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

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

We describe a graph-convolutional neural network (GCN) model whose reaction prediction capable as potent as the transformer model on sufficient data, and adopt the Baeyer-Villiger oxidation to explore their performance differences on limited data. The top-1 accuracy of GCN model (90.4%) is higher than that of transformer model (58.4%).

The process of predicting a suitable product given reactants is one of the central pillars in the chemical synthetic route. Chemists generally predict reaction outcomes with experience, heuristics, and rules of thumb, which is expensive and time-consuming. Recently, the rapid advances made in deep learning encourage the application of computer-aided methods in chemical research and the use of those algorithms has become a new trend in the field of molecular synthesis.[1-6]

In the past few years, there are many methods have been proposed to address the problem of reaction prediction.[7-9] A neural machine translation (NMT) model was introduced by Nam and Kim in 2016.[10] In their work, a molecule was represented as the simplified molecular-input line-entry system (SMILES),[11] a presentation of text sequences containing the information about atoms and chemical bonds in a compound, and reaction prediction was treated as a translation problem. In the same year, Wei *et al.* utilized fingerprints to describe chemical compounds, and further linked reactants and products as reaction fingerprints.[12] Trained on training samples with limited types (only include alkyl halide reactions and olefin reactions), the neural model could identify the suitable reaction types for input reactants. In addition, Coley *et al.* combined rigid reaction templates with a neural network to predict the majority of possible products from a given set of reactants.[13] In their study, the reaction templates were used to generate a series of possible candidate products that complied with chemical regulations, and then the neural network model selected the most possible product from the candidate products.

However, most of the previous studies have not revealed the chemical information of the reaction. In contrast to those algorithms, work by Coley *et al*. innovatively proposed a graph-convolutional neural network (GCN) model to describe the detailed features of compounds and regarded the problem of predicting chemical reactions as a graph-based task rather than a language translation task.[14,15] In point of fact, graph theory has been widely applied in many aspects of chemistry.[16-22] In this context, a molecule can be regarded as a graph comprising features of atoms and bonds, and chemical information such as aromaticity can be taken into consideration. More significantly, the application of the attention mechanism[23](further information about attention mechanism in Section S2.2 of the ESI†) enables the model to capture the features of the atom itself and other atoms and bonds in a reaction, and the prediction power on the basis of ample reaction data has been confirmed by Coley *et al.*[15]

The research of Philippe *et al.* further verified the ability of GCN model.[24] They investigated that the GCN model can lead to commensurable performance compared to transformer model on sufficient training data. It's noting that the transformer model, a fully attention-based NMT model, is a powerful tool in chemical reaction prediction. What's more, Wang *et al.*[25] indicated that the predictive performance of transformer model is greatly affected by the amount of available data and the same flaw was also pointed out by Zhang *et al.*[26] In other words, the transformer needs to be trained on ample training data before it is put into reaction prediction. Hence, we adopt the GCN model to make predictions with limited training data and the Baeyer-Villiger oxidation (further information about Baeyer-Villiger oxidation in Section S1 of the ESI†) is applied to explore the performance of GCN model trained on the limited samples.

In this study, we focus on the small-scale reaction prediction and show the performance comparison between the GCN model and transformer model in the case of scant data. In contrast to the transformer model which is text-based, the GCN

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

† Electronic Supplementary Information (ESI) available: See DOI:

‡ Yejian Wu and Chengyun Zhang contributed equally to this work.

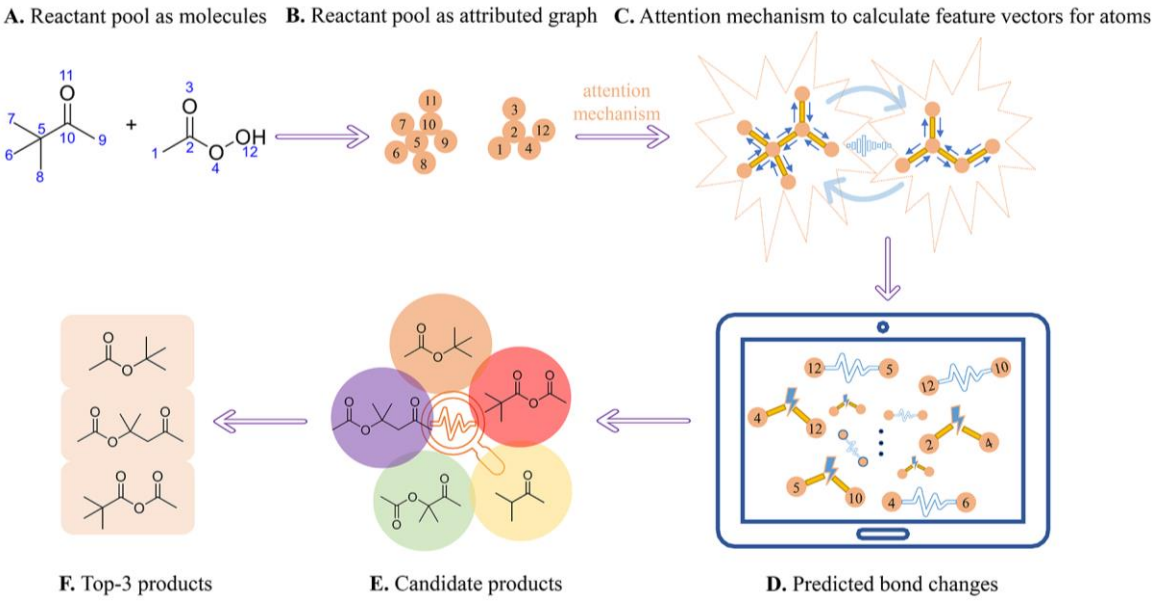

Fig. 1 Schematic diagram of the method for predicting reaction products. We represent the reactant molecules (A) as an attributed graph (B). A graph convolutional neural network learns to calculate both local features and context vectors (C) to predict the likelihood of bond changes for each pair of atoms (D). Generate a series of candidates (F) by arranging and combining the bonds that are most likely to change, and rank these candidates through another graph convolutional network. We extract top-3 products and compare them with the true product.

model treats a chemical molecule as a graph where the bonds are equal to edges and atoms correspond to nodes. We clearly mention that our work aims to show the predictive power of the GCN model in the face of the lack of training data rather than pursue state-of-the-art performance.

The prediction task in our work is divided into two steps and the overall schematic diagram is summarized in Fig. 1. Firstly, the model equipped with the Weisfeiler-Lehman Network (WLN) (further information about WLN in Section S2.1 of the ESI†) learns to analyze the likely reaction centers of an overall reaction - finding out where bonds break and where bonds form. Next, according to changing bonds, the Weisfeiler-Lehman Difference Network (WLDN) (further information about WLN in Section S2.4 of the ESI†) enumerate and rank product candidates with the constraints of the chemical valence rules. Within the perspective of the entire network, the process of the model predict reactions is parallel to the way where humans describe chemical reactions. To achieve high performance in predicting possible products with the scarcity of training examples, the model needs to not only accurately find the reaction center in a reaction but also compute the possibility of predicted products. And it should be noted that the reactants are input as SMILES in the process of predicting reaction although the GCN model extracts relevant properties of molecular in the formulation of graph. In other words, the model translates the reactants SMILES into the graph and incorporate the knowledge of chemistry to make reaction prediction.

We employ the Baeyer-Villiger oxidation reaction, a representative small-scale name reaction, to show the ability of the GCN model faced with the low-data problem. The data originally derived from Reaxys is collected by Zhang *et al.* They extracted those reactions from the commercial databases and filtered irrelevant information (e.g., temperature, time, yields). Then, the reaction dataset is further preprocessed to eliminate the duplicate and error reactions, and the simplified reaction dataset contains reactants and products only. What's more, those reactions SMILES are canonicalized by RDKit[27] and arbitrarily divided into training, validation and test datasets at a ratio of 8:1:1. In our work, we further apply the RXN Mapper[28], an atom-mapping tool, to map the reaction data so that each atom across a reaction has a unique label. With the atom mapping information, the correspondence between reactants and products in a reaction can be presented.

A conclusion that can be drawn from previous work is that the GCN model can compete against the transformer model on the large-scale data. In Philippe's work[24], they compared the performance of GCN and transformer model in a dataset called USPTO_MIT. The USPTO_MIT data set was processed by Jin *et al.*, which derived from USPTO granted patents. In order to show intuitive results and convenient comparison, top-n accuracy is adopted to evaluate model performance. It can represent the ratio of the target product that exists within the top-n candidates predicted by the model.[29] As listed in Table S1, the accuracy difference between GCN and transformer model is slight, which indicates that both the GCN model and transformer model can learn enough reaction knowledge from the sufficient data set.

However, the transformer model fails to absorb plenty of chemical knowledge from training data of limited size owing to the data-driven nature[7]. Hence, we apply the GCN model to explore the challenging small-scale reaction prediction. To manifest the performance of the GCN model on limited data, we

Table 1 Comparison of the top-n accuracy of the transformer and GCN models on the Baeyer-Villiger oxidation reaction dataset

Model	Top-N accuracy (%)		
	Top-1	Top-2	Top-3
Transformer[a]	60.9	68.0	72.1
GCN	90.4	93.4	93.9

[a]The top-n accuracy of transformer on the Baeyer-Villiger oxidation reaction dataset is originally derived from Zhang *et al.*'s work.

compare it with Zhang *et al.*'s work where the results of transformer model are revealed. Hence, there are two models involving in our experiment, one is the transformer which is a baseline model and the other is the GCN model. Those models both are trained and tested on the Baeyer-Villiger oxidation reaction dataset and the top-n results of them are shown in Table 1. It's worth noting that the results of transformer on the Baeyer-Villiger oxidation are originally derived from Zhang *et al.*'s work.

We observe that the GCN model performs well with a top-1 accuracy of 90.4%, which is much higher than the 60.9% accuracy of the transformer model, even for such small-scale data reaction involving the regioselectivity. In the case of top-3 accuracy, the GCN model is also significantly higher than the transformer model by 21.8%. The higher accuracy of the GCN model (>90%) shows that the GCN model has better applicability on limited data than the transformer model.

As depicted in Fig. 2, the transformer model, a data-driven model, can achieve similar performance to the GCN model on large-scale data sets. More importantly however, the gap of accuracy between the GCN model and transformer model becomes hard to ignore as the available data volume decreases. In the following section, we chose top-1 results to demonstrate the great value of the GCN model on the small-scale reaction prediction and detailly analyze the performance of this model to further improve its ability in the task of predicting the outcomes of reactions.

Fig. S3 represents some examples of group migration error that occur in the transformer model but not in the GCN model. Take Fig. S3(a) as an example, there are two chlorines are attached to the right α-carbon of the carbonyl group in S-methyl 6,6-dichloro-7-oxobicyclo[3.2.0]hept-2-ene-2-carbothioate.The electron-withdrawing effect of chlorine greatly reduces the electron cloud density of the carbon and hinders the migration of the substituent.[30] Hence, the reaction tends to generate the product in which the oxygen atom is inserted at the left position

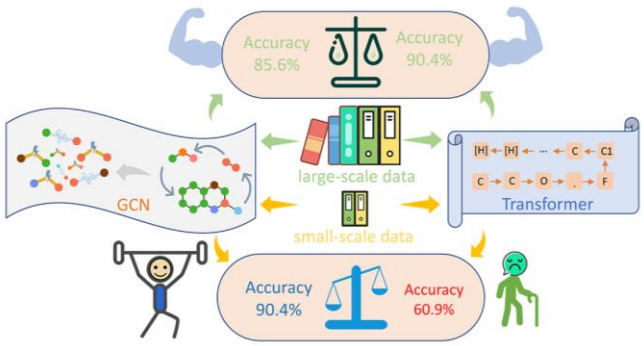

Fig. 2 Performance comparison between GCN model and transformer on large-scale data and small-scale data respectively.

of the carbonyl group. This is indeed what is logical and what the GCN model predicts. Besides, if an alkoxy group is attached to the adjacent carbon of the carbonyl group, the lone pair electrons on the oxygen attached to the α-carbon may facilitate the migration of the group.[31] There is a representative reaction example shown in Fig. S3(c). The 6-methoxy-2,2,5,7-tetramethyltetrahydrobenzo[*d*][1,3]dioxol-4(3a*H*)-one can be oxidized to form the 7-methoxy-2,2,6,8-tetramethyltetrahydro-[1,3]dioxolo[4,5-b]oxepin-4(3a*H*)-one.The prediction made by the GCN model is in line with the reported ground truth. However, the transformer model does not capture similar chemical rules. From Fig. S3, we can observe the GCN model can acquire more information about the rule of group migration compared to the transformer model.

Some examples of other error types predicted by the transformer model are listed in Fig. S4. We can find that a reactant contains two carbonyl groups that can be attacked by peroxide. An additional level of complexity for the model is to identify where is the real reactive site when the reactant is equipped with two carbonyl groups. As is depicted in Fig S4(a), 5-acetyl-3-((tert-butyldimethylsilyl)oxy)-2-methylcyclohexan-1-one contains two carbonyl groups and 4-acetyl-6-((tert-butyldimethylsilyl)oxy)-7-methyloxepan-2-one is the ground truth. However, the transformer model blindly treats both carbonyl groups in the reactant and the prediction made by this model is 6-((tert-butyldimethylsilyl)oxy)-7-methyl-2-oxooxepan-4-yl acetate. In contrast to the transformer model, the predictions of the GCN model are consistent with the reported ground truths, which indicating powerful predictive ability of the GCN model on the insufficient training samples. Furthermore, those examples reveal that the GCN model gains a deeper insight into the knowledge about the Baeyer-Villiger oxidation reaction than the transformer model after training on the same data set.

A side effect of phrasing reaction prediction as a translation problem is that the alteration of a single character in the SMILES may lead to the structure of a molecule to change or even render it invalid.[32] Due to the text-based characteristic, it's inevitable that the transformer model appears the error type of SMILES invalid in the prediction products. There are representative examples of SMILES error in Fig. S5. The common feature of those reactions is the complex ring structures, which may induce the SMILES invalid in the prediction process. In turn, the GCN model performs better on these reactions. In the Fig S5(c), when input methyl 2-hydroxy-9-oxo-4-phenylbicyclo[3.3.1]nonane-2-carboxylate, the ground truth product of this reaction is methyl 4-hydroxy-10-oxo-2-phenyl-9-oxabicyclo[3.3.2]decane-4-carboxylate. Due to the complexity of ring structure in this reaction, the wrong SMILES given by the transformer model cannot be converted to a molecular structure. However, the GCN gains the chemical information in a graph which can avoid SMILES error and performs better in terms of compounds with complex structures.

Despite the reliable predictive performance of the GCN model on the small-scale reaction, there still exists space for future improvement. To further improve the accuracy, we make a deeper analysis of results of the GCN model. The mistakes in our experiment are divided into four classes and the distribution of

prediction errors can be found in Table S2. The error type with the largest proportion is group migration error, accounting for 45.5% in wrong predictions. Because both induction effect and conjugation effect affect the migration ability of group in a reaction, it remains a challenge to judge the group which is more likely to migrate. And we further carried out a detailed analysis of the four categories errors of the GCN model. (detailed information is available in Section S5 of the ESI†).

As stated earlier, the GCN and transformer models can achieve comparable performance on a large set of data. However, when scaling down to a small size of data, the transformer is hard to learn enough knowledge for making accurate reaction prediction.[25,26] Hence, we describe the GCN model in our work and apply it to predict the most likely products given reactants. The key difference between GCN and transformer model is that the former relies on the graph theory but the latter is text-based.

In contrast to the transformer model, our model views reaction prediction as a transformation of graphs rather than a language translation task. By dividing the reaction prediction into two steps, the interpretability of the model is aligned with the chemist's description of a chemical reaction. Here, we chose the Baeyer-Villiger oxidation reaction, a classic small-scale name reaction, to verify the predictive performance of the GCN model. The top-1 accuracy of the GCN model is 90.4%, which is significantly high than the transformer model. It elucidates that the GCN model not only distills specific chemical principles of the Baeyer-Villiger oxidation reaction but also has a profound chemical understanding from limited-data. In addition to compare the results of the GCN and transformer model, we further list some error types of the GCN model to learn more about this model. Overall, the aim of our work is to show that the GCN model is a better-suitable method for small-scale reactions than the transformer model.

We were grateful to the National Natural Science Foundation of China (no. 81903438) for financial help.

Conflicts of interest

There are no conflicts to declare.

Notes and references

1 F. Peiretti and J. M. Brunel, *ACS Omega*, 2018, **3**, 13263-13266.
2 O. Engkvist, P. O. Norrby, N. Selmi, Y. H. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discov. Today*, 2018, **23**, 1203-1218.
3 X. Yang, Y. Wang, R. Byrne, G. Schneider and S. Yang, *Chemical Reviews*, 2019, **119**, 10520-10594.
4 H. Ozturk, A. Ozgur, P. Schwaller, T. Laino and E. Ozkirimli, *Drug Discov. Today*, 2020, **25**, 689-705.
5 W. A. Warr, *Mol. Inform*., 2014, **33**, 469-76.
6 J. Jiménez-Luna, F. Grisoni and G. Schneider, *Nature Machine Intelligence*, 2020, **2**, 573-584.
7 P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9** (28), 6091-6098.
8 M. H. S. Segler and M. P. Waller, *Chemistry*, 2017, **23**, 5966-5971.
9 J. Bradshaw, M. J. Kusner, B. Paige, M. H. Segler and J. M. Hernández-Lobato, 2018, arXiv:1805.10970.
10 J. Nam and J. Kim, 2016, arXiv: 1612.09529.11.
11 D. Weininger, *J. Chem. Inf. Model*, 1988, **28**, 31-36.
12 J. N. Wei, D. Duvenaud and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2016, **2**, 725-732.
13 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434-443.
14 W. Jin, C. W. Coley, R. Barzilay and T. Jaakkola, 2017, arXiv:1709.04555.
15 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.
16 A. T. Balaban, *J. Chem. Inf. Comput. Sci*., 1985, **25**, 334–343.
17 S. Fujita, *J. Chem. Inf. Model*, 1986, **26**, 212-223.
18 C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola and K. F. Jensen, *J. Chem. Inf. Model*, 2017, **57**, 1757-1772.
19 W. Torng and R. B. Altman, *J. Chem. Inf. Model*, 2019, **59**, 4131-4149.
20 A. M. Schweidtmann, J. G. Rittig, A. König, M. Grohe, A. Mitsos and M. Dahmen, *Energy & Fuels*, 2020, **34**, 11395-11407.
21 J. You, B. Liu, R. Ying, V. Pande and J. Leskovec, *Adv. Neural Inform Process Syst.*, 2018: 6410-6421.
22 S. Ishida, K. Terayama, R. Kojima, K. Takasu and Y. Okuno, *J. Chem. Inf. Model*, 2019, **59**, 5026-5033.
23 D. Bahdanau, K. Cho and Y. Bengio, 2014, arXiv:1409.0473.
24 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572-1583.
25 L. Wang, C. Zhang, R. Bai, J. Li and L. H. Duan, Chem. Commun., 2020, **56**, 9368-9371.
26 Y. Zhang, L. Wang, X. Wang, C. Zhang, J. Ge, J. Tang, A. Su and H. Duan, https://doi.org/10.26434/chemrxiv.13383275.v1.
27 G. Landrum, RDKit: Open-source cheminformatics, 2006, http://rdkit.org.
28 P. Schwalle, B, Hoover, JL.Reymond, H.Strobelt and T. Laino, https://doi.org/10.26434/chemrxiv.12298559.
29 C. Zhang, L. Wang, Y. Wu, Y. Zhang, A. Su and H. Duan, https://doi.org/10.26434/chemrxiv.13173674.v1.
30 E. E. Smissman and J. V. Bergen, *J. Org. Chem.*, 1962, **27**, 2316-2318.
31 N. Chida, T. Tobe and S. Ogawa, *Tetrahedron Lett.*, 1994, **35**, 7249.
32 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103-1113.