

Transformer: Linking Atom Mapping and Neural Machine Translation

Chengyun Zhang,^[a] Ling Wang,^[a] Yejian Wu,^[a] Yun Zhang,^[a] An Su,^{*[b]} and Hongliang Duan^{*[a]}

Chengyun Zhang and Ling Wang contributed equally to this work.

[a] C. Zhang, L. Wang, Y. Wu, Y. Zhang and Prof. H. Duan
Artificial Intelligent Aided Drug Discovery Lab, 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

Supporting information for this article is given via Supporting information.pdf.

Abstract: Atom mapping reveals the corresponding relationship between reactant and product atoms in chemical reactions, which is important for drug design, exploration for underlying chemical mechanism, reaction classification and so on. Here, we present a new method that links atom mapping and neural machine translation using the transformer model. In contrast to the previous algorithms, our method runs reaction prediction and captures the information of corresponding atoms in parallel. Meanwhile, we use a set of approximately 360K reactions without atom mapping information for obtaining general chemical knowledge and transfer it to atom mapping task on another dataset which contains 50K atom-mapped reactions. With manual evaluation, the top-1 accuracy of the transformer model in atom mapping reaches 91.4%. We hope our work can provide an important step toward solving the challenge problem of atom mapping in a linguistic perspective.

Introduction

Atom mapping (AM) numbers the atoms across a chemical reaction to indicate the one-to-one correspondence between an atom in a reactant and an atom in a product molecule.^[1] In chemical reactions, AM can be used to describe the arrangement and distribution of the atoms. Due to this labelling nature, AM can be applied to a variety of areas.^[2-9] First, AM is an important part of automatically extracting reaction cores from big reaction databases.^[2-5] In addition, AM can be applied to assign the reaction rules from given reactions to specific molecules.^[6] AM is also helpful in describing the mechanisms of enzymatic-catalysed reactions and identifying the feasibility of computationally derived metabolic pathways.^[7-9]

Despite the importance of AM in the understanding of the chemical composition of reactions, the AM information for known reactions in most chemical databases are insufficient. Considering that matching corresponding atoms for a reaction requires deep knowledge of chemistry while manual curation of AM is time-consuming, it is desirable to develop computational methods that can automatically predict the AM.

There have been increasing efforts to predict the AM by graph theory-based algorithms.^[10] In graph theory, a chemical molecule is represented as a graph in which atoms and bonds are represented as nodes and edges, respectively. Within this

context, AM becomes a graph matching procedure with NP-hard^[11] (detailed information about related work based on graph theory is available in the Section S1 of Supporting Information). Currently, common substructure-based method and optimization-based method are the two widely-used graph theory-based methods.^[7,12-19] Figure S1 shows the schematic of a typical AM approach that combines the characteristics of the common substructure and optimization. The algorithm captures the features of atoms and bonds in a reaction and determines what atoms are in the common structure between substrates and products. After that, the remaining atoms are labelled with numbers using optimization-based method.

Molecules can be represented as text sequences in addition to graphs. One of the most common text representations of molecules is simplified molecular-input line-entry system (SMILES).^[20] Similarly, molecules with AM information can be transformed into SMILES. From a linguistic perspective, SMILES can be treated as a kind of language. In this way, AM becomes a translation process which takes reactants with mapping numbers as inputs and outputs the mostly likely products with corresponding AM information.

Recently, multiple efforts have been made to apply neural machine translation (NMT) models to chemical reaction prediction.^[21-26] One of the popular NMT models is the transformer, an entirely attention-based NMT architecture eschewing recurrence.^[27] In our previous work, the transformer model has shown its good performance in the field of forward reaction and retrosynthetic reaction predictions.^[24-26] Compared to other NMT models, the main architectural characteristic of the transformer model is that it has completely eliminated any recurrences or convolution and solely relies on the attention mechanism to compute the representations of its inputs and outputs, which allows the transformer model to achieve better performance for chemical reaction prediction.

Meanwhile, introducing transfer learning, a method that takes neural network developed for one task and reused it to a related but different task, to the transformer can significantly improve the performance of the transformer model.^[28] Using transfer learning, the knowledge that solves one problem can be applied to another problem. For example, general chemical knowledge, such as the chirality of compounds, from a large chemical dataset can be applied to a related but different reaction prediction task with comparatively small chemical dataset.

In this paper, we describe our work aiming at solving the challenging AM computation task. We link the AM with NMT using a transfer learning-based transformer model. It is worth noting that our model can conduct the reaction prediction and the AM simultaneously. In other words, for a given reactant with labelling numbers, the transformer model predicts the most likely products with the corresponding labelled numbers, which presents an additional challenge compared to other atom matching predictions. Interestingly, at the end of our work in AM task, Schwaller *et al.*^[29] prove the fact that atom mapping can be learned on the self-supervised task in a reactions sequence. We note that our work is not simply an AM application. In fact, it not only maps the corresponding atoms across a reaction, but it also predicts the most likely product when given reactants. Other atom-matching study, inputs complete reactions (including reactants and products) and output reactions in which the atoms in the reactants and products are linked to a corresponding number in the prediction process. However, our work links the atom-matching work and neural machine translation work. This prediction requires not only the correct product structure, but also the correct atom numbers, and only if these two conditions are met can a prediction be judged to be correct.

Results and Discussion

Table 1 shows the performances of the pretrained (pretrained on the USPTO-transfer learning-360K dataset and tested on the USPTO-AM-50K dataset), transformer-baseline (trained and tested on the USPTO-AM-50K dataset) and transformer-transfer learning (pretrained and trained on the USPTO-transfer learning-360K and USPTO-AM-50K datasets respectively, and tested on the USPTO-AM-50K dataset) models in the five experiments. The average top-1 accuracy of the transformer-baseline model is 87.5% and the average top-1 accuracy of transformer-transfer learning model is 90.4%. Both the transformer-baseline and transformer-transfer learning models achieve top-1 accuracy higher than 80%, which proves that transformer model can be applied to AM tasks. The higher accuracy of the transformer-transfer learning model (>90%) shows that transfer learning can provide strong boosts over transformer-baseline models and successfully applied to chemical reaction prediction tasks. In contrast to other previous works, our task innovatively regards the AM as a language translation task and applies the transformer model to this task. Crucially, the results are significantly improved by introducing the transfer learning strategy.

On the other hand, the five top-1 accuracies of the pre-training model are all 0%. In other words, none of the five experiments achieve any predictive power on this task, providing evidence that the pretraining models have difficulty directly applying the general chemical knowledge obtained from the pretraining process to predict reactions. However, the transformer-transfer learning model performs well on this task. That said, the pretrained model needs further training on the USPTO-AM-50K dataset to process the chemical information from the USPTO-transfer learning-360K dataset so that it can handle the AM task in USPTO-AM-50K dataset. These findings highlight the importance of the complete transfer learning procedure in our work, which provides constructive insight into the application of the transfer learning method.

In the following of this section, we chose experiments 1 to present the evaluation (detailed information about top-n accuracies of transformer-transfer learning model is available in Table S1) of our method in AM task.

Comparison between the transformer-baseline and transformer-transfer learning models

The transformer-transfer learning model outperforms the transformer-baseline model on the processed patent data set classified into 10 classes. The detailed top-1 results for the transformer-baseline and transformer-transfer learning models broken down by the reaction classes are shown in the Figure 1. The transformer-transfer learning model performs significantly better in reaction class 1 (heteroatom alkylation and arylation), class 2 (acylation and related processes) and class 3 (C-C bond formation) (reaction examples that transformer-baseline model predicts incorrectly but transformer-transfer learning model predicts correctly are available in Section S3 of Supporting Information).

The difference in the accuracy between the transformer-baseline and transformer-transfer learning models is mainly because the transformer-baseline model displays limited understanding of chirality structure and SMILES, but the transformer-transfer learning model performs well. As illustrated in Table 2, when the reactants contained one or more chirality center structures, the transformer-baseline predict the raw structure but it cannot understand the stereo configuration (e.g. R or S), which compromises the quality of the products predicted by the model. For example, in the prediction of reaction class 1 in Table 2,

Table 1. The comparison results of the transformer models with different training steps in 5 different subsets of the USPTO-AM-50K dataset.

Entry	Top-1 accuracy (%)		
	Pretrained model	Transformer-baseline model	Transformer-transfer learning model
1	0	87.1%	90.4%
2	0	88.7%	91.0%
3	0	86.9%	90.0%
4	0	86.4%	90.2%
5	0	88.4%	90.4%
average	0	87.5%	90.4%

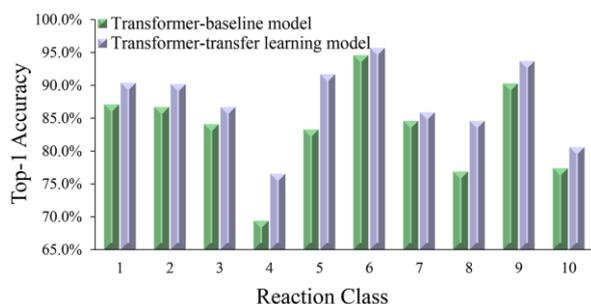


Figure 1. Comparison of the top-1 accuracy of the transformer-baseline and transformer-transfer learning models by reaction classes.

the carbon atom in position 16 is originally in the S configuration. However, the transformer-baseline model incorrectly projects the compound with the R configuration. For reaction classes 2 and 3, the transformer-baseline model also makes the same mistake. Chirality is a crucial property of asymmetry in several fields of chemistry, since it is conducive to understanding the theoretical and physical drives behind the formation and structures of a large number of chemical compounds. Chirality was the cause of the thalidomide disaster in the 1960s. Therefore, it's important to predict the chirality structure correctly.

On the contrary, the transformer-transfer learning model is able to identify the stereo configuration. In all cases shown in Table 2, the predictions by the transformer-transfer learning model for the target molecule are atom-mapped and can be linked to the correct stereo configurations. Consistent with our previous research finding, the common deficiency of the transformer model in reaction prediction is misunderstanding of the

chirality. Compared to the transformer-baseline model, the transformer-transfer learning model incorporates more chemical chirality knowledge since the model learns from the pretraining process. Additionally, both the transformer-baseline and the transformer-transfer learning models can correctly project the AM tasks and the atom-mapped compounds in the predicted molecules can be correctly linked to corresponding reactant atoms.

In addition to recognizing the chirality of compounds, the transfer learning pushes forward an immense influence in recognizing the underlying grammar of SMILES. As displayed in the Table 3, the predictions of reaction classes 1, 2 and 3 are represented by invalid SMILES. Consistent with our previous research results, the model is prone to produce invalid SMILES if there are complex ring structures in the compounds. This is a typical error in which the transformer-baseline model cannot produce grammatically valid SMILES. The practical effect of the transformer-baseline model has been limited since the grammar of SMILES is extremely sophisticated. However, the capability of obtaining the inner core SMILES of the transformer-transfer learning model is underpinned, thus decreasing the unfavorable effect of SMILES translation for the prediction results and improving the accuracy. For all the examples in Table 3, the transformer-transfer learning model not only successfully predicted the correct products, but also correctly predicted the mapping of the atoms.

Error analysis of the transformer-transfer learning model

To further improve the performance of transformer-transfer learning model, we perform an analysis on the incorrect predictions. We divide the errors in the procedure of reaction pre-

Table 2. Comparisons and representative examples of the transformer-baseline and transformer-transfer learning models in the prediction of class 1, 2 and 3 reactions with chiral carbon atoms.

Class	Reactants	Synthetic analysis	
		Transformer-transfer learning model (correct prediction)	Transformer-baseline model (wrong prediction)
1			
2			
3			

Table 3 Comparisons and representative examples of the transformer-baseline and transformer-transfer learning models in the prediction of class 1, 2 and 3 reactions with complex ring structures.

Synthetic analysis			
Class	Reactants	Transformer-transfer learning model (correct prediction)	Transformer-baseline model (wrong prediction)
1			SMILES error
2			SMILES error
3			SMILES error

dition into two categories: the results that are not consistent with the ground truth, and the predictions that match the ground truth in the form of the chemical structure but do not match SMILES representation of the ground truth.

An example of the former error type is shown in Figure 2. The alkylation reaction between 4-amino-2-chlorophenol and 4-chloro-5-fluoroquinazoline is possible with two different groups at the reaction site. Because the lone pair on nitrogen is higher in energy than that on oxygen, the amino group is more reactive towards nucleophilic attack than the hydroxyl group. In this context, the ground truth product is 2-chloro-4-((5-fluoroquinazolin-4-yl) amino) phenol rather than 3-chloro-4-((5-fluoroquinazolin-4-yl) oxy) aniline. It is worth mentioning that the AM information of the corresponding atoms between reactants

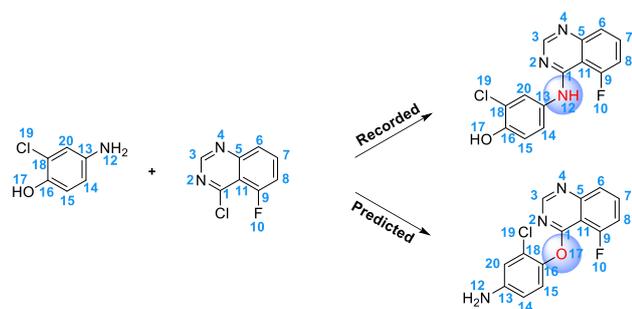
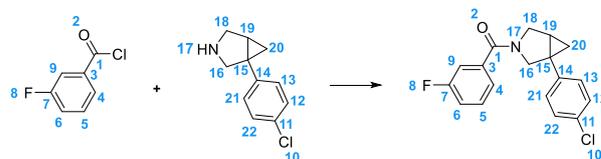


Figure 2. A representative example of the reaction where the prediction of the product is not consistent with the ground truth in the SMILES representation or in the graph representation. The top row is the targeting product and the bottom row is the result from the transformer-transfer learning model.

and products are correctly given by the transformer-transfer learning model, which shows remarkable performance in capturing of the AM information in the form of SMILES.

Figure 3 shows an example of the case where the product SMILES predicted by the transformer-transfer learning model is not consistent with the target SMILES, but the predicted SMILES can be converted to the same chemical structure as the ground



Ground truth:

```
Cl[C:1](=[O:2])[c:3]1[cH:4][cH:5][cH:6][c:7]([F:8])[cH:9]1.[Cl:10][c:11]1[cH:12][cH:13][c:14]([C:15]23[CH2:16][NH:17][CH2:18][CH:19]2[C H2:20]3)[cH:21][cH:22]1>>[C:1](=[O:2])([c:3]1[cH:4][cH:5][cH:6][c:7]([F:8])[cH:9]1)[N:17]1[CH2:16][C:15]2([C:14]3[cH:13][cH:12][c:11]([Cl:10])[cH:22][cH:21]3)[CH:19]([CH2:18]1)[CH2:20]2
```

Prediction:

```
Cl[C:1](=[O:2])[c:3]1[cH:4][cH:5][cH:6][c:7]([F:8])[cH:9]1.[Cl:10][c:11]1[cH:12][cH:13][c:14]([C:15]23[CH2:16][NH:17][CH2:18][CH:19]2[C H2:20]3)[cH:21][cH:22]1>>[C:1](=[O:2])([c:3]1[cH:4][cH:5][cH:6][c:7]([F:8])[cH:9]1)[N:17]1[CH2:16][C:15]2([c:14]3[cH:13][cH:12][c:11]([Cl:10])[cH:22][cH:21]3)[CH2:20]([CH:19]2)[CH2:18]1
```

Figure 3. A representative reaction example expressed by different SMILES, and the transformer-transfer learning model predicted chemical structure is as same as that of the ground truth.

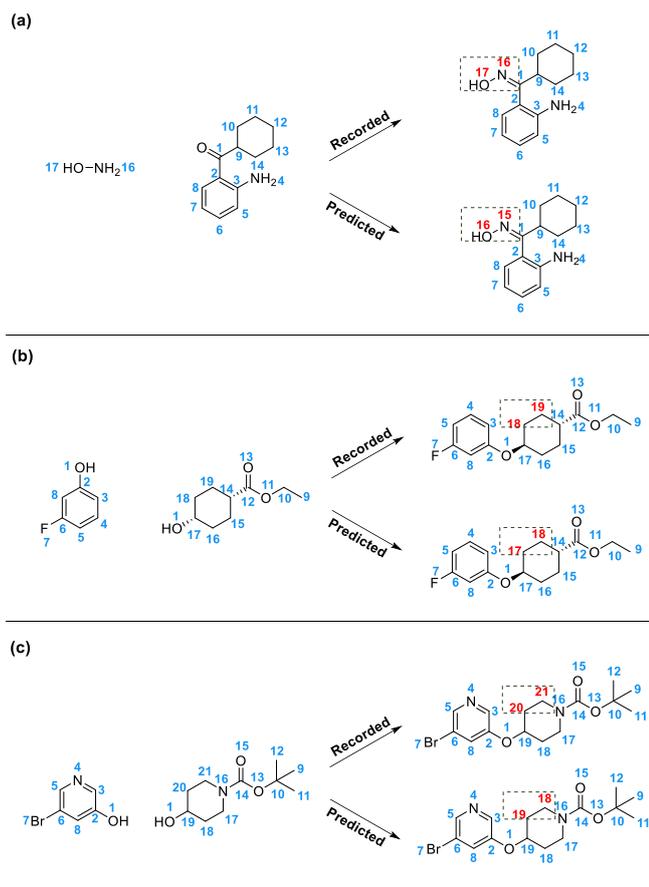


Figure 4. Example reactions that do not match the truth due to the failure of capturing the corresponding relationship between reactant and product atoms.

truth structure. It is an acylation reaction of two reactant molecules, 3-fluorobenzoyl chloride and 1-(4-chlorophenyl)-3-azabicyclo [3.1.0] hexane, and the product of this reaction is 1-(4-chlorophenyl)-3-azabicyclo [3.1.0] hexan-3-yl) (3-fluorophenyl) methanone.

As depicted in Figure 3, there is a subtle difference between the predicted SMILES and the ground truth SMILES. Due to the non-uniqueness of SMILES, a compound may be represented by several nonstandard SMILES. The (1-(4-chlorophenyl)-3-azabicyclo [3.1.0] hexan-3-yl) (3-fluorophenyl) methanone can be represented by the predicted and ground truth SMILES, and the two SMILES correspond to the identical chemical structure. Therefore, the prediction by the transformer-transfer learning model is chemically correct. Due to the accuracy metric's limitation, the predicted SMILES is judged to be wrong.

Incorrect atom numbers predictions are generally caused by complex compound structures. Several examples are shown in Figure 4. We notice that the incorrect mapping usually happens in the ring structure such as cyclohexane and piperidine. In Figure 4(b), ethyl (1r,4r)-4-(3-fluorophenoxy) cyclohexane-1-carboxylate has two different atoms equipped with same AM numbers. Similarly, tert-butyl 4-((5-bromopyridin-3-yl) oxy) piperidine-1-carboxylate has incorrect mapping between reactants and products. Due to the structural complexity in the molecules across a reaction, labelling atoms between reactants and products with numbers become a serious challenge.

However, there is a special AM error in the atom numbers prediction error. Figure 5 illustrates several wrong AM prediction examples involving a symmetric structure. The molecular symmetry factors mainly include symmetric plane, symmetric center and symmetric axle. Example (a) is a representative reaction in which the reactant involves a symmetric plane (Figure 5). N, N'-(5-chloro-1,3-phenylene) diformamide is a symmetric molecule, and the groups at the positions 9 and 14 on the benzene ring are equivalent, therefore, they have an equivalent effect when the reaction occurs.

Therefore, the answer predicted by the transformer-transfer learning model is that the reaction at the position 9 of the benzene ring is the same as the correct answer on position 14 of the benzene ring. However, in a superficial sense, the reaction sites for the predicted answer are different from those of the correct answer, which leads to a different atom number prediction, but it is actually chemically feasible. Example (b) is a typical etherification reaction that involves molecules with a symmetric center (Figure 5). Because of the symmetry of naphthalene-2,6-diol, the two hydroxyl groups in the molecules are chemically indistinguishable during reactions. A further example

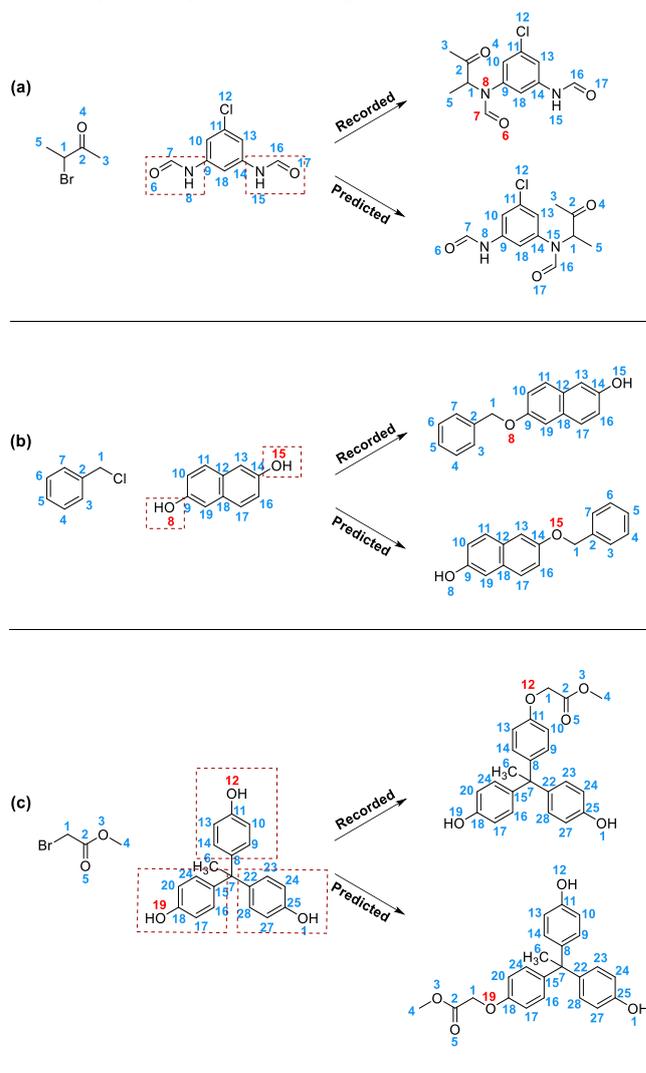


Figure 5. Example reactions that do not match the ground truth due to the symmetry factors in the molecules. Example (a) is a reaction containing molecules with a symmetric plane; example (b) is a reaction consisting of molecules with a symmetric center; and example (c) is a reaction involving molecules with a symmetric axle.

reaction is described in the Figure 5(c), which demonstrates the influence of a symmetric axis in a molecule. 4,4',4''-(ethane-1,1,1-triyl) triphenol contains three phenol groups and the reactivity of those groups are equal in the reaction. With the characteristics of molecular symmetry, multiple atom mappings are possible in the case of that reactions corresponded to symmetry factors. In our study, the predicted mapping and the reference mapping are compared by comparing the reaction SMILES. Therefore, a plausible mapping given by the transformer-transfer learning model may be regarded as a wrong result as a consequence of the limited mapping in the recorded dataset.

Evaluation of predictions by human expert chemists

There are a number of chemically plausible predictions made by the transformer-transfer learning model that are considered to be wrong answers by the algorithm, which results in lower accuracy. In addition to algorithmically checking the consistency between the predictions of the transformer-transfer learning model and the recorded predictions, we also invited expert chemists to evaluate the predicted results. To quantitatively appraise the results, the chemists identified the wrong answers of experiment 1. There were 67 chemically plausible predictions with correct corresponding atom number predictions, which account for 1.0% of the total test dataset (the detailed information about reactions that human expert chemists consider to be chemically plausible are in Section S4 of Supporting Information). As a result, the actual accuracy of the transformer-transfer learning model in experiment 1 can reach 91.4%.

Conclusion

In our work, we innovatively apply the transformer model to address the problem of capturing the corresponding relationship between reactants and products. In contrast to prior work which depend on the graph theory, our work considers AM as a translation task so the AM is not limited in the principle of MCS or MCD. By running the reaction prediction and atom matching simultaneously, the transformer model can not only show the relationship between the substrate and product molecules, but it also gives the top-n candidates to explore the chemical reactions with AM information. Furthermore, the introduction of transfer learning strategy improves the product prediction ability using AM information, and the results further verify the flexibility of the transfer learning method. With multiple efforts, the top-1 accuracy of the transformer-transfer learning model evaluated by expert chemists can reach 91.4%, which provides proof of concept that AM can be successfully treated as translation task and that transfer learning method is powerful when the size of training data is relatively smaller. We hope our work can offer some inspiring insight to tackling with the problem of AM and reaction prediction.

Experimental Section

Transformer model

The architecture of the transformer model is depicted in Figure 6. As an encoder-decoder based model,^[30,31] the transformer

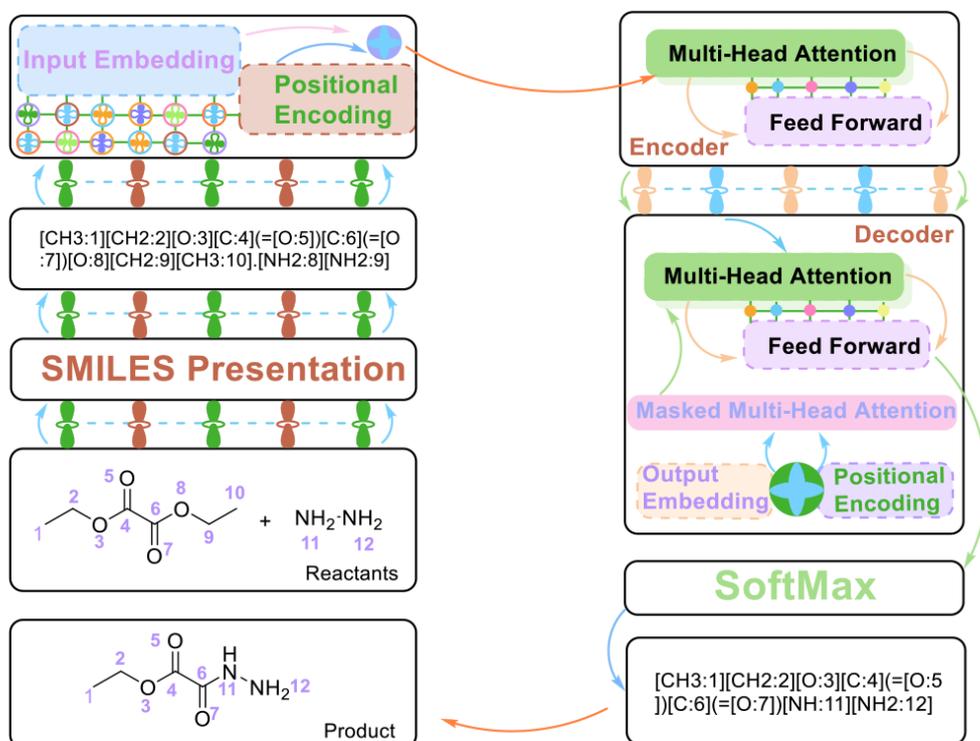


Figure 6. The transformer model architecture. The reactants with AM information are first converted into SMILES and input into the model. Then, the information about the reactants is processed by the encoder and decoder, respectively, in the form of the vector and the product with the corresponding AM information is given. It is worth noting that the encoder and the decoder contain a stack of N identical layers. N is the number of layers.

model is comprised of multi-head attention layers and positional feed forward layers. The encoder and decoder are the main parts of the transformer model. Several identical layers that contain two different sublayers constitute the encoder. The first sublayer is the multi-head attention mechanism, which consists of several parallel attention layers. The second layer is a connected feed forward network. Before the layer normalization,^[32] a residual connection^[33] around each of the sublayers is added to the encoder. In the decoder, there are three critical sublayers. Apart from the two sublayers that are identical to those of the encoder part, an additional layer, a masked multi-head attention mechanism that corresponds to the encoder's output, is introduced to the decoder. Furthermore, the residual connection still plays an important role around each of the sublayers.

Notably, the introduction of multi-head attention is a key feature of the transformer model. With the multi-head attention containing parallel attention layers, the model can concurrently attend to different versions of values. Therefore, the performance of the transformer model is superior to that of the models with a single attention mechanism. However, the information about the relative or absolute positions of the tokens in a string may be missing because the same attention is applied to the element of sequences no matter the length of the distance between tokens. To solve this problem, a positional encoding matrix^[34] is proposed. With this function, the model can obtain the information about the elements of sequences and make full use of the order of the sequences. In addition, the parameters which affect the model discussed in our prior work are used to explore the performance of the transformer model in a chemical task.^[24]

Atom mapping data: USPTO-AM-50K

The atom mapping data used in this study is the USPTO-AM-50K dataset. The dataset originated from the Lowe's data mining work, which obtained these reaction examples from the United States Patent and Trademark Office (USPTO) patents.^[35]

Schneider *et al.*^[6] further processed these reactions and extracted approximately 50K atom-mapped reactions spanning 10 broad reaction types, which can represent the common reaction types in a medicinal chemist's toolkit. The reaction classes are depicted in Table 4. The contextual information such as the reagents and temperature were eliminated to obtain reactions with only reactant and products, and then those reaction examples were canonicalized. Furthermore, Liu *et al.*^[22] further processed this data in order to split the reactions with multiple products into multiple distinct reactions and discarded the products with a SMILES length less than five characters, such as by-products and salts. Finally, those reactions were divided into training, validation and test sets at a ratio of 8:1:1.

Transfer learning data: USPTO-transfer learning-360K

The USPTO-transfer learning-360K dataset, which contains approximately 360K reaction examples, is applied to the transfer learning procedure. The source of these data was also derived from the Lowe's work.^[35] Additional processing of this dataset was implemented in this study. First, duplicate and incomplete reactions and their corresponding reagents are removed. It is

Table 4. Distribution and description of the major reaction classes within the processed reaction dataset.

Class	Description	Percentage of dataset (%)
1	heteroatom alkylation and arylation	30.3
2	acylation and related processes	23.8
3	C-C bond formation	11.3
4	heterocycle formation	1.8
5	protection	1.3
6	deprotection	16.5
7	reduction	9.2
8	oxidation	1.6
9	functional group interconversion (FGI)	3.7
10	functional group addition (FGA)	0.5

worth noting that the reactions belonging to the USPTO-AM-50K dataset are all removed from the USPTO-transfer learning-360K dataset, which prevents direct access to the AM information of the USPTO-AM-50K dataset. The difference between USPTO-AM-50K and USPTO-transfer learning-360K datasets is shown in the Figure 7. Compared to the USPTO-transfer learning-360K, the reactions in the USPTO-AM-50K dataset include the AM information. With the numbers that are attached to the atoms, the corresponding relationships between reactant and product atoms are represented in graphs as well as SMILES. Despite the difference between the USPTO-AM-50K and USPTO-transfer learning-360K datasets, the transformer model can capture the chemical knowledge from those datasets and combine their characteristics via transfer learning to solve the problem of atom matching.

Evaluation metric

We adopt the top-n accuracy as a key metric to evaluate the method performance. The top-n accuracy refers to the percentage of correct predictions found within the top-n predictions by our model. What's more, we have randomly split the USPTO-AM-50K dataset into training, validation and test datasets (with a ratio of 8:1:1) five times to avoid the results being dependent on the outcomes of a particularly favorable or unfavorable splitting.

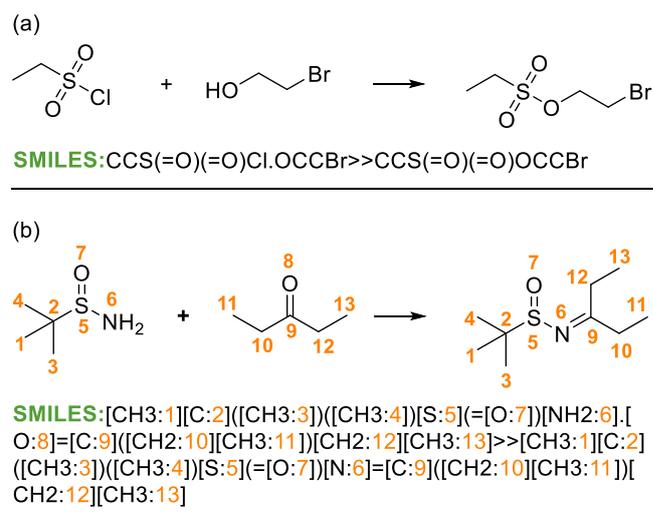


Figure 7. The difference between the USPTO-AM-50K and USPTO-transfer learning-360K datasets. The top is a reaction example from the USPTO-transfer learning-360K dataset. The bottom is a reaction with AM information from the USPTO-AM-50K dataset.

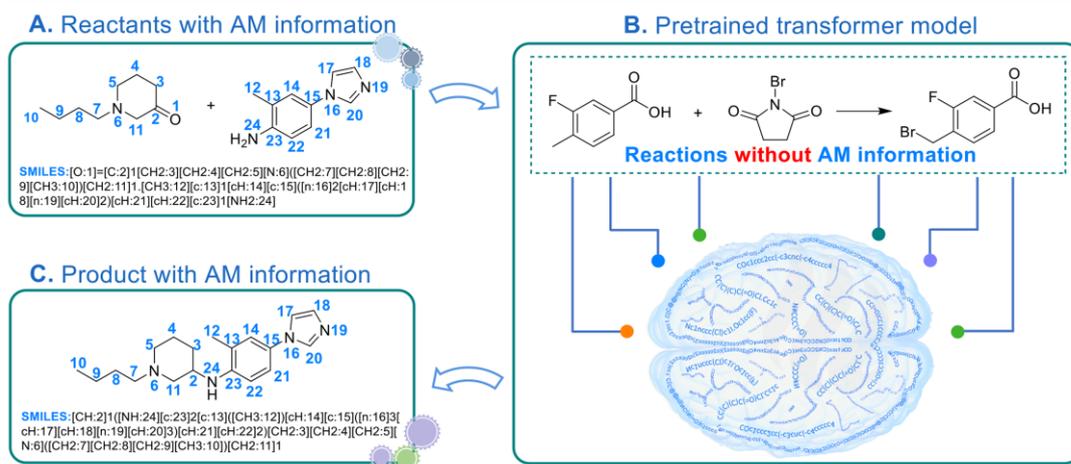


Figure 8. Schematic of the method regarding the AM problem as a translation task. Given an input SMILES that express reactant molecules with AM information, the machine translation model that training on the chemical reaction base outputs a SMILES of the predicted products with corresponding AM information. In this approach, the tasks of predicting reaction and AM can be solved simultaneously.

AM and reaction predictions

In this paper, we adapt the open source transformer model for our AM task and reaction prediction. The schematic of the approach to AM and reaction predictions is given in Figure 8. First, the transformer model is pretrained on the USPTO-transfer learning-360K dataset without atom mapping reactions. Each reaction is divided into reactant and product SMILES. The embedding of the reactant and product SMILES are passed to the encoder. In the pretraining process, the transformer model absorbs a good deal of basic chemical knowledge. Second, the pretrained model acts as a starting point to further train the model on the USPTO-AM-50K training dataset to learn to predict the products with corresponding atom mapping information. Furthermore, we apply the parameters of the pretrained model to initialize the model trained on the USPTO-AM-50K dataset. Finally, the model is tested on the USPTO-AM-50K testing dataset. In the testing process, we only feed the reactants with atom mapping information into the encoder. The model outputs corresponding products with atom mapping information.

Acknowledgements

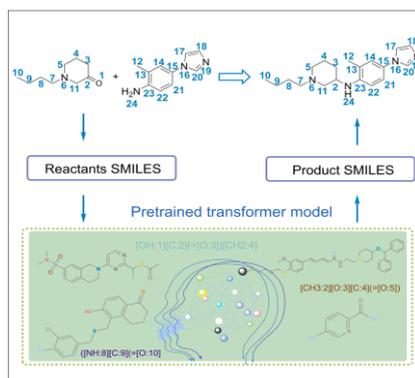
This project was supported by the National Natural Science Foundation of China, NSFC (Grant No.81903438).

Keywords: artificial intelligence • machine learning • transformer • reaction prediction • atom mapping

- [1] E. E. Litsa, M. I. Peña, M. Moll, G. Giannakopoulos, G. N. Bennett, L. E. Kavragi, *Chem. Inf. Model.* **2019**, *59*, 1121-1135.
- [2] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434-443.
- [3] W. Jin, C. W. Coley, R. Barzilay, T. S. Jaakkola, **2017**, arXiv:1709.04555.
- [4] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604-610.
- [5] A. Bøgevig, H. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev.* **2015**, *19*, 357-368.
- [6] N. Schneider, N. Stiefl, G. A. Landrum, *J. Chem. Inf. Model.* **2016**, *56*, 2336-2346.
- [7] M. Arita, *Genome Res.* **2003**, *13*, 2455-2466.
- [8] A. P. Heath, G. N. Bennett, L. E. Kavragi, *Bioinformatics* **2010**, *26*, 1548-1555.
- [9] S. M. Kim, M. I. Peña, M. Moll, G. N. Bennett, L. E. Kavragi, *J. Cheminform* **2017**, *9*, 51.
- [10] D. Bonchev, D. Rouvray, *Chemical Graph Theory: Introduction and Fundamentals*, Abacus Press, New York, London, **1991**.
- [11] S. A. Cook, *Proc. Third Annu. ACM Symp. Theory Comput., STOC'* **1971**, *71*, 151-158.
- [12] A. Kumar, C. D. Maranas, *J. Chem. Inf. Model.* **2014**, *54*, 3417-3438.
- [13] S. A. Rahman, G. Torrance, L. Baldacci, S. M. Cuesta, F. Fenninger, N. Gopal, S. Choudhary, J. W. May, G. Holliday, C. Steinbeck, J. Thornton, *Bioinformatics* **2016**, *32*, 2065-2066.
- [14] E. L. First, C. E. Gounaris, C. A. Floudas, *J. Chem. Inf. Model.* **2012**, *52*, 84-92.
- [15] M. Latendresse, J. P. Malerich, M. Travers, P. Karp, *J. Chem. Inf. Model.* **2012**, *52*, 2970-2982.
- [16] D. Fooshee, A. Andronico, P. J. Baldi, *Chem. Inf. Model.* **2013**, *53*, 1549-9596.
- [17] ChemAxon: <http://www.chemaxon.com>, Accessed on 2020-02-15.
- [18] H. Kraut, J. Eiblmaier, G. Grethe, P. Löw, H. Matuszczyk, H. J. Saller, *Chem. Inf. Model.* **2013**, *53*, 2884-2895.
- [19] W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Ryzdewski, A. Gambin, B. A. Grzybowski, *Nat Commun.* **2019**, *10*, 1434.
- [20] D. Weininger, *J. Chem. Inf. Model.* **1988**, *28*, 31-36.
- [21] J. Nam, J. Kim, **2016**, arXiv:1612.09529.
- [22] 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.
- [23] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. Bekas, A. A. Lee, **2018**, arXiv: 1811.02633v1.
- [24] H. Duan, L. Wang, C. Zhang, L. Guo, J. Li, *RSC Adv.* **2020**, *10*, 1371-1378.
- [25] R. Bai, C. Zhang, L. Wang, C. Yao, J. Ge, H. Duan, *Molecules* **2020**, *25*, 2357.
- [26] L. Wang, C. Zhang, R. Bai, J. Li, H. Duan, *Chem. Commun.* **2020**, 56, 9368-9371.
- [27] A. Vaswani, N. Shazeeret, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, **2017**, arXiv: 1706.03762.

- [28] E. S. Olivas, J. D. M. Guerrero, M. M. Sober, J. R. M. Benedito, A. J. S. Lopez, *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques-2 Volumes*, Information Science Reference, New York, London, **2009**.
- [29] P. Schwaller, B. Hoover, J. Reymond, H. Strobelt, T. Laino, **2020**, DOI: <https://doi.org/10.26434/chemrxiv.12298559.v1>.
- [30] D. Bahdanau, K. Cho, Y. Bengio, **2014**, arXiv: 1409.0473.
- [31] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, **2014**, arXiv:1406.1078.
- [32] J. L. Ba, J. R. Kiros, G. E. Hinton, **2016**, arXiv:1607.06450.
- [33] K. He, X. Zhang, S. Ren, J. Sun, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* **2016**, 1, 770-778.
- [34] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. D. Dauphin, **2017**, arXiv:1705.03122.
- [35] D. M. Lowe, *Extraction of Chemical Structures and Reactions from the Literature*, Thesis, University of Cambridge, **2012**.

The transformer model is applied to reaction prediction and the task of capturing the information of corresponding atoms between reactants and products. Using transfer learning strategy, the top-1 accuracy of our model in atom mapping task can reach 91.4%, which reveals the fact that atom mapping can be regarded as a translation task.



Chengyun Zhang, Ling Wang, Yejian Wu,
Yun Zhang, An Su,* and Hongliang Duan*

Page No. – Page No.
**Transformer: Linking Atom Mapping
and Neural Machine Translation**