Virtual data augmentation method for reaction prediction in small dataset scenario

Xinyi Wu,[a]+ Yun Zhang,[a]+ Jiahui Yu,[a] Chengyun Zhang,[a] Haoran Qiao,[b] Yejian Wu,[a] Xinqiao Wang,[a] Zhipeng Wu,[a] and Hongliang Duan*[a]

[a]. X. Wu, Y. Zhang, J. Yu, C. Zhang, Y. Wu, X. Wang, Z. Wu, Prof. H. Duan
Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences
Zhejiang University of Technology
Hangzhou, 310014 (P. R. China)
E-mail: hduan@zjut.edu.cn

[b]. H. Qiao
College of Mathematics and Physics
Shanghai University of Electric Power
Shanghai, 201203 (P. R. China)

Abstract

To improve the performance of data-driven reaction prediction models, a new data augmentation method for augmenting data volumes is presented that aims to add fake data in training dataset. This method is only pay attention to small-scale reactions, and manually generates fake data which is chemical and credible molecules by replacing functional groups in reaction sites. And we call this method as virtual data augmentation. Additionally, the transformer model is introduced to explore the effectiveness of virtual data augmentation method in the task of reaction prediction based on small data sets. We apply our method to five classic coupling reactions, the results show that the overall performance of the transformer-baseline model and transformer-transfer learning model combined with virtual data augmentation method is obviously improved, compared to raw datasets. Especially for Suzuki reaction, combining transfer learning strategy and virtual data augmentation method, reaches top-1 accuracy of 97.8%. To sum up, virtual data augmentation can be used as a measure to face up the problem of insufficient data and significantly improves the performance of reaction prediction.

Introduction

Organic synthesis is not only occupying a core position in organic chemistry field, but also support the research and development of others fields, such as material science, environment science and drug discovery. With the maturity of artificial intelligence techniques, there are many successful applications of integrated organic chemistry and artificial intelligence,[1-4] such as reaction prediction.[5-8] As the name implies, reaction prediction task is that inferred products from the given reactants or reagents by learning chemical rules based on deep learning methods. Since Schwaller *et al.* creatively used sequence-to-sequence model to assist predict the generation of organic chemical reactions,[9] reaction prediction has been a heated topic. Indeed, this new mode can reduce the cost of material and human resources, even guide chemists and help design new molecules by reducing the number of synthesis attempts compared with traditional organic experiments depending on the professional chemical knowledge of organic

chemistry experts.

However, it's well known that the deep learning methods are determined by large datasets and previous researches have demonstrated that focusing on massive reaction dataset can achieve considerable efforts.[10,11] While when it comes to a particular reaction type, the data volumes are insufficient to support related applications, due to high cost and time-consuming experiments. For example, Wang *et al.* applied 9959 heck reactions to the transformer-baseline model for reaction prediction, and only got 66.3% accuracy. [11] As a result, deep learning methods must be able to comprehensively deal with small datasets to solve project-tailored task in cross domain with chemistry.

Luckily, many strategies are designed for the poor performance in small dataset of deep learning methods.[12-16]One of the effective methods is transfer learning, which transfer prior-knowledge learned from abundant data to another domain task with less data available in similar scenario.[10,11,17-19] Reymond *et al.* had performed transfer learning on carbohydrate reactions and showed better performance than a model trained on carbohydrate reactions only.[19] Apart from transfer learning method, data augmentation strategies are crucial for deep learning pipelines aiming at reaction prediction tasks, as model's performance increases with the amount of training data. Data augmentation is the process of modifying, or "augmenting" a dataset with additional data, which is a powerful strategy used in image processing.[20-22] Also, previous research by Tetko *et al.* had proved that simultaneously augmenting input and target data can improve the performance for prediction of new sequences.[23] Broadly speaking, augment training set sequences makes deep learning methods achieve better accuracy according to the characteristic of simplified molecular-input line-entry system (SMILES).[24,25] It's worth noting that all the augmented SMILES are valid structure without changing chemical meaning. Unlike these methods, synthetic data augmentation is another powerful data augmentation. Maimaiti *et al.* manually created a batch of fake data to increase the target training set by deleting words, randomly sampling or replacing some words in the text in the task of text generation.[26] Through this way, synthetic data augmentation was realized by transforming the text in the low-resource language scenarios. As a result, inspired by text replacement and similarity between SMILES representation and text, we attempt to adding fake data instead of "random SMILES" into training dataset to improve models' accuracy and called it as virtual data augmentation. The fake data are generated by replacing substituent with equivalent functional groups in reactants, which do not change reaction site and atom valence of reactant molecules. It can be expected that this method can better improve the performance of the data-driven mode.

In this paper, we apply and securitize virtual data augmentation regime and show that fake data lead to better performance on transformer model, which is a state-of-art natural language process model.[23,27,28] Although the transformer model shows excellent performance in various reaction tasks, the data-driven model is still powerless when facing low data resources. We also clearly mention that our study is to predict the outcomes of reaction, and the detailed process can be seen in Fig. 1. For the datasets used in this study are coupling reactions, an organic chemical reaction in which two chemical entities (or units) combine to form one molecule. When the virtual data

augmentation method is trained on the transformer-baseline model, compared with the raw data, the accuracy of reaction prediction is improved from 2.74% to 25.8%. Also, integrated with transfer learning methods, the transformer model increased from 1% to 53%, which prove that this virtual data augmentation can help to improve model's performance. All in all, this virtual data augmentation aims to expand the density of sample data points in the chemical space already covered by the existing documentary data set. And we believe that this method can be a useful tool in low resource scenarios to tackle small dataset task with deep learning methods.

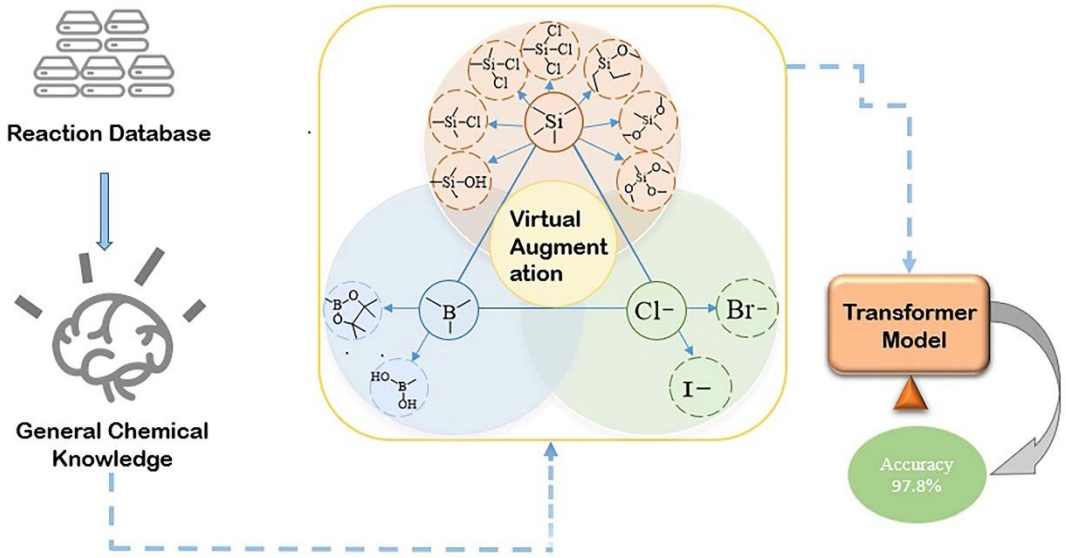


Fig. 1 Schematic illustration of the virtual data augmentation method.

2 METHODS

2.1Data

2.1.1Dataset Preparation

In this work, we exported Buchwald-Hartwig, Chan-Lam, Kumada, Hiyama and Suzuki five coupling reaction datasets based on the name and structure search from the 'Reaxys' database.[29] Each data set is preprocessed with following procedures. First, these data sets are deleted irrelevant information (e.g., pressure, temperature and yield et al) and remained reaction entries and reagent entries. Secondly, the reaction SMILES were canonized and all the duplicated reaction entries were removed. Finally, these five reaction datasets were filtered using template screening followed respective reaction rule.

Next up, the virtual data augmentation method was divided into two types according to general characteristics of reactants in these five coupling reactions. The first augmentation method is augmenting one of the reactants and we defined it as single augmentation. As Fig.2(a) shows, to the Chan-Lam reaction, the virtual data augmentation was carried out in the reactants with halogen functional group, while the Buchwald-Hartwig reaction augmented the reactants with boron functional group. The

other one virtual data augmentation method is simultaneously augmented multiple reactants. As shown in Fig. 2(b), the Hiyama reaction is simultaneously augmented the reactants with silicon functional group and reactants with halogen functional group respectively. The Kumada reaction simultaneously augmented the reactants with halogen functional groups respectively and Grignard reagents containing halogen functional groups; Suzuki reaction simultaneously increases reactants with halogen functional groups and reactants with boron functional groups.

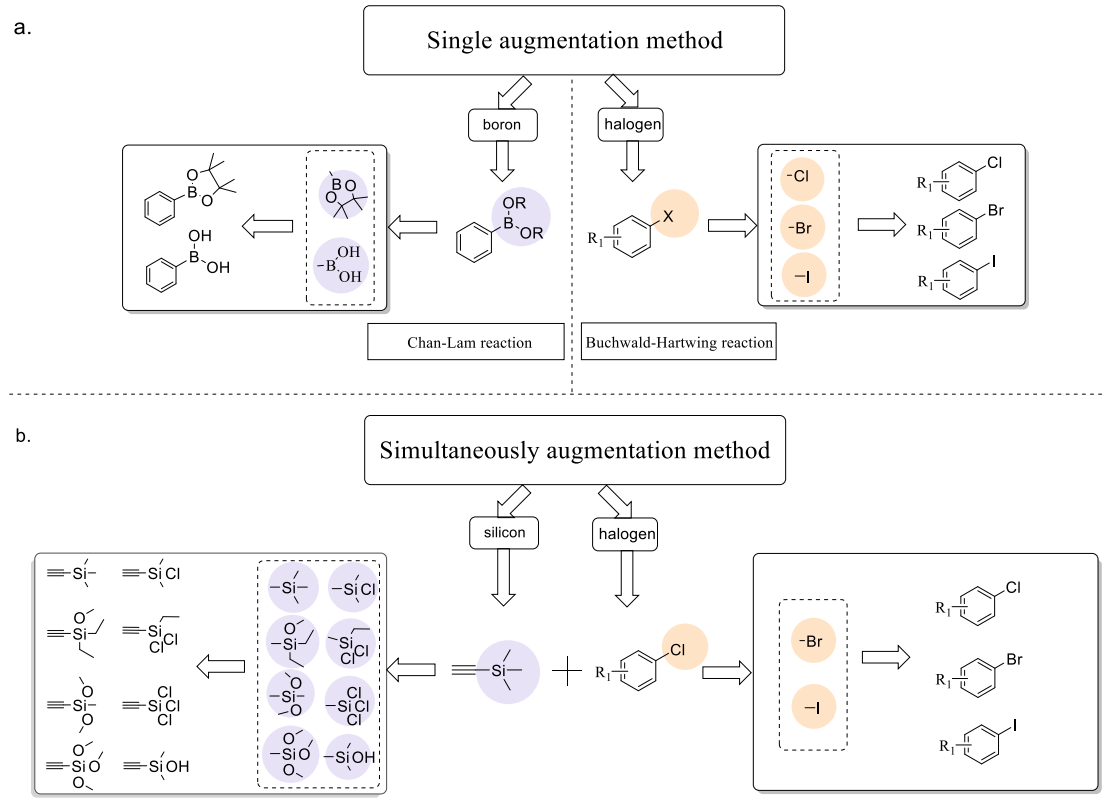
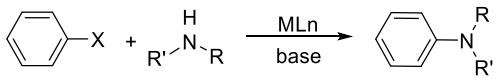
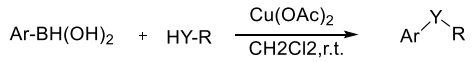
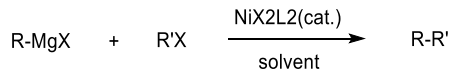
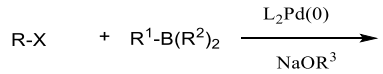


Fig. 2 The schematic diagram of virtual data augmentation. (a) The single augmentation method of Buchwald-Hartwig and Chan-Lam coupling reactions. (b) The representative example of simultaneously virtual data augmentation method of Hiyama coupling reaction.

Table 1 Comparison of data quantity of five coupling reactions before and after using virtual data augmentation method.

name	depiction	raw dataset	virtual dataset
Hiyama	R-X + R^1Si-R2 ⟶ R-R^2	2067	19011
Buchwald-Hartwig		4419	7640

Chan-Lam		5276	9170
Kumada		9657	54062
Suzuki		92399	424194

In addition, Table 1 shows the data quantity of five coupling reactions before and after data augmentation, and describes the corresponding reaction formulas. It can be seen that the amount of data is obviously increased after using the virtual data augmentation method, which is 2-6 times of the raw data.

Ultimately, for the augmented dataset and raw dataset, we randomly divided the datasets into training, validation, and test datasets at a ratio of 8: 1: 1. To avoid contingency, we augmented the training dataset of these five reactions without augmenting validation and test datasets. It is worth noting that the repetitive reaction produced by virtual data amplification method has been deleted by us. In addition, all scripts are written in Python (version 3.7) and using RDKit for processing.[30]

2.1.2 USPTO dataset:

The data we used to pre-train the model were derived from the U.S. Patent and Trademark Office (USPTO), and this data set comes from Lowe's patent mining work, which extracted instances of USPTO patent reactions granted between 1976 and 2016 as available public data.[31] Coley *et al.* extracted 480k reactions from the USPTO authorized patents.[32] We processed the data of Coley *et al.*, deleted reaction reagents and chirality, filtered out incomplete or wrong reactions, and then after pretreatment such as standardization and repeat removal, recovered chirality from the USPTO for each reaction originally containing chirality. In particular, we should emphasize that there is only one single product of the reaction we extracted, and the raw data set and the augmented data set have also been deleted from the data set of the USPTO. Finally, about 41W single product reactions were obtained as pre-training data set.

2.2Model

During the work, the transformer baseline, transformer transfer learning was adopted to verify the augmented method's validation.

The model we used is entirely based on the Transformer model, which is a powerful model for handling Natural Language Processing (NLP) task and was proposed by Google in 2017.[14] The model was originally designed and applied to the neural machine translation task. When the reaction prediction task is regarded as a language translation task, the Transformer model can be applied to this task. This model relies entirely on attention mechanism, and it can handle text tasks without using RNNs and convolution, and it also avoids the recursion problem in encoder-decoder architectures. It contains

several identical encoder-decoder layers. In addition, the application of Multi-Headed Attention (MHA) in the decoder makes the calculation speed faster and improves the performance of the model. The reaction prediction process of transformer model is that the reactants are transmitted to the encoder in the form of SMILE code as input, and then transferred to the next encoder until the last encoder is transmitted to the decoder, which outputs the predicted results. The model used in our work for reaction prediction originated from Zhang *et al.*[33]

We also introduced transfer learning strategy into the Transformer model. During the pretraining process, a large chemical reaction dataset USPTO-41W was used to pretrain the model. The model transfers the general chemistry information learned from pretraining to the target task of predicting the outcomes of these five coupling reactions. In addition, the transfer learning is combined with virtual data augmentation to further improve model's performance. With this new strategy, the model can abundantly learn chemical information from USPTO-41Wdataset and fake data adding to the training dataset.

Result and discussions

In this work, the transformer model was used to predict the outcomes of several coupling reactions. To avoid the occurrence of overfitting or underfitting, the accuracy of the transformer-baseline model was carried 10-fold cross validation. The virtual data augmentation method was first tested on the Hiyama coupling data set, the smallest dataset of five reaction dataset. Table 2 shows the accuracies of Hiyama, Burcharld-Harwig, Chan-Lam, Kumada, Suzuki reaction based on the transformer-baseline model. The development of this model with raw datasets as training sets provided 23.67%, 41.63%, 64.71%, 78.99%, 95.05% accuracy, respectively. An attempt to use the transformer model to predict the test set through augmented training sets resulted in much higher top-1 predictions of 49.47%, 49.32%, 68.50%, 85.40 ,97.79% respectively. Overall, though the model got a poor performance in predicting outcomes of these five reactions, an obvious increasement can be found before and after augmented training dataset. This result can be expected, for that adding fake data to expand data volume can assist model to improve predictive performance.

Table 2. Accuracy comparison of several coupling reactions between raw data and augmented data based on the transformer-baseline model and transformer-transfer model

Model	Dataset	Reaction Types				
		Hiyama	Buchwald-Hartwig	Chan-Lam	Kumada	Suzuki
Transformer-baseline model	Raw data	23.67	41.63	64.71	78.99	95.05

	Augmented data	49.47	49.32	68.50	85.40	97.79
Transformer-transfer model	Raw data	60.87	94.57	96.39	96.48	97.84
	Augmented data	69.57	95.93	96.77	97.00	98.63

To better verify whether this virtual data augmentation is effective, transfer learning method was integrated with transformer-baseline model. Table 2 also summarized the accuracies of these five reactions in the transformer-transfer learning model. On the one hand, with introduced the transfer learning strategy, compared with the augmented datasets, the overall accuracy of transformer-transfer learning model improved nearly 20% on average compared with the transformer-baseline model. On the other hand, the gaps between raw datasets and augmented datasets had great improvement after combined to transfer learning method, especially in Buchwald-Hartwig dataset. These results indicated that adding fake data can be combined with transfer learning method to jointly solve the problem of data scarcity. Another important point is that adding fake data to the training dataset had better performance compared to raw datasets, which indicated this virtual data augmentation method is effective and had generality that be used in different scenarios.

Since the virtual data augmentation method made a great difference in transformer models, it is desirable to visualize the relation location of datasets in chemical space and how the model allow for interpretation. We first visualized raw data and augmented data among these five reactions using Uniform manifold approximation and projection (UMAP) and the tree-map (TMAP).[34,35] In this section, we generated the plots of reactant molecules belonging to raw datasets and augmented datasets using UMAP, which represent molecules as Morgan fingerprints to create a two-dimensional representation of high-dimensional data distributions. Taking Hiyama reaction and Chan-Lam reactions as examples, As Figure3(a) demonstrated that the silicon-containing molecules generated by virtual data augmentation occurring in the training set of Hiyama reaction (light pink) are closely to the Hiyama raw datasets (pink), and the halogen-containing molecules generated by virtual data augmentation occurring in the training set of Hiyama reaction (light blue) are closely to the Hiyama raw datasets(blue). Also, we generated the UMAP plot of Chan-Lam reactions in Figure3(b), which has only boron-containing molecules been augmented singly. The boron-containing molecules generated by virtual data augmentation (blue) are still very close to the raw datasets (light blue). This graphical analysis confirms that the effectiveness of virtual data augmentation based on text replacement for maneuvering in chemical space from the source to the objective. It is believed that this theory is also applicable in other reaction data sets.

Additionally, to further explore the relationship of the datasets we used, all the datasets except USPTO-41W was visualized by the TMAP. TMAP is another powerful visualization tool to represent large high-diversional datasets as two-diversional

connected tree. In this TMAP plot, both raw reactions and virtual data augmentation reactions are represented a point according to the reaction fingerprint RXNFP, for which dates from a neural network trained to classify patent chemical reactions. It is worth noting that the data we put into TMAP are 5,000 reactions randomly selected from the data of five coupling reactions before and after amplification, and if there are less than 5,000 reactions in the raw data set, we put all the data into the visualization tool. As Fig.3(c) shows, the raw datasets and augmented datasets derived from the same type reactions were well overlapping, and the reactions belongs to different reaction types can be obviously separately, illustrating that the fake data produced by this virtual data augmentation method is relatively similar to the raw data. Taken together, from the perspective of training model, adding the fake data derived from virtual data augmentation according to text replacement is effective and create positive effort in improving deep learning models' performance.

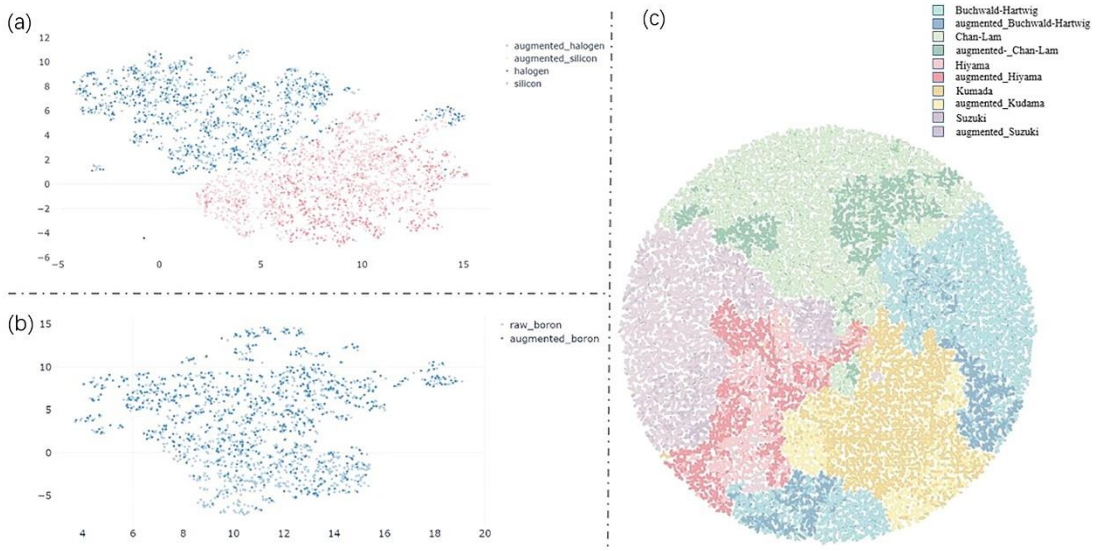


Fig. 3 UMAP plot of molecules from raw data and virtual augmented data and TMAP plot ofrxnfp of reactions from raw data and virtual augmented data. (a)UMAP map of Hiyama coupling reaction before and after virtual data augmentation. (b)UMAP map before and after virtual data augmentation of Chan-Lam coupling reaction. (c)TMAP before and after virtual data augmentation of five classic coupling reaction.

Finally, we use attention weight to visualize the learning process of model for transformer model.[36] Attention weights is the key to take into account long-distance dependencies and has been used in reaction predictions and other fields. For predicting the outcomes of variable coupling reaction, specific reagents have certain impact on the output of model. Especially, using attention weights can provide us a straight form of how the model learning the molecules SMILES input and output. Fig.4(a) shows the visualization of attention weight of a group of raw Hiyama reactions. The darker the token, the more noticeable it is in this particular layer or output step. It can be seen from the figure that [F-] in the reagent activates the Si-R bond with low polarization in silicone, so as to exchange with R-X, resulting in cross-coupling reaction. Fig.4(b) corresponds to a group of reactions after Hiyama augmentation, and the weight of attention in reaction is almost the same, focusing on the position where cross-coupling occurs. The results show that there is no difference between the reaction sites of the augmented Hiyama data and the raw data. This means that the fake data we put forward is meaningful in training the model.

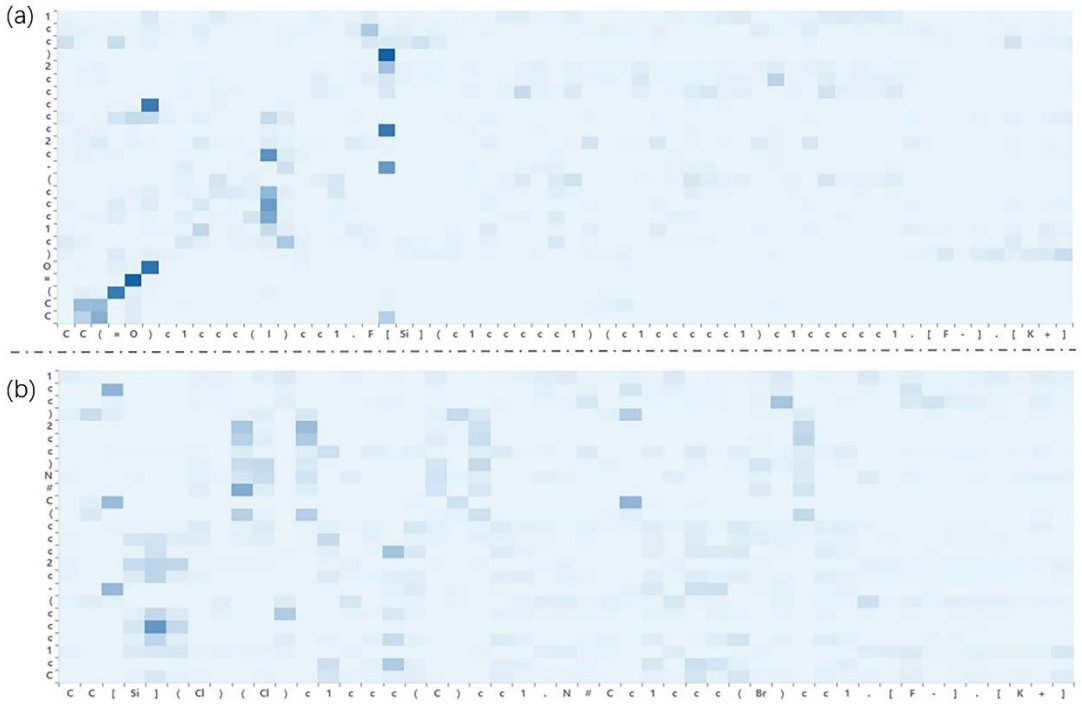


Fig. 4 Visualization of attention weight before and after Hiyama reaction augmentation. The horizontal axis contains two reactants and reagents, and the vertical axis is the product.
(a)SMILES:CC(=O)c1ccc(I)cc1.F[Si](c1ccccc1)(c1ccccc1)c1ccccc1.[F-].[K+]>>CC(=O)c1ccc(c2ccccc2)cc1
(b)SMILES:CC[Si](Cl)(Cl)c1ccc(C)cc1.N#Cc1ccc(Br)cc1.[F-].[K+]>>Cc1ccc(-c2ccc(C#N)cc2)cc1

Table 3. The comparisons of different augmented reactants.

Reaction types	Accuracy (%)		
	Augmented halogen	Augmented silicon (or boron)	Simultaneously Augmented
Hiyama	44.44	48.31	49.47
Kumada	80.85	84.68	85.40
Suzuki	96.82	95.26	97.79

In addition to augmenting both reactants in simultaneously augmentation (As shown in Table 1 above), we conduct augmenting experiments based on either reactant in reactions that can be augmented simultaneously. For these reactions, we augmented the one reactant and the other one reactant in the order. The results are shown in Table 3. For Hiyama reactions, augmented reactant with halogen, the transformer-baseline model achieves 44.44% accuracy, and augmented reactants with silicon, the accuracy is on par with halogen augmented. While the two reactants are augmented at the same time, the transformer-baseline model's performance increased nearly 5%. This similar phenomenon can be observed in Kumada reactions and Suzuki reactions. Especially for Kumada reaction, after simultaneously augmented reactants, the performance of transformer-baseline model increases from 80.85% to 85.40%. For the Suzuki reaction, it may be that its own data sets are more than other data sets, so the accuracy rate is not improved much, but the overall accuracy rate is still improved to nearly 98%. After augmented all the reactants, the transformer model can learn more chemical information about the reaction, and thus achieving a higher performance in reaction prediction. It can be attribute to transformer model's ability to encoder and decoder text sequence.

Table 4. Compare the accuracy of Suzuki reaction datasets of different sizes in the transformer-baseline model before and after data augmentation.

Suzuki Dataset	Accuracy (%)						
	1k	3k	5k	7k	1w	3w	6w
raw data	1.20	45.04	69.50	78.34	83.57	91.17	93.13
Augmented data	1.58	60.38	74.52	81.11	85.87	93.84	95.90

To better understand the model training, we conduct several experiments where different number of training data sets were randomly chosen to monitor its prediction performance in the transformer-baseline model. Different from others tests, this test is based on Suzuki reaction, which is the largest dataset in all of datasets we self-built. We aimed at the one-fold set of Suzuki reaction, and randomly sampled 1k, 3k, 5k, 7k, 1w, 3w, 6w training dataset in raw data. Then these datasets were augmented according to the reactants containing halogen and boron. On the one hand, the transformer model's performance is affected by the size of training set. For instance, As the table 4 shows,

sampled 1k reactions from Suzuki dataset, the transformer-baseline model only provided the smallest performance of 1.20% based on the row dataset, even if integrated with virtual data augmentation, this model reached 1.58%. In contrast, we sampled 1w row data, the model calculated top-1 performance of 83.57%, and increased nearly 2.3% after applying virtual data augmentation method. With the augmentation of the training set, the performance of the model has been significantly improved, demonstrated that direct data augmentation can be great for improving model's performance. On the other hand, the effect of virtual data augmentation does not always increase accompanying with the increasement of training set amount. For example, when training 1k raw data and relevant augmented data, the transformer-baseline model increased 0.4% before and after applying virtual data augmentation. The gap of augmented 3k training dataset reach the largest increasement of 15%. While sampled dataset surpass 5k dataset, the virtual data augmentation method keeps steady gradually, and there is no obvious upward trend. Therefore, despite the fact that the transformer-baseline model is faced with more samples, the transformer-baseline model may be overused, and no new knowledge is acquired from chemical reactions.

Table 5. The number of error types in reaction prediction for five coupling reactions.

Wrong type	Hiyama lift rate (%)	Suzuki lift rate (%)	Buchwald-Hartwig lift rate(%)	Cham-Lam lift rate (%)	Kumada lift rate (%)
Chirality error	1.00	5.50	1.63	1.71	4.05
Group isomerism error	10.67	9.50	15.51	10.86	13.51
Number of carbon error	16.50	11.00	11.43	10.29	16.22
SMILES error	11.65	33.50	33.06	28.57	28.38
Other's error	60.19	40.50	38.37	48.57	37.84

We further analyzed wrong predictions predicted by the transformer-baseline model of these five reactions before and after adding fake data to the training set to evaluate the validity of virtual data augmentation. As Table 5 shows, there are mainly four errors: invalid SMILES errors, the number of atom errors, chirality errors and functional group isomerism errors. Fig. 5 and Fig. 6 respectively list several representative examples in the reaction of Hiyama and Suzuki. In these four errors, no matter training on the raw dataset or the augmented dataset, the most common error besides other messy errors is the SMILES error. The error rate of functional group isomerization is second only to that of smiles error, especially for Hiyama reaction. It can be attributed to the data augmentation method we applied is based on replace variable functional group, which makes the functional groups confuse the predictive performance of the model.

For the other errors, they were observed in multiple reaction predictions project. The result for these errors is that the transformer-baseline model's modest ability in tacking small reactions. Furthermore, compared the amounts of wrong predictions of raw data to the errors of augmented dataset, we found that applied virtual data

augmentation method, the ratio of each error is reduced nearly 20%. That means the method we proposed actually improved the model's performance from source.

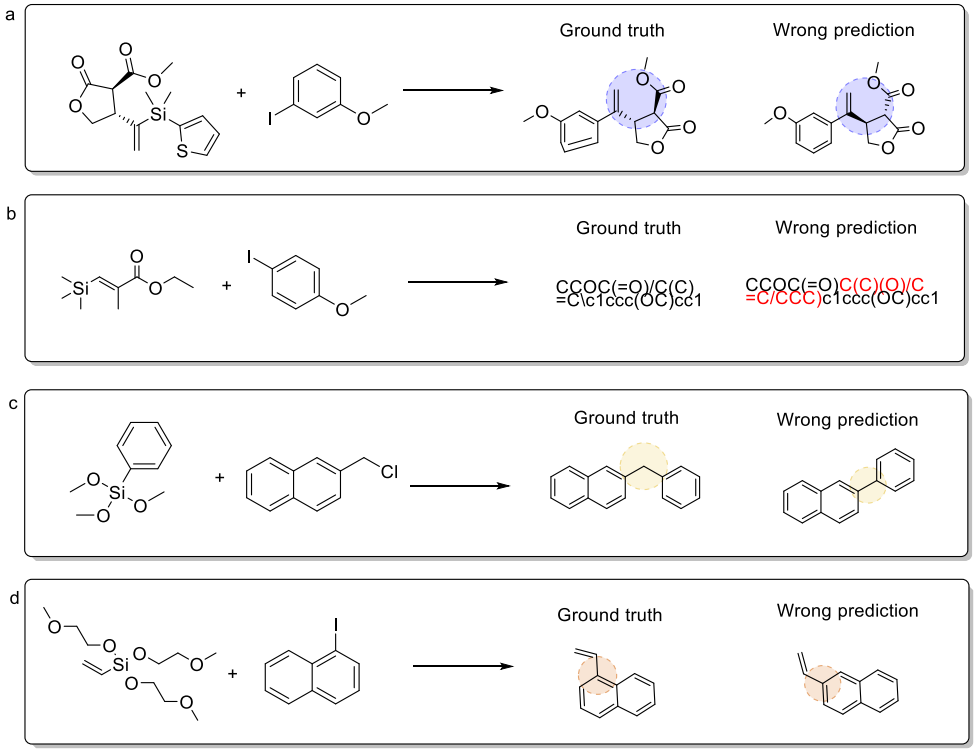


Fig. 5 Typical error analysis of Hiyama coupling reactions. (a) chirality errors (b)SMILES errors (c) the number of atom errors (d) functional group isomerism errors

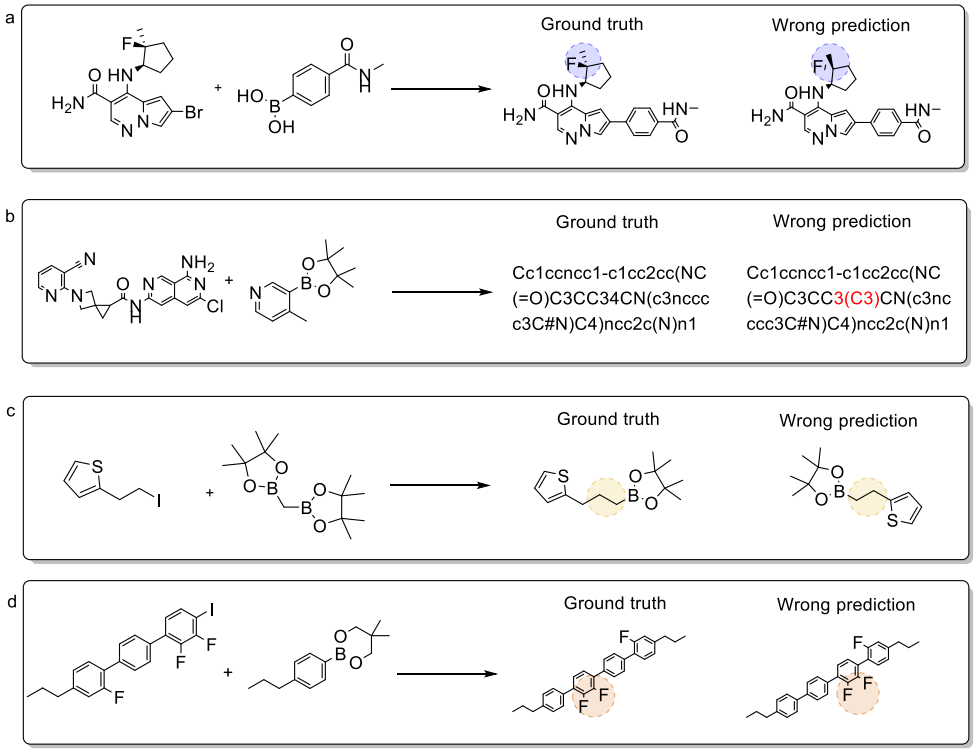


Fig. 6 Typical error analysis of Suzuki coupling reactions. (a) chirality errors (b)SMILES errors (c) the number of atom errors (d) functional group isomerism errors

Conclusion:

This study exhibited that an innovative data augmentation method can improve the performance of transformer model by augmenting data size of training set. Training the model to learn more latent chemical information in organometallic coupling reactions by equivalently replacing groups in reactants corresponding to the raw datasets expanded the training set and increased the predictive performance of models. This concept has been intensively used in image recognition field, while in solving chemical problem has not been reported. [37-39] For the first time, we showed the application of virtual data augmentation in chemical reaction field and found that adding fake data from chemical level boost predictive performance in reaction prediction. We also found that virtual data augmentation method combined with transfer learning strategy can achieve a better accuracy of prediction. Additionally, we used visualization tools to represent the effectiveness of virtual data augmentation method and applied attention weight to visualize the prediction process. The accurate visualization demonstrated that virtual data augmentation is meaningful in chemical level and showed this model became more sensitive to the selection of reaction sites. To sum up, our work shows that the transformer model is suitable for small-scale reaction and what we have done provides a new possibility for data augmentation methods. Also, it is an important step to improve the reaction prediction performance in small data sets. Due to the lack of available institutional data, the development of integrating deep learning methods with chemical filed may be limited. However, the above results all confirm that this virtual data augmentation strategy can contribute to reaction prediction based on small data sets. We believe that this method can be applied to other tasks that with limited data sets by augmenting training dataset.

Availability of data and materials

The dataset (pretraining and self-built) and the code are available from: https://github.com/hongliangduan/Virtual-data-augmentation-methood-for-reaction-prediction-in-small-dataset-scenario

Author contributions

These authors contributed equally: X. W. and Y. Z. and H. D. designed the research project. X. W., Y. Z., J. Y., Y. W., X. W., Z. W. collected literature and established a self-built dataset. H. Q., C. Z. designed and trained the models. X. W. and Y. Z. analyzed data and wrote the manuscript. All authors discussed the results and approved the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This project was supported by the National Natural Science Foundation of China, (No.81903438) and Natural Science Foundation of Zhejiang Province (LD22H300004).

Reference

1. Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature, 2018, 555(7698): 604-610.
2. Liu B, Ramsundar B, Kawthekar P, et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. ACS central science, 2017, 3(10): 1103-1113.
3. Baylon J L, Cilfone N A, Gulcher J R, et al. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. Journal of chemical information and modeling, 2019, 59(2): 673-688.
4. Coley C W, Thomas D A, Lummiss J A M, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science, 2019, 365(6453).
5. Nam J, Kim J. Linking the neural machine translation and the prediction of organic chemistry reactions. arXiv preprint arXiv:1612.09529, 2016.
6. Coley C W, Barzilay R, Jaakkola T S, et al. Prediction of organic reaction outcomes using machine learning. ACS central science, 2017, 3(5): 434-443.
7. Ahneman D T, Estrada J G, Lin S, et al. Predicting reaction performance in C–N cross-coupling using machine learning. Science, 2018, 360(6385): 186-190.
8. Schwaller P, Laino T, Gaudin T, et al Molecular transformer for chemical reaction prediction and uncertainty estimation. 2019.
9. Schwaller P, Gaudin T, Lanyi D, et al. “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chemical science, 2018, 9(28): 6091-6098.
10. Wang L, Zhang C, Bai R, et al. Heck reaction prediction using a transformer model based on a transfer learning strategy. Chemical Communications, 2020, 56(65): 9368-9371.
11. Zhang Y, Wang L, Wang X, et al. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. Organic Chemistry Frontiers, 2021, 8(7): 1415-1423.
12. Dao T, Gu A, Ratner A, et al. A kernel theory of modern data augmentation. International Conference on Machine Learning. PMLR, 2019: 1528-1537.
13. Yang Q, Sresht V, Bolgar P, et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. Chemical Communications, 2019, 55(81): 12152-12155.
14. Moret M, Friedrich L, Grisoni F, et al. Generative molecular design in low data regimes. Nature Machine Intelligence, 2020, 2(3): 171-180.
15. Schwaller P, Vaucher A C, Laino T, et al. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. 2020.
16. Tetko I V, Karpov P, Bruno E, et al. Augmentation Is What You Need!. International Conference on Artificial Neural Networks. Springer, Cham, 2019: 831-835.
17. Bai R, Zhang C, Wang L, et al. Transfer learning: making retrosynthetic predictions based on a small chemical reaction dataset scale to a new level. Molecules, 2020, 25(10): 2357.
18. C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, O. Qi, L. Lai and J. Pei, Transfer learning for drug

discovery,J. Med. Chem., 2020,63, 8683–8694.
19. Pesciullesi G, Schwaller P, Laino T, et al. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. Nature communications, 2020, 11(1): 1-8.
20. Simard P Y, Steinkraus D, Platt J C. Best practices for convolutional neural networks applied to visual document analysis. Icdar. 2003, 3(2003).
21. Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE, 2018: 117-122.
22. Alexey D, Fischer P, Tobias J, et al. Discriminative, unsupervised feature learning with exemplar convolutional, neural networks. IEEE TPAMI, 2016, 38(9): 1734-1747.
23. Tetko I V, Karpov P, Van Deursen R, et al. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nature communications, 2020, 11(1): 1-11.
24. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 1988, 28(1): 31-36.
25. Weininger D, Weininger A, Weininger J L. SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of chemical information and computer sciences, 1989, 29(2): 97-101.
26. Maimaiti M, Liu Y, Luan H, et al. Improving Data Augmentation for Low-Resource NMT Guided by POS-Tagging and Paraphrase Embedding. Transactions on Asian and Low-Resource Language Information Processing, 2021, 20(6): 1-21.
27. Xie Z, Wang S I, Li J, et al. Data noising as smoothing in neural network language models. arXiv preprint arXiv:1703.02573, 2017.
28. Zheng S, Rao J, Zhang Z, et al. Predicting retrosynthetic reactions using self-corrected transformer neural networks. Journal of chemical information and modeling, 2019, 60(1): 47-55.
29. http://www.elsevier.com/online-tools/reaxys.
30. http://www.rdkit.org.
31. Lowe D M. Extraction of chemical structures and reactions from the literature. University of Cambridge, 2012.
32. Jin W, Coley C W, Barzilay R, et al. Predicting organic reaction outcomes with weisfeiler-lehman network. arXiv preprint arXiv:1709.04555, 2017.
33. Zhang C, Cai X, Qiao H, et al. Self-supervised molecular pretraining strategy for reaction prediction in low-resource scenarios. 2021.
34. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.
35. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology, 2019, 37(1): 38-44.
36. Schwaller P, Probst D, Vaucher A C, et al. Mapping the space of chemical reactions using attention-based neural networks. Nature Machine Intelligence, 2021, 3(2): 144-152.
37. Cireşan D C, Meier U, Gambardella L M, et al. Deep, big, simple neural nets for handwritten digit recognition. Neural computation, 2010, 22(12): 3207-3220.
38. Dosovitskiy A, Springenberg J T, Riedmiller M, et al. Discriminative unsupervised feature learning with convolutional neural networks. Advances in neural information processing

systems, 2014, 27.
39. Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE, 2018: 117-122.