Electronic Supplementary Material (ESI) for *Chem. Commun.*.

Supporting Information

A graph-convolutional neural network for addressing small-scale reaction prediction

Yejian Wu,‡[a] Chengyun Zhang,‡[a] Ling Wang[a] and Hongliang Duan*[a]

[a]Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, China.
*E-mail: hduan@zjut.edu.cn

Table of content

Section S1 Baeyer-Villiger oxidation

Baeyer-Villiger oxidation is a classic small-scale reaction where ketones or aldehydes can be transformed into esters by peroxyacid or peroxide.[1] The detailed description of the Baeyer-Villiger reaction is shown schematically in Fig. S1. As a typical 1,2-hydrocarbyl migration reaction, the mechanism is similar to pinacol rearrangement. Firstly, the oxygen of the carbonyl group is protonated by the hydrogen ion of peroxyacid, which increases the electrophilicity of carbonyl group and makes it more vulnerable to be attacked by the peroxyacid. Next, peroxyacid attacks the carbonyl carbon to form the Criegee intermediate. In a concerted process, a substituent of the carbonyl group moves to the electron-deficient oxygen of the peroxide group as the carboxylic acid leaves away from the intermediate. Finally, an ester is formed with the deprotonation of the oxocarbenium ion. What of particular note, it is generally believed that the hydrocarbyl group that can stabilize positive charge the best would be most likely to migrate.[2] The Baeyer-Villiger reaction follows the rule of group migration and the migratory ability is approximately ranked tertiary > secondary > aryl > primary > methyl.[3] Therefore, the Baeyer–Villiger oxidation can be viewed as a regioselective reaction, which remains an additional level of challenge in prediction task. The GCN model needs to not only focus on the stability of bonds but also determine which hydrocarbon group will migrate.

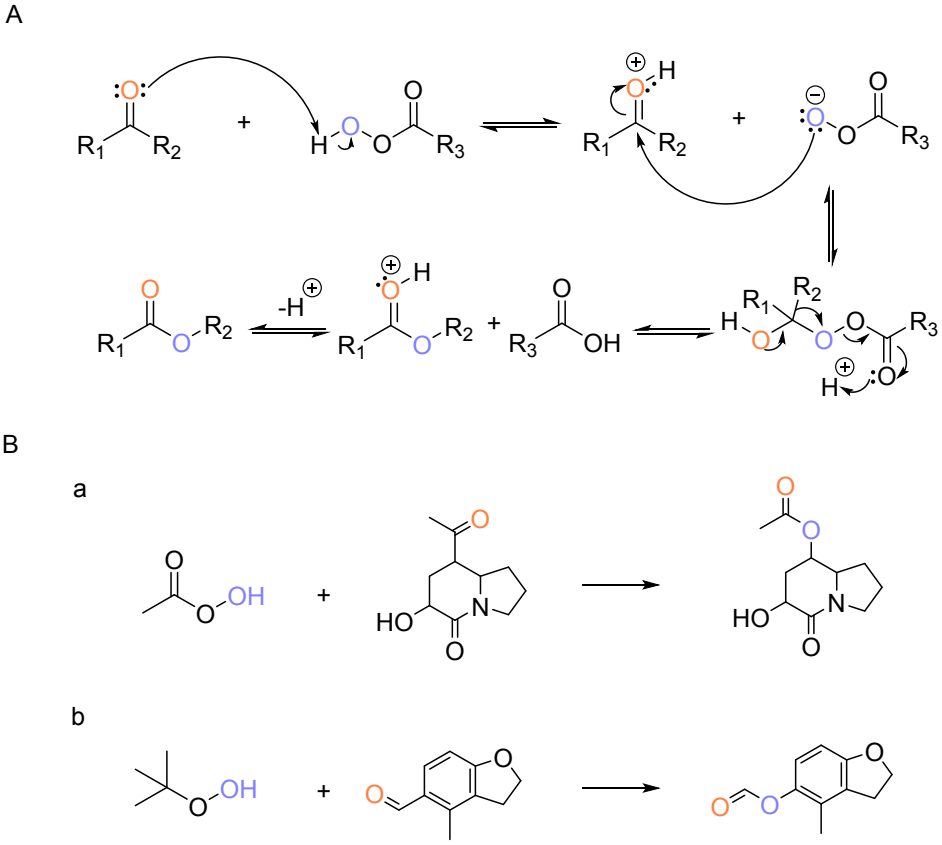


Fig. S1. A detailed description of the Baeyer-Villiger reaction. A. General mechanism of Bae-yer-Villiger reaction. B. Examples of Baeyer-Villiger reaction in which (a) is a ketone reactant and (b) is an aldehyde reactant, both of which are oxidized to esters.

Section S2 Model

S2.1 Weisfeiler-Lehman Network (WLN)

As a category of graph convolutional networks, the WLN[4] is adopted to learn the molecular graph isomorphism by embedding the Weisfeiler-Lehman (WL) algorithm[5]. In the GCN model, a chemical reaction consists of a pair of molecular graphs where Gr and Gp represent the reactant and the product respectively. A molecular graph G = (V, E) is composed of an atom set (V) and a bond set (E). Rely on the WL algorithm that aggregates flowing information between neighboring atoms, the information about state features of the atom and structures of adjacent atoms could be up-dated. In each iteration, the GCN model can obtain information about local atom environment.

S2.2 Attention Mechanism

To further improve the capability of the WLN model, attention mechanism, a powerful algorithm in machine learning field,[6] has been adopted to the GCN model for propagating information among disconnected molecules. Taking into account the fact that the atoms that are not involved in reaction center may have an effect on a chemical reaction, the attention mechanism allows chemical information to flow more widely. By doing the weighted sum of the atom to all other atoms, the global chemical environment is recorded. Finally, the likelihood of atomic bond pair changes is calculated with the weightings of local and global attention strength. An overview of the

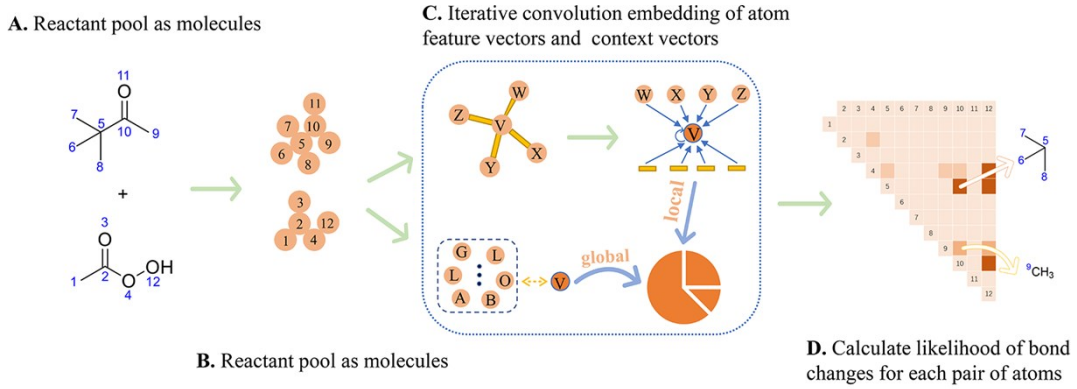


Fig. S2 Wisfeller-Lehman Network (WLN) model embeds the attention mechanism to predict the reaction center. Start to represent the molecule (A) as a graph (B), and iteratively update the feature vector by merging passing information between neighboring atoms and calculate the local feature vector for each atom (C). The global attention mechanism generates a context vector for each atom by extracting the information of all other atoms. Finally, calculate the (D)likelihood of atom bond pair changes based on the weighting of local and global attention strength.

attention mechanism can be found in Fig. S2. As is depicted in Fig. S2, the tert-butyl group has a higher attention weight, which indicates the tert-butyl group is more likely to migrate in the tert-butyl methyl ketone. It's worth remembering that the information about reactions obtains not only the connectivity between the atoms but also the atom properties such as atomic number, formal charge.

S2.3 Reaction Center Prediction

From Gr to Gp, the chemical reaction sites are considered as the change of atoms and bonds in graph connectivity.[7] In the training procedure of reaction center prediction, local and global atomic features from WLN are transferred to another neural network where the cross-entropy loss function is chosen to minimize differences between product labels and predicted scores. Only with this step, can the atom pairs that may change be determined and scored by the GCN model. And the top-k atom pairs are ranked according to their scores.

S2.4 Weisfeiler-Lehman Difference Network (WLDN)

In the former phase, the top-k atom pairs with the highest score are selected and ranked in an atom pair set. Furthermore, the candidate set is gained by enumerating all possible bond pair changes in the atom pair set. To highlight the difference between reactants and products, the WLDN[7] is divided into two components. The first component called Siamese WLN is used to learn the atom feature vectors of reactants and candidate products that combine with local and remote information. And then the vector difference of the homologous atoms is fed into another WLN to calculate the candidate scores. The candidate products are ranked and we compare the top-3 candidates with the true product.

Section S3 Performance of the transformer and GCN models on the USPTO_MIT Data Set

Table S1 Comparison of the top-n accuracies of the transformer and GCN models on the USPTO_MIT Data Set

Model	Top-N accuracy (%)[a]		
	Top-1	Top-2	Top-3
Transformer	90.4	93.7	94.6
GCN	85.6	90.5	92.8

[a]The top-n accuracies of transformer and GCN models on the USPTO_MIT Data Set are originally derived from Philippe *et al.*'s work.[8]

Section S4 Comparison between GCN model and transformer model

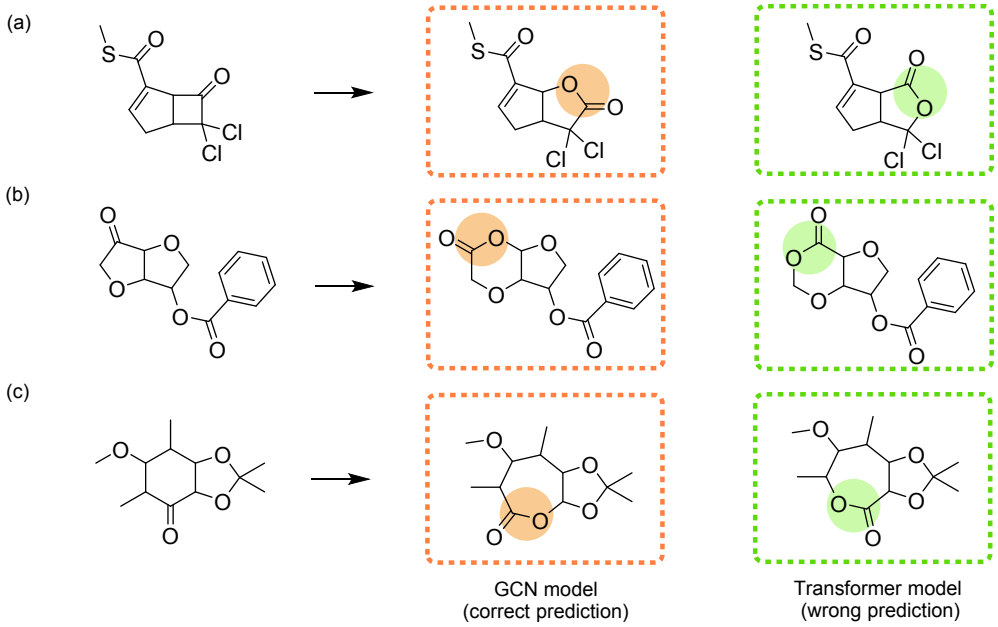


Fig. S3 Comparisons between the GCN model's top-1 correct predictions and the transformer model's top-1 wrong predictions, in which the transformer model makes group migration error.

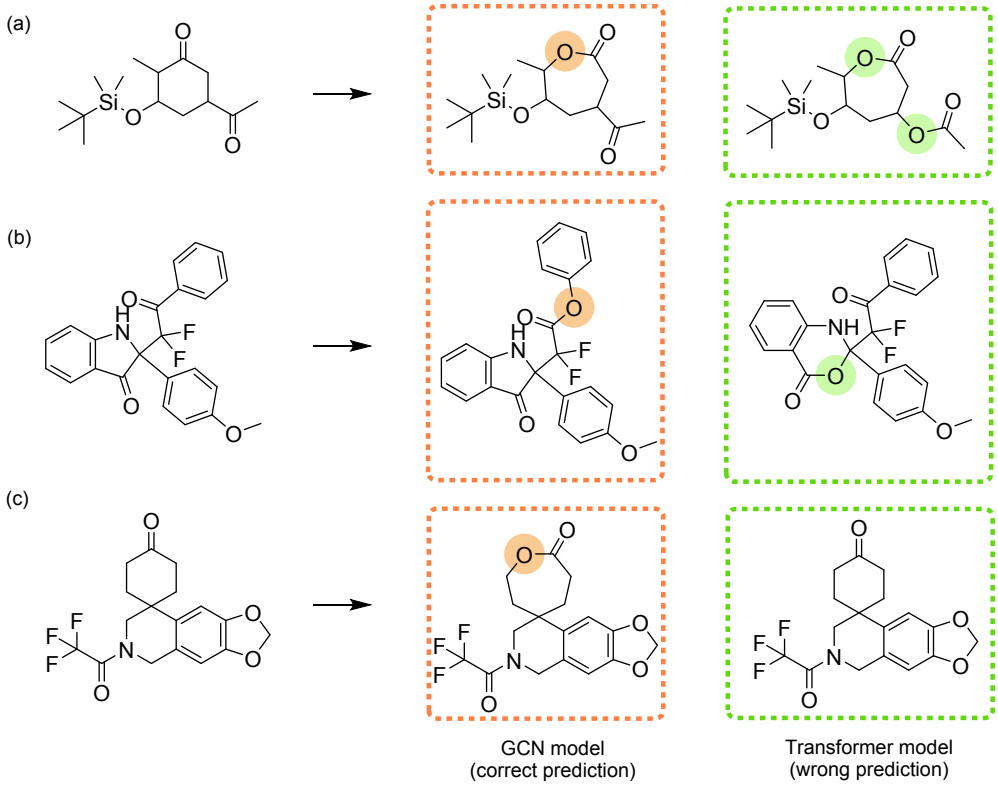


Fig. S4 Comparisons between the GCN model's top-1 correct predictions and the transformer model's top-1 wrong predictions, in which the transformer model makes other errors.

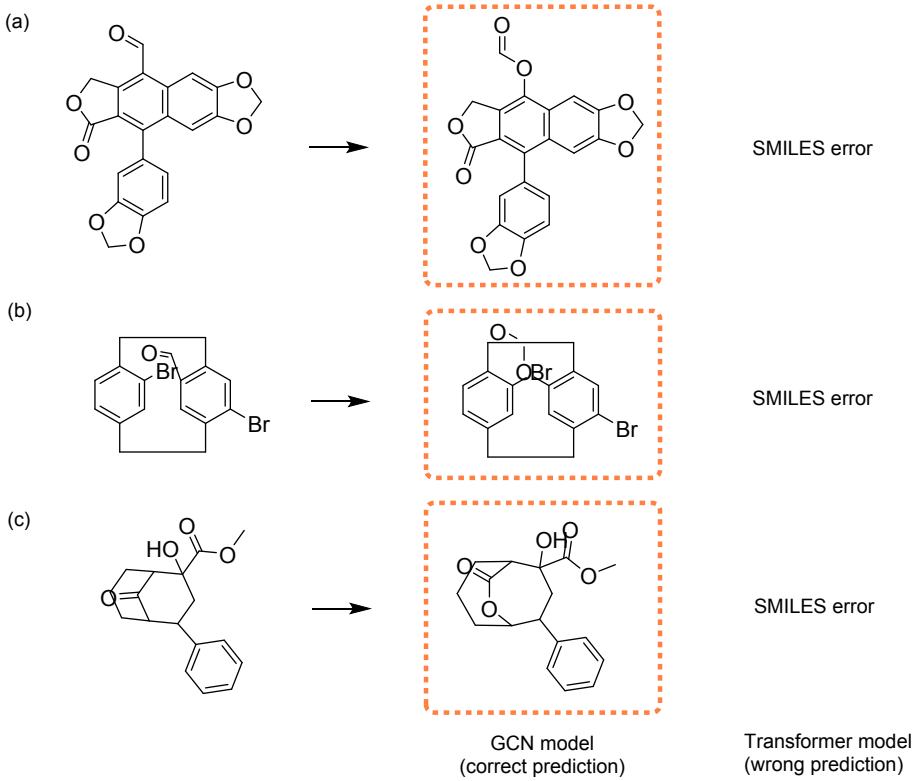


Fig. S5 Comparisons between the GCN model's top-1 correct predictions and the transformer model's top-1 wrong predictions, in which the transformer model makes invalid SMILES code error.

Section S5 Error analysis of the GCN model

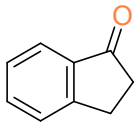
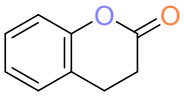
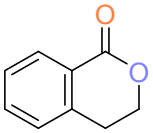
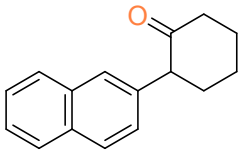
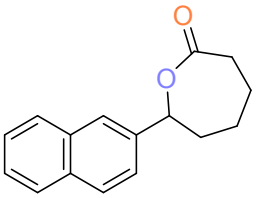
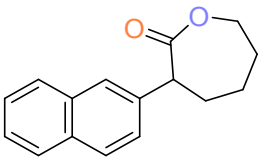
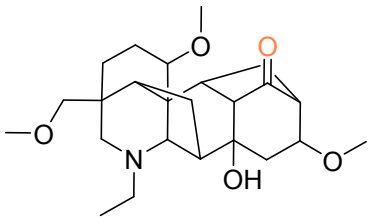
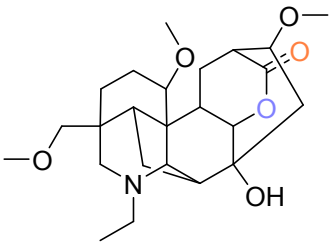
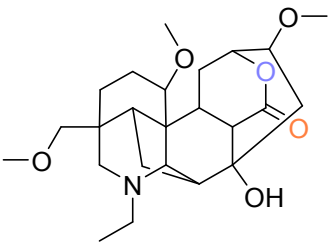
Table S2 Distribution and description of the major predicted errors of the GCN model in the Baeyer-Villiger reaction top-1 predictions.

Types of top-1 predicted errors	Rate (%)	Count
Group migration error	45.5	10
^{2}H recognition error	9.1	2
No chemical reaction	13.6	3
Other error	31.8	7
Total	100.0	22

Examples of group migration error category are displayed in Table S2. Take the Table S2(1) as an example, since the conjugation effect of the aryl group, the positive charge is dispersed and tends to be stable. The aryl group displays an enhanced migration ability compared to primary alkyl group. Moreover, the substitution of the electron-donating group on the benzene ring further strengthens its migratory aptitude. As a result, the ground truth product is chroman-2-one rather than isochroman-1-one. This error type is also common in compounds with other complex structures. Such as the reaction in Table S2(3), we can observe that the reactant

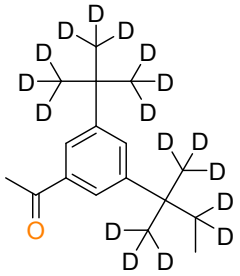
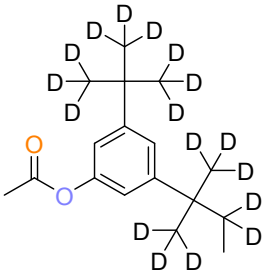
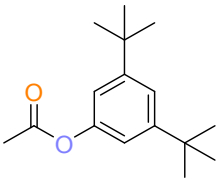
containing complex bridged ring structure makes the task more complicated. Even that the complex compounds may add to complexity of task, the GCN model is equipped with the ability to precisely predict the preferential migration groups in most reactions and gains better results than the transformer model in the face of small shortage of training samples.

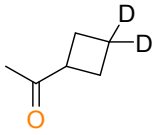
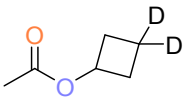
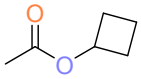
Table S2 Examples of group migration error in GCN model top-1 predictions.

	Reactant	Reported Ground Truth	Top-1 Prediction
1			
2			
3			

Analyzing the predictions made by the GCN model, we noticed that this model ignores the identification of isotope atoms. As listed in Table S3(1), the reported ground truth is 3,5-bis(2-(methyl-*d*3)propan-2-yl-1,1,1,3,3,3-*d*6)phenyl acetate when 1-(3,5-bis(2-(methyl-*d*3)propan-2-yl-1,1,1,3,3,3-*d*6)phenyl)ethan-1-one is given. However, the prediction product provided by the GCN model is 3,5-di-tert-butylphenyl acetate, which shows the error of the identification of isotope atoms. In spite of the fact that the predictions given by the GCN conform to the group migration rule of Baeyer-Villiger oxidation reaction, this model fails to distinguish the difference between ^{1}H and ^{2}H.

Table S3 Examples of ^{2}H recognition error in GCN model top-1 predictions.

	Reactant	Reported Ground Truth	Top-1 Prediction
1			

2			

Several examples of non-reactive error type and other unexplainable mistakes are depicted in Table S4 and Table S5, respectively. The prediction product 2-fluoro-1-phenylethan-1-one (Table S4(3)) is the same as the reactant, which can be considered as a non-reactive error. Compare to 3-((benzyloxy)methyl)-2-methylpentan-3-yl formate (Table S5(2)) that is the reported ground truth, the ((3-((benzyloxy)methyl)-2-methylpentan-3-yl)oxy)methanol is a prediction product that violates the basic chemical rules. We guess that these errors are related to the atom mapping. In our work, the data need to be preprocessed by RXN Mapper for obtaining the atom mapping information. By tracing the atom mapping process, we find that atom mapping of some reactions is wrong. An example is listed in Fig. S6, carbonyl group of the product 2-methyl-1-phenylbutan-2-yl acetate should come from 3-benzyl-3-methylpentan-2-one, but the RXN Mapper attribute it to peracetic acid. The proper one-to-one correspondence of the atom labels plays a critical role in the model's training process. The atom mapping error makes the GCN model misjudge the reaction center as well as affect the predictive power of the model.

Table S4 Examples of non-reactive error in GCN model top-1 predictions.

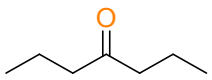
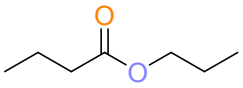
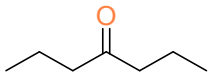
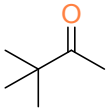
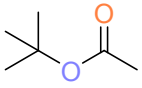
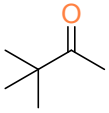
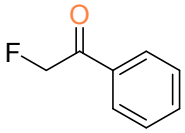
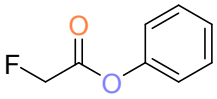
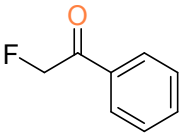
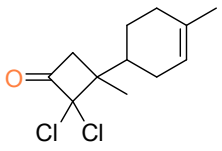
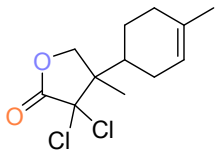
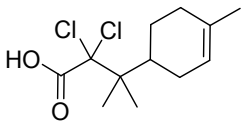
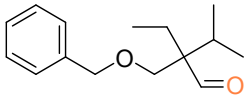
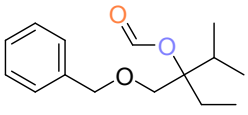
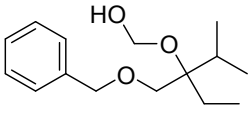
	Reactant	Reported Ground Truth	Top-1 Prediction
1			
2			
3			

Table S5 Examples of other errors in GCN model top-1 predictions.

	Reactant	Reported Ground Truth	Top-1 Prediction
1			
2			

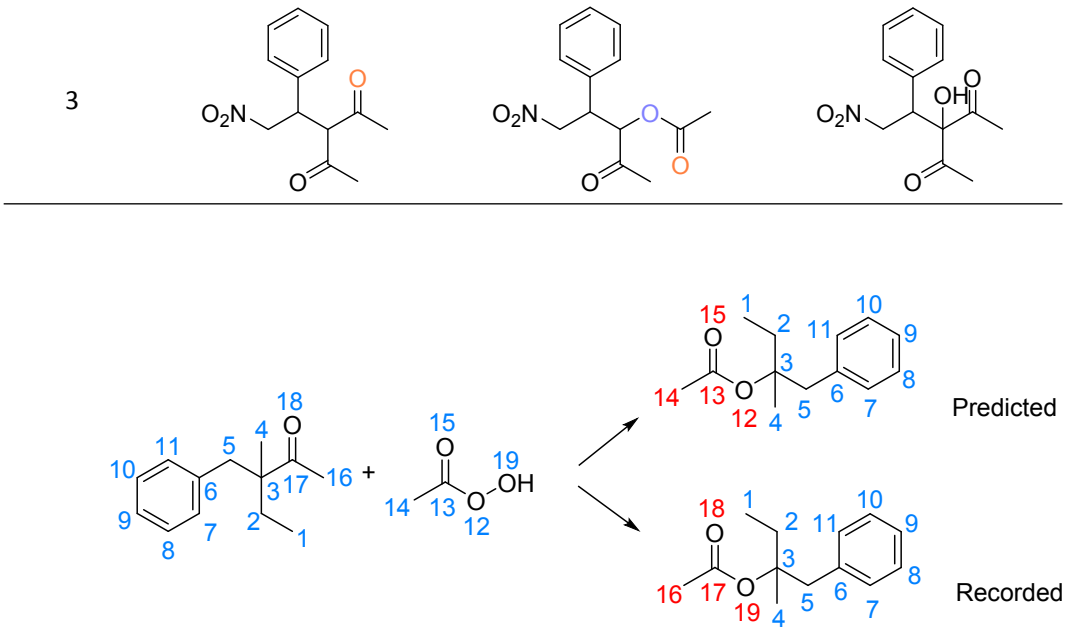


Fig. S6 An example in which the predicted result of atom mapping does not match the recorded result.

Section S6 Prediction of Suzuki reaction

Suzuki reaction is a reaction where the organoboron species is cross-coupled with the halide by using a palladium catalyst and a base to form a carbon-carbon single bond. To manifest the power of GCN model on small-scale reaction, we further adopt the Suzuki reaction to show the predictive performance of GCN model. Meanwhile, we have explored the prediction accuracies comparison between transformer model and GCN model on different size of Suzuki reaction. As displayed in Fig. S7, when the size of data is small, the top-1 accuracy of GCN is significantly higher than that of transformer model. As the size of data increases, the top-1 accuracy of the transformer is gradually improved, while the GCN model can maintain an excellent prediction result.

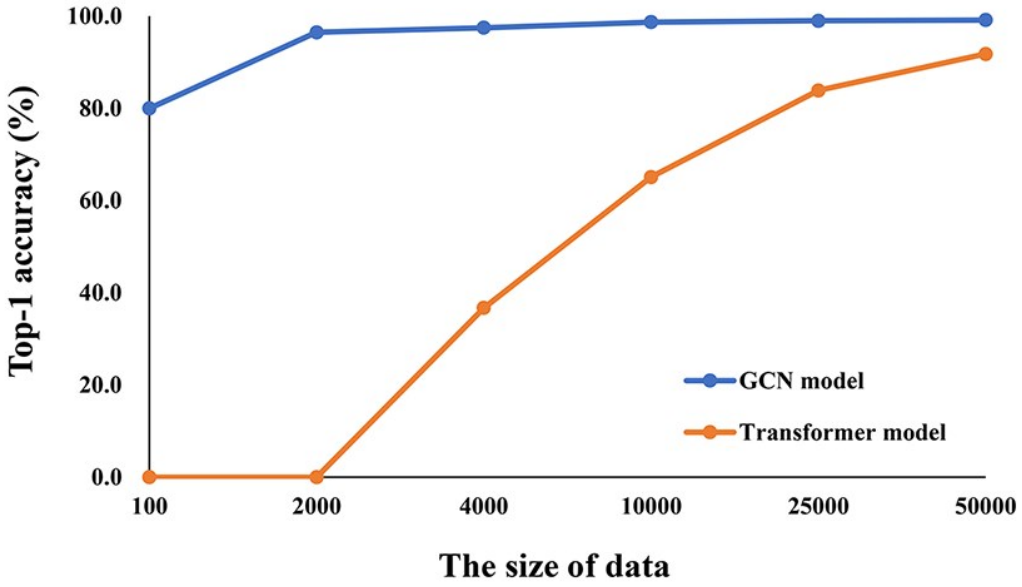


Fig. S7 The top-1 accuracies of GCN model and transformer model on different size of Suzuki reaction.

We further extract some reactions containing more than two reactants from the Suzuki data set, and use the GCN model for training and testing. The top-n accuracy is shown in Table S6. The top-1 accuracy reaches 98.2%, which is similar to that of the Suzuki reaction containing only two reactants.

Table S6 The top-n accuracy of GCN model on Suzuki reaction containing more than two reactants.

Model	Top-n accuracy (%)		
	Top-1	Top-2	Top-3
GCN	98.2	98.8	99.3

Section S7 Application of GCN model in other fields

Apart from reaction prediction, the GCN shows the potential power in other fields. Torng et al.[9] apply a model based on GCN to make drug-target interactions prediction and their work demonstrated that the model can effectively capture the information about protein-ligand binding interactions. By utilizing molecular graphs, the GCN model can be helpful for finding the structure-property relationships and this method outperforms state-of-the-art quantitative structure-property relationship (QSPR) models.[10] What's more, the GCN approach makes a contribution to revealing molecular property by learning relevant knowledge from graphs of molecules.[10] The GCN model becomes an important tool in chemical research gradually.

The neural network model, processed data sets, and evaluation code will be made available at https://github.com/hongliangduan/A-graph-convolutional-neural-network-for-addressing-small-scale-reaction-prediction.

Section S8 References

1. J. R. Alvarez-Idaboy, L. Reyes and N. Mora-Diez, *Org. Biomol. Chem.*, 2007, **5**, 3682-3689.
2. G.-J. ten Brink, I. W. C. E. Arends and R. A. Sheldon, *Chem. Rev*., 2004, **104**, 4105–4123.
3. T. Lei, W. Jin, R. Barzilay and T. Jaakkola, 2017, arXiv:1705.09037.
4. M. F. Hawthorne, W. D. Emmons and K. S. Mccallum, *J. Am. Chem. Soc.*, 1958, **80**, 6393-6398.
5. N. Shervashidze, P. Schweitzer, v. E. J. Leeuwen, K. Mehlhorn and K. M. Borgwardt, *J. Mach. Learn. Res.*, 2011, **12**, 2539– 2561.
6. D. Bahdanau, K. Cho and Y. Bengio, 2014, arXiv:1409.0473.
7. C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, *Chem. Sci.*, 2019, **10**, 370-377.
8. P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS. Cent. Sci.*, 2019, **5**, 1572-1583.
9. W. Torng and R B. Altman, *Journal of Chemical Information and Modeling*, 2019, **59**,4131-4149.
10. J. Lim, S. Ryu, K. Park, Y. J. Choe, J. Ham and W. Y. Kim, *Journal of Chemical Information and Modeling*, 2019, **59**, 3981-3988.
11. X. Wang, Z. Li, M. Jiang, S. Wang, S. Zhang and Z. Wei, *Journal of chemical information and modeling*, 2019, **59**, 3817-3828.