GraphRXN: A Novel Representation for Reaction Prediction

Baiqing Li
Guangzhou Laboratory
Shimin Su
Guangzhou Laboratory
Chan Zhu
Guangzhou Laboratory
Jie Lin
Guangzhou Laboratory
Xinyue Hu
Guangzhou Laboratory
Lebin Su
Guangzhou Laboratory
Zhunzhun Yu
Guangzhou Laboratory
Kuangbiao Liao
Guangzhou Laboratory https://orcid.org/0000-0001-9089-0569
Hongming Chen (Chen_hongming@gzlab.ac.cn)
Guangzhou Laboratory

Article

Keywords:

Posted Date: May 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1665893/v1

Abstract

In recent years, it has been seen that artificial intelligence (AI) starts to bring revolutionary changes to chemical synthesis. However, the lack of suitable ways of representing chemical reactions and the scarceness of reaction data has limited the wider application of AI to reaction prediction. Here, we introduce a novel reaction representation, GraphRXN, for reaction prediction. It utilizes a universal graph-based neural network framework to encode chemical reactions by directly taking two-dimension reaction structures as inputs. The GraphRXN model was evaluated by three publically available chemical reaction datasets and gave on-par or superior results compared with other baseline models. To further evaluate the effectiveness of GraphRXN, wet-lab experiments were carried out for the purpose of generating reaction data. GraphRXN model was then built on high-throughput experimentation data and a decent accuracy (R^2 of 0.713) was obtained on our in-house data validation. This highlights that the GraphRXN model can be deployed in an integrated workflow which combines robotics and AI technologies for forward reaction prediction.

Introduction

Organic synthesis is the foundation for the development of life science, such as pharmaceutics and chemical biology[1,2]. For decades, the discovery of chemical reaction was driven by serendipitous intuition stemming from expertise, experience and mechanism exploration[3]. However, professional chemists sometimes have hard time to predict whether a specific substrate can indeed go through a desired reaction transformation, even for some well-established reactions[4,5]. When optimizing reaction yield or selectivity, small changes in reaction factors, including catalysts, temperature, ligands, solvents and additives, may result in outcomes that deviate from the intended target. Thus, scientists intend to develop computational model[6–8] to explore the relationships between reaction factors and outcomes.

One of the crucial parts in modelling is finding an appropriate representation of chemical reaction. Quantum-mechanics (QM) based descriptors, representing electrostatic or steric characterizations, calculated by density functional theory (DFT) or other semi-empirical methods [9–12] are frequently used for modeling. Doyle *et al.*[13] utilized QM derived descriptors to build a random forest model, which achieved good prediction performance of the Buchwald-Hartwig cross-coupling of aryl halides with 4-methylaniline. Sigman *et al.* [14] defined four important DFT parameters to capture the conformational dynamics of the ligands, which were fed into multivariate regression modelling for the correlation of ligand properties and relative free energy. Denmark *et al.* [15] generated a set of three-dimension QM descriptors to develop an AI-based model for enantioselectivity prediction. Applying QM descriptors to modeling offers the advantage of model interpretability, but it usually requires a deep understanding of reaction mechanisms, which may be difficult to transfer to other reaction prediction tasks. Another kind of popular descriptors is the so-called reaction fingerprints. Glorius and co-workers[16] developed a multiple fingerprint features (MFFs) as molecular descriptors, by concatenating 24 different fingerprints, to predict the enantioselectivities and yields for different experimental datasets. Although good results were

observed, this method can be a time and resource intensive process, as a single molecule was represented in a 71,374-bit array. Reymond *et al.* [17] reported a molecular fingerprint called differential reaction fingerprint (DRFP), by taking reaction SMILES as input which were embedded into an arbitrary binary space via set operations for subsequent hashing and folding, to perform reaction classification and yield prediction. Though the reaction fingerprints are easily built, the reaction fingerprint may lose certain chemical information due to the limited predefined substructures, and thus a task-specific representation which could learn from dataset is needed.

One possible solution to the issue of universal reaction descriptors is to apply graph neural networks (GNNs) on reaction prediction tasks[18,19]. Owing to the powerful capacity for modelling graph data, GNNs have recently become one of the most popular AI methods and have achieved remarkable prediction performance on several tasks[20–23]. Various graph-based models, such as graph conventional network(GCN)[21,24], GraphSAGE[25], graph attention network(GAT)[26] and message passing neural network(MPNN)[27], have been proposed to learn a function of the entire input graph over molecular properties, by either directly applying a weight matrix on the graph structure or using a message passing and aggregation procedure to update node features iteratively. A molecule is regarded as a graph, where atoms are treated as nodes and bonds are treated as edges. Node and edge features are influenced by proximal ones, and these features are learned and aggregated to form the embedding of entire molecule graph[28,29]. In this work, we proposed a modified communicative message passing neural network (GraphRXN), which was used to generate reaction embeddings for reaction modelling without using predefined fingerprints. For chemical reactions comprised of multiple components, reaction features can be built up by aggregating embeddings of these components together and correlated to the reaction output via a dense layer neural network.

Another major challenge for reaction prediction is the access of high-quality data[30,31]. Though numerous data were accumulated, bias toward positive results in the literatures led to unbalanced datasets. What's more, extracting valid large-scale data from literature requires substantial human intervention[32,33]. High-throughput experimentation (HTE) is a technique that can perform a large number of experiments in parallel[34,35]. HTE could serve as a powerful tool for advancing AI chemistry as it has the capability to significantly increase experiment throughput, and ensure data integrity and consistency. With this technology, several high-quality reaction datasets were reported[30], including Buchwald-Hartwig amination[13,36,37], Suzuki coupling[38–40], photoredox-catalyzed cross coupling[41]. These datasets contain both successful and failed reactions, which is critical for building forward reaction prediction models. Three public HTE datasets were used as proof of concept studies for our method and encourage results were demonstrated. As further verification, we used our in-house HTE platform to generate data of Buchwald-Hartwig cross-coupling reaction. The GraphRXN methodology was then applied on the in-house dataset and a decent prediction model was obtained (R^2 of 0.713), which highlights that our method can be integrated with reaction robotics system for reaction prediction We expect that deep

learning based methods like GraphRXN, combined with the data-on-demand reaction machine, could potentially push the boundary of reaction methodology development[42,43] .

Graphrxn Development

GraphRXN for representing chemical reaction

A deep-learning graph framework, GraphRXN, was proposed to be capable of learning reaction features and predicting reactivity. The architecture of GraphRXN is shown in Fig. 1 and its detailed implementation was described in Algorithm 1.

The input of GraphRXN is the reaction SMILES where each reaction component (either reactants or products) is represented by the directed molecular graph $G(V, E)$. Firstly, all graph components of the reaction were encoded into graph embedding individually. The graph encoding part of GraphRXN's is a modified MPNN (message passing neural network) algorithm, which was previously reported communicative Message Passing Neural Network (*CMPNN*[44]).

For each individual graph of the reaction, it learns through three steps: (1) message passing; (2) information updating; and (3) read out. All node features ($X_v, \forall v \in V$) and edge features ($X_{e_{v,w}}, \forall e_{v,w} \in E$) are propagated in the message passing and updating stage as shown in Algorithm 1:

1. for the v node at step *k*, its intermediate message vector $m^k(v)$ is obtained by aggregating the hidden state of its neighboring edges at the previous step $h^{k-1}\left(e_{u,v}\right)$, then the previous hidden state $h^{k-1}(v)$ is concatenated with its current message $m^k(v)$ and fed into a communicative function to obtain current node hidden state $\mathrm{h}^k(v)$;
2. for the edge $e_{v,w}$ at step *k*, its intermediate message vector $m^k\left(e_{v,w}\right)$ is obtained by subtracting the previous edge hidden states $h^{k-1}\left(e_{v,w}\right)$ from hidden state of its starting node $h^k(v)$, then the initial edge state $h^0\left(e_{v,w}\right)$ and weighted vector $\mathrm{W} \bullet m^k\left(e_{v,m}\right)$ are added up and fed into an activation function ($ReLU$) to form current edge state $h^k\left(e_{v,w}\right)$;
3. After K steps iteration, the message vector ($m(v)$) is obtained by aggregating hidden states $h^K\left(e_{u,v}\right)$ of its neighbouring edges. The node message vector $m(v)$, current node hidden state $h^K(v)$ and initial node information $x(v)$ are fed into a communicative function to form the final node embedding $h(v)$.

Gated Recurrent Unit (GRU) is chosen as the readout operator to aggregate the node vectors into a graph vector. Thus, through CMPNN module, 2D molecular structures of the reaction components are encoded

into molecular feature vectors. The length of molecule feature vector is adjustable (here it is set to 300 bit).

These molecular feature vectors are then aggregated into one reaction vector by either summation or concatenation operation (named as GraphRXN-sum and GraphRXN-concat respectively). The length of GraphRXN-sum vector is set to 300 bit and GraphRXN-concat is multiple times of 300 (depending on the maximal reaction components). If we take a two-components reaction (A + B → P) for example, when summation operation is selected to aggregate features of A, B and P, the length of reaction vector is 300 bit; when concatenation operation is selected to aggregate molecular features, the length of reaction vector is 900 bit. In addition, for some reaction components which are inappropriate to be depicted by graph structure, such as inorganic reagents or catalysts, one-hot embedding will be used to characterize them. Finally, a dense layer is used to fit reaction outcomes, including reaction yield and selectivity.

Algorithm 1

for $graph \in reaction\ components\ graphs$ **do**
  $\boldsymbol{h}^0(e_{v,w}) \leftarrow \boldsymbol{x}_{e_{v,m}}, \forall e_{v,m} \in E; \boldsymbol{h}^0(v) \leftarrow \boldsymbol{x}_v, \forall v \in V$
  for $k = 1...K$ **do**
    for $v \in V$ **do**
      $\boldsymbol{m}^k(v) \leftarrow \text{AGGREGATE}\ (\{\boldsymbol{h}^{k-1}(e_{u,v}), \forall u \in N(v)\});$
      $\boldsymbol{h}^k(v) \leftarrow \text{COMMUNICATE}\ (\boldsymbol{m}^k(v), \boldsymbol{h}^{k-1}(v))$
    end for
    for $e \in E$ **do**
      $\boldsymbol{m}^k(e_{v,w}) \leftarrow \boldsymbol{h}^k(v) - \boldsymbol{h}^{k-1}(e_{w,v});$
      $\boldsymbol{h}^k(e_{v,w}) \leftarrow \sigma(\boldsymbol{h}^0(e_{v,w}) + \boldsymbol{W} \cdot \boldsymbol{m}^k(e_{v,w}))$
    end for
  end for
  $\boldsymbol{m}(v) \leftarrow \text{AGGREGATE}\ (\{\boldsymbol{h}^K(e_{u,v}), \forall u \in N(v)\});$
  $\boldsymbol{h}(v) \leftarrow \text{COMMUNICATE}\ (\boldsymbol{m}(v), \boldsymbol{h}^K(v), \boldsymbol{x}(v))$
  $\boldsymbol{z} \leftarrow \text{Readout}\ (\{\boldsymbol{h}(v), \forall v \in V\})$
end for
$\boldsymbol{case1}: GraphRXN\text{-}sum\ vector = SUMMATION(\boldsymbol{z}_{graph})$
$\hat{y} = FC(GraphRXN\text{-}sum\ vector)$
$\boldsymbol{case2}: GraphRXN\text{-}concat\ vector = CONCATENATION(\boldsymbol{z}_{graph})$
$\hat{y} = FC(GraphRXN\text{-}concat\ vector)$

Modelling datasets

As shown in Table 1, in total, four reaction datasets were used to validate the performance of our GraphRXN model. Three of them are open-source HTE datasets and one of them is generated from in-house HTE platform.

Table 1
Description of three reaction datasets used as benchmark.

Entry	Description	Size	Source
Dataset 1	Yield for Buchwald-Hartwig coupling reaction	4,608	Doyle et al.[13]
Dataset 2	Yield for Suzuki-Miyaura coupling reaction	5,760	Perera et al. [38]
Dataset 3	Stereo-selectivity for asymmetric N, S-acetal formation reaction	1,075	Denmark et al.[15]
Dataset 4	Ratio for Buchwald-Hartwig coupling reaction	1,558	In-house HTE dataset

Modelling process

The original outcome value x was treated with z-score normalization, where μ is the mean of all samples, σ is the standard deviation of all samples.

$$\hat{x} = \frac{x-\mu}{\sigma} \quad \text{Eq. (1)}$$

Regarding the performance measures, three evaluation metrics on the test set were used, including correlation coefficient (R^2), mean absolute error (MAE) and root mean squared error (RMSE). For model evaluation, *k* fold cross-validation was done on all datasets. To make a strict comparison, ten folds cross-validation was adopted on dataset 1–2 which was consistent with the reported Yield-BERT study by Reymond *et al.* [45,46], and dataset 3 which was consistent with the reported study by Perera *et al.*[38]. Five folds cross-validation was adopted in the in-house dataset. For GraphRXN and DeepReac + models, 20% of the training data were used to make validation set. To avoid overfitting, an early-stop mechanism was introduced, *i.e.* when the model performance on the validation set became stable, the training process would stop.

GraphRXN method was applied on all four datasets and in addition, two previously published reaction prediction methods Yield-BERT and DeepReac + were also used for comparison. A sequence-based model, Yield-BERT[45,46], developed by Reymond *et al.* achieved excellent prediction performance. Another graph-based neural network, DeepReac + [47], was recently reported by Liu *et al.* to deliver impressive results on three chemical reaction datasets. The source codes were downloaded from corresponding GitHub repositories. Hyper-parameters search and minor modifications were conducted for resolving some incompatibility issues of python environment (see Supplementary Materials Table S10-S12).

Model Performance On Public Datasets

Four models, including Graph-concat, Graph-sum, Yield-BERT and DeepReac+, were built on three public datasets and the results were shown in Table 2. Dataset 1 and 2 are collections of reaction yield from

coupling reactions, while Dataset 3 is a collection of stereo-selectivity from asymmetric reactions. The average R^2, MAE and RMSE values for respective test set throughout the 10-fold cross-validation procedure were listed in Table 2.

Table 2
Model Performance over ten-fold CV on test set

Dataset	Methods	R^2	MAE	RMSE
Dataset 1	GraphRXN-concat	**0.951**	4.30	5.98
	GraphRXN-sum	0.937	4.85	6.80
	Yield-BERT	0.951	4.00	6.03
	DeepReac+	0.922	5.25	7.54
Dataset 2	GraphRXN-concat	**0.844**	7.94	11.08
	GraphRXN-sum	0.838	8.09	11.29
	Yield-BERT	0.815	8.13	12.08
	DeepReac+	0.827	8.06	11.65
Dataset 3	GraphRXN-concat	**0.892**	0.16	0.23
	GraphRXN-sum	0.881	0.18	0.24
	Yield-BERT	0.886	0.16	0.24
	DeepReac+	0.853	0.18	0.25

For Dataset 1, the performance of GraphRXN-concat model was similar to the baseline method Yield-BERT but better than the GraphRXN-sum and DeepReact + model. For Dataset 2, both GraphRXN-concat and GraphRXN-sum outperformed the Yield-BERT and DeepReact + method. For Dataset 3, the R^2 of GraphRXN-concat was 0.892, which was better than GraphRXN-sum (0.881), Yield-BERT (0.886) and DeepReac+ (0.853). Among these three metrics, we believe that MAE is more meaningful for chemists, as it gives a possible error between the observed and predicted values. MAE/RMSE may better serve as reference value for chemists to decide whether to conduct the experiment or not. Our GraphRXN-concat model gave better MAE and RMSE values than Yield-BERT and DeepReac+, which demonstrated that our GraphRXN model can provide on-par or slightly better performance over the baseline models. Details of model prediction on each fold were included in Supplementary Materials Table S13-S15.

Model Performance On In-house Datasets

HTE platform

In recent years, HTE, operated under standard codes, has been used to perform parallel experiments for rapid screening arrays of reactants or conditions, which generated large amounts of high-quality reaction data[48,49]. We have developed an in-house HTE platform by assembling various state-of-the-art automated workstations/modules.

All experiments in this study were carried out using HTE, including solid dispensing, liquid dispensing, heating and agitation, reaction workup, sample analysis and data analysis (Fig. 2). Exquisite design of experiment was required before HTE [50] (see supplementary information).

- Solid dispensing: Solid samples were stored in the dispensing containers. Then an overhead gravimetric dispensing unit delivered target amounts of samples from dispensing containers to the designated 4 mL vials.
- Liquid dispensing: Liquid samples were stored in uniform bottles. Then the liquid-handling robot transferred target volume of samples to the designated 4 mL vials in a programmed manner. With the amounts of solid and liquid samples dispensed in 4mL vials, the liquid-handling robot was used again to make stock solution accordingly. All stock solutions were mixed thoroughly using vortex mixer. Stock solutions were transferred into the designated glass tubes of 96-well aluminium blocks for reaction setup using the liquid-handling robot.
- Heating and agitation: The 96-well aluminium blocks were placed on orbital agitators under pre-set temperature and time.
- Reaction workup: After the reactions were stopped and cooled down, pipetting workstation was used to process the reaction mixtures in batches, including quenching, dilution and filtration. Then samples were prepared in 96-well plates for UPLC-MS analysis.
- Sample analysis: Samples were sequentially injected into UPLC-MS for expected substance determination and quantification.
- Data analysis: Raw data generated by UPLC-MS were fed into Peaksel[51], an analytical software developed by Elsci, which was capable of executing batch-level integration rendering us the UV response area of target substance.

Experimental workflow

Though good results were obtained on three public datasets, we sought to further evaluate GraphRXN on the in-house dataset as further verification (Fig. 3).

Buchwald-Hartwig coupling reaction was used as examined reaction in this study. For the standard condition, we used t-BuXPhos-Pd-G3 as catalyst, 7-Methyl-1,5,7-triazabicyclo[4.4.0]dec-5-ene (MTBD) as base, and DMSO as solvent(Fig. 3a). Firstly, the palladium precatalyst t-BuXPhos-Pd-G3 collocate with MTBD performed well with primary amines [52–54]. Secondly, the catalyst and base are DMSO soluble which would facilitate the HTE process [55,56,]. As of substrates, a series of *ortho*-, *meta*-, and *para*-substituted, including electron-donating and electron-withdrawing groups, aryl-Br and aryl-NH_2 were

selected (Fig. 3b). In total, 50 primary amines (26 Ph-NH_2, 24 Py-NH_2) and 48 bromides (24 Ph-Br, 24 Py-Br) were used in our dataset generation (see Supplementary Materials Table S1-S8).

Reaction setup. In this study, all reactions were carried out at 0.016 mmol scale in 96-well aluminum blocks using HTE platform. For reaction setup, all robots were embedded in a glovebox filled with N_2. The 96-well aluminum blocks were sealed under N_2 and then subjected to orbital agitators with the pre-set parameter of 850 rpm and 65°C. After 16 hours, the 96-well aluminum blocks were cooled down to room temperature. In total, 2,127 reactions were successfully conducted on HTE platform (detailed HTE layout Supplementary Materials Table S9).

Analysis. For each glass tube, 0.0625 equivalence of 4,4'-Di-tert-butyl-1,1'-biphenyl was added as internal standard (IS). Reaction solutions were then transferred to filter plates and the filtrates were collected by 96-well plates. The sample plates were then analyzed by UPLC-MS. The UV responses of product and IS were obtained using Peaksel. The ratios of UV response of product over IS ($ratio_{UV}$) were calculated using the following equation.

$$ratio_{UV} = \frac{A_{product} \times c}{A_{IS}} \times 100\% \quad \text{Eq. (2)}$$

where $A_{product}$ is the response area of the target product at the wave length of 254 nm,A_{IS} is the response area of the IS at the wave length of 254 nm, c is a constant (0.0625 eq.), which represents the mole ratio of IS and product at 100% theoretical yield.

During the course of data analysis, 569 reaction data derived from abnormal spectra were discarded. Eventually, 1,558 reaction data were obtained. For more details about the experiments, please see supplementary materials.

HTE results

According to the substituted aromatic amines/bromides, reactions can be grouped into four groups (G1-G4), *i.e.* diphenylamines derivatives (reactions between Ph-NH_2 and Ph-Br, G1), phenylpyridine amine derivatives (reactions between Ph-NH_2 and Py-Br, G2), phenylpyridine amine derivatives (reactions between Py-NH_2 and Ph-Br, G3) and 2,2'-dipyridylamide derivatives (reactions between Py-NH_2 and Py-Br, G4). G1 contains 317 reaction points, while G2, G3 and G4 group have 419, 401 and 421 reactions respectively. Hereby shows the $ratio_{UV}$distribution for all four groups, where the light color represents low value, and the dark color corresponds to high value, ranging from 0 to 1 (Fig. 5). The grey grids represent failed reactions or discarded data. For the entire dataset, half of the reaction ratio lies in the range from 0 to 0.2. The $ratio_{UV}$distribution was not balanced with heavy condense on low value which would be a challenging task for modeling. Among these, 13% of reactions in G1 gave ratio ≥ 0.5, while only 0.7%, 8% and 5% for G2, G3 and G4 respectively, which indicates the chosen reaction condition in HTE may be more suitable for reactions between Ph-NH_2 and Ph-Br.

Performance of GraphRXN on in-house HTE dataset

The performances of models built on separate reaction groups as well as the entire dataset are listed in Table 3. A five-fold cross validation without replacement was done for train-test split (the results of each CV fold on test set see Supplementary Materials S14-S18).

Table 3
The performance of GraphRXN on in-house dataset.

Group	Size	Methods	R^2	MAE	RMSE
Entire	1,558	GraphRXN-concat	**0.713**	**0.06**	**0.09**
		GraphRXN-sum	0.704	0.06	0.09
		Yield-BERT	0.645	0.10	0.07
		DeepReac+	0.610	0.07	0.10
G1	317	GraphRXN-concat	0.661	0.08	0.11
		GraphRXN-sum	0.462	0.11	0.14
		Yield-BERT	**0.718**	**0.07**	**0.10**
		DeepReac+	0.551	0.09	0.13
G2	419	GraphRXN-concat	**0.629**	**0.05**	**0.07**
		GraphRXN-sum	0.592	0.06	0.07
		Yield-BERT	0.512	0.06	0.08
		DeepReac+	0.528	0.06	0.08
G3	401	GraphRXN-concat	**0.802**	**0.06**	**0.08**
		GraphRXN-sum	0.775	0.06	0.08
		Yield-BERT	0.785	0.06	0.08
		DeepReac+	0.745	0.07	0.09
G4	421	GraphRXN-concat	0.459	0.08	0.12
		GraphRXN-sum	0.419	0.09	0.12
		Yield-BERT	**0.503**	**0.08**	**0.11**
		DeepReac+	0.23	0.10	0.14

Our GraphRXN-concat model obtained better performance on the entire dataset comparing with other baseline models. The test set plots over five-folds cross validation of GraphRXN-concat and GraphRXN-sum on the entire dataset are shown in Fig. 4.

For comparison, we also built models on individual groups of dataset. The performances of GraphRXN-concat were superior than other models on G2, G3 and the entire dataset but slightly worse on G1 and G4. It seems that R^2 on small datasets can fluctuate considerably, *e.g.* R^2 of four groups are rather different from each other, while values of MAE and RMSE are similar across all four groups. The results indicate that the smaller dataset with limited structural diversity that might deteriorate the prediction accuracy, while larger dataset with diverse structures can allow to learn a better model from a larger reaction space. Although the difference is minor, GraphRXN-concat performed consistently better than GraphRXN-sum model among all four groups and the entire dataset.

Conclusion

In this work, GraphRXN was proposed and proved to be an effective method for chemical reaction encoding, which gave good prediction performance over four reaction datasets. For the three public dataset, GraphRXN provided on-par or slightly better performance over the baseline models. In addition, we used HTE platform to build standardized dataset, and GraphRXN also delivered good correlations. It has demonstrated that deep learning model could yield moderate to good accuracy in reaction prediction regardless of limited size of the datasets and many complex influencing variables. Although a chemical reaction goes through certain transitional states, it seems that the model can directly predict reaction outcome using structural features of reaction components without the guidance of mechanism. These results have motivated us to apply this HTE + AI strategy on more reaction types in the future. The source code of GraphRXN and our in-house reaction dataset are available at https://github.com/jidushanbojue/GraphRXN.

Declarations

Acknowledgment

The authors thank Nanshan Zhong, Tao Xu, Duanqing Pei and Hongming Hou for their support in building HTE platform. Additionally, we thank Waters, Tecan, Chemspeed and Peaksel for technical support.

Author Contributions

B.L., S.S., and J.L. conducted the computational study. S.S. and C.Z. conducted the high-throughput experiments. X.H, L.S. and Z.Y. conducted the data analysis. B.L. and S.S. contributed in the manuscript writing. K.L. and H.C. conceived the idea, and directed the research.

Funding

Funding was provided by the National Natural Science Foundation of China (22071249 and 22002169).

Code availability

The source code of GraphRXN and our in-house reaction dataset are available at https://github.com/jidushanbojue/GraphRXN.

References

1. Campos, K. R. *et al.* The importance of synthetic chemistry in the pharmaceutical industry. Science **363** (2019).
2. Whitesides, G. M. Reinventing chemistry. Angew Chem Int Ed Engl **54**, 3196–3209 (2015).
3. Davies, I. W. The digitization of organic synthesis. Nature **570**, 175–181 (2019).
4. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. Nature **533**, 73–76 (2016).
5. Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. Science **361** (2018).
6. Gromski, P. S., Henson, A. B., Granda, J. M. & Cronin, L. How to explore chemical space using algorithms and automation. Nature Reviews Chemistry **3**, 119–128 (2019).
7. Szymkuć, S. *et al.* Computer-Assisted Synthetic Planning: The End of the Beginning. Angewandte Chemie International Edition **55**, 5904–5937 (2016).
8. Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. Accounts of Chemical Research **51**, 1281–1289 (2018).
9. Werth, J. & Sigman, M. S. Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. Journal of the American Chemical Society **142**, 16382–16391 (2020).
10. Werth, J. & Sigman, M. S. Linear Regression Model Development for Analysis of Asymmetric Copper-Bisoxazoline Catalysis. ACS Catalysis **11**, 3916–3922 (2021).
11. Zahrt, A. F., Rose, B. T., Darrow, W. T., Henle, J. J. & Denmark, S. E. Computational methods for training set selection and error assessment applied to catalyst design: guidelines for deciding which reactions to run first and which to run next. Reaction Chemistry & Engineering **6**, 694–708 (2021).
12. Henle, J. J. *et al.* Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. Journal of the American Chemical Society **142**, 11578–11592 (2020).
13. Ahneman Derek, T., Estrada Jesús, G., Lin, S., Dreher Spencer, D. & Doyle Abigail, G. Predicting reaction performance in C–N cross-coupling using machine learning. Science **360**, 186–190 (2018).
14. Zhao, S. *et al.* Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. Science **362**, 670–674 (2018).
15. Zahrt Andrew, F. *et al.* Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science **363**, eaau5631 (2019).
16. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. Chem **6**, 1379–1390 (2020).

17. Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. Digital Discovery **1**, 91–97 (2022).
18. Zhou, J. *et al.* Graph neural networks: A review of methods and applications. AI Open **1**, 57–81 (2020).
19. Wu, Z. *et al.* A Comprehensive Survey on Graph Neural Networks. IEEE Trans Neural Netw Learn Syst **32**, 4–24 (2021).
20. Schwaller, P. *et al.* Machine intelligence for chemical reaction space. *WIREs Computational Molecular Science* **n/a**, e1604 (2022).
21. Coley, Connor W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. Chemical Science **10**, 370–377 (2019).
22. Louis, S.-Y. *et al.* Graph convolutional neural networks with global attention for improved materials property prediction. Physical Chemistry Chemical Physics **22**, 18141–18148 (2020).
23. Feinberg, E. N. *et al.* PotentialNet for Molecular Property Prediction. ACS Central Science **4**, 1520–1530 (2018).
24. Torng, W. & Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. Journal of Chemical Information and Modeling **59**, 4131–4149 (2019).
25. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. Advances in neural information processing systems **30** (2017).
26. Sacha, M. *et al.* Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. Journal of Chemical Information and Modeling **61**, 3273–3284 (2021).
27. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. in *International conference on machine learning.* 1263–1272 (PMLR).
28. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. Nature Communications **8**, 13890 (2017).
29. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70 (eds Precup Doina & Teh Yee Whye) 1263–1272 (PMLR, Proceedings of Machine Learning Research, 2017).
30. Kearnes, S. M. *et al.* The Open Reaction Database. Journal of the American Chemical Society **143**, 18820–18826 (2021).
31. Baldi, P. Call for a Public Open Database of All Chemical Reactions. Journal of Chemical Information and Modeling (2021).
32. Swain, M. C. & Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling **56**, 1894–1904 (2016).
33. Vaucher, A. C. *et al.* Automated extraction of chemical synthesis actions from experimental procedures. Nature Communications **11**, 3601 (2020).

34. Shevlin, M. Practical High-Throughput Experimentation for Chemists. ACS Med Chem Lett **8**, 601–607 (2017).
35. Krska, S. W., DiRocco, D. A., Dreher, S. D. & Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. Acc Chem Res **50**, 2976–2985 (2017).
36. Kashani, S. K., Jessiman, J. E. & Newman, S. G. Exploring Homogeneous Conditions for Mild Buchwald–Hartwig Amination in Batch and Flow. Organic Process Research & Development **24**, 1948–1954 (2020).
37. Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. Nature Reviews Drug Discovery **17**, 709–727 (2018).
38. Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science **359**, 429–434 (2018).
39. Reizman, B. J., Wang, Y.-M., Buchwald, S. L. & Jensen, K. F. Suzuki–Miyaura cross-coupling optimization enabled by automated feedback. Reaction Chemistry & Engineering **1**, 658–666 (2016).
40. Beker, W. *et al.* Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. Journal of the American Chemical Society **144**, 4819–4827 (2022).
41. Kariofillis, S. K. *et al.* Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. Journal of the American Chemical Society **144**, 1045–1055 (2022).
42. Keith, J. A. *et al.* Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. Chemical Reviews **121**, 9816–9872 (2021).
43. Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. Nature Reviews Methods Primers **1**, 23 (2021).
44. Song, Y. *et al.* in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* Article 392 (Yokohama, Yokohama, Japan, 2021).
45. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. Machine Learning: Science and Technology **2**, 015016 (2021).
46. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. Nature Machine Intelligence **3**, 144–152 (2021).
47. Gong, Y., Xue, D., Chuai, G., Yu, J. & Liu, Q. DeepReac+: deep active learning for quantitative modeling of organic chemical reactions. Chemical Science **12**, 14459–14472 (2021).
48. Isbrandt, E. S., Sullivan, R. J. & Newman, S. G. High Throughput Strategies for the Discovery and Optimization of Catalytic Reactions. Angewandte Chemie International Edition **58**, 7180–7191 (2019).
49. Tu, N. P. *et al.* High-Throughput Reaction Screening with Nanomoles of Solid Reagents Coated on Glass Beads. Angewandte Chemie International Edition **58**, 7987–7991 (2019).

50. Cook, A., Clément, R. & Newman, S. G. Reaction screening in multiwell plates: high-throughput optimization of a Buchwald–Hartwig amination. Nature Protocols **16**, 1152–1169 (2021).
51. *PEAKSEL*, <https://elsci.io/peaksel/index.html> (
52. Surry, D. S. & Buchwald, S. L. Dialkylbiaryl phosphines in Pd-catalyzed amination: a user's guide. Chemical Science **2**, 27–50 (2011).
53. Baumgartner, L. M., Dennis, J. M., White, N. A., Buchwald, S. L. & Jensen, K. F. Use of a Droplet Platform To Optimize Pd-Catalyzed C–N Coupling Reactions Promoted by Organic Bases. Organic Process Research & Development **23**, 1594–1601 (2019).
54. Bruneau, A., Roche, M., Alami, M. & Messaoudi, S. 2-Aminobiphenyl Palladacycles: The "Most Powerful" Precatalysts in C–C and C–Heteroatom Cross-Couplings. ACS Catalysis **5**, 1386–1396 (2015).
55. Brocklehurst, C. E., Gallou, F., Hartwieg, J. C. D., Palmieri, M. & Rufle, D. Microtiter Plate (MTP) Reaction Screening and Optimization of Surfactant Chemistry: Examples of Suzuki–Miyaura and Buchwald–Hartwig Cross-Couplings in Water. Organic Process Research & Development **22**, 1453–1457 (2018).
56. Gesmundo, N. J. *et al.* Nanoscale synthesis and affinity ranking. Nature **557**, 228–232 (2018).

Figures

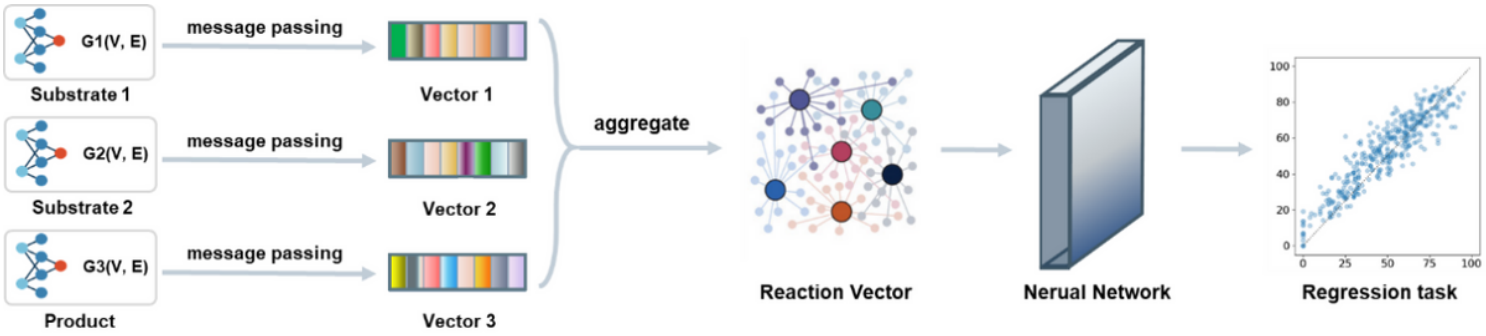

Figure 1

Model architecture of GraphRXN.

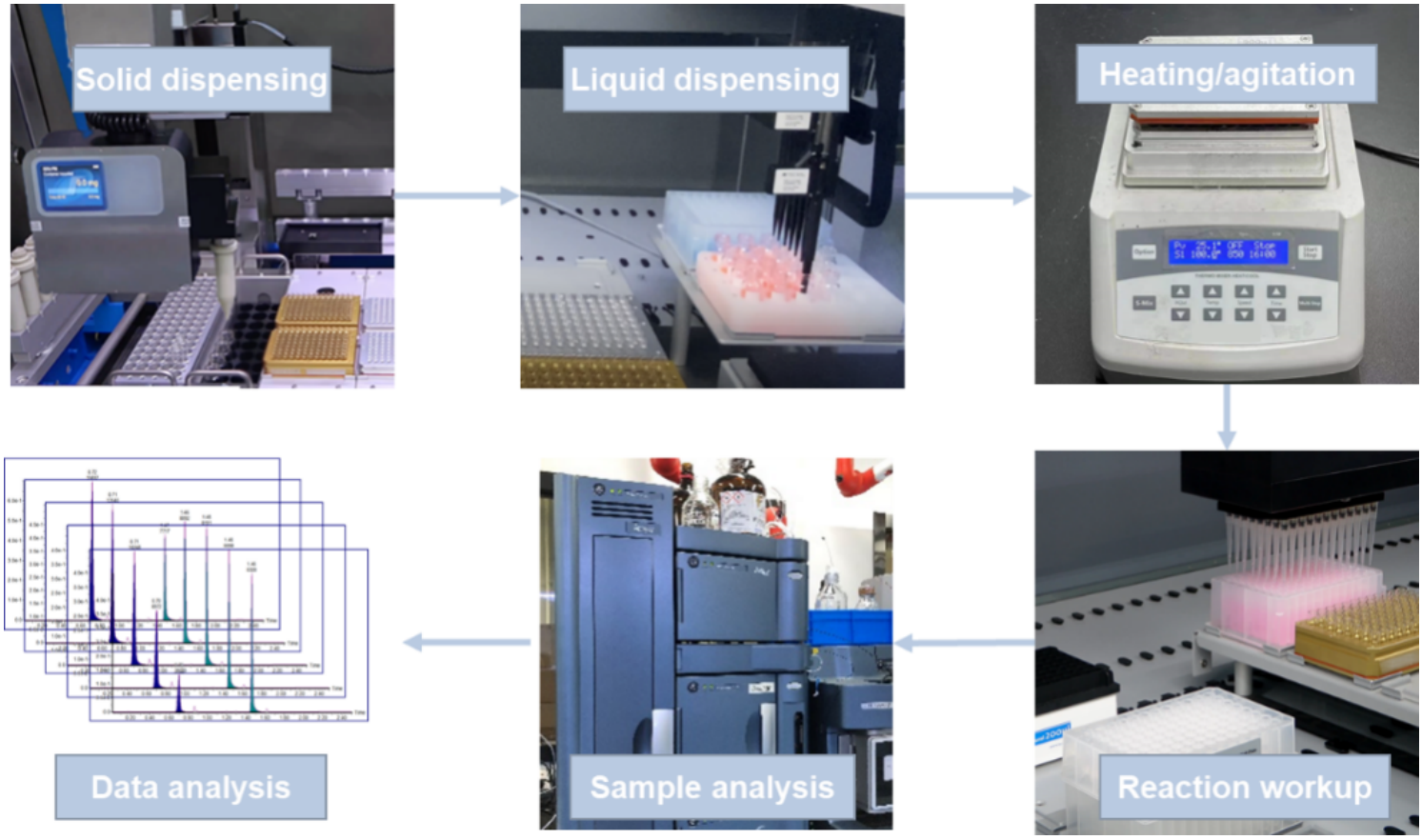

Figure 2

General workflow of HTE process.

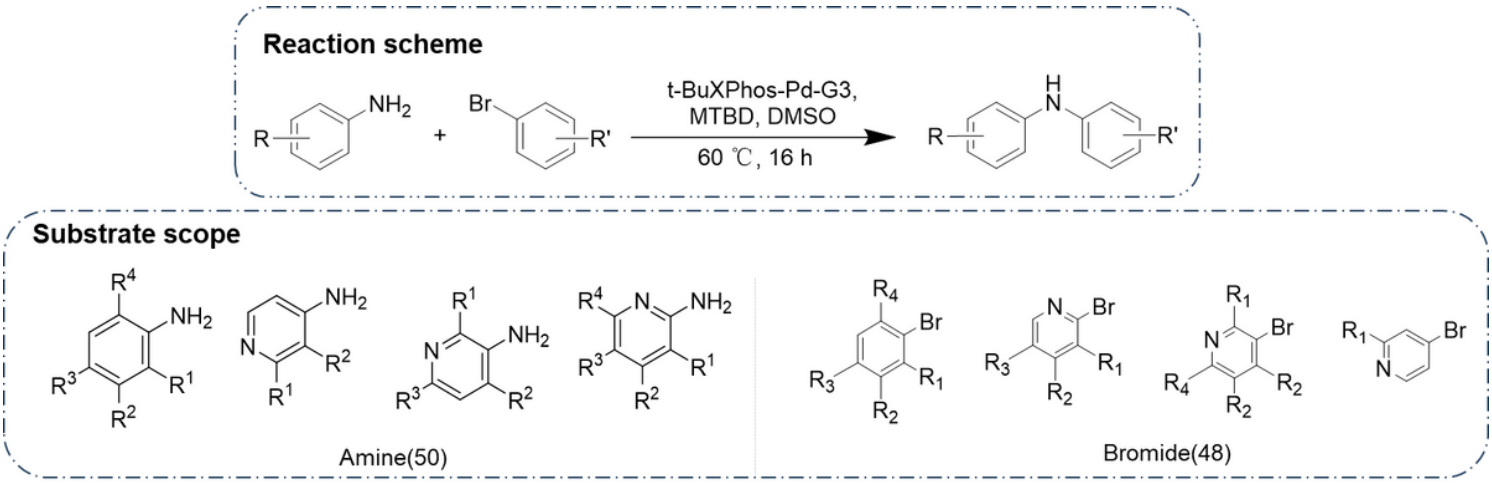

Figure 3

Reaction scheme and substrate scope.

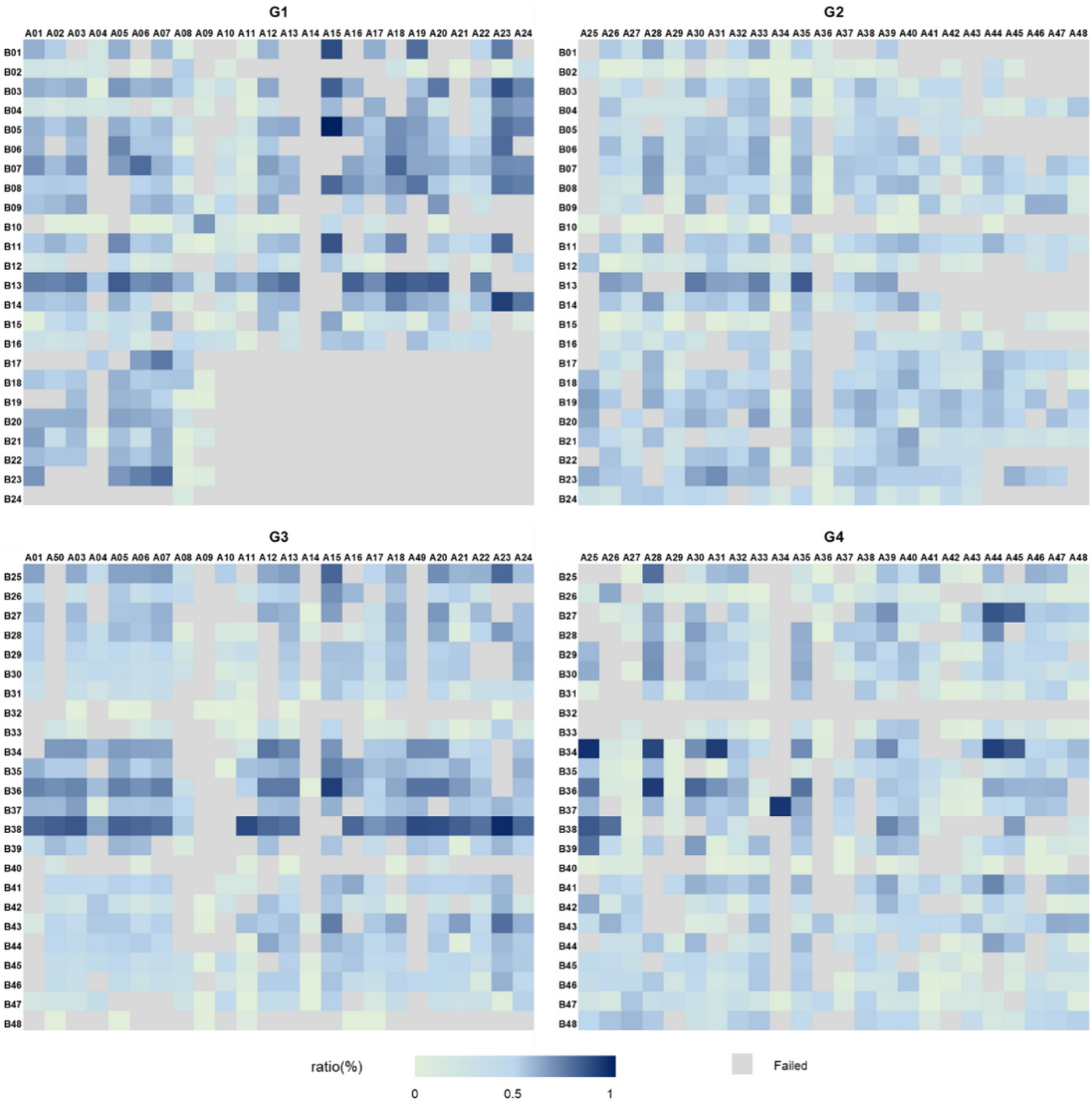

Figure 4

Distribution of $ratio_{UV}$, where A represents amine, and B represents

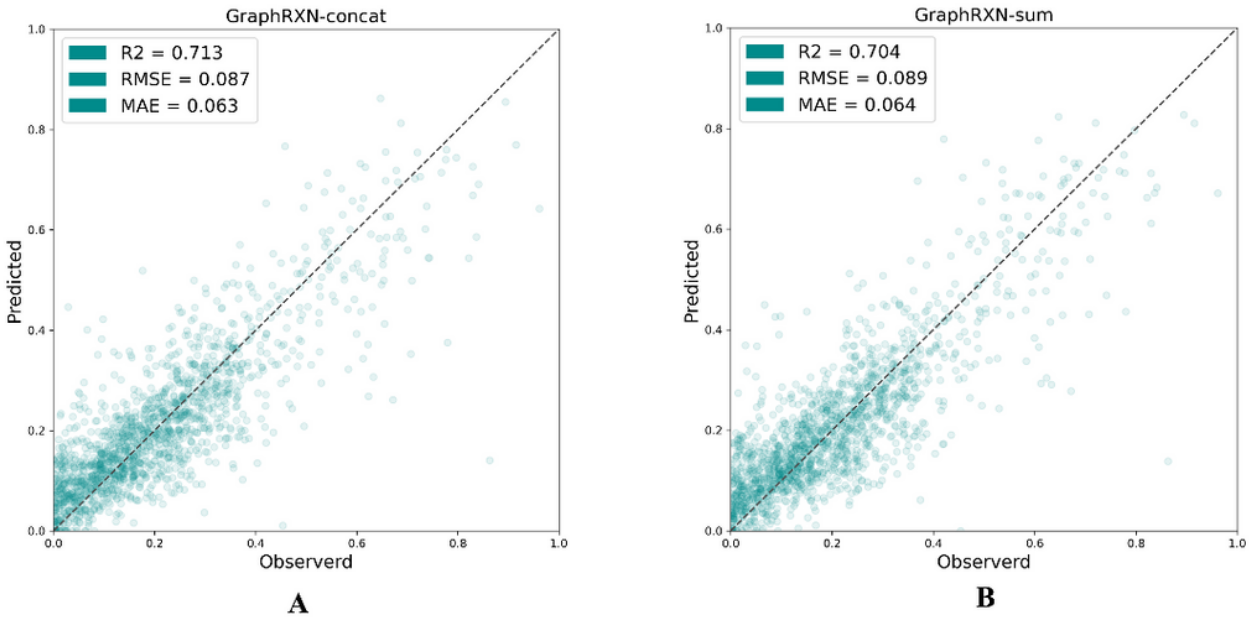

Figure 5

The scatter plots of GraphRXN on the entire dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData1.xlsx
- SupplementaryMaterials.docx