org/10.48550/arXiv.2203.04810 (2023).

34. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

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

36. Li, X., Zhang, S., Xu, L. & Hong, X. Predicting Regioselectivity in Radical C−H Functionalization of Heterocycles through Machine Learning. *Angew. Chem. Int. Ed.* **59**, 13253–13259 (2020).

37. Schneider, N., Stiefl, N. & Landrum, G. A. What’s What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **56**, 2336–2346 (2016).

38. Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).

39. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).

40. Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).

41. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).

42. Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **55**, 39–53 (2015).

43. Schleinitz, J. *et al.* Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C–O Couplings. *J. Am. Chem. Soc.* **144**, 14722–14730 (2022).

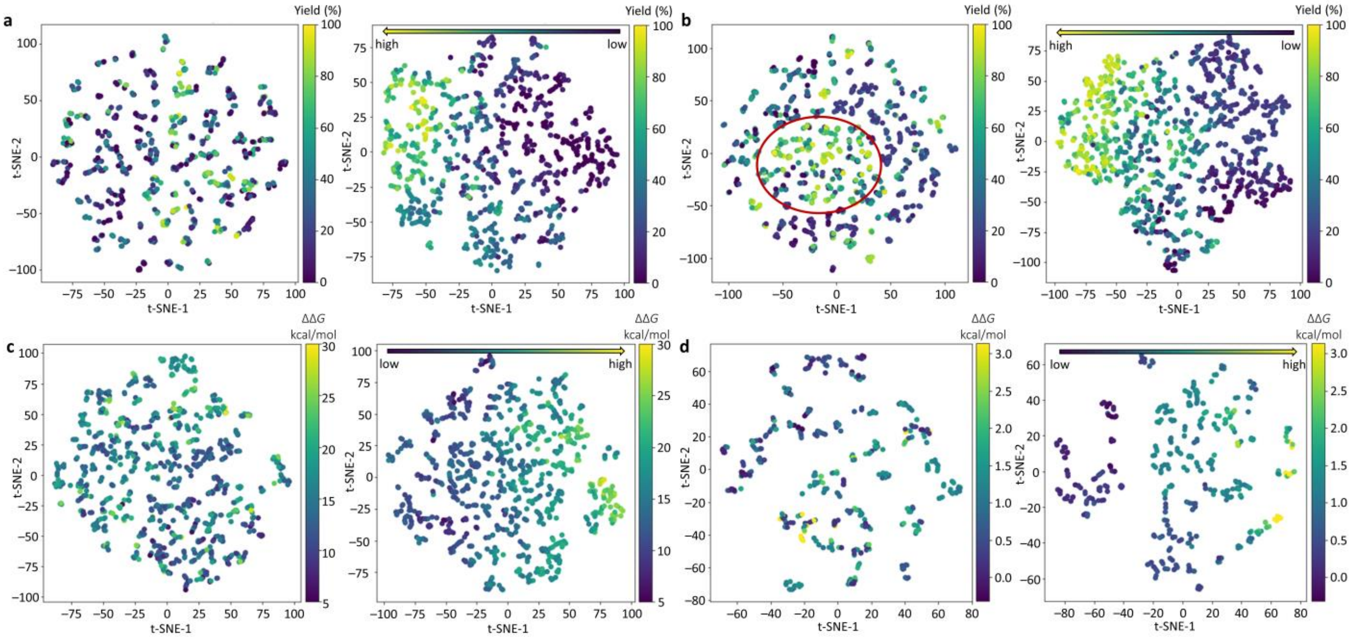
44. Xu, L. *et al.* Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem. Int. Ed.* **60**, 22804–22811 (2021).

45. Xu, L.-C. *et al.* Enantioselectivity prediction of pallada-electrocatalysed C–H activation using transition state knowledge in machine learning. *Nat. Synth.* **2**, 321–330 (2023).

46. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

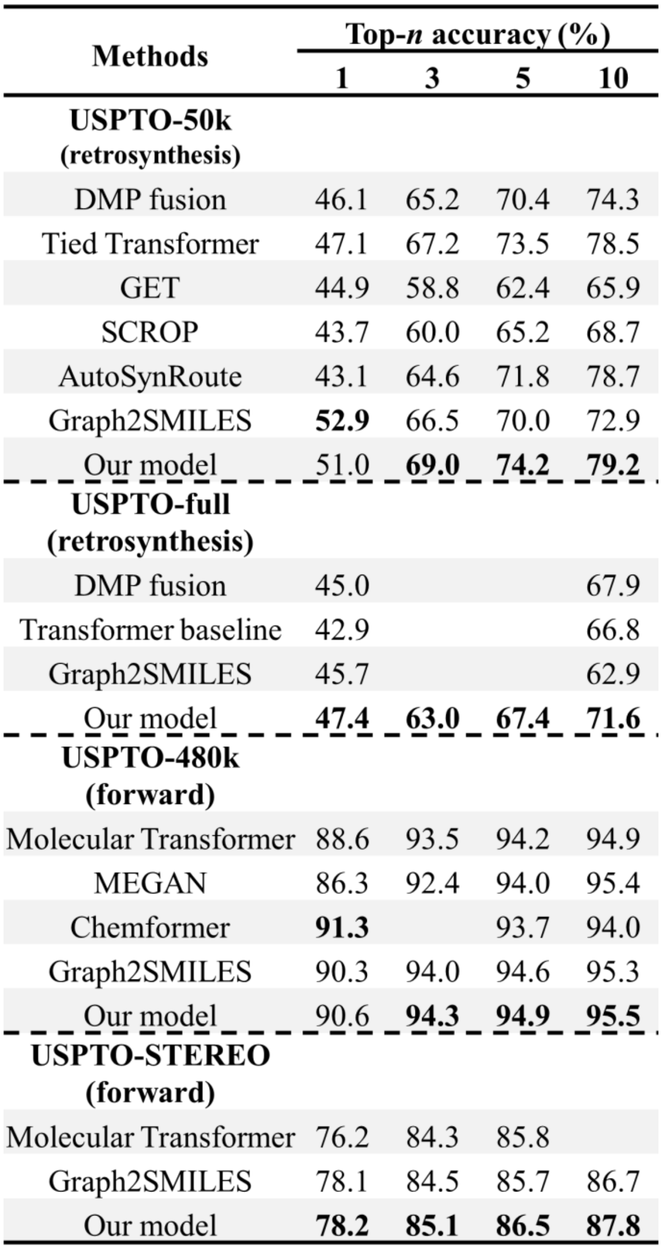
47. Newman-Stonebraker, S. H. *et al.* Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374**, 301–308 (2021).

Extended data



Extended Data Fig. 1: t-SNE visualization of test set reaction embeddings across four benchmark datasets. a, Buchwald-Hartwig; **b,** Suzuki-Miyaura dataset; **c,** Radical C–H functionalization dataset; **d,** asymmetric thiol addition dataset. For each subfigure: left panel, from pretrained model; right panel, from model fine-tuned on corresponding training set.

Extended Data Table 1 Comparative analysis of RXNGraphormer and other representative models on USPTO-50k, USPTO-full, USPTO-480k, and USPTO-STEREO datasets. The best performance of each task is shown in bold.



Methods	**Top-*n* accuracy (%)**			
	1	**3**	**5**	**10**
USPTO-50k (retrosynthesis)				
DMP fusion	46.1	65.2	70.4	74.3
Tied Transformer	47.1	67.2	73.5	78.5
GET	44.9	58.8	62.4	65.9
SCROP	43.7	60.0	65.2	68.7
AutoSynRoute	43.1	64.6	71.8	78.7
Graph2SMILES	**52.9**	66.5	70.0	72.9
Our model	51.0	**69.0**	**74.2**	**79.2**
USPTO-full (retrosynthesis)				
DMP fusion	45.0			67.9
Transformer baseline	42.9			66.8
Graph2SMILES	45.7			62.9
Our model	**47.4**	**63.0**	**67.4**	**71.6**
USPTO-480k (forward)				
Molecular Transformer	88.6	93.5	94.2	94.9
MEGAN	86.3	92.4	94.0	95.4
Chemformer	**91.3**		93.7	94.0
Graph2SMILES	90.3	94.0	94.6	95.3
Our model	90.6	**94.3**	**94.9**	**95.5**
USPTO-STEREO (forward)				
Molecular Transformer	76.2	84.3	85.8	
Graph2SMILES	78.1	84.5	85.7	86.7
Our model	**78.2**	**85.1**	**86.5**	**87.8**