Self-supervised molecular pretraining strategy for low-resource reaction prediction scenarios

Chengyun Zhang[1#], Xiang Cai[2#], Haoran Qiao[3#], Yun Zhang[1], Yejian Wu[1], Xinqiao Wang[1], Haiying Xie[4], Feng Luo[4] & Hongliang Duan[1*]

[1]Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences, Zhejiang University of Technology, Hangzhou 310014, P. R. China

[2]PyWise Biotech, Suzhou 215000, P. R. China

[3] College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 201203, P. R. China

[4]PUROTON Gene Medical Institute Co., Ltd., Chongqing 400700, P. R. China

[#]These authors contributed equally: Chengyun Zhang, Xiang Cai and Haoran Qiao

*Corresponding Author:

Hongliang Duan

Email: hduan@zjut.edu.cn

Abstract

In the face of low-resource reaction training samples, we construct a chemical platform for addressing small-scale reaction prediction problem. By using a self-supervised pretraining strategy called MASS, the transformer model can absorb the chemical information about 1 billion molecules and then finetunes on small-scale reaction prediction, which is different from previous works that only rely on reaction samples. To demonstrate the broad applicability of our approach, we adopt three different name reactions in our work. In the Baeyer-Villiger, Heck and Sharpless asymmetric epoxidation reaction prediction tasks, the average accuracies increase by 5.7%, 10.8%, 4.8% respectively, marking an important step to low-resource reaction prediction.

Introduction

Chemical synthesis is of great importance to the discovery of novel molecules for medicine and materials design, and one of the fundamental elements of this research is reaction prediction. Traditionally, predicting accurate products when given reactants is a rather rigorous task relying on experiments and professional chemistry experience, which inevitably demands a great deal of time and investments. Thanks to the advancements in computer science, the task of determining products of organic reactions can be aided by developing algorithms recently.

The existing algorithms for predicting reactions typically fall into three categories including template-based, physical chemistry and template-free methods. Approaches based on the so-called templates make reaction prediction by applying chemical rules that are encoded by expert chemists or automatically extracted from reaction databases.[1-6] Due to the template-based nature, those methods are incapable of predicting precise reactions that beyond the scope of their knowledge bases. As for physical chemistry algorithms, those methods depend on the calculation of energy barriers, which computational cost is prohibitively expensive.[7-9] Unlike template-based or physical chemistry algorithms, the template-free approaches are generally based on deep learning and bypass the need to require curated chemical reaction rules or laborious calculation of energy barriers.[10-15]

Recent innovations in deep learning represent great opportunities for template-free approaches in reaction prediction. Nam and Kim first applied a data-driven approach to reaction prediction task and treated it as a neural machine translation (NMT) task in 2016.[10] The next year, the success of a fully attention-based transformer model in language translation gathers interest from chemists and accelerates the development of computer-aided reaction prediction.[16] In 2019, Schwaller *et al* innovatively leveraged this model to make reaction prediction and the predictive accuracy can increase remarkably in United States Patent and Trademark Office (USPTO) dataset.[11] What's more, in a recent study by Tetko *et al*, authors reported an augmented transformer model, further showing the prominent capability of this model.[12] Until now, the transformer architecture remains a powerful approach for addressing the challenge of prediction reaction.

Although showing outstanding performance on various reaction tasks, template-free methods such transformer model meet their bottleneck in the face of low chemical data regimes, due to the data-driven feature. In order to solve such problem, many studies concentrate on transferring knowledge from a large-scale reaction dataset to a specific reaction prediction task.[13-15] For instance, Schwaller *et al* transferred the general chemistry knowledge of a USPTO dataset to a carbohydrate reaction prediction task, and the accuracy in this task can increase by 30.0%.[13] This reaction transfer learning strategy can be regarded as a practical workaround for some reaction prediction but the fundamental issue remains: it's difficult to construct or obtain such large-scale chemical reaction dataset like USPTO in reality. Even several chemical databases have been created by the effort of chemists[17,18], reactions are not easy to access in bulk, for commercial or technical reasons. In addition, building an applicable large reaction dataset also calls for an enormous effort in a variety of complex preprocessing. As a result, the carefully designed large reaction datasets are still scare now, which prevents data-driven models from tackling with the problem of some particular reaction predictions, especially in small-scale reactions.

Similar to this case in reaction prediction, language translation modeling also suffers from the lack of training samples. In natural language domain, a common approach for dealing with this problem is self-supervised training where a model pretrains on large numbers of monolingual data then finetunes on limited parallel data.[19] As we mentioned, the reaction prediction can be regarded as an NMT task where the reactant is a language and the product is another language. That said, reactants and products of reactions constitute a bilingual corpus. Following this idea, molecules can be treated as a monolingual data. Inspired by the success of self-supervised training in natural language processing (NLP), we cast our eyes to molecular data that is more accessible compared to reactions and utilize the corresponding chemical information about it to strengthen the predictive ability of our model when faced with limited reaction training examples.

To valid the effectiveness of our work, we adopt the popular encoder-decoder transformer framework and a self-supervised pretraining method called MAsked Sequence to Sequence (MASS)[20] for delivering molecular information to the prediction of products. The MASS method originally was proposed by Microsoft and achieves outstanding improvements in NLP. Inspired by BERT[21] but unlike it, this self-supervised pretraining method is introduced to pretrain encoder and decoder jointly rather than only pay attention to the encoder. The pretraining procedure is consist of two steps. Firstly, the encoder is forced to understand the meaning of the unmasked tokens. Second, MASS drives the decoder paying more attention to the source presentation by masking the input token of the decoder side that are not masked in the encoder. With the process of the self-supervised pretraining, the feature of molecules can be learned by transformer and then transferred to downstream reaction prediction tasks.

Furthermore, we organize a large-scale molecular dataset to pretrain transformer model and apply this pretrained model to a variety of small-scale reaction prediction tasks. This molecular dataset contains 1 billion compounds (some examples are displayed in Fig. S1) that are derived from two popular open-source compound databases called ZINC[22], ChEMBL[23]. In the term of downstream reaction prediction tasks, the Baeyer–Villiger[15] and Heck reaction datasets[14], two classic small-size reactions, are utilized to our experiments. In addition, we construct a new name reaction dataset involving chiral challenge for further demonstrating the universal applicability of our method. The schematic of our work can be founded in the Fig. 1.

To sum up, we innovatively combine the self-supervised molecular pretraining with transformer architecture to build a chemistry platform for low-resource reaction

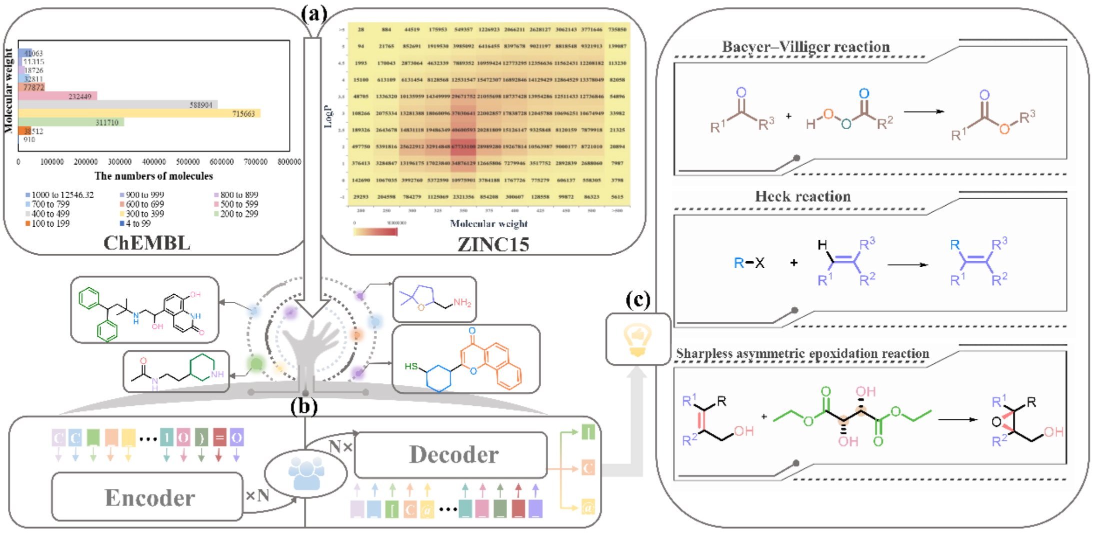

Fig. 1 Schematic diagram of the method for low-resource reaction prediction scenarios. The chemical molecular information (a) is absorbed by transformer architecture in self-supervised molecular pretraining procedure(b). Then, the pretrained model finetunes on a variety of downstream small-scale reaction prediction tasks (c).

prediction scenarios. Compare to studies that rely only on reaction dataset, our contribution is to demonstrate that it is possible to combine chemical molecular information with small-scale reaction prediction task. To the best of our knowledge, this is the first attempt to leverage a self-supervised pretraining with billions of molecules to provide a boost in the predictive ability of transformer model on determining the products of reactions.

Results

In our work, we implement three small-scale reaction prediction tasks to demonstrate the abroad applicability of our method. The table1, 2 show the performances of molecule-pretrained-Mass model (pretrains on molecular dataset and finetunes on downstream small-scale reactions) and baseline model (only trains on small-scale reactions) respectively, in the face of various small-scale reaction datasets with different splitting. Definitely, the knowledge of 1 billion molecules lays a strong foundation for guiding transformer model to make it more competitive to small-size reaction predic-

Table 1 The results of the molecule-pretrained-Mass model in 10 subsets of different reactions.

Entry	Reaction classes		
	Baeyer-Villiger reaction	**Heck reaction**	**Sharpless asymmetric epoxidation reaction**
1	74.3%	80.7%	61.8%
2	75.7%	81.4%	61.2%
3	75.7%	82.9%	65.4%
4	73.0%	81.9%	65.4%
5	67.7%	81.3%	61.8%
6	74.3%	79.4%	61.2%
7	77.0%	79.6%	63.0%
8	76.5%	81.6%	59.5%
9	74.3%	80.7%	63.0%
10	83.6%	82.7%	63.3%

Table 2 The results of the baseline model in 10 subsets of different reactions.

Entry	Reaction classes		
	Baeyer-Villiger reaction	**Heck reaction**	**Sharpless asymmetric epoxidation reaction**
1	70.4%	71.0%	55.6%
2	66.8%	70.3%	57.7%
3	72.1%	72.7%	60.7%
4	69.0%	73.5 %	62.1%
5	62.8%	73.2%	55.9%
6	69.5%	69.0%	56.2%
7	72.6%	65.7%	60.1%
8	71.7%	68.5%	56.2%
9	67.3%	71.3%	55.9%
10	73.0%	69.4%	57.1%

tion. Within each task, the molecule-pretrained-Mass model outperforms baseline model. In the experiment 1of Heck reaction, this model gains an 80.7% accuracy, which is 9.7% higher than baseline model (71.0%). What's more, the performance difference between those two models becomes hard to ignore in Sharpless asymmetric epoxidation reaction. The model's accuracies most over 60% while the results of baseline model most below it, in 10 experiments.

The improvements in different reaction prediction are represented in the Fig. 2. The accuracy increases in the heck reaction prediction all over 8.0%. Especially in the experiment 7, the accuracy of our model arises by 13.9%. Furthermore, the accuracy improvement in experiment 9 of Sharpless asymmetric epoxidation reaction can reach

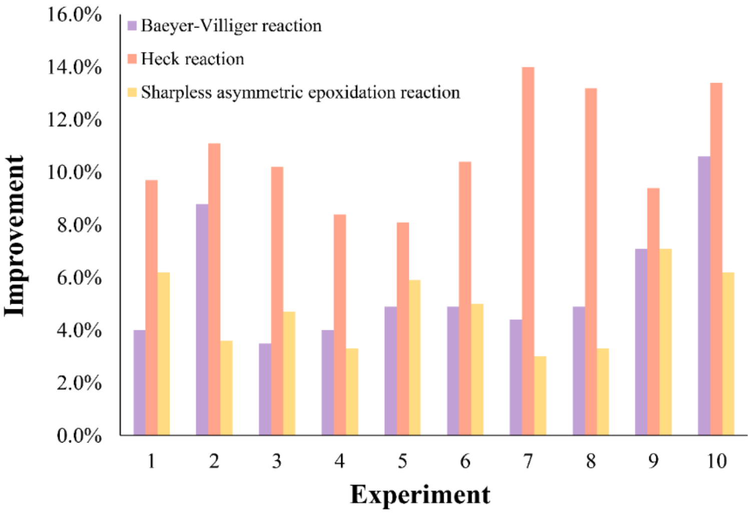

Fig. 2 The improvement in accuracy in different experiment of different reaction prediction.

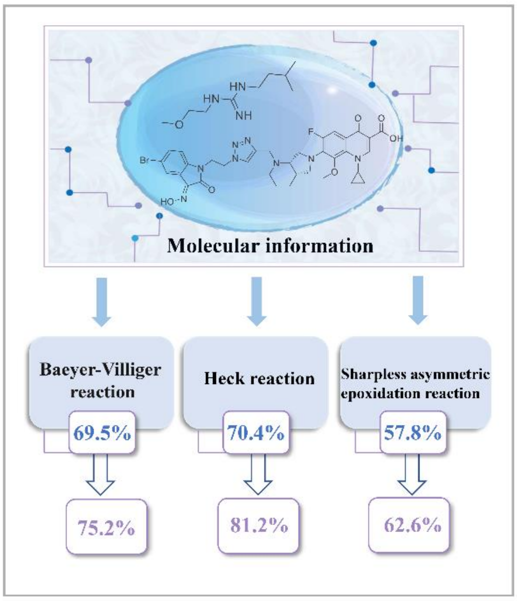

Fig. 3 The average accuracy comparison between molecule-pretrained-Mass model and baseline model in different reaction prediction tasks.

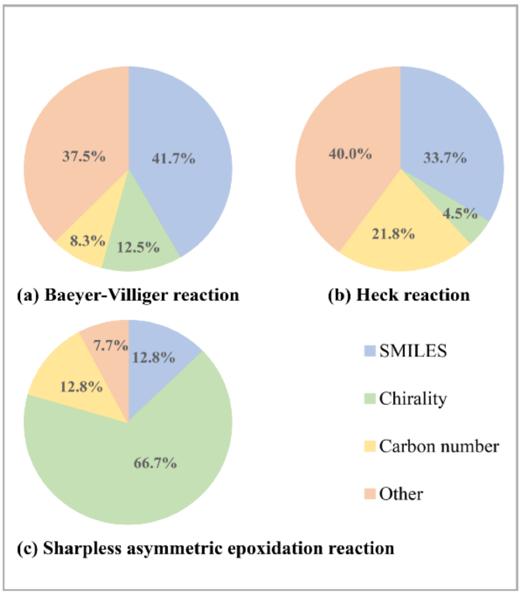

Fig. 4 The distribution of different improvement types in experiment 1 of three reactions prediction. (a) is the distribution of improvement types in Baeyer-Villiger reaction prediction. (b) is the distribution of improvement types in Heck reaction prediction. (c) is the distribution of improvement types in Sharpless asymmetric epoxidation reaction prediction.

7.1%.

To illustrate clearly the accuracy gap between molecule-pretrained-Mass and baseline models, the average accuracies in different tasks are displayed in Fig. 3. In the task of predicting Baeyer-Villiger, Heck and Sharpless asymmetric epoxidation reactions, the average accuracies of molecule-pretrained-Mass model are 5.7%, 10.8%, 4.8% respectively higher than baseline model. The significant increase in accuracy indicates that chemical molecular information indeed can help improve the model's understanding on reactions, even though molecular information doesn't contain the corresponding relationship between reactant and product molecules.

We chose experiment 1 to future analyze the improvement of model's performance in different reaction prediction tasks after applying the self-supervised molecular pretraining strategy. The distribution of improvement types is shown in the Fig. 4.

The improvement in general errors. Compared to the baseline model, the molecule-pretrained-Mass model have a better understanding of general knowledge such as SMILES presentation, the count of carbon number and chirality. In the following, we make analysis around those three aspects.

SMILES invalidation is common in the text-based predictive model. Due to the fragile nature of SMILES text presentation, a change of a single character may lead to a grammatically invalid SMILES that is not able to be translated into a chemical structure. Such problem can be reduced by forcing models to learn meaningful SMILES text presentation from a mass of molecular training samples. Pretrained on a large number of different SMILES strings, the molecule-pretrained-Mass model learns more grammatical knowledge compared to baseline model. In Baeyer-Villiger, Heck and Sharpless asymmetric epoxidation reaction prediction tasks, the progress of SMILES presentation can account for 41.7%, 33.7% and 12.8% of the total enhanced

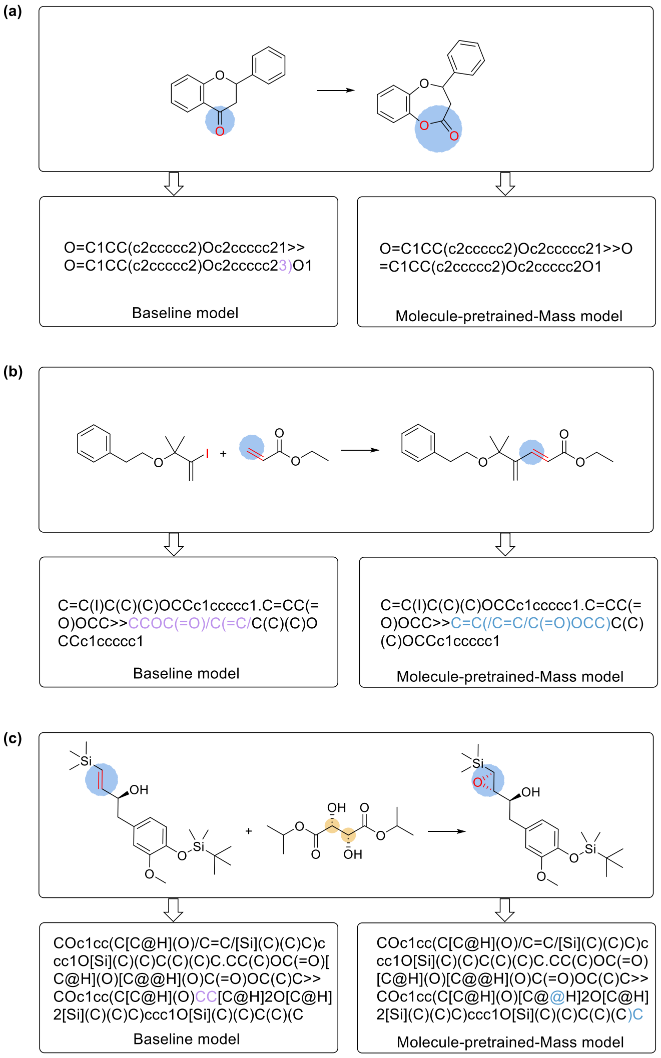
(a)
O
O
O
O
O=C1CC(c2ccccc2)Oc2ccccc21>>
O=C1CC(c2ccccc2)Oc2ccccc23)O1
Baseline model
O=C1CC(c2ccccc2)Oc2ccccc21>>O
=C1CC(c2ccccc2)Oc2ccccc2O1
Molecule-pretrained-Mass model
(b)
O
I
+
O
O
O
O
O
C=C(I)C(C)(C)OCCc1ccccc1.C=CC(=
O)OCC>>CCOC(=O)/C(=C/C(C)(C)O
CCc1ccccc1
Baseline model
C=C(I)C(C)(C)OCCc1ccccc1.C=CC(=
O)OCC>>C=C(/C=C/C(=O)OCC)C(C)
(C)OCCc1ccccc1
Molecule-pretrained-Mass model
(c)
Si
OH
O
Si
O
+
O
OH
O
O
O
OH
Si
O
OH
O
Si
O
COc1cc(C[C@H](O)/C=C/[Si](C)(C)C)c
cc1O[Si](C)(C)C(C)(C)C.CC(C)OC(=O)[
C@H](O)[C@@H](O)C(=O)OC(C)C>>
COc1cc(C[C@H](O)CC[C@H]2O[C@H]
2[Si](C)(C)C)ccc1O[Si](C)(C)C(C)(C
Baseline model
COc1cc(C[C@H](O)/C=C/[Si](C)(C)C)c
cc1O[Si](C)(C)C(C)(C)C.CC(C)OC(=O)
[C@H](O)[C@@H](O)C(=O)OC(C)C>>
COc1cc(C[C@H](O)[C@@H]2O[C@H]
2[Si](C)(C)C)ccc1O[Si](C)(C)C(C)(C)C
Molecule-pretrained-Mass model

Fig. 5 SMILES errors that occur in baseline model but not in molecule-pretrained-Mass model. (a) is a Baeyer-Villiger reaction, (b) is a Heck reaction and (c) is a Sharpless asymmetric epoxidation reaction.

accuracy respectively (Fig. 4).

Fig. 5 shows some presentive SMILES mistakes that predicted by baseline model but not present in molecule-pretrained-Mass model. Take Fig. 5(a) as an example, the two SMILES strings are pretty similar, but if you look closely, you'll see that the SMILES in the left has an extra '3)' more than that in the right. And that subtle change of alphabets causes the failure of baseline model in the predicting the product of the reactant 2-phenylchroman-4-one. Similarly, the 'C=C(/C=C/C(=O)OCC)' in the product SMILES of Fig. 5(b) is replaced by 'CCOC(=O)/C(=C/' strings, leading the baseline model's prediction grammatically incorrect. Different to baseline model, the molecule-pretrained-Mass model successful produces the valid SMILES strings.

Furthermore, we observe a phenomenon that the more complex the structure of compound, more likely the SMILES error occurs. Fig. 5(c) is a reaction that contains rings and chirality. In the form of text presentation, the length of its SMILES sequence is longer than that of other reactions (Fig. 5(a) and (b)). With such long SMILES, the model become more possible to make mistakes, allowing the improper characters to appear in its prediction. Although faced with such complex SMILES sequence, the molecule-pretrained-Mass model captures the feature of it and produces a correct product SMILES.

The second significant improvement is in the chirality. In the small-scale reaction, the lack of corresponding knowledge makes data-driven model more prone to chirality error. In the Fig. 6(a), both two predictions follow the migratory rule of Baeyer-Villiger reaction. Due to the insufficient chirality information, the baseline model mistakes a configuration of a carbon in the product and gives (1S,3S,3aS,3a^{1}R,6aR,11bR)-1-hydroxy-10-methoxy-3,3a^1-dimethyl-2,3,3a,3a^1,4,6a,7,11boctahydrodibenzo[de,g]chromen-5(1H)-one as product. In contrast, the molecule-pretrained-Mass model successfully predicts the configuration of this carbon atom.

Interestingly, we notice that this improvement type can account for 66.7% in the improved accuracy of the Sharpless asymmetric epoxidation reaction (Fig. 4(c)). This case can be attributed to its unique characteristic of this reaction: chirality changes. In

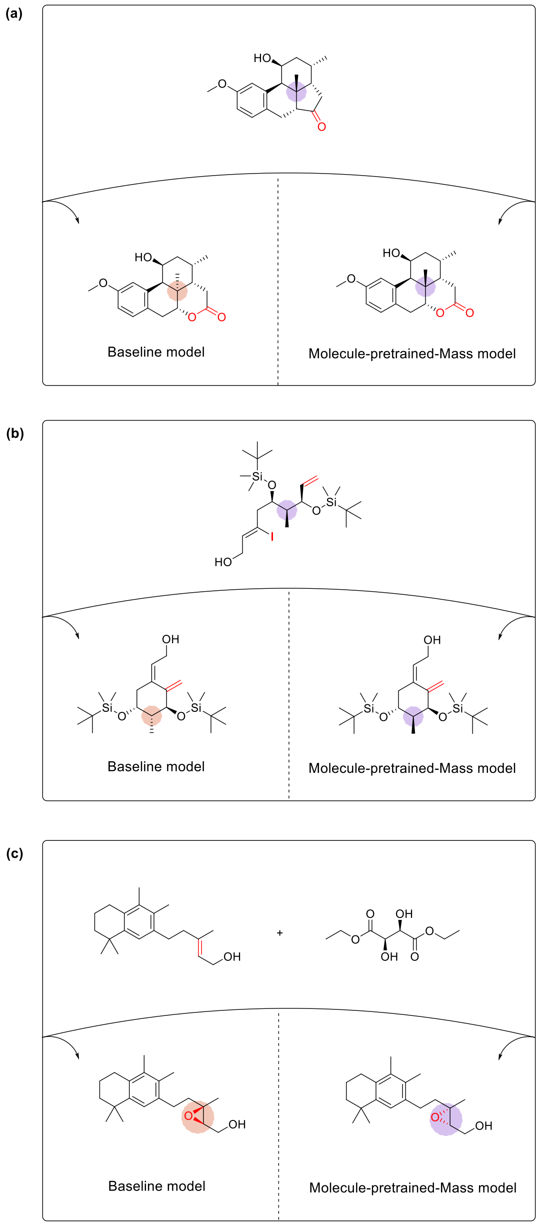
(a)
Baseline model
Molecule-pretrained-Mass model
(b)
Baseline model
Molecule-pretrained-Mass model
(c)
Baseline model
Molecule-pretrained-Mass model

Fig. 6 Chirality errors that occur in baseline model but not in molecule-pretrained-Mass model. (a) is a Baeyer-Villiger reaction, (b) is a Heck reaction and (c) is a Sharpless asymmetric epoxidation reaction.

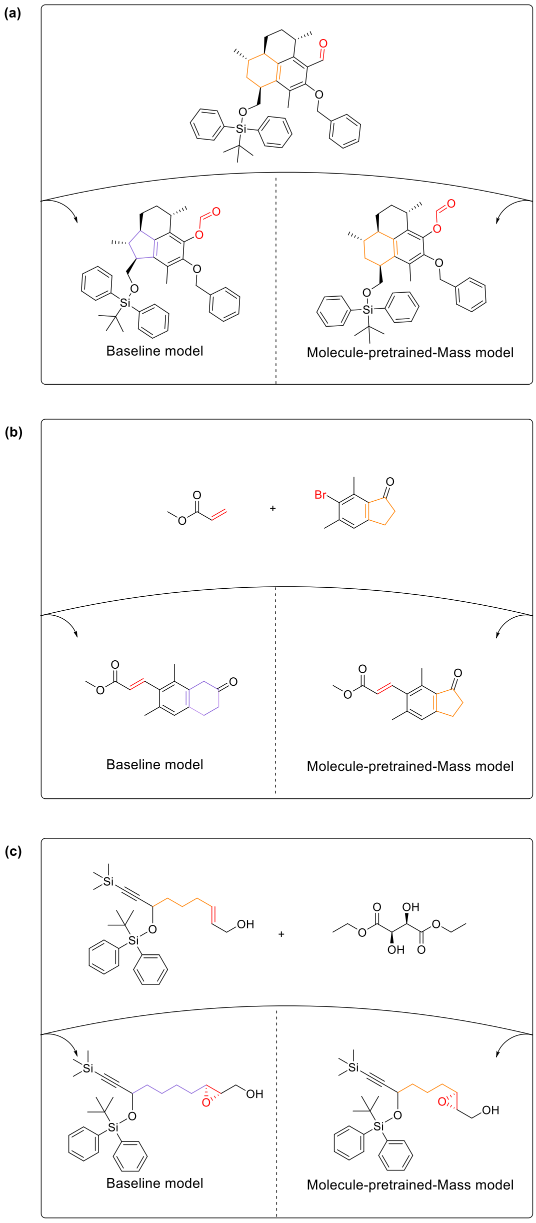
(a)
Baseline model
Molecule-pretrained-Mass model
(b)
Baseline model
Molecule-pretrained-Mass model
(c)
Baseline model
Molecule-pretrained-Mass model

Fig. 7 Carbon number errors that occur in baseline model but not in molecule-pretrained-Mass model. (a) is a Baeyer-Villiger reaction, (b) is a Heck reaction and (c) is a Sharpless asymmetric epoxidation reaction.

the Sharpless asymmetric epoxidation reaction, the enantiomer of the product is determined by the optical property of tartrate ester. Therefore, predicting a correct product of this reaction demands models to obtain a deep insight into stereochemistry. Thanks to the support of chirality compounds from the tremendous molecular dataset, the molecule-pretrained-Mass model achieves a better performance, which is different from the baseline model. Take the Fig. 6(c) as an example, the molecule-pretrained-Mass model successful predicts the product of (E)-3-methyl-5-(3,4,8,8-tetramethyl-5,6,7,8-tetrahydronaphthalen-2-yl) pent-2-en-1-ol in the presence of L-(+)-diethyl tartrate.

The last is an illustration of the improvement in counting carbon number of compounds. This carbon number error often occurs in a reaction containing complex or large structures. The multiple carbon atoms are included in such structures, causing models become difficult in keeping the numbers of those atoms in mind. In our work, the molecule-pretrained-Mass model is more sensitive to the numbers of carbon atoms in a compound, because this model has trained vast diverse compound structures.

The Fig. 7 shows examples of different reactions that the baseline model counts carbon number wrong but the molecule-pretrained-Mass model predicts right. Take the Fig. 7 (a) as an example, a hexatomic ring of reactant (3S,7S,9S,9aR)-5-(benzyloxy)-7-(((tert-butyldiphenylsilyl) oxy) methyl)-3,6,9-trimethyl-2,3,7,8,9,9a-hexahydro-1H-phenalene-4-carbaldehyde is incorrectly predicted to a five ring by the baseline model. In addition, this model adds a carbon atom in the reactant (E)-7-((tert-butyldiphenylsilyl) oxy)-9-(trimethylsilyl) non-2-en-8-yn-1-ol (Fig. 7(c)). Conversely, the molecule-pretrained-Mass model gains a better mathematical ability for counting carbon numbers of those complex structure and accurately predicts corresponding products.

Furthermore, the improvement of counting carbon number is very remarkable in the task of predicting Heck reaction. This improvement can account for 21.8% of the overall growth in accuracy. The reactants of this reaction type commonly contain multiple or large ring structure, which means that models need to have a good memory on the constituents of those reactants. The structure knowledge from our molecular dataset assists the molecule-pretrained-Mass model in remembering the carbon atoms of a Heck reaction. An example of how the molecular structure knowledge benefits our model is shown in the Fig. 7(b). The molecule-pretrained-Mass model accurately predicts the product methyl (E)-3-(4,6-dimethyl-3-oxo-2,3-dihydro-1H-inden-5-yl) acrylate, when given the reactants methyl acrylate and 6-bromo-5,7-dimethyl-2,3-dihydro-1H-inden-1-one.

The improvement in specific chemical challenges. Having analyze the improve

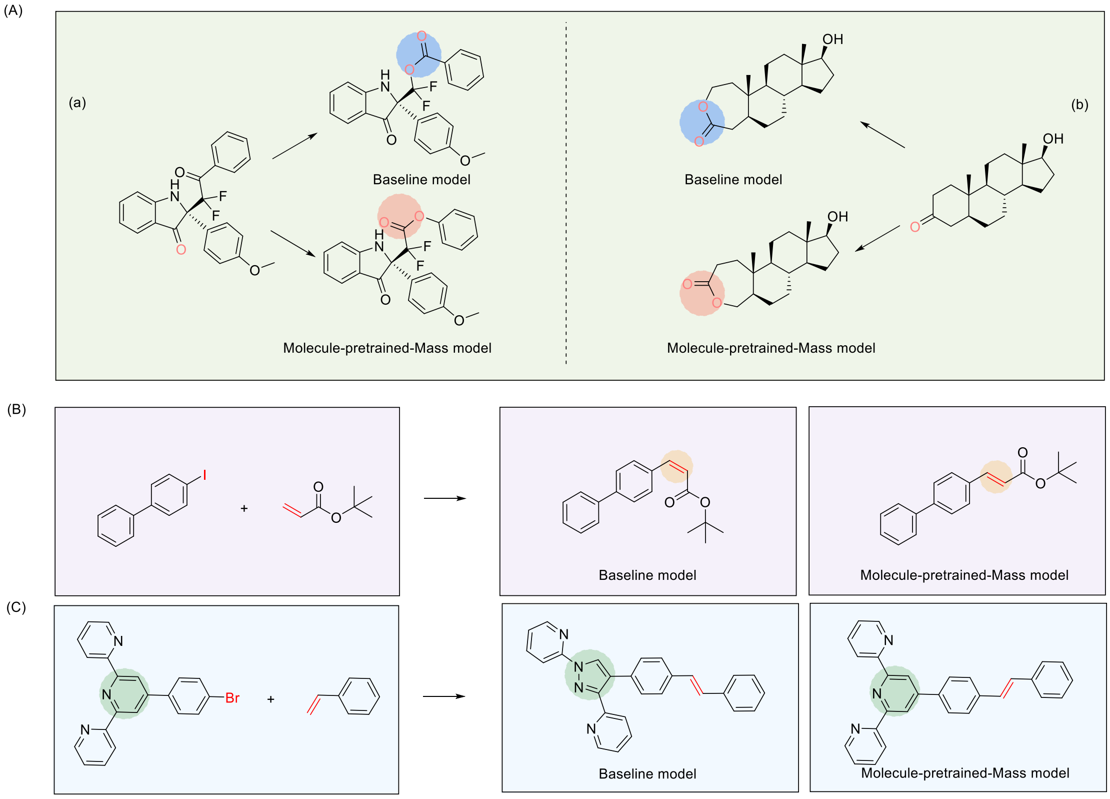
(A)
(a)
Baseline model
Molecule-pretrained-Mass model
(b)
Baseline model
Molecule-pretrained-Mass model
(B)
Baseline model
Molecule-pretrained-Mass model
(C)
Baseline model
Molecule-pretrained-Mass model

Fig. 8 Special errors that occur in baseline model but not in molecule-pretrained-Mass model. (A) are two Baeyer-Villiger reaction where (a) involves a ketone and (b) involves an aldehyde, (B) is a stereochemistry error in Heck reaction and (C) is a functional group error in heck reaction.

ment of the molecule-pretrained-Mass model in general errors, we next to shows the advantage of molecular information in some specific chemical challenges of different reactions.

The first chemical challenge is the group migration in the Baeyer-Villiger reaction. In this rearrangement reaction, the migratory capacity of group can impact the regiochemistry. In the exploration of the Baeyer–Villiger reaction prediction, the ratio of migration group improvement can reach 25.0% in the total growth.

There are two examples in Fig. 8(A) for displaying the difference between the molecule-pretrained-Mass and baseline models in understanding the migration rule of Baeyer–Villiger reaction. The first example is a transformation from (R)-2-(1,1-difluoro-2-oxo-2-phenylethyl)-2-(4-methoxy-phenyl) indolin-3-one to phenyl (R)-2,2-difluoro-2-(2-(4-methoxyphenyl)-3-oxoindolin-2-yl)-acetate. In this reaction, phenyl group can stabilize more positive charge. As a result, the oxygen is inserted at the right of carbonyl group. However, the baseline model seems to not have a fully understanding of the characteristic of Baeyer–Villiger reaction and mistakes that (R)-difluoro(2-(4-methoxyphenyl)-3-oxoindolin-2-yl) methyl benzoate is the product of this reaction, when the molecule-pretrained-Mass model captures the difference between the migratory groups of the reactant. The molecule-pretrained-Mass model also successful recognizes the migration group of a reactant containing complex ring constructure in Fig. 8(b) and gives correct product (5aR,5bS,7aS,8S,10aS,10bR,12aR)-8-hydroxy-5a,7a-dimethylhexadecahydro-3H-cyclopenta [5,6] naphtho [2,1-c] oxepin-3-one.

Second, we discuss the improvement in the Heck reaction after bringing in a large amount of molecular information. The stereoselectivity is one of chemical traits of this reaction. In general, the product of heck reaction prefers to be *E*- isomerism that is more thermodynamic stable. However, the lack of relevant information hampers data-driven models to predict an appropriate isomerism of products.

An example is displayed in Fig. 8(B). Because of the large ring structure of 4-iodo-1,1'-biphenyl group, the product of this reaction tends to be tert-butyl (E)-3-([1,1'-biphenyl]-4-yl) acrylate. The baseline model not take the effect of [1,1'-biphenyl]-4-yl group into the consideration and the wrong product called tert-butyl (Z)-3-([1,1'-biphenyl]-4-yl) acrylate is given by this model. In contract to baseline model, the molecule-pretrained-Mass model that quipped with mass stereochemistry knowledge, gives the right prediction with an appropriate isomerism.

At last, we want to demonstrate that the molecule-pretrained-Mass model obtains more knowledge about the functional group compared to baseline model in Heck reaction prediction. In our work, 26 functional group errors are predicted by baseline model but not given by the molecule-pretrained-Mass model. A representative example is displayed in the Fig. 8(C), the baseline model replaces the pyridine ring group of 4'-(4-styrylphenyl)-2,2':6',2''-terpyridine with a pyrazole ring group, inducing the wrong product (E)-2,2'-(4-(4-styrylphenyl)-1H-pyrazole-1,3-diyl) dipyridine. This situation may be as a result of the complex structures of reactants or long SMILES reactant representation in Heck reaction prediction. However, the corresponding chemical information from self-supervised molecular pretraining procedure can effectively reduce the frequency of this error type and lead the increase in accuracy.

The Visualization of TMAP in different reactions. As we mentioned, the three reaction datasets that we apply are quite different and each of these reactions has its own characteristics. For further proving the difference between them, we adopt a dimensionality reduction algorithm called TMAP in our work. With TMAP, reactions that belonging to same type form a cluster which is separate from other types and different reaction clusters are highlighted by color encoding. For example, the cluster is labeled with yellow in the bottom of the picture, showing that those reactions are Sharpless asymmetric epoxidation reactions. However, the cluster of Baeyer-Villiger marked with purple is in the left of this picture. Obviously, those two reaction types are succ-

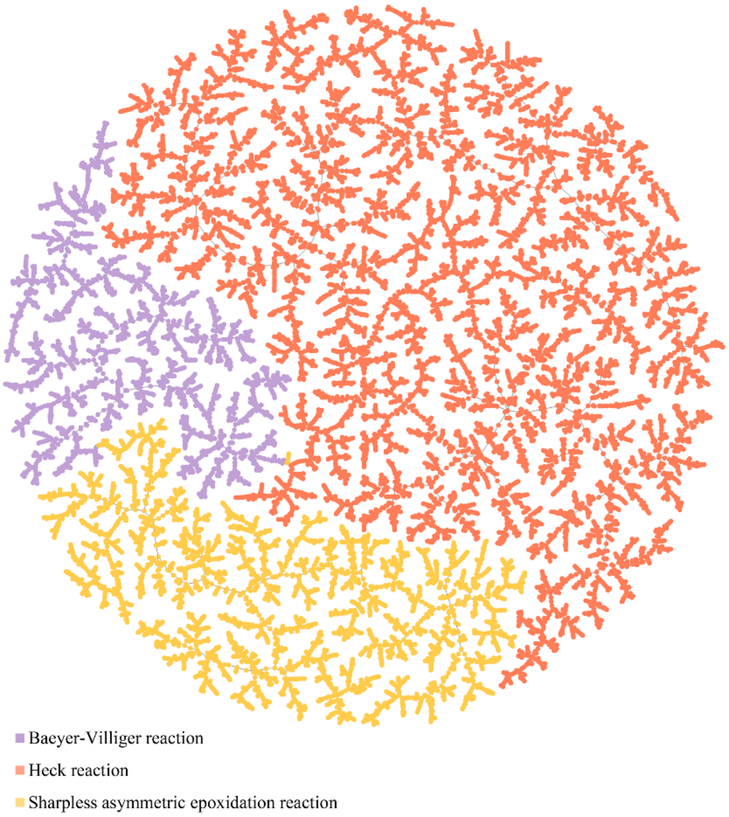

Fig. 9 The TMAP of three reactions. The Baeyer-Villiger reactions are color coded by purple, the Heck reactions are color coded by orange, the Sharpless asymmetric epoxidation reactions are color coded by yellow.

essfully distinguished by TMAP. With the visualization of TMAP, the different reactions applied in our work can be classified, showing that our model is not merely used for dealing with a specific small-scale reaction prediction problem.

The effect of reaction training examples size. We construct different training subsets of Heck reaction to investigate the influence of reaction training examples size on our approach. As is illustrated in Fig. 10, the power of our pretraining method in Heck reaction is affected by the size of training samples. With 500 reaction training samples, the improvement of performance in Heck reaction is 20.2%, which is similar to the performances when the size of reaction training samples is 6000 and 7000. However, the improvement of accuracy can up to 34.5% when the pretrained model finetunes on 2000 reactions. Namely, the self-supervised molecular pretraining can benefit different size reaction prediction and the effect of this approach is influenced by the size of the training data.

Discussion

In our work, we adopt the self-supervised molecular pretraining to absorb corres-

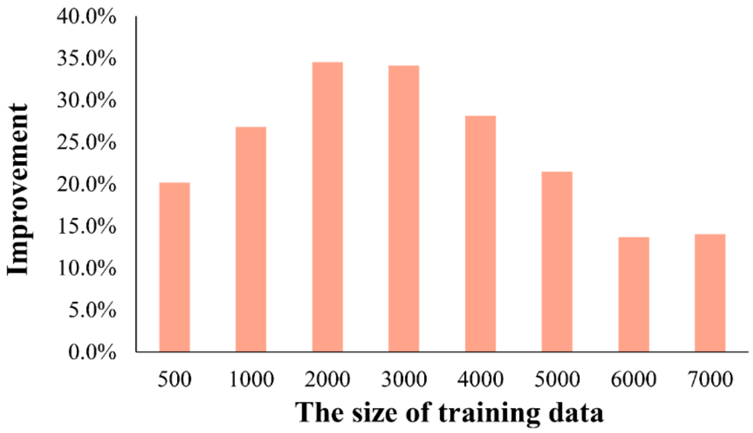

Fig. 10 The improvement in accuracy with different sizes of reaction training data in Heck reaction prediction.

ponding knowledge from 1 billion molecules and deliver it to three small-scale reaction predictions. With the aid of the self-supervised molecular pretraining, our model's performance increased by 5.7%, 10.8%, 4.8% on average in Baeyer-Villiger, Heck and Sharpless asymmetric epoxidation reaction tasks, respectively. More importantly, the molecule-pretrained-Mass model not only has a better understanding on general prediction challenges such as SMILES and chirality, but also gains more comprehension about some specific chemical features of different small-scale reactions. In addition, we implement the TMAP algorithm to clearly demonstrate the difference between reactions. We also investigate that the effect of training dataset size in our work. It should be clearly mentioned that our method can combine with other strategies such data augmentation for addressing the problem of reaction prediction in low-resource scenarios. In other word, our method is a complement rather than an alternative, and the synergy between our method and other applications may offer a strong boost to deal with small-scale reaction prediction.

Methods

In our work, we apply three different datasets to valid the universal ability of our method in small-scale reaction prediction. The representative examples of those reactions can be found in Fig. 11.

Baeyer-Villiger reaction dataset. Baeyer–Villiger reaction is a typical example of small-scale reaction.[24] With a peroxyacid or peroxide, an ester can be formed from a ketone or an aldehyde (detailed information about Baeyer–Villiger reaction is available in Section S2). Fig. 11 shows some presentative examples of this reaction type. As a rearrangement reaction, the key feature of this reaction is that the regiochemistry relies on the migratory capacity of group. Usually, the migratory aptitude of group is ranked as followed: tertiary alkyl > secondary alkyl > aryl > methyl. Take Figure 2(b) as example, the methyl group of the 1-(bicyclo [3.3.1] nonan-1-yl) ethan-1-one is more reluctant to undergo migration, and the bicyclo [3.3.1] nonan-1-yl acetate is formed as a product.

The Baeyer-Villiger reaction dataset is originally from the work of Zhang *et al.*[15] Those reactions are extracted from a commercial database called Reaxys[17] by using the name of this reaction types and reaction templates. Then, the incomplete, repeated and other error reaction samples are further eliminated from this raw dataset. The filtered data is further processed for only obtaining reactants and products. Finally, there

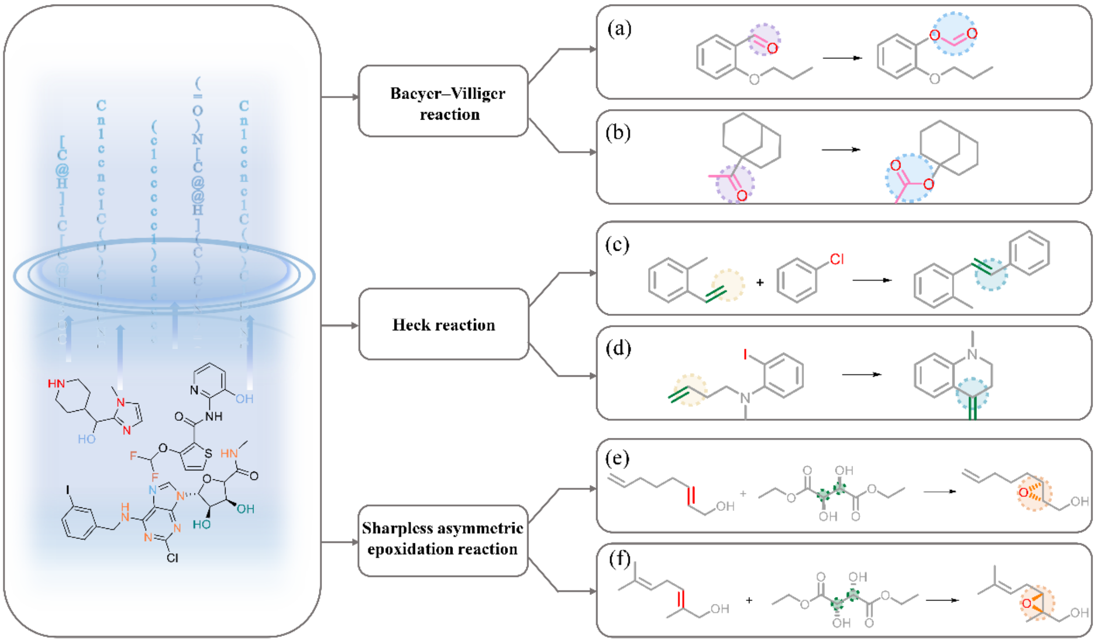

Fig. 11 Presentative examples of reactions in three datasets. (a) is a Baeyer–Villiger reaction where reactant is an aldehyde and (b) is a Baeyer–Villiger reaction where reactant is a ketone. (c) is a Heck reaction that occurs in a reactant and (d) is a Heck reaction that occurs between molecules. (e) is a Sharpless asymmetric epoxidation reaction involving L-(+)- diethyl tartrate and (f) is a Sharpless asymmetric epoxidation reaction involving D-(-)- diethyl tartrate.

are 2254 Baeyer-Villiger reactions samples in the dataset.

Heck reaction dataset. The palladium-catalysed reaction where a new alkene is formed by coupling an alkene with an organic halide or triflate, is referred to as the Heck reaction (detailed information about Heck reaction is available in Section S3).[25] As an effective tool, the reaction gives great boost in constructing carbon-carbon bonds. According to the reactant types, the heck reaction can be divided into two classes: intermolecular and intramolecular Heck reactions (Fig. 11(c) and (d)). In general, the more substituted groups alkene of reactant obtains, the slower the reaction occurs. Therefore, the order of reaction rates roughly follows CH_2=CH_2 > CH_2=CHOAc > CH_2=CHMe > CH2=CHPh > CH2=C(Me)Ph in the Heck reaction.

The heck dataset is constructed in Wang *et al* 's study.[14] Similar to the procedure of preparing Baeyer-Villiger reaction dataset, this reaction dataset is derived from the Reaxys and processed by expurgating duplicated and wrong reactions. In order to gain reactions that only involves reactants and products, the contextual information such as temperature and time are removed from the filter Heck dataset. Ultimately, a dataset containing 9405 intermolecular and 554 intramolecular Heck reactions is organized.

Sharpless asymmetric epoxidation reaction dataset. In the presence of a chiral tartrate ester, prochiral or chiral allylic alcohols can be oxidated to an enantiopure 2,3-epoxy alcohols, which is called Sharpless asymmetric epoxidation reaction (detailed information about Sharpless asymmetric epoxidation reaction is available in Section S4).[26] This reaction is reagent controlled: the optical property of tartrate ester affects the enantiomer of the product 2,3-epoxy alcohol. Here are some examples shown in the Figure 11(e) and (f). Under different chiral tartrate esters, the configuration of products is corresponding different. As an enantioselective reaction, the Sharpless asymmetric epoxidation reaction involves chirality changes, adding additional challenge for models.

The preprocessing steps of Sharpless asymmetric epoxidation reaction dataset follows the procedure that Zhang and Wang *et al* did[14,15]. The raw Sharpless asymmetric epoxidation reaction dataset is come from Reaxys and throughs a cleaning procedure. Because of the reagent-controlled trait of this reaction type, the reagent information is preserved in our dataset. In total, 3060 Sharpless asymmetric epoxidation reactions are applied to our work.

In our work, all data are presented by simplified molecular-input line-entry system[27] (SMILES) text representation. What's more, the reaction datasets are all split for testing, validation, training respectively, at a ratio of 1:1:8. It's worth mentioned that we adopt 10-fold cross validation to split reaction datasets. This strategy ensures that the overfitting problem can be avoided and a particularly favorable or unfavorable splitting not make influence on the prediction. In addition, the accuracy is used as a key metric for quantitatively measuring performance of our approach. The accuracy means the percentage of right predictions found within the results given by model.

Model. We employ the powerful transformer architecture for learning knowledge about molecules and reaction in our work. This model is based solely on an attention mechanism and eschews recurrence most commonly used in encoder-decoder architectures.[16] At present, the transformer remains a popular architecture for wide variety of problems, including NLP, computer vision (CV) and reaction prediction.

The key component of transformer is Multi-Head Attention (MHA). For each head, the equation is defined as

$$Attention(Q, K, V) = softmax\,(\frac{Q^T K}{\sqrt{dk}})V$$

where Q, K, V are input embedding matrices and d_k is the embedding dimension. With the MAH, the information can be handled parallelly on different subspaces (detailed information about our model is available in Section S5).

MASS pretraining method. MASS is a self-supervised pretraining method proposed for sequence-to-sequence learning framework[20]. Unlike BERT and other pretraining model, this approach focuses on both the encoder and decoder, which is perfectly suitable to our reaction prediction task. With this self-supervised pretraining method, the transformer model can study on a large-size molecular dataset, then transmits that chemical molecular knowledge to different downstream reaction prediction task.

The detail mechanism of this self-supervised pretraining method is shown in Fig. 1(b). A fragment of input sequence will be masked, and the task of decoder is to predict corresponding masked information. The objective function can be described as follows

$$L(\chi) = \frac{1}{|\chi|}\sum_{x\in\chi} \mathrm{logP}\left(x^{u:v}\middle|x^{\backslash u:v}\right)$$

where $x^{\backslash u:v}$ is denoted as fragment sequence x that are masked from position u to v. The values u and v are randomly chosen and restrict to $0 < u < v < N$ and N is number of tokens of sequence x.

TMAP. TMAP is a data visualization method which can capable of dealing with data points and reduce arbitrary high dimensionality to a two-dimensional tree.[28] Due to the tree-like nature, this method is more suitable in the exploration and interpretation of dataset compared to other dimensionality reduction algorithms. This method relying on a combination of locality sensitive hashing, graph theory and modern web technology, consists of four main components: a) LSH forest indexing, b) the generation of a d-approximate l-nearest neighbor graph, c) of a minimum spanning tree (MST) of the d-approximate l-nearest neighbor graph d) construction of a lay-out for the resulting MST.

However, the TMAP is originally designed for the visualization of molecular dataset. Thanks to the efforts of Schwaller *et al*, this method is extended for displaying chemical reactions.[29]

Data availability

The reaction datasets used in our study are available at https://github.com/hongliangduan/Self-supervised-molecular-pretraining-strategy-for-low-resource-reaction-prediction-scenarios/tree/master/data

Code availability

The code for our model is available at https://github.com/hongliangduan/Self-supervised-molecular-pretraining-strategy-for-low-resource-reaction-prediction-scenarios

References

1. E. J. Corey, W. T. Wipke, R. D. Cramer III & W. J. Howe. Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192 (1969).

2. D. A. Pensak & E. J. Corey. LHASA—Logic and heuristics applied to synthetic analysis, *J. Am. Chem. Soc*. **61**, 1–32 (1977).

3. W. L. Jorgensen, E. R. Laird, A. J. Gushurst, J. M. Fleischer, S. A. Gothe, H. E. Helson, G. D. Paderes & S. Sinclair. CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl. Chem* **62**, 1921–1932 (1990).

4. H. Satoh & K. Funatsu. SOPHIA, a knowledge base-guided reaction prediction system - utilization of a knowledge base derived from a reaction Database. *J. Chem. Inf.Comput. Sci*. **35**, 34–44 (1995).

5. M. H. S. Segler & M. P. Waller. Modelling chemical reasoning to predict and invent reactions. *Chem. – Eur. J.* **23**, 6118–6128 (2017).

6. T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuc´,M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska,A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A.Adamski, B. Sieredzin´ska, M. Mrksich, S. L. J. Trice & B. A. Grzybowski. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 1–11 (2018).

7. W. R. Dolbier Jr., H. Korniak, K. N. Houk & C. Sheu. Electronic Ccontrol of stereose-lectivities of electrocyclic reactions of cyclobutenes: A triumph of theory in the prediction of organic reactions. *Acc. Chem. Res.* **29**, 471–477 (1996).

8. L.-P. Wang, R. T. McGibbon, V. S. Pande & T. J. Martinez. Automated discovery and refinement of reactive molecular dynamics pathways. *J. Chem. Theory Comput.* **12**, 638–649 (2016).

9. O. Engkvist, P.-O. Norrby, N. Selmi, Y.-h. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard & L. A. Smyth. Computational prediction of chemical reactions: current status and outlook. *Drug*

Discovery Today **23**, 1203–1218 (2018).

10. J. Nam & J. Kim. Linking the neural machine translation and the prediction of organic chemistry reactions. Preprint at https://arxiv.org/abs/1612.09529 (2016).
11. P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas & A. A. Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
12. I. V. Tetko, P. Karpov, R. V. Deursen & G. Godin. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat. Commun.* **11**, 5575 (2020).
13. G. Pesciullesi, P. Schwaller, T. Laino, & J. Reymond. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
14. L. Wang, C. Zhang, R. Bai, J. Li & H. Duan. Heck reaction prediction using transformer model based on transfer learning strategy. *Chem. Commun.* **56**, 9368–9371 (2020).
15. Y. Zhang, L. Wang, X. Wang, C. Zhang, J. Ge, J. Tang, A. Su & H. Duan. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Org. Chem. Front.* , **8**, 1415 (2021).
16. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser & I. Polosukhin. Attentionis all you need. *In Advances in Neural Information Processing Systems* 5998-6008 (2017).
17. Reaxys, https://www.reaxys.com/
18. Scifinder , https://scifinder-n.cas.org/
19. A. Siddhant, A. Bapna, Y. Cao, O. Firat, M. Chen, S. Kudugunta, N. Arivazhagan & Y. Wu,

Leveraging monolingual data with self-supervision for multilingual neural machine translation. Preprint at https://arxiv.org/abs/2005.04816 (2020).

20. K. Song, X. Tan, T. Qin, J. Lu & T. Liu, MASS: masked sequence to sequence pre-training for language generation. Preprint at https://arxiv.org/abs/1905.02450 (2019).

21. J. Devlin, M.-W. Chang, K. Lee & K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2018).

22. T. Sterling & J. J. Irwin. ZINC 15 − Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337(2015).

23. D. Mendez, A. Gaulton, A P. Bento, J. Chambers, M. D. Veij, E. Félix, M. P. Magariños, J. F Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J Radoux, A. Segura-Cabrera, A. Hersey & A. R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, 930–940 (2019).

24. G.-J. Ten Brink, I. W. C. E. Arends & R. A. Sheldon. The Baeyer–Villiger reaction: new developments toward greener procedures. *Chem. Rev.* **104**, 4105–4123 (2004).

25. R. F. Heck. Acylation, methylation & carboxyalkylation of olefins by Group VIII metal derivatives. *J. Am. Chem. Soc.* **90**, 5518–5526(1968).

26. T. Katsuki & K. B. Sharpless. The first practical method for asymmetric epoxidation. *J. Am. Chem. Soc.***102**, 5974–5976 (1980).

27. D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).

28. D. Probst & J. Reymond, Visualization of very large high-dimensional data sets as minimum

spanning trees. Preprint at https://arxiv.org/abs/1908.10410 (2020).

29. P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino & J.-L. Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).

Acknowledgements

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

Author contributions

These authors contributed equally: C. Z., X. C and H. Q., C. Z., X. C. and H.D. designed the research project. C. Z., X. C and H. Q. designed and trained models. C. Z. analyzed data and wrote the manuscript. All authors discussed the results and approved the manuscript.

Competing interests

The authors declare no competing interests.