nature machine intelligence

Supplementary information

https://doi.org/10.1038/s42256-025-01098-4

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

In the format provided by the authors and unedited

Table of Contents

1. The details of reaction dataset

1.1 Detailed list of open-source datasets for pre-train data collection

The data used for model pre-training comprises multiple open-source chemical reaction datasets, in addition to over one million chemical reaction records collected from the WIPO database. A detailed list of these open-source datasets, along with specific access links, is presented in Table S1.

Table S1| List of open-source datasets with corresponding access links.

Dataset	Access
USPTO[1]	https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873
Open reaction database[2]	https://github.com/open-reaction-database/ord-data
IBM reaction dataset[3]	https://ibm.ent.box.com/v/uspto-yields-data/folder/115917544053
Chemical reaction database[4]	https://kmt.vander-lingen.nl/
CJHIF[5]	https://github.com/jshmjs45/data_for_chem/tree/e51071b52cad54088745ee8b08cc1e798f51a28d

1.2 Data format alignment and standardization

Pre-train reaction data from diverse sources undergo molecular SMILES verification and canonicalization using RDKit[6], with alignment of the reaction dataset formats. Any extraneous auxiliary information, such as atom-mapping details, is removed. The specific alignment process involves concatenating the SMILES strings of reactants, solvents, and additives—components present before the reaction—using a period (".") as a delimiter, followed by canonicalization to produce a unified SMILES representation for the reactants. For products, which appear post-reaction, multiple compound SMILES are also concatenated using a period and canonicalized to generate a unified SMILES representation for the products. Reaction data is then deduplicated to create a standardized pre-trained reaction dataset. For downstream tasks such as reaction performance prediction, the dataset is prepared following the aforementioned alignment and standardization preprocessing steps to generate dataset for model training and validation.

1.3 Fictitious reaction dataset generation

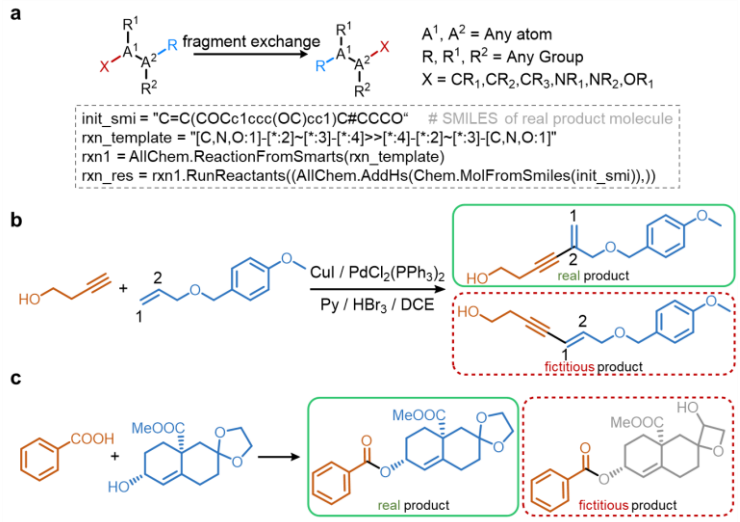


Fig. S1 | The details of fictitious product generation. a, Reaction templates and Python implementation; **b,** Conventional substituent-shift example; **c,** Scaffold-modification example.

To generate negative samples for pre-trained classification tasks, RDKit is utilized to perform fragment exchange reactions on the products of recorded chemical reactions, thereby creating synthetic chemical reaction data. Specifically, as illustrated in Fig. S1a, substituents such as alkyl, alkenyl, alkynyl, amino, and alkoxy groups are transferred from their original attached atom (A^1) to an adjacent atom (A^2) within the same molecule to produce fictitious products. While the algorithm can perform conventional group shift (Fig. S1b), it can also alter molecular scaffold structures (Fig. S1c). If multiple potential fictitious products can be derived from a single molecule, one is randomly selected to match the number of fictitious reactions to the number of real reactions.

1.4 The details of downstream dataset

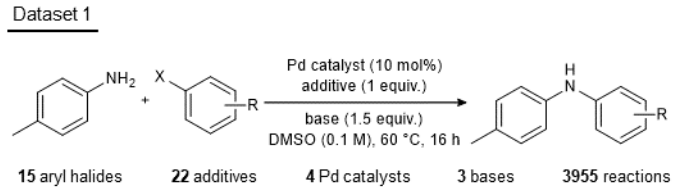


Fig. S2| Dataset 1 comprises Pd-catalyzed C–N cross-coupling reactions between 4-methylaniline and aryl halides, reported by Doyle et al[7].

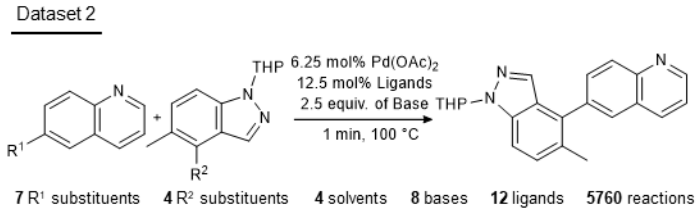


Fig. S3 | Dataset 2 features Pd-catalyzed C–N cross-coupling reactions of quinoline derivatives with indazole derivatives, documented by Perera et al[8].

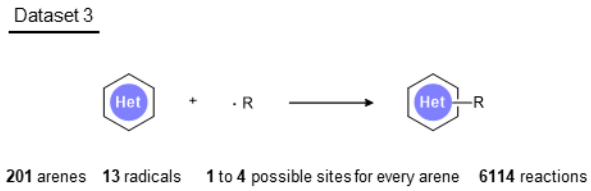


Fig. S4 | Dataset 3 includes the DFT-calculated regioselectivity of heterocycle radical C−H functionalization reactions, reported by Hong et al[9].

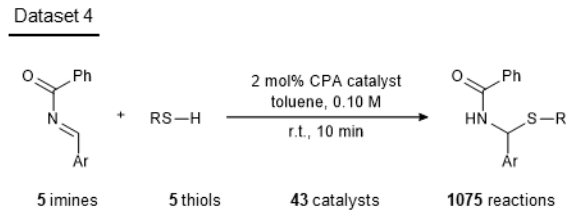


Fig. S5 | Dataset 4 contains data on chiral phosphoric acid-catalyzed thiol addition to *N*-acylimines, reported by Denmark et al[10].

The model was tested on four datasets spanning orders of magnitude to evaluate its performance in regression predictive tasks for reactivity (Buchwald–Hartwig and Suzuki–Miyaura reaction datasets), regioselectivity (the radical C–H functionalization dataset), and enantioselectivity (the asymmetric thiol addition dataset). The radical C–H functionalization dataset, which provides regioselectivity data, was generated using DFT calculations, while the remaining three datasets were derived from chemical experiments. As depicted in Fig. S2, the Buchwald–Hartwig reaction dataset (Dataset 1) encompasses reaction data involving 15 types of aryl halides, 22 additives, 4 Pd catalysts, and 3 bases, totaling 3955 reaction entries. As showed in Fig. S3, the Suzuki–Miyaura reaction dataset (Dataset 2) includes data on 7 quinoline derivatives, 4 indazole derivatives, 4 solvents, 8 bases (including one blank), and 12 ligands (including one blank), totaling 5760 reaction entries. The radical C–H functionalization dataset (Dataset 3) comprises 201 types of arenes and 13 types of radicals, with arenes featuring 1 to 4 potential regioselective sites, resulting in a total of 6114 reaction entries (Fig. S4). These regioselectivity data were obtained through high-precision DFT calculations, with geometry optimizations of all minima and transition states performed using the B3LYP[11,12] functional with a 6-311+G(2d,p) basis set. Single-point energies were computed using the M06-2X[13] functional and aug-cc-pVTZ basis set[14]. Solvation energy corrections in water were evaluated using a self-consistent reaction field (SCRF) method with the SMD model[15]. The asymmetric thiol addition dataset involves 5 imines, 5 thiols, and 43 catalysts, comprising a total of 1075 experimental entries (Fig. S5). The pre-processed datasets are accessible via https://github.com/licheng-xu-echo/RXNGraphormer.

The model's predictive performance was evaluated on downstream tasks of reaction retrosynthesis prediction using the USPTO-50k and USPTO-full datasets, and on forward synthesis prediction tasks using the USPTO-480k and USPTO-STEREO datasets. The data partitioning followed the same ratios reported by Coley[16], as detailed in Table S2. Access to these datasets can be obtained via https://github.com/licheng-xu-echo/RXNGraphormer.

Table S2| The details of datasets employed for evaluating the model's retrosynthesis and forward synthesis prediction capabilities.

Dataset	Synthesis planning type	Training set size	Validation set size	Test set size
USPTO-50k	retrosynthesis	40,008	5,001	5,007
USPTO-full	retrosynthesis	810,496	101,311	101,311
USPTO-480k	forward synthesis	409,035	30,000	40,000
USPTO-STEREO	forward synthesis	902,581	50,131	50,258

1.5 The details of reaction type for USPTO-50k dataset

To evaluate the pre-trained model's ability to discern similarities and differences among various reaction types, we utilized the USPTO-50k dataset reported by Schneider et al.[17], which is labeled with 50 distinct reaction categories. Each category in this dataset contains 1,000 reactions, totaling 50,000 reaction entries. The class IDs and class names for these reactions are listed in Table S3. The order of reaction indices in this table corresponds to the orientation of the reaction type axis in the heatmap displayed in Fig. 3 of the main text. A detailed methodology for calculating the distances between reaction classes is provided in section 3.1.

Table S3 | Class ID, major category, and class name of USPTO-50k dataset.

Index	Class ID	Major category defined in main text	Class name
1	1.2.1	1	Aldehyde reductive amination
2	1.2.4	1	Eschweiler-Clarke methylation
3	1.2.5	1	Ketone reductive amination
4	1.3.6	1	Bromo N-arylation
5	1.3.7	1	Chloro N-arylation
6	1.3.8	1	Fluoro N-arylation
7	1.6.2	1	Bromo N-alkylation
8	1.6.4	1	Chloro N-alkylation
9	1.6.8	1	Iodo N-alkylation
10	1.7.4	1	Hydroxy to methoxy
11	1.7.6	1	Methyl esterification
12	1.7.7	1	Mitsunobu aryl ether synthesis
13	1.7.9	1	Williamson ether synthesis

14	1.8.5	1	Thioether synthesis
15	2.1.1	2	Amide Schotten-Baumann
16	2.1.2	2	Carboxylic acid + amine reaction
17	2.1.7	2	N-acetylation
18	2.2.3	2	Sulfonamide Schotten-Baumann
19	2.3.1	2	Isocyanate + amine reaction
20	2.6.1	2	Ester Schotten-Baumann
21	2.6.3	2	Fischer-Speier esterification
22	2.7.2	2	Sulfonic ester Schotten-Baumann
23	3.1.1	3	Bromo Suzuki coupling
24	3.1.5	3	Bromo Suzuki-type coupling
25	3.1.6	3	Chloro Suzuki-type coupling
26	3.3.1	3	Sonogashira coupling
27	3.4.1	3	Stille reaction
28	5.1.1	4	N-Boc protection
29	6.1.1	5	N-Boc deprotection
30	6.1.3	5	N-Cbz deprotection
31	6.1.5	5	N-Bn deprotection
32	6.2.1	5	CO2H-Et deprotection
33	6.2.2	5	CO2H-Me deprotection
34	6.2.3	5	CO2H-tBu deprotection
35	6.3.1	5	O-Bn deprotection
36	6.3.7	5	Methoxy to hydroxy
37	7.1.1	6	Nitro to amino
38	7.2.1	6	Amide to amine reduction
39	7.3.1	6	Nitrile reduction
40	7.9.2	6	Carboxylic acid to alcohol reduction

41	8.1.4	7	Alcohol to aldehyde oxidation
42	8.1.5	7	Alcohol to ketone oxidation
43	8.2.1	7	Sulfanyl to sulfinyl
44	9.1.6	8	Hydroxy to chloro
45	9.3.1	8	Carboxylic acid to acid chloride
46	10.1.1	9	Bromination
47	10.1.2	9	Chlorination
48	10.1.5	9	Wohl-Ziegler bromination
49	10.2.1	9	Nitration
50	10.4.2	9	Methylation

2. The details of RXNGraphormer

2.1 The encoder of RXNGraphormer

For the molecular graph, nodes were annotated with eight atomic features as depicted in "node features" of Fig. S6a, and edges were annotated with five bond features as shown in "edge features" of the same sub-figure. These annotations facilitate the capture of the local stereochemical and electronic environments of molecules. These features can be rapidly generated using SMILES via RDKit, without the necessity for 3D molecular structures and quantum chemical calculations.

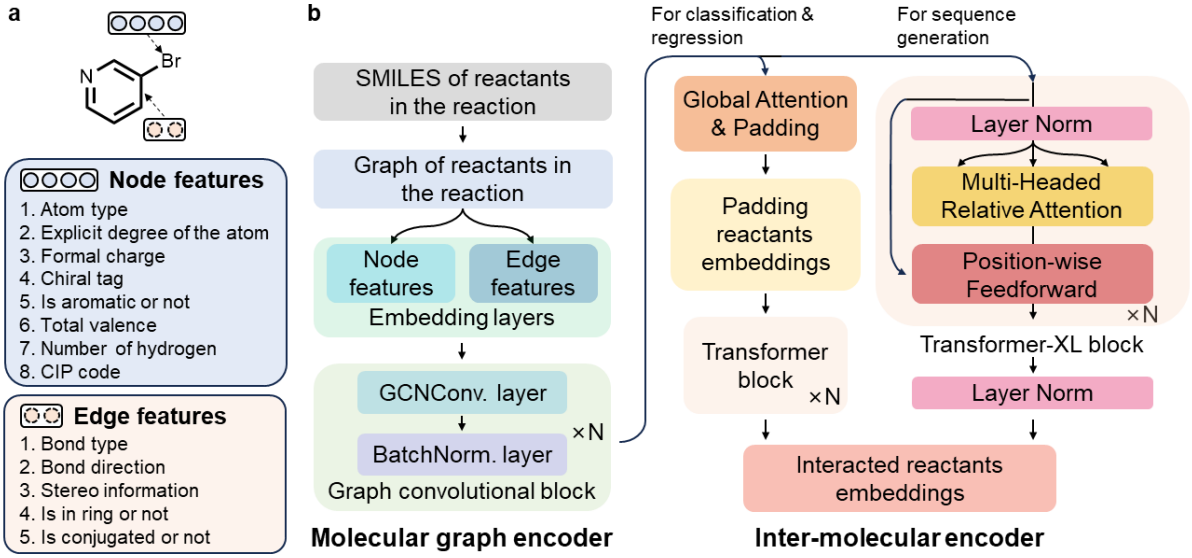


Fig. S6 | a, The node features and edge features generated for each molecule. b, The detailed modules in molecular graph encoder and inter-molecular encoder for RXNGraphormer.

For clarity, Fig. S6b illustrates only the forward propagation process of the reactants' SMILES through the molecule encoder and inter-molecule encoder within the RXNGraphormer architecture. The processing of the products and the delta–mol graph in regression tasks is handled by separate but identical molecular graph encoder and inter-molecular encoder modules, mirroring the approach used for reactants. Initially, the input SMILES strings are converted into molecular graphs using RDKit, and atomic and bond features shown in Fig. S6a are computed. These features are then transformed into node and edge embeddings through embedding layers. Subsequently, the embeddings undergo multiple graph convolutional blocks, each consisting of a GCN convolutional layer and a batch normalization layer, to generate embeddings for each molecule involved in the reaction. For classification and regression tasks, these molecular embeddings are processed through a global attention module and aligned via padding, followed by successive classical Transformer blocks designed to capture intermolecular interactions among reactants, resulting in interacted reactants embeddings. For sequence generation tasks, modified Transformer-XL modules, as reported by Coley et al.[16], are employed to capture interactions between reactants' molecules, with layer normalization yielding the final interacted reactants embeddings.

2.2 Activated modules of RXNGraphormer in various tasks

We employed the unified RXNGraphormer model architecture for the classification pre-training of both real and fictitious reactions, as well as the fine-tuning for downstream tasks, including reaction performance prediction and synthesis planning. Different modules in the model will be activated depending on the type

of task being executed. As depicted in Fig. S7, during the reaction pre-training classification task, the reactant graph encoder processes information from the reactants group, which includes reactants, solvents, and additives, generating embeddings for each reactant molecule. Subsequently, the inter-reactant encoder captures intermolecular interactions among the reactants to produce interacted reactants embeddings, just as mentioned above. Similarly, for the products, the product graph encoder and the inter-product encoder generate embeddings for each product molecule and capture intermolecular interactions among the products, respectively, resulting in interacted products embeddings. Subsequently, the concatenated feature vector - comprising (1) the element-wise absolute difference between interacted reactant and product embeddings (|R - P|), (2) the interacted reactant embeddings (R), and (3) the interacted product embeddings (P) - is fed into a classification module composed of fully connected layers, which outputs a binary classification indicating whether the reaction is real.

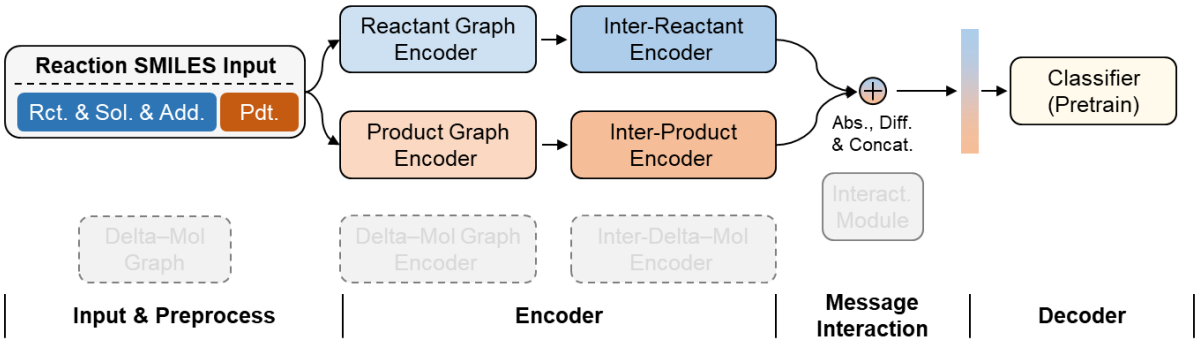


Fig. S7 | Modules activated in the RXNGraphormer during the classification task. Modules not in use are shaded grey.

In the downstream fine-tuning tasks for reactivity and selectivity regression prediction, the reactant graph encoder, inter-reactant encoder, product graph encoder, and inter-product encoder within the RXNGraphormer are all pre-trained. The reaction SMILES input is processed using the delta–link method to generate delta–mol graphs, which includes information on bond formations and cleavages (for details on the implementation of the delta–link method, see section 2.3). Similar to the processing of reactants and products, embeddings for each delta–mol graph are generated by the delta–mol graph encoder. The inter-delta–mol encoder captures more complex interactions between delta–mol graphs, resulting in interacted delta–mol graph embeddings. These embeddings are then concatenated with those of the reactants and products, forming an integrated embedding that incorporates information about reactants, products, and mechanism-related bond-change intermediates. This integrated embedding is processed by an interaction module composed of fully connected layers to facilitate extensive interactions during the reaction process, producing final reaction embeddings that capture both the intrinsic reaction characteristics and intermediary processes. These embeddings are subsequently processed through a regression module, also composed of fully connected layers, to give the reactivity and selectivity of the reaction (as shown in Fig. S8).

During the forward synthesis prediction task for synthesis planning (Fig. S9), the model only receives a concatenated SMILES string of reactants, solvents, and additives, without differentiation. This input is processed solely through the RXNGraphormer's reactant graph encoder and inter-reactant encoder, and the resulting interacted reactant embeddings are directly utilized by the sequence decoder to generate the SMILES of the products. Conversely, in the retrosynthesis prediction task (Fig. S10), the model's input consists of the product's SMILES. This input is processed through the RXNGraphormer's product graph encoder and inter-product encoder, and the interacted product embeddings produced are directly used by the sequence decoder to generate the SMILES of the reactants.

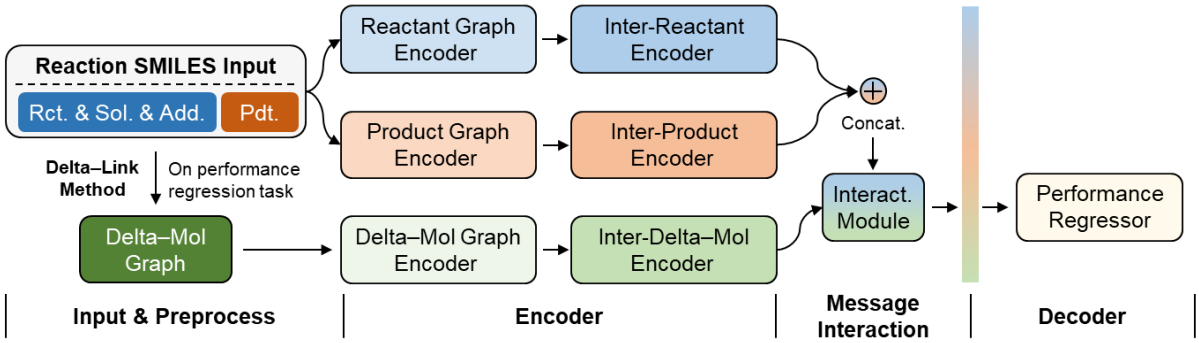


Fig. S8 | Modules activated in the RXNGraphormer during the regression task.

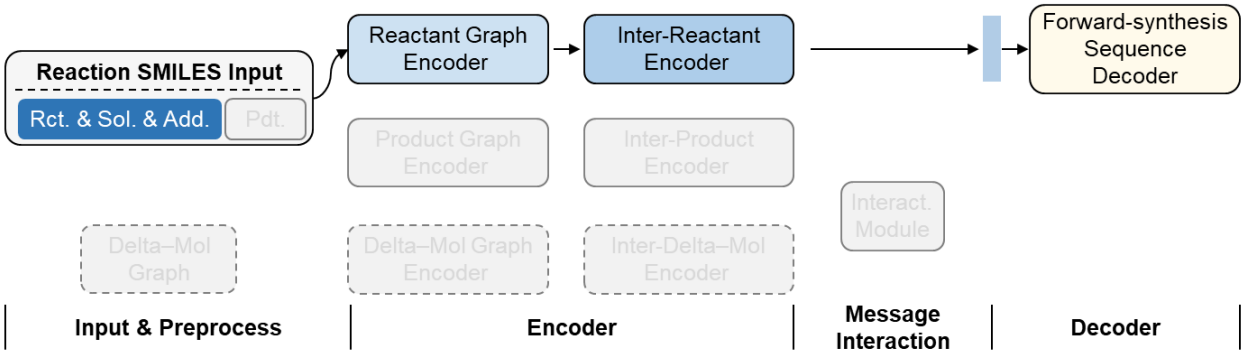


Fig. S9 | Modules activated in the RXNGraphormer during the forward synthesis prediction task. Modules not in use are shaded grey.

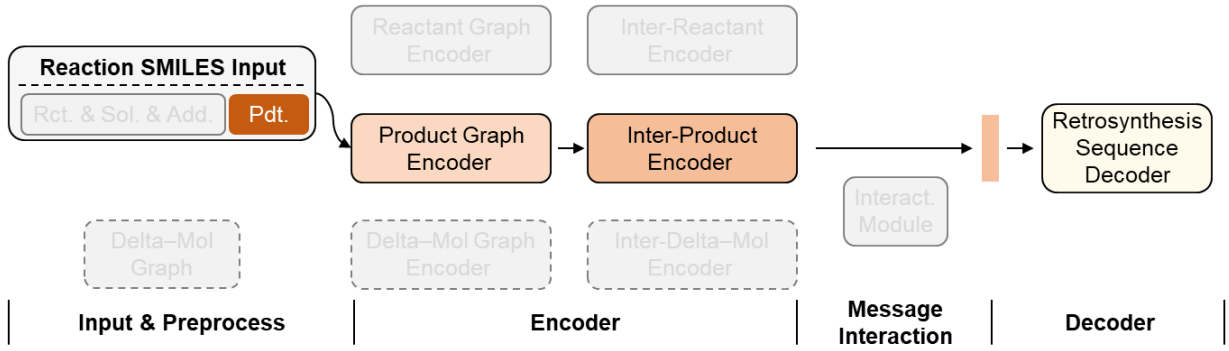


Fig. S10 | Modules activated in the RXNGraphormer during the retrosynthesis prediction task. Modules not in use are shaded grey.

2.3 The details of delta–link method

To enrich the model's information of reaction mechanisms in the regression prediction tasks for reaction performance, we developed a delta–link method inspired by reaction mechanisms. This method captures the changes in chemical bonds between reactants and products before and after the chemical reaction. The specific workflow is illustrated in Fig. S11. The input for the delta–link method comprises "mixed SMILES of reactant and other reagents" and "product SMILES". The first step in the process involves distinguishing the reactant SMILES from other reagent SMILES, pinpointing the actual reactants involved in the reaction[18–20]. The second step compares the chemical bond differences between the reactant and product SMILES to identify formed and broken bonds. Subsequently, the third step involves adding or removing bonds to interpolate the changes in chemical bonds before and after the reaction, thereby generating various "frames"

of the reaction, known as delta–mol graphs. The final step simplifies the delta–mol graphs by deduplicating SMILES and removing hydrogen atom SMILES, resulting in a concise graph that contains detailed information about the reaction process. For details on the algorithm implementation, please visit https://github.com/licheng-xu-echo/RXNGraphormer.

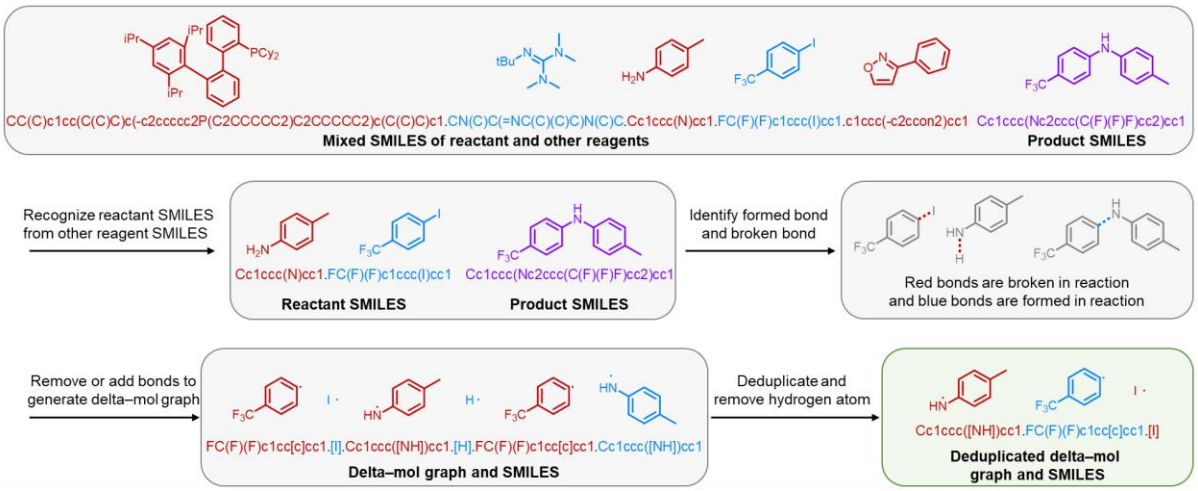


Fig. S11 | The workflow of delta–link method.

2.4 Hyperparameters setting for RXNGraphormer

The hyperparameters for the pre-trained classification model, the downstream reaction performance regression model, and the synthesis planning model are detailed in Table S4. Detailed hyperparameter setting can be found in https://github.com/licheng-xu-echo/RXNGraphormer.

Table S4 | Hyperparameter setting used in the experiments for different task.

	Parameter	Value
Shared parameters	GNN type	GCN
	GNN layer number	4
	Embedding dimension size	256
	Jk-concat	last
Pre-train model	Node readout type	mean
	GNN aggregation type	mean
	Inter-molecular layer type	Transformer
	Inter-molecular layer number	4
	Number of attention heads in inter-molecular layer	2
	Output layer number	3

Regression model	Node readout type	mean
	GNN aggregation type	mean
	Inter-molecular layer type	Transformer
	Inter-molecular layer number	4
	Number of attention heads in inter-molecular layer	2
	Inter-molecular readout type	mean
	Interaction module layer type	Fully-connect
	Interaction module layer number	1
	Output layer number	1
Synthesis planning model	Node readout type	sum
	GNN aggregation type	add
	Inter-molecular layer type	Transformer-XL-derived
	Inter-molecular layer number	6
	Number of attention heads in inter-molecular layer	8
	Inter-molecular encoder filter size	2048
	Decoder layers	6
	Number of attention heads in decoder layer	8

3. Model performance and analysis

3.1 Fitting results of pre-trained model and reaction distance analysis

The model's pre-training task was designed to differentiate between real and fictitious reactions within a dataset comprising 13 million entries. The dataset was split into training and test sets at a ratio of 9:1, with the model achieving an accuracy of 0.8932 on the test set.

To assess the model's ability to discern similarities and differences among various reaction types, we used the USPTO-50k dataset as a case study. After pre-training, the RXNGraphormer was utilized to generate reaction embeddings for these reactions, corresponding to the outputs from the model's penultimate layer. We calculated the average Euclidean distance between each pair of reaction classes within this dataset using the formula provided below.

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

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. Here, N and M denote the number of reactions in each class, with both N and M equaling 1,000 for the USPTO-50k dataset. The dimension d represents the number of features in each reaction embedding generated by the RXNGraphormer, which is 768.

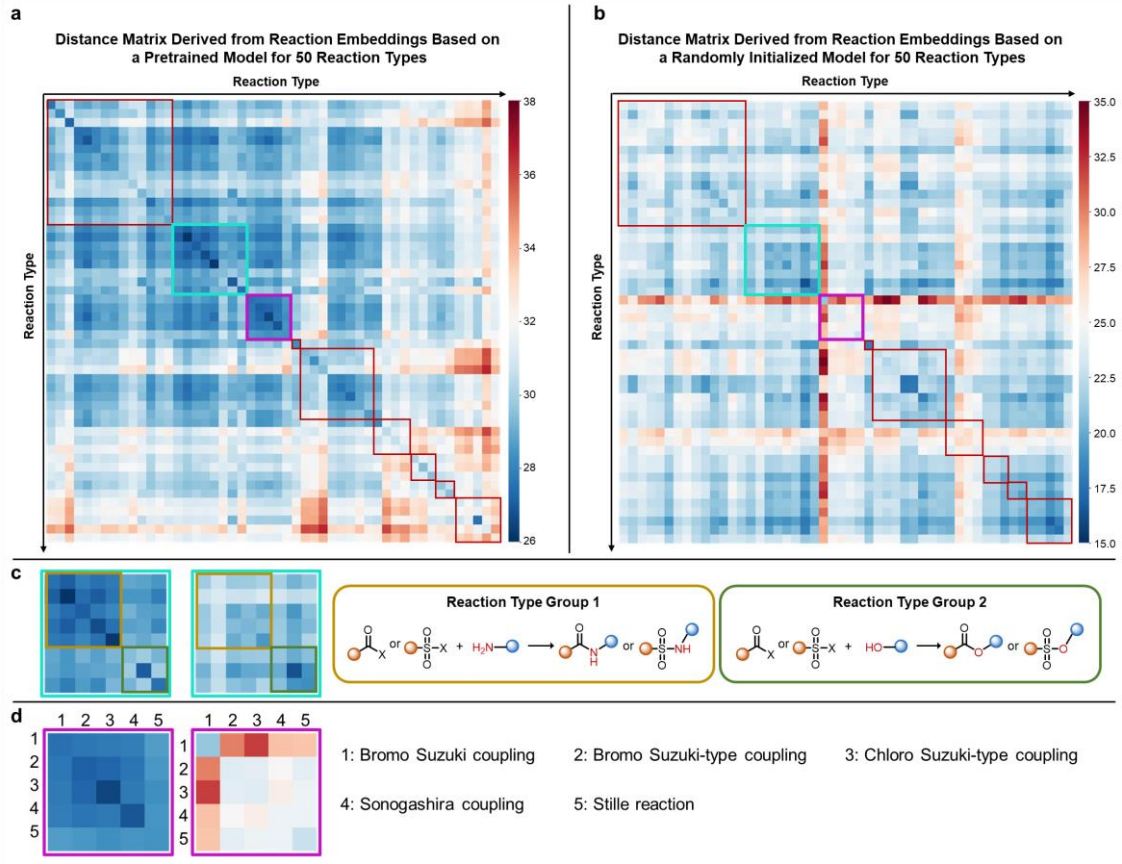


Fig. S12 | The analysis for distance matrices that generated by pre-trained RXNGraphormer and randomly initialized RXNGraphormer.

The distance matrices generated from 50 reaction classes are presented as heatmaps in Fig. S12a and S12b. Fig. S12a displays the reaction distance heatmap calculated from reaction embeddings generated by the pre-trained RXNGraphormer, while Fig. S12b shows the heatmap derived from reaction embeddings generated by a randomly initialized RXNGraphormer. The randomly initialized model, not having been trained at the reaction level, produces embeddings that essentially represent average molecular graph characteristics of the reactions. By comparing Fig. S12a and S12b, it is evident that the pre-trained RXNGraphormer can significantly distinguish between different types of reactions, despite the pre-training task not involving any reaction type information specifically.

Specifically, Fig. S12c illustrates the distance matrices between reaction type group 1 (amides and sulfonamides formation reactions) and reaction type group 2 (ester and sulfonic ester formations) generated by both the pre-trained and randomly initialized models. It is clear that the matrix from the pre-trained model (the left one) shows closer distances and more similar reaction features within reaction type group 1, forming a distinct subgroup, with reaction type group 2 forming another. In contrast, the matrix from the randomly initialized model (the right one) does not display such characteristics.

Fig. S12d presents the distance matrices for a series of coupling reactions. In the latent reaction space generated by the pre-trained model, these reactions are closely clustered, as they all generate new C–C bond; however, the randomly initialized model does not reflect this characteristic. Specifically, the distance matrix from the randomly initialized model shows significant differences between 'bromo Suzuki coupling' and four other types of reactions. Analysis reveals that in the USPTO-50k dataset, every entry under 'bromo Suzuki coupling' (class ID 3.1.1) includes information about a Pd catalyst, whereas 'bromo Suzuki-type coupling' (class ID 3.1.5) and 'chloro Suzuki-type coupling' (class ID 3.1.6) rarely include Pd catalyst information, and 'Sonogashira coupling' (class ID 3.3.1) and 'Stille reaction' (class ID 3.4.1) only occasionally do. Thus, as previously mentioned, the phenomenon observed in Fig. S12d arises because the embeddings generated by the randomly initialized model represent the average molecular graph characteristics without any reaction-level recognition. In contrast, the model pre-trained on distinguishing whether reactions are real chemical reactions learned higher-dimensional information at the reaction level and developed the ability to differentiate between different reactions.

Beyond analyzing relationships within the same reaction types, we also examined an off-diagonal region outlined by orange dashed lines in the heatmap (Fig. S13b), which reflects the pre-trained model's ability to distinguish different reaction types. The horizontal axis of this sub-heatmap displays five reaction types, include bromination, chlorination, Wohl-Ziegler bromination, nitration, and methylation (all belonging to the major category of functional group addition), while the vertical axis presents three reductions and two oxidations. These reaction classes differ fundamentally in their bond-change mechanisms: C–X bond formation versus bond-order changes involving hydrogen/oxygen gain/loss. The substantial differences in chemical bond transformations among these ten reaction types result in the overall warm-toned appearance of this region in the heatmap.

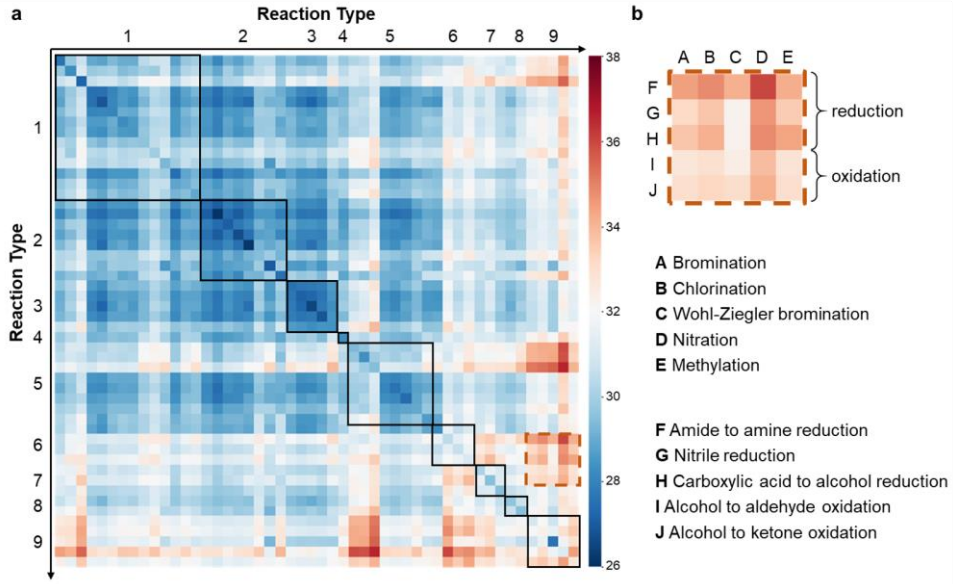


Fig. S13 | Distance matrix analysis of off-diagonal reaction regions representing different reaction types.

3.2 The detailed results of regression models

We conducted ten trials of model training and inference testing on four distinct chemical reaction datasets, using ten different random seeds for random division of training and testing sets. Specifically, these datasets include the Buchwald–Hartwig reaction dataset (dataset 1), the Suzuki–Miyaura reaction dataset (dataset 2), the radical C–H functionalization dataset (dataset 3), and the asymmetric thiol addition dataset (dataset 4). The division ratios for the training and test sets were set at 7:3 for datasets 1 and 2, 8:2 for dataset 3, and 600:475 for dataset 4. Due to the nature of reaction yields, which range from 0% to 100%, we imposed constraints on the model outputs for datasets 1 and 2: predictions below 0 were set to 0, and those above 100 were capped at 100. The results of the ten regression predictions on the training sets and test sets for these datasets are displayed in Fig. S14 to S17.

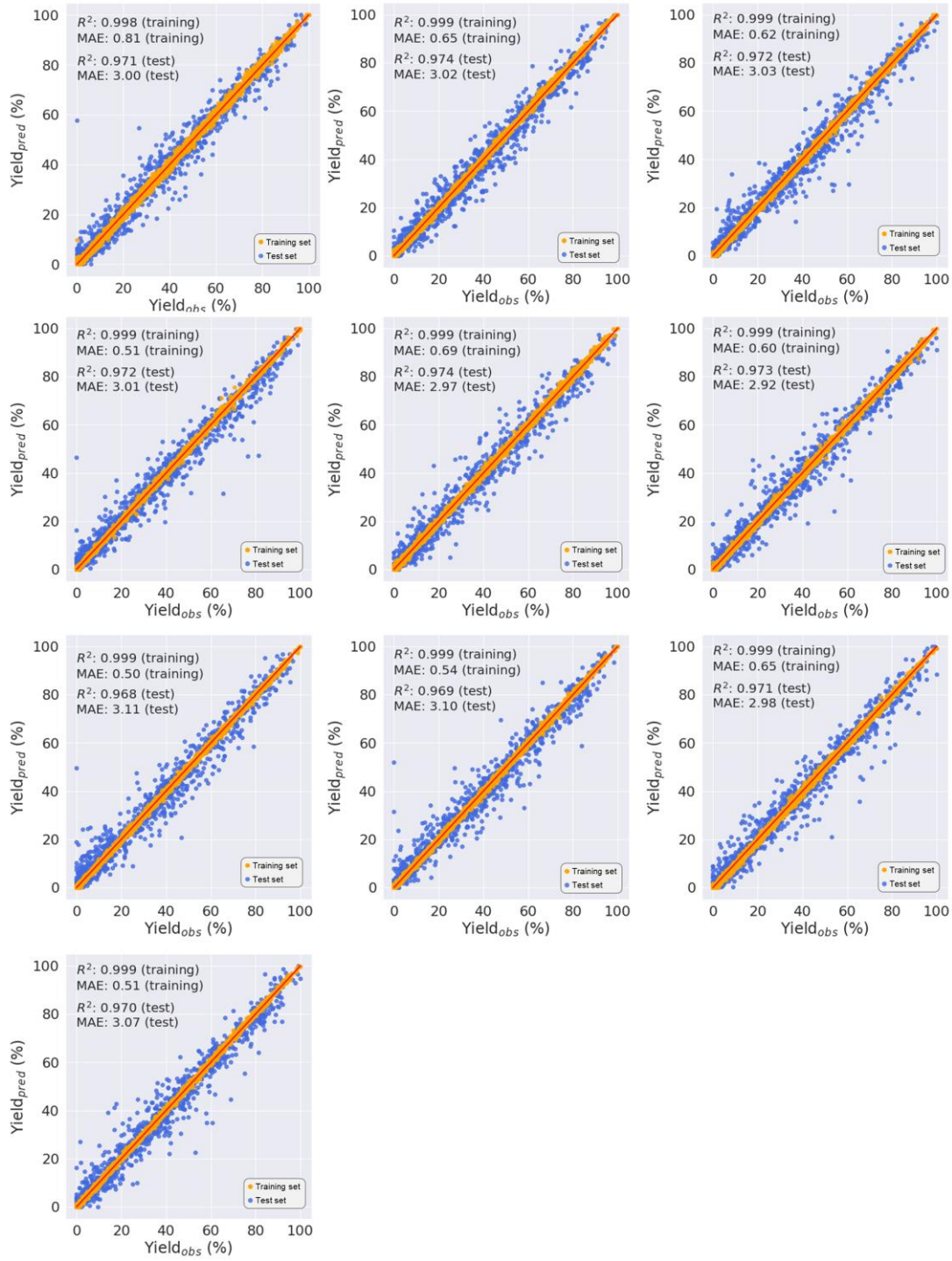


Fig. S14 | Ten regression prediction trials for the Buchwald–Hartwig reaction dataset.

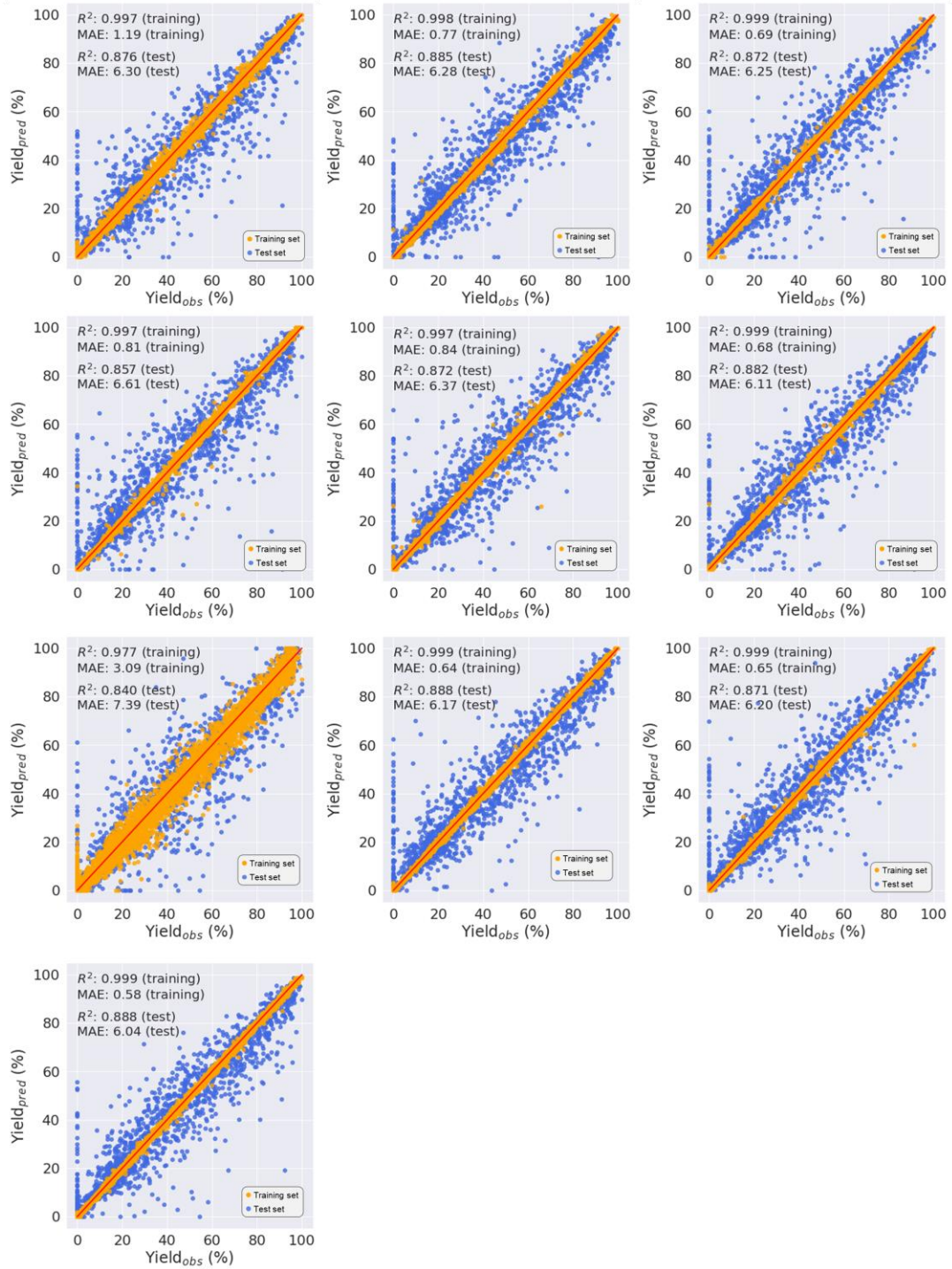


Fig. S15 | Ten regression prediction trials for the Suzuki–Miyaura reaction dataset.

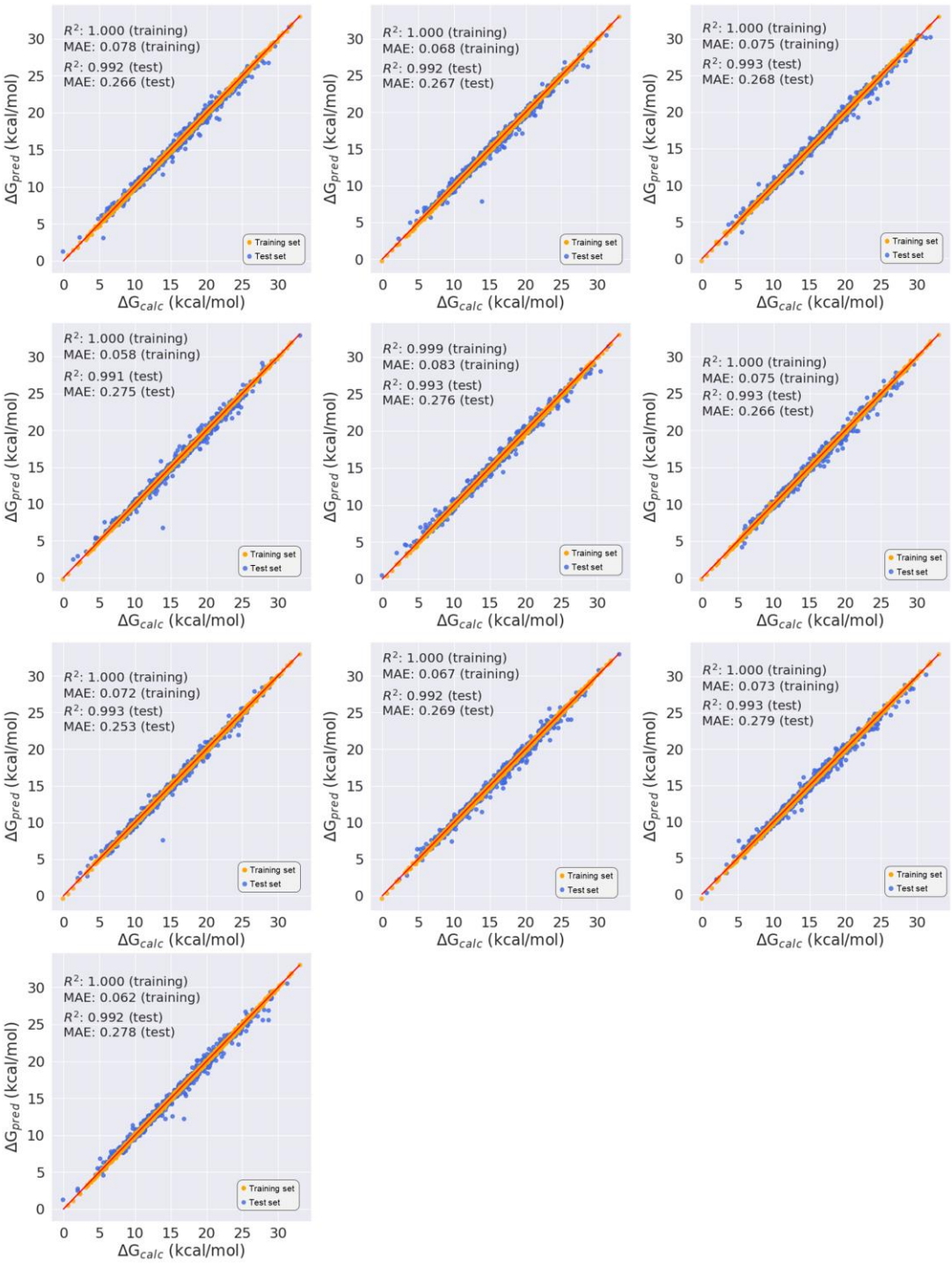


Fig. S16 | Ten regression prediction trials for the radical C–H functionalization dataset.

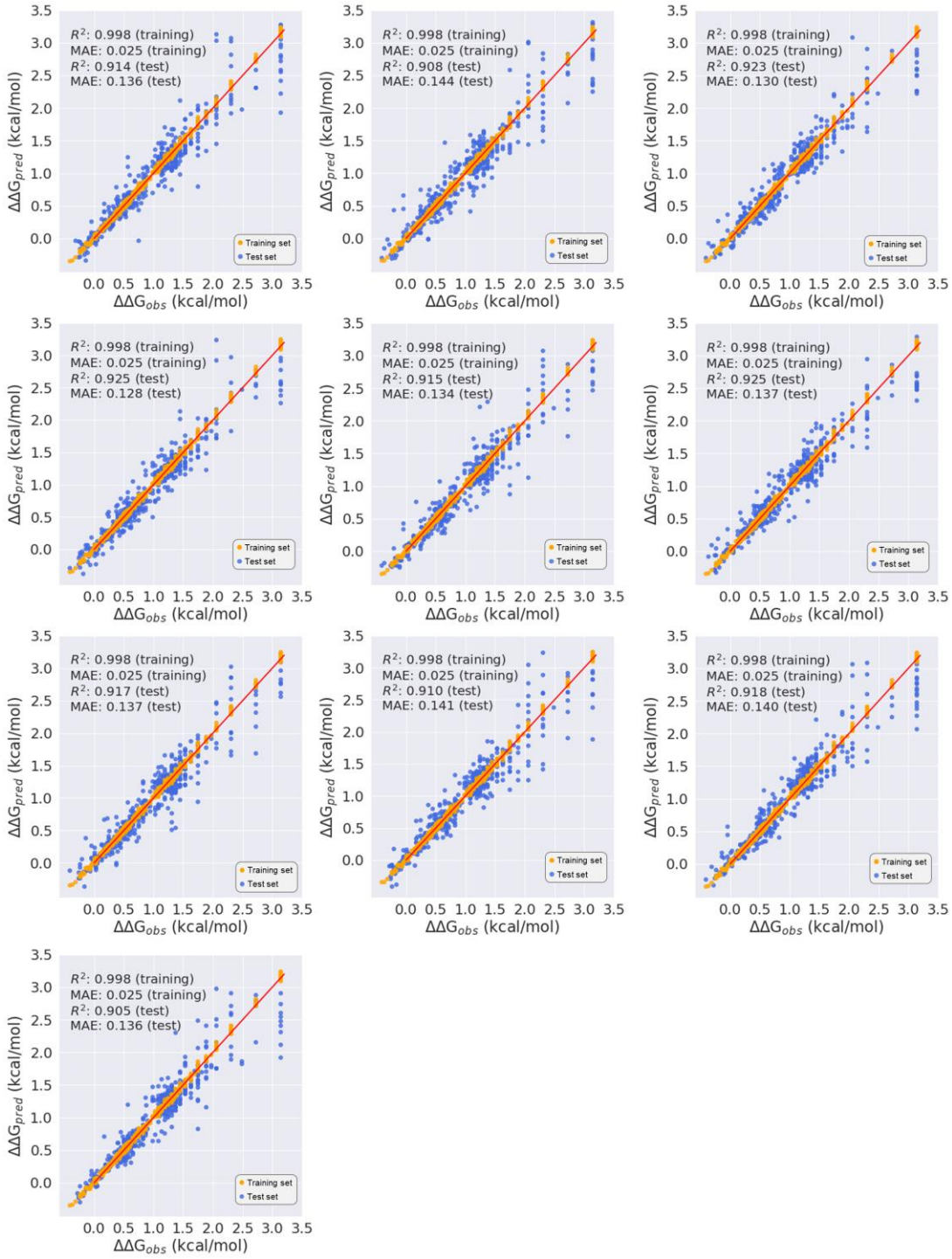


Fig. S17 | Ten regression prediction trials for the asymmetric thiol addition dataset.

We conducted a rigorous comparison of RXNGraphormer with several cutting-edge methods that do not rely on additional DFT calculations as well as the original method reported along with these datasets. These methods include MFF+RF by Glorius[21], which employs multiple molecular fingerprints combined with Random Forest; YieldBERT by Schwaller et al.[3], based on text-based reaction representations; GraphRXN by Chen and Liao et al.[22], which leverages a GNN architecture adapted from chemprop[23]; and ReaMVP by Yang et al.[24], integrating both 2D and 3D molecular graphs for yield prediction. As shown in Table S5, RXNGraphormer substantially outperformed these approaches on datasets 2, 3, and 4—covering reactivity, regioselectivity, and stereoselectivity—achieving R^2 values of 0.873, 0.992, and 0.916, and mean absolute

errors (MAE) of 6.37, 0.270 kcal/mol, and 0.136 kcal/mol, respectively. On the Buchwald–Hartwig reaction yield prediction dataset (dataset 1), performances of RXNGraphormer and ReaMVP were comparable, with R^2 scores of both 0.971, and MAEs of 3.02 and 3.11, respectively.

Table S5 | Comparative analysis of reactivity, regioselectivity, and enantioselectivity predictions by RXNGraphormer and other representative models.

		Dataset 1 (7:3)	Dataset 2 (7:3)	Dataset 3 (8:2)	Dataset 4 (675:400)
MFF-RF	MAE	4.78±0.13	7.49±0.15	0.831±0.017	0.144±0.007
	R^2	0.928±0.006	0.850±0.007	0.939±0.003	0.907±0.011
YieldBERT	MAE	3.99±0.15	8.13±0.34	0.377±0.016	0.164±0.004
	R^2	0.951±0.005	0.815±0.013	0.987±0.002	0.894±0.005
GraphRXN	MAE	4.33±0.12	7.94±0.15	0.339±0.010	0.164±0.007
	R^2	0.951±0.003	0.844±0.007	0.988±0.001	0.892±0.008
ReaMVP	MAE	3.11±0.07	6.59±0.20	0.826±0.026	0.494±0.012
	R^2	0.971±0.002	0.864±0.010	0.932±0.003	-0.253±0.053
Original method	MAE	-	-	0.79[9]	0.152[10]
	R^2	0.92[7]	-	0.939	-
Our model	MAE	**3.02±0.06**	**6.37±0.37**	**0.270±0.007**	**0.136±0.005**
	R^2	**0.971±0.002**	**0.873±0.014**	**0.992±0.001**	**0.916±0.007**

Furthermore, we constructed three aryl halide-based out-of-sample (OOS) test sets (chloride, bromide, iodide) and four additive-based test sets within the Buchwald-Hartwig dataset. The aryl halides in this dataset were categorized into three groups: aryl chlorides, aryl bromides, and aryl iodides, as illustrated in Fig. S18a. The OOS regression predictions for these three categories of aryl halides are displayed in Fig. S18b. The additives in this dataset were partitioned into four distinct OOS test sets, as illustrated in Fig. S19a. The OOS regression predictions for these four categories of additives are displayed in Fig. S19b. RXNGraphormer significantly outperformed other methods in each of these subsets (as detailed in Table S6), with R^2 scores of -0.053, 0.890, 0.823, 0.883, 0.906, 0.792, and 0.736 and MAE values of 15.12, 5.81, 7.54, 6.43, 6.00, 8.50, and 9.94 for the chloride, bromide, iodide, additive 1, additive 2, additive 3, and additive 4 test sets, respectively.

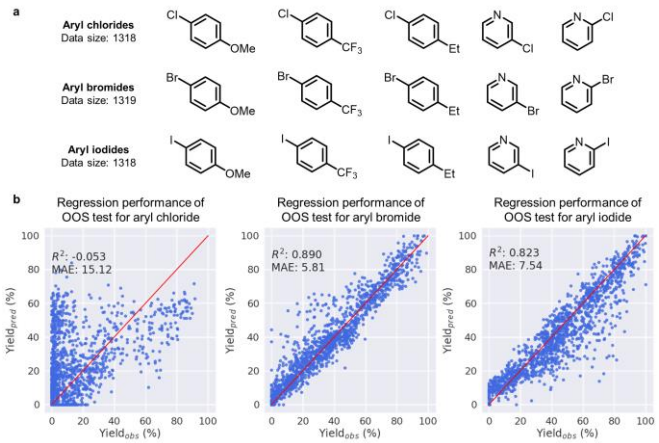


Fig. S18 | The details of aryl halides-based OOS test sets. a, The composition of OOS test sets. **b,** Regression results on these test sets.

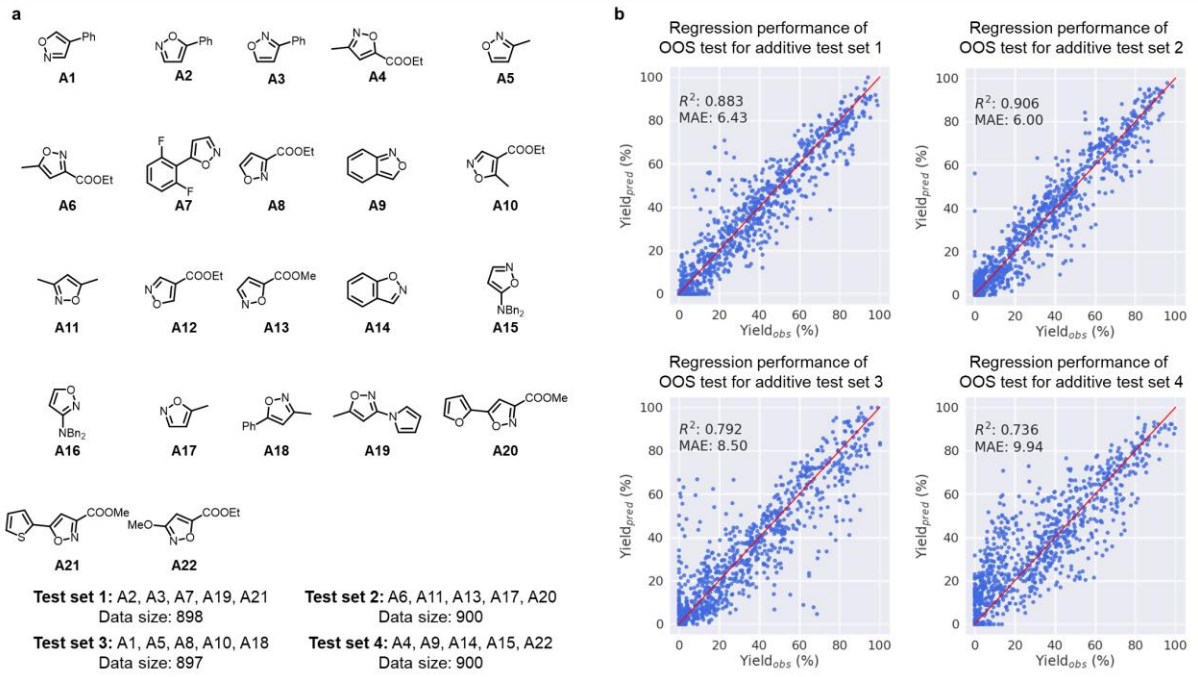
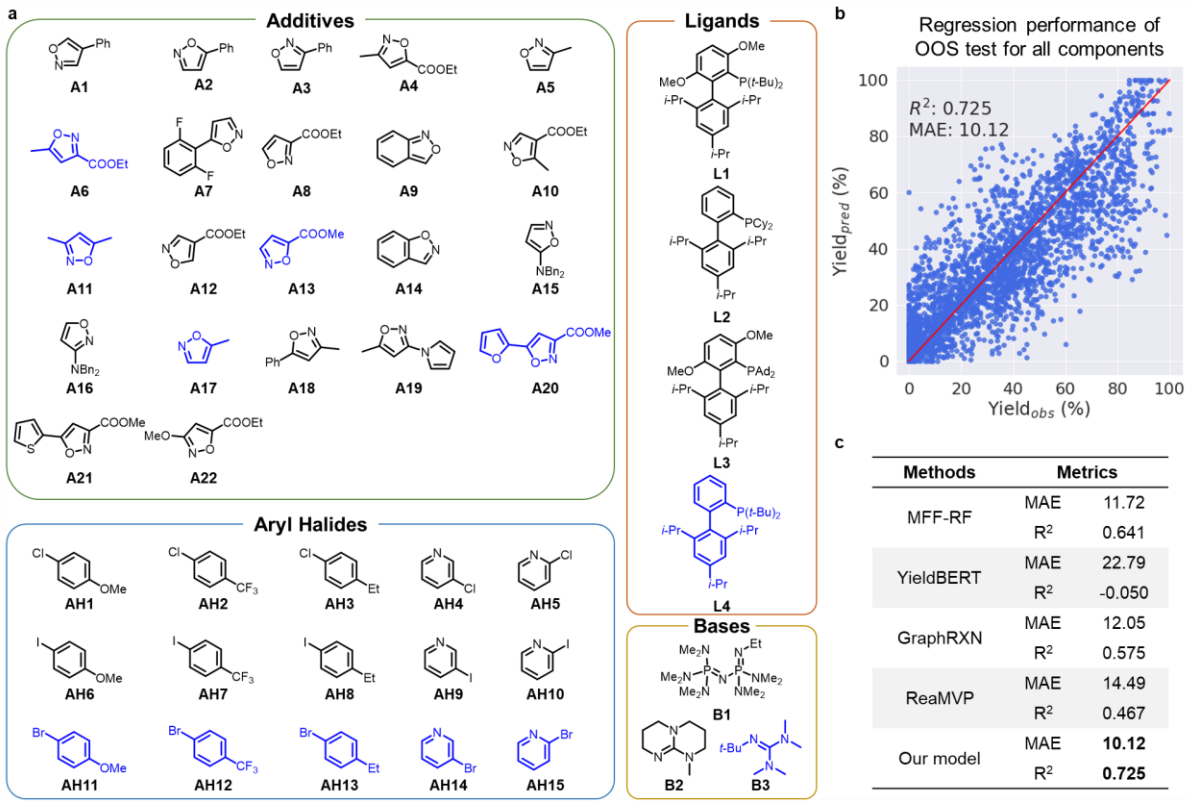


Fig. S19 | The details of additive-based OOS test sets. a, The composition of OOS test sets. **b,** Regression results on these test sets.

Table S6 | Comparative analysis of reactivity predictions for single component OOS test sets on Buchwald-Hartwig dataset by RXNGraphormer and other representative models.

		Chloride	Bromide	Iodide	Additive test 1	Additive test 2	Additive test 3	Additive test 4
MFF-RF	MAE	25.98	8.45	12.29	7.85	8.85	11.80	16.84
	R^2	-1.463	0.799	0.605	0.815	0.780	0.603	0.240
YieldBERT	MAE	26.82	5.91	10.48	7.35	7.27	9.13	13.67
	R^2	-1.605	0.877	0.665	0.824	0.829	0.741	0.444
GraphRXN	MAE	25.94	12.52	18.05	9.69	8.90	9.65	15.65
	R^2	-1.490	0.535	0.159	0.732	0.774	0.730	0.263
ReaMVP	MAE	21.66	7.12	8.88	7.28	6.08	8.97	10.61
	R^2	-0.498	0.840	0.732	0.844	0.896	0.792	0.693
Our model	MAE	**15.12**	**5.81**	**7.54**	**6.43**	**6.00**	**8.50**	**9.94**
	R^2	**-0.053**	**0.890**	**0.823**	**0.883**	**0.906**	**0.792**	**0.736**

Additionally, we created a component-combination test set by selecting unseen compounds across all four dimensions (as shown in Fig. S20a). Test components were defined as: additives from the additive test set 2, aryl bromides, ligand **L4**, and base **B3** (blue-highlighted in Fig. S20a). Any reaction containing $\geqslant$1 test component was allocated to the test set (2,935 reactions), with the remaining 1,020 reactions forming the training set. This partitioning ensures complete exclusion of test components from training data. Under these stringent conditions, RXNGraphormer achieved an R^2 of 0.725 and MAE of 10.12 (Fig. S20b), outperforming benchmark methods (Fig. S20c).



Methods	Metrics	
MFF-RF	MAE	11.72
	R^2	0.641
YieldBERT	MAE	22.79
	R^2	-0.050
GraphRXN	MAE	12.05
	R^2	0.575
ReaMVP	MAE	14.49
	R^2	0.467
Our model	MAE	**10.12**
	R^2	**0.725**

Fig. S20 | The details of leave component-combinations out OOS test on Buchwald-Hartwig dataset. **a,** OOS dataset partitioning details; **b,** Regression results of RXNGraphormer on combination OOS sets; **c,** Comparative performance of RXNGraphormer and other benchmark models.

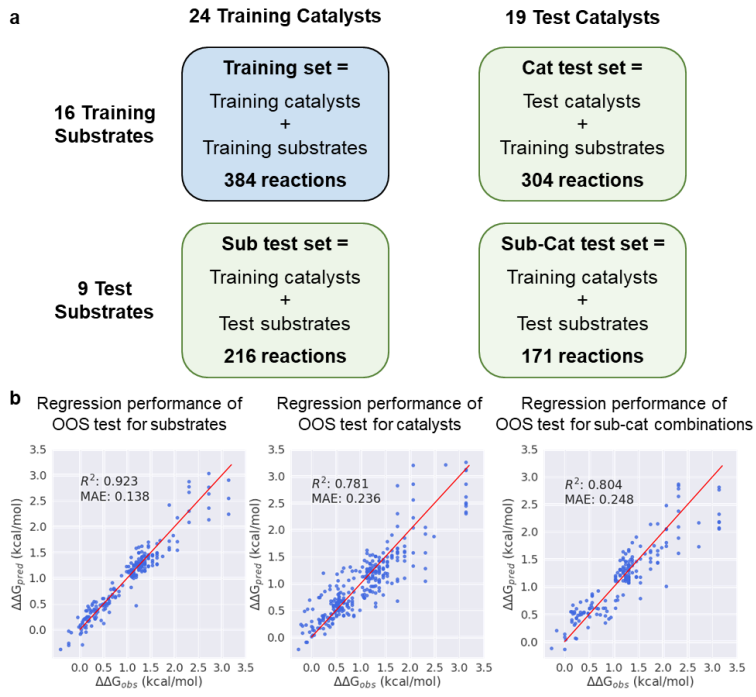


Fig. S21 | The details of substrate, catalyst, and substrate-catalyst combination OOS predictions on asymmetric thiol addition dataset. **a,** OOS dataset partitioning details. **b,** Regression results of RXNGraphormer on OOS sets.

For the asymmetric thiol addition dataset, we performed OOS enantioselectivity predictions for unseen substrates (sub), catalysts (cat), and substrate-catalyst combinations (sub-cat), which is same with the Denmark's original work[10] (Fig. S21a). The OOS regression predictions for these three OOS test sets are displayed in Fig. S21b. Since YieldBERT and ReaMVP were specifically designed for yield prediction (with special handling of 0-100 range labels) and performed poorly in random split tasks, we mainly compared with MFF-RF and GraphRXN methods (as shown in Table S7). 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. The enantioselectivity of this reaction is primarily governed by the chiral phosphoric acid catalyst. In the original work[10], Denmark et al. employed the steric descriptor average steric occupancy (ASO), which captures subtle catalyst structural variations. Consequently, the ASO-based method outperforms our model on cat and sub-cat test sets involving unseen catalyst prediction.

Table S7 Comparative performance of RXNGraphormer and other benchmark models for OOS predictions on sub, cat, and sub-cat test set on asymmetric thiol addition dataset. (MAE values in kcal/mol).

		Sub test set	Cat test set	Sub-cat test set
MFF-RF	MAE	**0.137**	0.247	0.268
	R^2	**0.925**	0.735	0.719
GraphRXN	MAE	0.163	0.246	0.286
	R^2	0.909	0.776	0.711
Original method	MAE	0.161	**0.211**	**0.238**
	R^2	-	-	-
Our model	MAE	0.138	0.236	0.248
	R^2	0.923	0.781	0.804

To obtain a comprehensive evaluation of RXNGraphormer's synthesis planning capabilities, we systematically assessed both forward and retrosynthetic performance across all four USPTO-derived datasets. Beyond demonstrating single-step retrosynthesis prediction on USPTO-50k and USPTO-full datasets, and single-step reaction outcome prediction on USPTO-480k and USPTO-STEREO datasets (as presented in the main text), we conducted reciprocal validation experiments. Specifically, we evaluated reaction outcome prediction performance on USPTO-50k and USPTO-full, while testing single-step retrosynthesis capability on USPTO-480k and USPTO-STEREO. As shown in Table S8, comparative analysis consistently demonstrates superior performance in forward prediction versus retrosynthetic prediction across all datasets.

Table S8 Additional evaluation of RXNGraphormer on USPTO-derived datasets. Forward prediction: USPTO-50k and USPTO-full. Retrosynthesis prediction: USPTO-480k and USPTO-STEREO.

Dataset	Top-*n* accuracy (%)			
	1	3	5	10
USPTO-50k (forward)	82.4	90.2	91.4	92.6
USPTO-full (forward)	70.0	78.5	80.3	82.0
USPTO-480k (retrosynthesis)	19.9	27.9	31.0	34.4
USPTO-STEREO (retrosynthesis)	12.0	16.5	18.2	20.4

3.3 The details of reaction embedding visualization using *t*-SNE

We utilized the outputs from the penultimate layer of models that were only pre-trained, without further fine-tuning for reaction performance prediction tasks, and those fine-tuned on corresponding datasets (Buchwald–Hartwig reaction dataset, Suzuki–Miyaura reaction dataset, the radical C–H functionalization dataset, and the asymmetric thiol addition dataset) as reaction embeddings. We projected these reaction embeddings onto a 2D plane using the *t*-SNE[25] algorithm from scikit-learn[26], where each point on the plane represents a reaction. Fig. S22 presents *t*-SNE projections of reaction embeddings generated by both pre-trained and fine-tuned models for test sets across four reaction performance prediction tasks, with points colored by their target labels. Comparison between left-right panels in each subfigure reveals visually discernible stronger correlation between embeddings and reaction performance after task-specific fine-tuning.

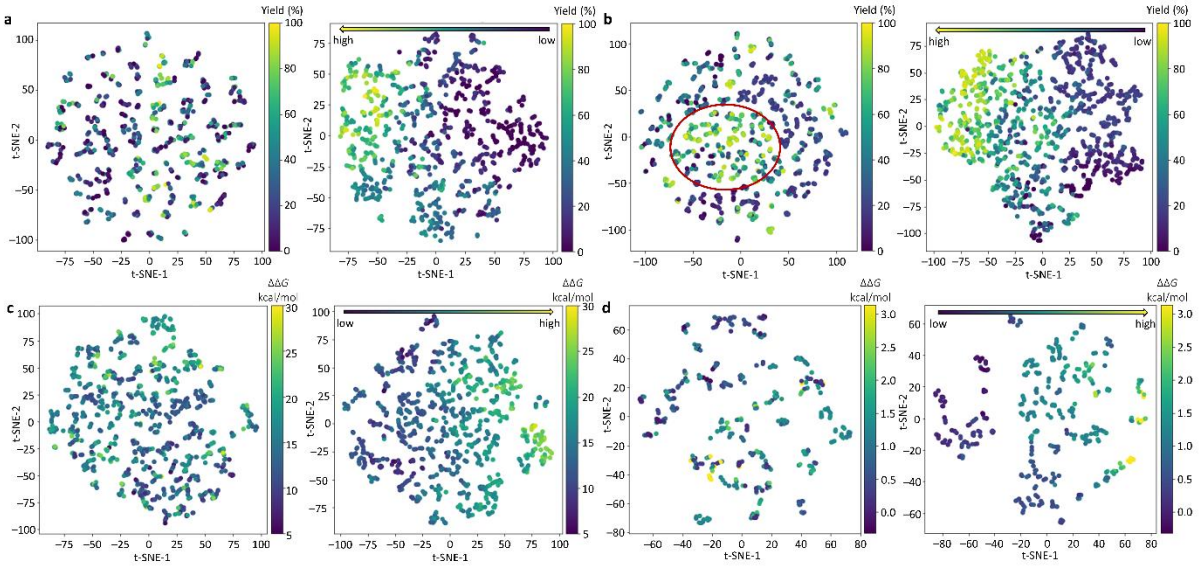


Fig. S22 | *t*-SNE visualization of test set reaction embeddings across four benchmark datasets. a, Buchwald-Hartwig dataset; **b,** Suzuki-Miyaura dataset; **c,** Radical C–H functionalization dataset; **d,** asymmetric thiol addition dataset. For each subfigure: left panel, from pretrained model; right panel, from model fine-tuned on corresponding training set.

Additionally, using RXNGraphormer fine-tuned on the asymmetric thiol addition dataset, we generated reaction embeddings and corresponding *t*-SNE projections (Fig. S23a) for the sub-cat OOS test set. Fig. S23a displays the same data distribution through three complementary perspectives: performance view, catalyst view, and product view. The performance view in Fig. S23a 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 $\Delta\Delta G$ > 2.2 kcal/mol, Fig. S23b 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. S23b right). Product view analysis shows no significant clustering patterns correlated with product types. These patterns collectively demonstrate how our model achieves excellent prediction accuracy in reaction performance tasks.

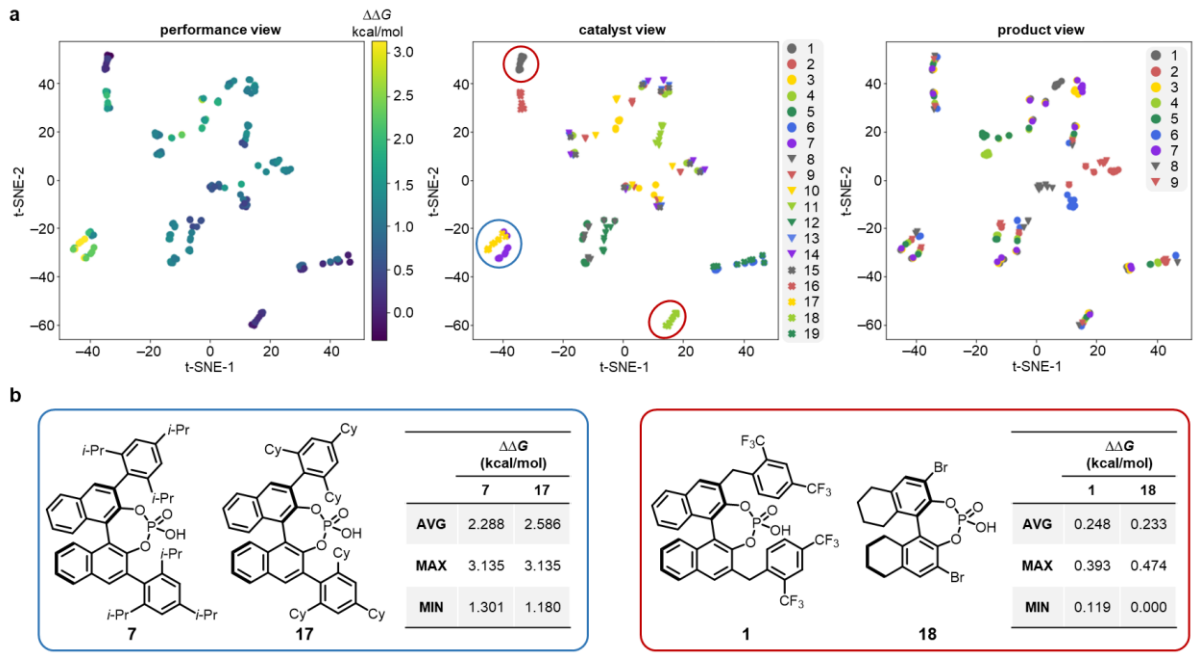


Fig. S23 | *t*-SNE analysis on sub-cat test set of the asymmetric thiol addition dataset a, Performance, catalyst, and product view projections; **b,** Representative catalysts for high and low selectivity.

4. The details of model performance metrics of other benchmarked models

In our downstream tasks, we conducted a comparative analysis of RXNGraphormer with other representative state-of-the-art models. In the reaction performance regression prediction task, we compared the methodologies of MFF-RF[21], YieldBERT[3], GraphRXN[22], and ReaMVP[24]. For ReaMVP, we directly used performance metrics reported in the original publications for the Buchwald-Hartwig and Suzuki-Miyaura random split predictions, as well as for OOS predictions. All other results were obtained by training the model and predicting using the official recommended parameters. In the synthesis planning tasks, we compared RXNGraphormer with ten template-free and atom mapping-free models, using performance metrics directly cited from the original reports.

5. Code and data availability

All preprocessed reaction datasets, as well as the code used for data preprocessing, model training, and model validation, are accessible through https://github.com/licheng-xu-echo/RXNGraphormer.

6. Reference

1. Lowe, D. M. Extraction of chemical structures and reactions from the literature. (Apollo - University of Cambridge Repository, 2012). doi:10.17863/CAM.16293.
2. Kearnes, S. M. *et al.* The Open Reaction Database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
3. 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).
4. The chemical reaction database. https://kmt.vander-lingen.nl/.
5. 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).
6. RDKit: open-source chemoinformatics and machine learning. http://www.rdkit.org.
7. 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).
8. Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018).
9. 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).
10. Zahrt, A. F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
11. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
12. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).

13. Zhao, Y. & Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06 functionals and 12 other functionals. *Theor. Chem. Acc.* **119**, 525–525 (2008).

14. Kendall, R. A., Dunning, T. H. & Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **96**, 6796–6806 (1992).

15. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **113**, 6378–6396 (2009).

16. 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).

17. 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).

18. Chen, S., An, S., Babazade, R. & Jung, Y. Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. *Nat. Commun.* **15**, 2250 (2024).

19. Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, eabe4166 (2021).

20. Chen, S., Babazade, R., Kim, T., Han, S. & Jung, Y. A large-scale reaction dataset of mechanistic pathways of organic reactions. *Sci. Data* **11**, 863 (2024).

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

22. 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).

23. Heid, E. *et al.* Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **64**, 9–17 (2024).

24. 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).

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

26. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).