Data Augmentation and Transfer Learning Strategies for Reaction Prediction in Low Chemical Data Regimes

Yun Zhang,[a] Ling Wang,[a] Xinqiao Wang,[a] Chengyun Zhang,[a] Jiamin Ge,[a] Jing Tang,[a] An Su,*[b] and Hongliang Duan*[a]

Yun Zhang, Ling Wang and Xinqiao Wang contributed equally to this work.

[a] Y. Zhang, L. Wang, X. Wang, C. Zhang, J Ge, J. Tang and Prof. H. Duan
Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences
Zhejiang University of Technology, Hangzhou 310014, China
E-mail: hduan@zjut.edu.cn

[b] Dr. A. Su
College of Chemical Engineering
Zhejiang University of Technology, Hangzhou 310014, China
E-mail: ansu0912@outlook.com

Abstract: Effective and rapid deep learning method to predict chemical reactions contributes to the research and development of organic chemistry and drug discovery. Despite the outstanding capability of deep learning in retrosynthesis and forward synthesis, predictions based on small chemical datasets generally result in low accuracy due to an insufficiency of reaction examples. Here, we introduce a new state art of method, which integrates transfer learning with transformer model to predict the outcomes of the Baeyer-Villiger reaction which is a representative small dataset reaction. The results demonstrate that introducing transfer learning strategy markedly improves the top-1 accuracy of the transformer-transfer learning model (81.8%) over that of the transformer-baseline model (58.4%). Moreover, we further introduce data augmentation to the input reaction SMILES, which allows for better performance and improves the accuracy of the transformer-transfer learning model (86.7%). In summary, both transfer learning and data augmentation methods significantly improve the predictive performance of transformer model, which are powerful methods used in chemistry field to eliminate the restriction of limited training data.

Introduction

With nearly 200 years history of documented-research, organic synthesis remains occupying the core position in many areas such as drug discovery and organic chemistry. There is a closely related issue in the synthesis of new molecules: reaction prediction. The task of the reaction prediction is to infer the potential products of a given set of reaction components (reactants, reagents and reaction conditions). Driven by improved computing power, data availability and algorithms, the artificial intelligence (AI) technology, which has the potential to simplify and automate reaction prediction, is emerging as a desirable strategy.[1-4]

Currently, the methods of computer-assisted chemical reaction prediction can be roughly divided into three categories. The first method is rule-based expert system with manually encoding or automatically deriving from a chemical reaction database.[5-12] However, this way couldn't be applied to project predictions out of its knowledge base and even be outdated. The second method is using physical chemistry to calculate energies of transition states from feasible reaction routes.[13-15] During the process of calculating energy barrier of a reaction, it requires expensive computational cost. As a result, some experts have been spared no efforts to develop new approaches.

The third method for predicting products is based-on deep learning technique, which seeks to mitigate and eliminate limitations of rule-based and physical chemistry methods.[16-22] For this new method, the key idea is to regard the reaction prediction task as a translation problem, where it aims to map reactant sequences to product sequences. The reactant, reagent and product molecules involved in a reaction are all represented as single line text sequences, such as the Simplified Molecular Input Line Entry System (SMILES).[23-24] Nam and Kim were the first to link the neural machine translation (NMT) model with the chemical reaction predictions, where the sequence-to-sequence (seq2seq) model was trained on chemical reactions and outputted a series of products SMILES.[25] Afterwards, Schwaller *et al.* further applied the seq2seq model to address the forward reaction prediction task.[26] This model can not only be applied to reaction prediction, but also to retrosynthesis. Liu *et al.* made the first steps toward using the seq2seq model in retrosynthetic analysis.[27]

Following the seq2seq model, the transformer model is another NMT model commonly used for chemical reactions, which was established by google company.[28] Experiments by multiple laboratories demonstrated that the transformer model achieved better performance in reaction prediction.[26] Compared to seq2seq model, the transformer model is a newly simple network architecture completely depend on self-attention mechanism without using recurrent and convolutional neural networks, which allows for more parallelization and improves the speed of training. It is precisely because of the unique architecture of the transformer model that it performs better than seq2seq model in processing reaction prediction tasks.

All these deep learning methods learn chemical knowledge from large data sets without human intervention and can be used in numerous real-world applications. However, these technologies are bogged down in the sceneries with sparse availability of labeled data. Transfer learning, an important tool in AI, can be utilized to surmount the restriction of limited amounts of data.[29-32] With transfer learning, the knowledge of solving one task

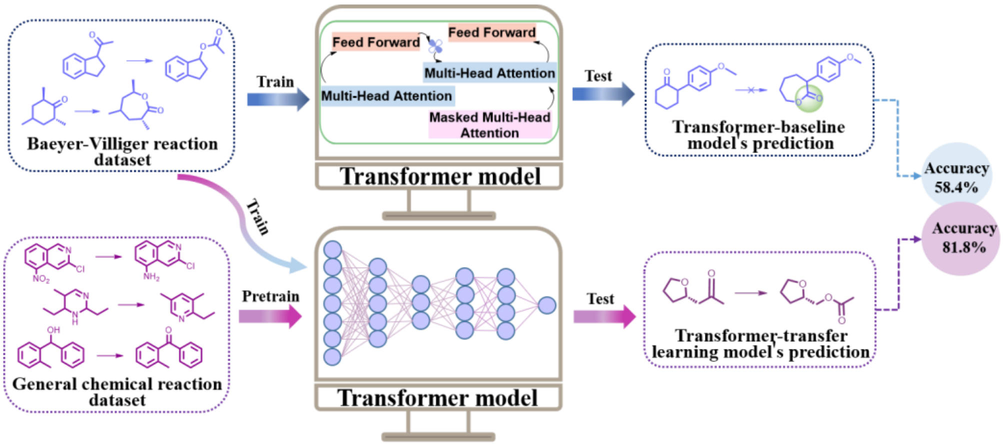

Baeyer-Villiger reaction dataset
Train
Feed Forward
Feed Forward
Multi-Head Attention
Multi-Head Attention
Masked Multi-Head Attention
Transformer model
Test
Transformer-baseline model's prediction
Accuracy 58.4%
Train
General chemical reaction dataset
Pretrain
Transformer model
Test
Transformer-transfer learning model's prediction
Accuracy 81.8%

Figure 1. Schematic diagram of the method for predicting Baeyer-Villiger reaction products. The prediction process involving the transformer-baseline model indicates that the transformer model is only trained and tested on the Baeyer-Villiger reaction. The prediction process involving the transformer-transfer learning model indicates that the Baeyer-Villiger reaction prediction is projected on the transformer model with the introduction of transfer learning.

can be applied to another task. For example, general chemical knowledge from the former large chemical dataset can be applied to the latter relative but different reaction prediction task with limited labeled data. To predict the target task based on limited data, a cluster of pretrained neural networks is collected to regarding its feature labels, which includes large enough dataset to be applied in pretraining. The pretrained model automatically obtains feature labels and stores these labels in the hidden layer. The label obtained by dealing with relevant tasks are transferred to the target task model, provided that these features are correlative. In the previous research of our group, it has been proved that transfer learning method can solve the problem of reaction prediction of Heck reaction with limited dataset.[33]

In this paper, we define our original studies aiming at resolving the challenging conundrum of Baeyer-Villiger reaction prediction with small dataset. We combine transformer architecture with transfer learning methodology to predict products of Baeyer-Villiger reaction using fully data-driven method (Figure 1). To further improve the predictive performance of transformer-transfer learning model on the limited data, the data augmentation is introduced to this experiment. This strategy is originally proposed to alleviate the low-data problem by presenting the same entity with different representations and recent work has shown the successful applications of data augmentation in various neural networks.[34-38] With data augmentation, a chemical reaction can be represented by multiple SMILES strings and the model can obtain more knowledge of a reaction using a batch of random SMILES strings. Despite that the augmented SMILES strings contain same chemical information; the model can absorb more implicit feature of data by constructing a reaction with different SMILES sequences.

As study case we focus on Baeyer-Villiger reaction, for which is a classic chemical reaction in organic chemistry and plays a pivotal role in the synthesis of natural products.[39] Furthermore, the Baeyer-Villiger reaction is high degree of regioselectivity and the detailed mechanism is shown in Figure 2. Also, the Baeyer-Villiger reaction is a typical example of small dataset. If this reaction can be correctly predicted by computer, it will bring great convenience for related synthesis and contribute to the research of catalyst and green procedures of chemical reactions.

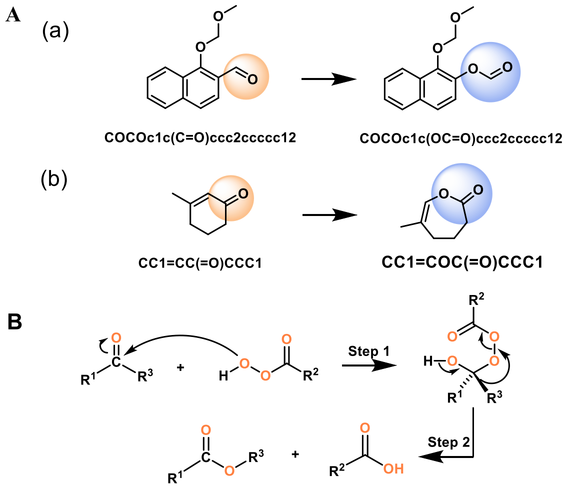


Figure 2. Detailed introduction about Baeyer-Villiger reaction. A. Examples of Baeyer-Villiger reaction in which the aldehyde reactant (a) and ketone reactant (b) are respectively oxidized to esters. B. General mechanism of Baeyer-Villiger reaction.

Results and Discussion

The detailed accuracies of the transformer-baseline, transformer-transfer learning and transformer-transfer learning models with several different level data augmentations are described in Table 1. The top-1 accuracies of the transformer-baseline and transformer-transfer learning models are 58.4% and 81.8% respectively. After the application of transfer learning strategy, the accuracy of the transformer-transfer learning model in reaction prediction shows a significant improvement over that of the

Table 1. Comparison of model's performance of the transformer-baseline, transformer-transfer learning and transformer-transfer learning with different number of data augmentation.

Model	Top-N accuracy (%)			
	Top-1	Top-2	Top-3	Top-5
Transformer-baseline model	58.4	66.7	68.4	71.1
Transformer-transfer learning model	81.8	86.2	89.3	90.7
Transformer-transfer learning model with data augmentation ×1	86.7	92.4	93.8	93.8
Transformer-transfer learning model with data augmentation ×2	84.0	92.4	94.2	94.2
Transformer-transfer learning model with data augmentation ×4	82.7	90.2	93.3	94.2

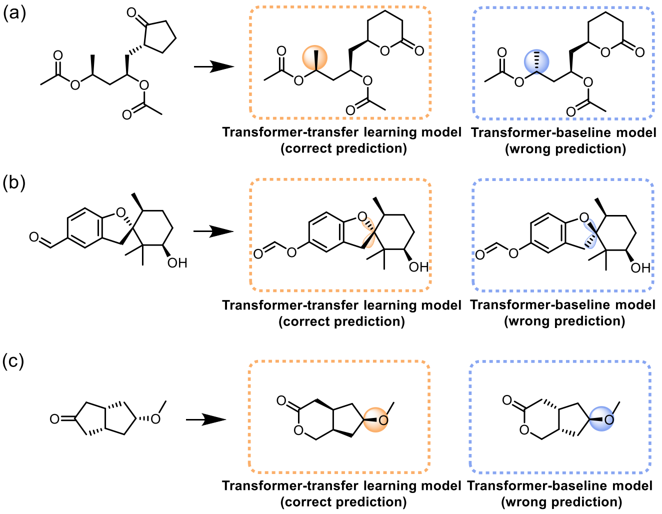

(a)
Transformer-transfer learning model (correct prediction)
Transformer-baseline model (wrong prediction)
(b)
Transformer-transfer learning model (correct prediction)
Transformer-baseline model (wrong prediction)
(c)
Transformer-transfer learning model (correct prediction)
Transformer-baseline model (wrong prediction)

Figure 3. Comparisons and representative examples of the transformer-baseline and transformer-transfer learning models in the prediction of Baeyer-Villiger reaction.

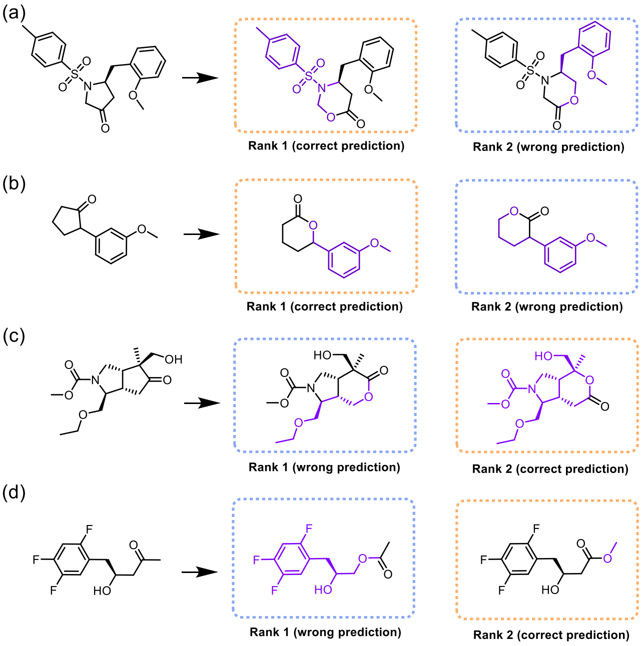

(a)
Rank 1 (correct prediction)
Rank 2 (wrong prediction)
(b)
Rank 1 (correct prediction)
Rank 2 (wrong prediction)
(c)
Rank 1 (wrong prediction)
Rank 2 (correct prediction)
(d)
Rank 1 (wrong prediction)
Rank 2 (correct prediction)

Figure 4. Representative examples of transformer-transfer learning model's top-2 predictions which the predicted results by the model scan stops as soon as the first two predictions are found.

transformer-baseline model. For example, the top-1, top-2, top-3 and top-5 accuracies increase between 19.5% and 23.4%, which demonstrates that both transformer-baseline and transformer-transfer learning models could be applied to reaction predictions, while the pretraining model couldn't achieve any predictive ability on this task in that the top-1 accuracy of the pretraining model is 0%. And several representative examples which are correctly predicted by the transformer-transfer learning model but wrong predicted by the transformer-baseline model are displayed in Figure 3. To some extent, with the introduction of pretraining knowledge, transfer learning can greatly promote transformer-baseline model's performance and can be well used to cope with Baeyer-Villiger reaction prediction.

In addition, a better performance of transformer-transfer learning model is observed because of integrating data augmentation method. The top-1 accuracy of transformer-transfer learning with onefold data augmentation is slighter higher than the transformer-transfer learning model, reaching 86.7%. Also, the top-2, top-3, top-5 accuracies of transformer-transfer learning and transformer-transfer learning model with onefold data augmentation approximately increase between 3.7% to 6.2%. What's more, the different levels of data augmentation make different influence on the performance of transformer-transfer learning model. Note that the trend in accuracies indicates that data augmentation doesn't improve performance of this model consistently with the number of augmented SMILES increasing and onefold augmentation achieves the best performance in the task of predicting target products. With the level of augmented SMILES increasing, the positive effect of this approach weakens gradually in the experiment. For example, when using twofold data augmentation, the top-1 accuracy of the model decreases from 86.7% to 84.0%. And similar situation also appears in larger data augmentation, which reveals a fact that an appropriate level of data augmentation is important for transformer-transfer learning model. Moreover, those results demonstrate that the powerful ability of representing a reaction with multiple SMILES strings and the transformer model indeed obtain additional information about chemistry from augmented training data. Therefore, these findings reveal a fact that an appropriate level of data augmentation makes a positive difference to the transformer-transfer learning model and enable the model to reach a higher accuracy.

Moreover, the model's accuracy improves as "n" increases. In particular, a significant improvement is found between top-1 and top-2 accuracy of transformer model with three strategies, which may be related to the characteristic of Baeyer-Villiger reaction's characteristics: under the condition of a given oxidant, this reaction is formed by the migration of one of the two groups on the ketone to form an ester. According to the possibility of group migration, the model may make predictions to produce two types of esters. And several representative examples of top-2 predictions are shown in Figure 4.

Performance comparison of the transformer-baseline and transformer-transfer learning models with onefold data augmentation

There are several error types continuously accompanying with transformer model to tackle Baeyer-Villiger reaction prediction task: group migration error, chirality error, SMILES error, carbon number errors and other errors. Group migration error is a unique error type caused by Baeyer-Villier reaction. However, with the introduction of transfer learning and data augmentation methods, the transformer-transfer learning model achieves a higher accuracy than transformer-baseline model. And the counts of major predicted errors for transformer-baseline and transformer-transfer learning model with onefold data augmentation are shown in Figure 5. In detail, the transformer-transfer learning model with onefold augmentation makes only 3 mistakes on carbon number error and 5 mistakes on chirality error. Among these errors predicted by this model, the largest reduction observed is SMILES error, which reduces by 15 mistakes compared to the transformer-baseline model. Besides, the ratio of group migration error decreases correspondingly.

Although the number of mistakes described above (Figure 5) has been widely reduced when makes a comparison between transformer-baseline and transformer-transfer learning models, these errors still has a certain impact on the performance of transformer-transfer learning with data augmentation model. As a result, our follow-up work is to conduct a detailed analysis of the major predicted errors which appears in the transformer-transfer learning model based on onefold data augmentation.

Error of group migration

The transformation of ketones into esters and cyclic ketones into lactones or hydroxy acids in Baeyer-Villger reaction undergoes alkyl migration, which is a decisive factor leading to the regioselectivity of this reaction. According to the mechanism of Baeyer-Villiger reaction, the regioselectivity relies on the migratory aptitude of different alkyl groups and this ability is influenced by electron density and steric bulk of groups. Generally, for the reactions of unsymmetrical ketones as reactant, the approximate order of migration is tertiary alkyl > secondary alkyl > aryl > primary alkyl > methyl. However, the transformer-transfer learning model with onefold augmentation can't recognize and distinguish migrating groups' electronic effects and steric properties compared to experienced synthetic scientists. During the predicting process, the transformer-transfer learning model with onefold augmentation mostly makes migration error, which accounting for 50.0 % among the total number of errors. Figure 6 displays several representative examples of comparisons between transformer-transfer learning with onefold augmentation model's top-1 wrong predictions and ground truth. Taking Figure 6 (a) as an example, dimethyl (R)-2-methyl-2-((R)-3-methyl-4-oxopentyl)succinate undergoes group migration and affords dimethyl (R)-2-((R)-3-acetoxybutyl)-2-methylsuccinate with dimethyl (R)-2-butyl-2-methylsuccinate group migrating in theory, while the wrong prediction of transformer-transfer learning model with onefold augmentation is that methyl group migrates. In fact, this error type can be avoided by imputing symmetrical ketones reactant or predicting general chemical reaction.

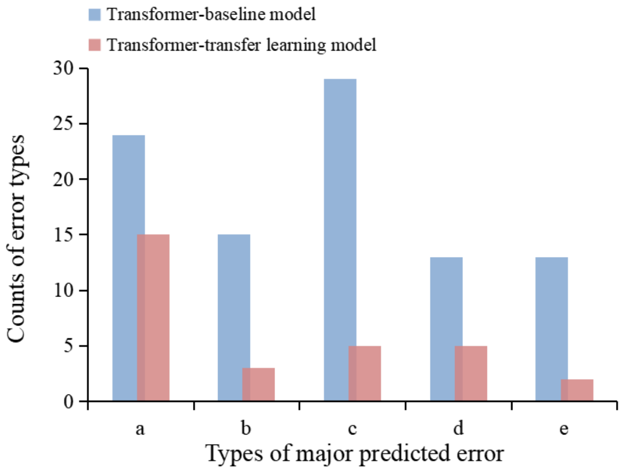


Figure 5. Comparisons of error types between transformer-baseline and transformer-transfer learning model with onefold augmentation. a. Group migration error b. Carbon number error c. SMILES error d. Chirality error e. Other errors.

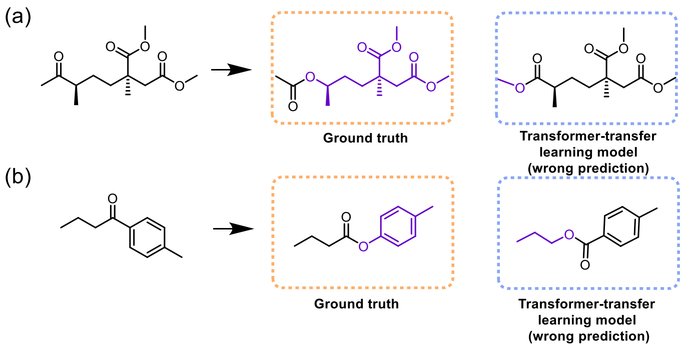

(a)
Ground truth
Transformer-transfer learning model (wrong prediction)
(b)
Ground truth
Transformer-transfer learning model (wrong prediction)

Figure 6. Comparisons between transformer-transfer learning with onefold augmentation model's top-1 wrong predictions and ground truth, which the transformer-transfer learning model with onefold augmentation makes groups migration error in the forward predictions.

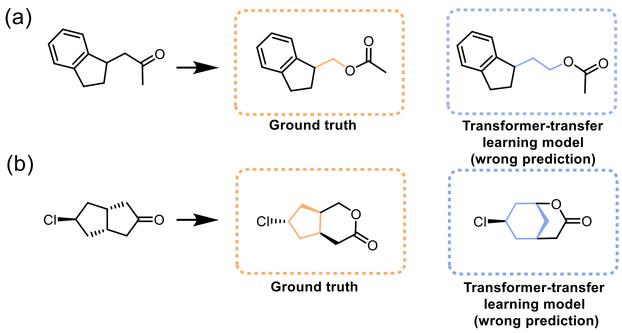
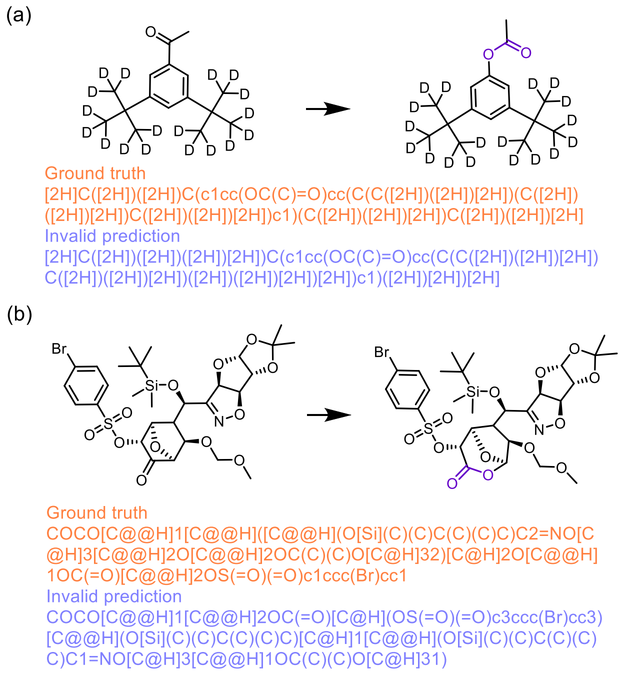

(a)
Ground truth
Transformer-transfer learning model (wrong prediction)
(b)
Cl
Cl
Cl
Ground truth
Transformer-transfer learning model (wrong prediction)

Figure 7. Comparisons between transformer-transfer learning with onefold augmentation model's top-1 wrong predictions and ground truth, which the transformer-transfer learning model with onefold augmentation makes carbon number error in the Baeyer-Villiger reaction predictions.



Ground truth
[2H]C([2H])([2H])C(c1cc(OC(C)=O)cc(C(C([2H])([2H])[2H])(C([2H])([2H])[2H])C([2H])([2H])[2H])c1)(C([2H])([2H])[2H])C([2H])([2H])[2H]
Invalid prediction
[2H]C([2H])([2H])([2H])[2H])C(c1cc(OC(C)=O)cc(C(C([2H])([2H])[2H])C([2H])([2H])[2H])([2H])([2H])[2H])[2H])c1)([2H])[2H])[2H]

Ground truth
COCO[C@@H]1[C@@H]([C@@H](O[Si](C)(C)C(C)(C)C)C2=NO[C@H]3[C@@H]2O[C@@H]2OC(C)(C)O[C@H]32)[C@H]2O[C@@H]1OC(=O)[C@@H]2OS(=O)(=O)c1ccc(Br)cc1
Invalid prediction
COCO[C@@H]1[C@@H]2OC(=O)[C@H](OS(=O)(=O)c3ccc(Br)cc3)[C@@H](O[Si](C)(C)C(C)(C)C)[C@H]1[C@@H](O[Si](C)(C)C(C)(C)C)C1=NO[C@H]3[C@@H]1OC(C)(C)O[C@H]31)

Figure 8. Comparisons and representative examples of transformer-transfer learning with onefold augmentation model's predictions, and the model makes SMILES error in the Baeyer-Villiger reaction predictions.

Error of carbon number

According to original documented literature, this kind of error occurs more when the reactant has large and complex chemical structure. These complex structures are basically composed of multiple carbon atoms.[33] Because the model lacks mathematical knowledge, the methylene in reactant 1-(2,3-dihydro-1H-inden-1-yl)propan-2-one is incorrectly predicted to ethyl group(Figure 7 (a)). Also, in Figure 7 (b), the five-membered ring with chlorine atom in reaction is incorrectly predicted to six-membered ring. Although the product structures predicted by the model are incorrect, the reaction sites are predicted correctly, which further proves that the model has a deeper understanding of the Baeyer-Villiger reaction.

Error of SMILES

Wrong prediction of SMILES is another deficiency of transformer-transfer learning model with onefold augmentation in the prediction of Baeyer-Villiger reaction. The SMILES error in our experiment refers to grammatically invalid SMILES, which couldn't be converted to plausible chemical structure. Since the transformer-transfer learning model with data augmentation couldn't clearly understand the meaning of the chemical entity represented by SMILES, and SMILES is a fragile text representation with a small character variation in the SMILES can lead to molecule chemical structure's transformation or even invalidation. Duan *et al.* had drawn a conclusion that transformer model always incorrectly predicted SMILES of target molecule due to the complexity of compound's structure and scarcity of training dataset.[40] Some representative examples of SMILES error are shown in Figure 8. Although transformer-transfer learning model with onefold augmentation achieves this task, the predicted result is not chemically meaningful. The reason can be attributed to that the benzene ring of 1-(3,5-bis(2-(methyl-d3)propan-2-yl-1,1,1,3,3,3-d6)phenyl)ethan-1-one is substituted by two complex alkyl group, and has few relevant researches in the existing literature about this chemical reaction. Furthermore, the same situation is also observed in Bai's report [25], and no feasible techniques have been proposed in this field up to now.

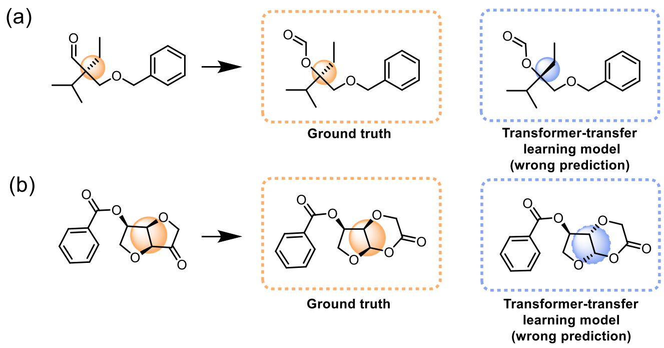


Figure 9. Comparisons between transformer-transfer learning with onefold augmentation model's top-1 wrong predictions and ground truth, which the transformer-transfer learning model with onefold augmentation makes chirality error in the Baeyer-Villiger reaction predictions.

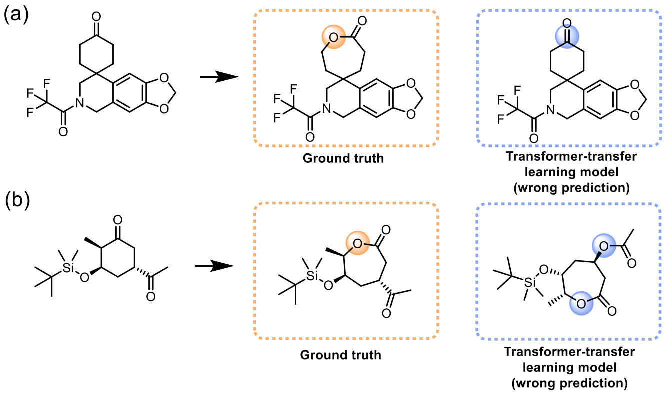


Figure 10. Comparisons between transformer-transfer learning with onefold augmentation model's top-1 wrong predictions and ground truth.

Error of chirality changes

Stereochemistry is one of the most fundamental components of organic chemistry, and the difficulty in learning stereoselectivity is to identify compounds' chirality changes. Yet, the experimental results prove that recognizing configuration of molecule remains an unconquerable obstacle for transformer-transfer learning model with onefold data augmentation. And the reason for this error is that SMILES text undergoes a complex transformation procedure such as canonicalization procedure when chemical reaction is imported or extracted. Thereby the probability of chirality problem is significantly increased. The difference between the predicted result and the ground truth is configuration, but the predicted product still follows Baeyer-Villiger reaction rule (Figure 9). In Figure 9 (a), there's no difficulty to notice the difference in ethyl-bonded carbon atom, which the (*R*)-2-((benzyloxy)methyl)-2-ethyl-3-methylbutanal is in *R* configuration, while the configuration of product predicted by the model is in *S* configuration.

Other errors

Several other mistakes with breaking Baeyer-Villiger reaction's general rules are observed in the outcomes of reaction prediction. We list few representative examples in regard to this error type in Figure 10. Reactants in example (a) is a complex compound with polycyclic hydrocarbons, and in example (b) is a compound with two carbonyl groups. Since transformer-transfer learning model with onefold augmentation is not extremely sensitive to the response of complicated cyclic compounds or multisite reactants, the transformer-transfer learning model with onefold augmentation outputs wrong predictions. This obvious disadvantage of transformer-transfer learning model with onefold data augmentation is also need to be solved urgently.

Predictions analysis of the transformer-baseline, transformer-transfer learning and transformer-transfer learning with onefold augmentation models on Baeyer-Villiger reaction without chirality

An evaluation of the transformer models is carried out through performing the forward reaction prediction of Baeyer-villiger reaction samples without chirality. In this experiment, all reactions' chiral configuration in Baeyer-Villiger reaction dataset is deleted, and are tested on transformer-baseline, transformer-transfer learning and transformer-transfer learning with onefold augmentation models. Table 2 shows the top-1 accuracies of these three models on the test dataset which are removed chirality. Both transformer-baseline, transformer-transfer learning and transformer-transfer learning model with onefold augmentation display better performance on the prediction of reactants without chirality compared to the predictions of reactants with chirality. It is worth mentioning that the top-1 accuracy of the transformer-transfer learning model improves 4.5% with removing chirality, while the transformer-baseline model only improves 2.5%. This result further proves that the introduction of transfer learning strategy could improve the prediction ability of the transformer model, but also reveal a shortcoming, where the transformer model seems to be more effective in addressing the task of achiral reaction prediction.

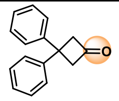
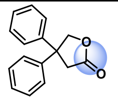
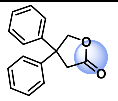
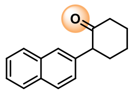
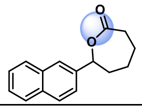
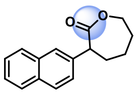
However, the introduction of data augmentation can't further improve the performance of transformer-transfer learning model on the data which doesn't contain the information about chirality. As shown in the Table 2, the top-1 accuracy of the transformer-transfer learning model is the same as the transformer-transfer learning model with onefold augmentation, which indicates that the augmented data without stereochemical information doesn't bring additional knowledge to our model.

To further understand the performance of transfer learning method in the transformer model, achiral reactions are divided into reactions involving symmetrical reactants and reactions involving unsymmetrical reactants. Results show that all of the wrong predictions happen to asymmetric compounds, which proves that transformer model is more effective in the applications of reactions that don't involve regiochemistry. And Table 3 displays representative examples of these two types predictions. In conclusion, the performance of the transformer model is significantly improved with the introduction of the transfer learning approach.

Table 2. Comparison of model's performance of the transformer-baseline, transformer-transfer learning and transformer-transfer learning with onefold data augmentation on Baeyer-Villiger reaction without chirality.

Model	Top-1 accuracy (%)	
	Reaction with chirality	Reaction without chirality
Transformer-baseline model	58.4	60.9
Transformer-transfer learning model	81.8	86.3
Transformer-transfer learning model with data augmentation ×1	86.7	86.3

Table 3. Representative examples of the transformer-transfer learning and transformer-baseline models' top-1 prediction.

Reactant	Transformer-transfer learning model (correct prediction)	Transformer-transfer learning model (wrong prediction)
		
		

Conclusion

In this work, we introduce two innovative methods to predict the outcomes of Baeyer-Villiger reaction that combines transfer learning strategy with the transformer model. We show that the transformer-transfer learning model outperforms the transformer baseline model (58.4%) and achieve an 81.8% top-1 accuracy, which increases approximately 23.4%. This indicates that our approach significantly improves the performance of the transformer model in processing reaction prediction task. Moreover, these novel methods leverage the benefits of transfer learning and data augmentation to capture sufficient chemical knowledge while tacking the deficiency of scarce data. It should be mentioned that the introduction of data augmentation can further improve the accuracy of the transformer-transfer learning model from 81.8% to 86.7%.

In addition, we also perform deeper analysis of error appeared in our experiment, such as SMILES error, chirality error and group migration error, which both appear in transformer-baseline and transformer-transfer learning with onefold data augmentation models. Despite the types of errors appearing in the transformer-transfer learning with onefold data augmentation model is the same as transformer-baseline model, the observed improvement in performance still proves that transfer learning and data augmentation methods are fruitful in tackling the task of chemical reaction prediction. And hoped that these errors can be progressed and changed in subsequent research.

More broadly, this study demonstrates the power of integrated transfer learning and transformer model, in addition to providing a useful tool for chemical reaction prediction of small data. We anticipate this approach could be applied to other similar reactions and combine with other algorithms to further accelerate the process of AI development in reaction predictions.

Experimental Section

Transformer model

The model used in our work is completely based on transformer model. Currently, the transformer is the basic model architecture in the field of natural language processing (NLP). It equips with encoder-decoder architecture that resembles with seq2seq model. In addition, the model entirely depends on self-attention mechanism and adds multi-head attention to allow more parallelization and feed-forward network to improve model's performance.

Performance evaluation

The top-n accuracy plays a key role in evaluating model's performance, and it's entirely justified for the evaluation of reaction prediction. The top-n accuracy represents the ratio of the total number of correct outcomes predicted by the model. In "top-n", the "n" is variable and can be all positive integers. Top-1 means that once the first prediction is found, the prediction results of the model scan will stop. Similarly, top-2 means that once the first and second predictions are found, the prediction results of the model scan will stop.

Data preparation

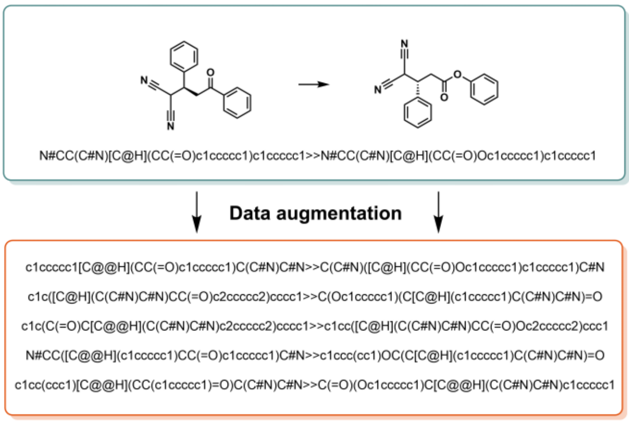
General chemical reaction dataset

The dataset used for transformer model's pretraining process is named as general chemical reaction dataset, which contains approximately 380,000 chemical reactions. These reaction examples were originally sourced from Lowe's data set,[41] which were extracted from United States Patent and Trademark Office (USPTO) patents, and then subjected to a collection of pretreatments in which all the reagents and conditions were deleted. We further filtered out duplicate, incorrect and incomplete reactions. It's worth noting that all of Baeyer-Villiger reactions are not in this dataset.

Baeyer-Villiger reaction dataset

The small dataset we used in this paper is Bayer-Villiger reaction. First, Baeyer-Villiger reactions are extracted from the "Reaxys" database based on reaction template and the name of the target reaction. Second, we further process the "raw" Baeyer-Villiger reactions by deleting reaction samples which are empty, repeated and erroneous. Third, all the reagents are removed in order to render the reaction examples only contain reactants and products. Following the reagents clean, the reaction samples need to be canonicalized, which it allows the efficient representation of molecular structure. Finally, 2225 Baeyer-Vlliger reactions are obtained in total, and the dataset is split into training, validation and test datasets (8:1:1).

Figure 11. An example of fivefold data augmentation. All SMILES strings represent the same reaction.


N
O
N
N
O
O
N#CC(C#N)[C@H](CC(=O)c1ccccc1)c1ccccc1>>N#CC(C#N)[C@H](CC(=O)Oc1ccccc1)c1ccccc1
Data augmentation
c1ccccc1[C@@H](CC(=O)c1ccccc1)C(C#N)C#N>>C(C#N)([C@H](CC(=O)Oc1ccccc1)c1ccccc1)C#N
c1c([C@H](C(C#N)C#N)CC(=O)c2ccccc2)cccc1>>C(Oc1ccccc1)(C[C@H](c1ccccc1)C(C#N)C#N)=O
c1c(C(=O)C[C@@H](C(C#N)C#N)c2ccccc2)cccc1>>c1cc([C@H](C(C#N)C#N)CC(=O)Oc2ccccc2)ccc1
N#CC([C@@H](c1ccccc1)CC(=O)c1ccccc1)C#N>>c1ccc(cc1)OC(C[C@H](c1ccccc1)C(C#N)C#N)=O
c1cc(ccc1)[C@@H](CC(c1ccccc1)=O)C(C#N)C#N>>C(=O)(Oc1ccccc1)C[C@@H](C(C#N)C#N)c1ccccc1

SMILES augmentation

Data augmentation is a technique for increasing the volume of data by the means of adding copies of existing data or newly created data from existing data (Figure 11). The chemical reactions data used in our study are represented in SMILES form. We only perform data augmentation on the training dataset of Baeyer-Villiger reaction dataset. The augmentation of the reaction SMILES is done with a Python script (version 3.7) utilizing the RDKit (version 2019.03).

Baeyer-Villiger reaction prediction

In transfer learning procedure, the transformer model is first pretrained on the general chemical reaction dataset. The pretrained model is exerted as an initialization or feature extractor to finish similar task. The basic chemical information and characteristics are transferred to address the target task of predicting the products of Baeyer-Villiger reaction by pretraining process. Then, the model is trained on Baeyer-Villiger reaction, which is called fine-tuning step. In the process of transformer-baseline model's prediction, the transformer model is only trained on the Baeyer-Villiger reaction dataset. It is worth noting that the transformer-transfer learning model with data augmentation mentioned in our article is pretrained on general chemical reaction dataset and trained on Baeyer-Villiger reaction with data augmentation.

Acknowledgements

This project was supported by the National Natural Science Foundation of China, NSFC (Grant No.81903438).

Keywords: artificial intelligence• transformer• transfer learning• data augmentation• reaction prediction• Baeyer-Villiger reaction

[1] W. Beker, E. P. Gajewska, T. Badowski, B. A. Grzybowski, *Angew.Chem. Int. Ed.* **2019**, 58, 4515-4519.

[2] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, 3, 434-443.

[3] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, 2, 725-732.

[4] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, T. Laino, *Chem. Sci.* **2018**, 9, 6091-6098.

[5] W.L. Jorgensen, E. R. Laird, A. J. Gushurst, J. M. Fleischer, S. A. Gothe, H. E. Helson, G. D. Paderes, S. Sinclair, *Pure &Appl. Chem.* **1990**, 62, 1921-1932.

[6] E. J. Corey, W. T. Wipke, R. D. Cramer III, W. J. Howe, *Science* **1969**, 166, 178-192.

[7] D. A. Pensak, E. J. Corey, *J. Am. Chem. Soc.* **1977**, 61, 1-32.

[8] H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1995**, 35, 34-44.

[9] J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model.* **2009**, 49, 593-602.

[10] M. H. S. Segler, M. P. Waller, *Chem. Eur. J.* **2017**, 23, 6118–6128.

[11] V. H. Nair, P. Schwaller, T. Laino, *Artificial Intelligence in Swiss Chemical Research.* **2019**, 73, 997-1000.

[12] E. J. Corey, W. T. Wipke, R. D. Cramer III, and W. J. Howe. J. Am. *Chem. Soc.* **1972**, 94, 431-439.

[13] L.-P. Wang, R. T. McGibbon, V. S. Pande, T. J. Martinez, *J. Chem. Theory Comput.* **2016**, 12, 638-649.

[14] O. Engkvist, P. O. Norrby, N. Selmi, Y. hong Lam, Z. Peng, E.C. Sherer, W. Amberg, T. Erhard, L.A. Smyth, *Drug Discov. Today.* **2018**, 23, 1203-1218.

[15] W. R. Dolbier, Jr., H. Korniak, K. N. Houk, C. Sheu, *Acc. Chem. Res.* **1996**, 29, 471-477.

[16] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, 555, 604-610.

[17] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, 10, 370-377.

[18] M. H. S. Seglerand, M. P. Waller, *Chem. Eur. J.* **2017**, 23, 5966-5971.

[19] A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeode, C. R. Butler, *Chem. Commun.* **2019**, 55, 12152-12155.

[20] P.Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C.Bekas, A. Iuliano , T.Laino, *Chem. Sci.* **2020**, 11, 3316-3325.

[21] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, E. Ozkirimli, *Drug Discov. Today.* **2020**, 25, 4 689-705.

[22] P. Schwaller, T. Gaudin, D. Lányi, C. Bekasa, T. Laino, *Chem. Sci.* **2018**, 9, 6091-6098.

[23] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, 28, 1, 31-36.

[24] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Model.* **1989**, 29, 97−101.

[25] J. Nam, J. Kim, **2016**. Available online: https://arxiv.org/abs/1612.09529

[26] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A.A. Lee, *ACS Cent. Sci.* **2019**, 5, 1572-1583.

[27] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci.* **2017**, 3, 1103-1113.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, arXiv:1706.03762.

[29] G. Pesciullesi, P. Schwaller, T. Laino, J.-L. Reymond, *Nat. Commun.* **2020**, 11,4874.

[30] C. Cai, S. Wang, Y. Xu, W. Zhang, K. Tang, O. Qi, L. Lai, J. Pei, *J. Med. Chem.* **2020**, 63, 8683-8694.

[31] S. J. Pan, Q. Yang, Fellow, IEEE, *Trans. Knowl. Data Eng.* **2010**, 22, 1345-1359.

[32] R. Bai, C. Zhang, L. Wang, C. Yao, J. Ge, H. Duan, *Molecules* **2020**, 25, 2357.

[33] L. Wang, C. Zhang, R. Bai, J. Li, L. H. Duan, *Chem. Commun.* **2020**, 56, 9368-9371.

[34] E. J. Bjerrum, **2017** arXiv:1703.07076v2.

[35] T. Dao, A. Gu, A. J. Ratner, V. Smith, C.D. Sa,.C. Ré, *Proc. Mach. Lern. Res.* **2019**, 97, 1528–1537.

[36] M. E. Fortunato, C. W. Coley, B. C. Barnes, K. F. Jensen, *J. Chem. Inf. Model.* **2020**, 60, 3398–3407.

[37] I. V. Tetko, P. Karpov, R. V. Deursen, G. Godin, *Nat commun.* **2020**, 11, 5575.

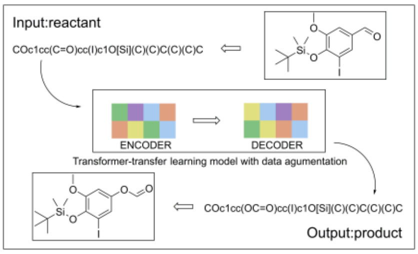
[38] M. Moret, L. Friedrich, F Grisoni, D. Merk, G. Schneider, *Nat. Mach. Intell.* **2020**, 2, 171-180.

[39] G. -J. ten Brink, I. W. C. E. Arends, and R. A. Sheldon, *Chem. Rev.* **2004**, 104. 4105−4123.

[40] H. Duan, L. Wang, C. Zhang, L. Guo, J. J. Li, *RSC Adv.* **2020**, 10, 1371-1378.

[41] D. M. Lowe, Extraction of Chemical Structures and Reactions from the Literature; University of Cambridge: **2012**.

Entry for the Table of Contents


Input:reactant
COc1cc(C=O)cc(I)c1O[Si](C)(C)C(C)(C)C
ENCODER
DECODER
Transformer-transfer learning model with data agumentation
COc1cc(OC=O)cc(I)c1O[Si](C)(C)C(C)(C)C
Output:product

A proof-of-concept methodology for predicting Baeyer-Villiger reaction using transfer learning and data augmentation is presented. Using transformer-transfer learning with data augmentation, the top-1 accuracy achieves 86.7% over that of the transformer-baseline model (58.4%%), which reveals the fact that transfer learning and data augmentation make a difference to the transformer model.