Learning Chemical Rules of Retrosynthesis with Pre-training

Yinjie Jiang[1], Ying Wei[2*], Fei Wu[1,3*], Zhengxing Huang[1], Kun Kuang[1], Zhihua Wang[3]

[1] Zhejiang University
[2] City University of Hong Kong
[3] Shanghai Institute for Advanced Study of Zhejiang University
{jiangyinjie, wufei, zhengxinghuang, kunkuang, zhihua.wang}@zju.edu.cn, yingwei@cityu.edu.hk,

Abstract

Retrosynthesis aided by artificial intelligence has been a very active and bourgeoning area of research, for its critical role in drug discovery as well as material science. Three categories of solutions, i.e., template-based, template-free, and semi-template methods, constitute mainstream solutions to this problem. In this paper, we focus on template-free methods which are known to be less bothered by the template generalization issue and the atom mapping challenge. Among several remaining problems regarding template-free methods, failing to conform to chemical rules is pronounced. To address the issue, we seek for a pre-training solution to empower the pre-trained model with chemical rules encoded. Concretely, we enforce the atom conservation rule via a molecule reconstruction pre-training task, and the reaction rule that dictates reaction centers via a reaction type guided contrastive pre-training task. In our empirical evaluation, the proposed pre-training solution substantially improves the single-step retrosynthesis accuracies in three downstream datasets.

Introduction

Firstly formulated by Corey and Wipke (1969), *Retrosynthesis* is the task to design synthesis routes of target products, which plays a significant role in chemistry and pharmacy. With the increasing number of chemical reactions, even experienced and professional chemists spend much time on retrosynthesis. Automated retrosynthesis machines are urgently needed. Thus, single-step retrosynthesis prediction which serves as the foundation of the retrosynthesis route planning becomes a critical juncture for the intersection of machine learning and chemistry.

Existing machine learning works on retrosynthesis prediction are mainly divided into three categories: template-based methods (Segler and Waller 2017; Dai et al. 2019; Sun et al. 2021; Seidl et al. 2021), semi-template-based methods (Yan et al. 2020; Shi et al. 2020; Somnath et al. 2021; Seo et al. 2021), and template-free methods (Schwaller et al. 2020; Zhu et al. 2021; Sun et al. 2021). Template-based and semi-template-based methods heavily depend on templates or Atom-Atom-Mappings, which are also unsolved challenges in chemistry. Meanwhile, the performance of template-free methods is barely satisfactory because of the lack of additional chemical information. After taking a closer look at the results of previous models, we pinpoint three challenges in retrosynthesis prediction, especially for template-free methods (see Figure **??**): 1) generated molecules are invalid. 2) generated reactants break *Law of Conservation of Atoms*. 3) generated reactants do not react or do not produce the target product.

To address the three challenges, we propose a **P**re-trained **M**odel for **S**ingle-step **R**etrosynthesis (PMSR) where we design three pre-training tasks. Besides auto-regression, we propose a molecule recovery task with regional masks for the first two challenges. The masked elements are recovered by other surrounding visible elements, which helps the model generate valid molecules. These masked elements are also expected to be predicted by the given product, encouraging the model to follow the conservation of atoms. Additionally, it is widely accepted that the reaction type as prior knowledge greatly improves the performance of retrosynthesis. Thus, we propose a supervised contrastive task in PMSR to force the model to focus more on reaction centers. Our main contributions can be summarized as follows.

- We summarize three challenges of single-step retrosynthesis prediction and propose three solutions to these challenges, i.e., masked element recovery, masked fragment recovery, and reaction classification.
- We design three pre-training tasks customized to retrosynthesis, including auto-regression, molecule recovery, and contrastive rection classification. The three pre-training tasks solve the three challenges and improve the performance of retrosynthesis. We also introduce the pointer-generator architecture and data augmentation in PMSR, both of which further benefit retrosynthesis.
- After fine-tuning on USPTO-50K (Schneider, Stiefl, and Landrum 2016), USPTO-FULL (Dai et al. 2019) and Pistachio (Mayfield, Lowe, and Sayle 2017), our model surpasses previous methods by a large margin. We also conduct experiments to generate all precursors, including reactants and reagents, on USPTO-50K (Schneider, Stiefl, and Landrum 2016) and USPTO-MIT (Jin et al. 2017). PMSR also achieves satisfactory results, which shows the power of our pre-training tasks.

[*] Corresponding authors.

Related Work

Single-Step Retrosynthesis Prediction

Template-Based Methods Template-based retrosynthesis prediction aims to prioritize different reaction templates in different ways. RetroSim (Coley et al. 2017) compares the molecular similarity to select templates. NeuralSym (Segler and Waller 2017) constructs a classification task to choose templates. GLN (Dai et al. 2019) maximizes the conditional joint probability of both templates and the reactants using their learned graph embeddings. Additionally, DualTB (Sun et al. 2021) introduces an energy-based model in GLN. MHN (Seidl et al. 2021) uses modern Hopfield networks to associate different molecules and templates, which improves the performance of template relevance prediction. LocalRetro (Chen and Jung 2021) tries to extract more general templates only with local information. All these methods suffer from the low generalization of templates as well as the huge and increasing number of templates.

Semi-Template-Based Methods Semi-template-based methods resort to the reaction centers identified by Atom-Atom-Mappings (AAM), in which atoms in a product is mapped to those in the corresponding reactants. RetroExpert (Yan et al. 2020), G2Gs (Shi et al. 2020) and GraphRetro (Somnath et al. 2021) predict reaction centers to generate synthons first and then complete synthons to reactants. MEGAN (Sacha et al. 2021) modifies the product step by step to generate reactants with a graph-to-sequence model. GTA (Seo et al. 2021) generates reactants via a transformer trained with the cross-attention MSE calculated by AAM. While all the above methods rely on the correctness of AAM, automated AAM remains an open problem in chemistry (Jaworski et al. 2019; Schwaller et al. 2021a).

Template-Free Methods Template-free methods train sequence-to-sequence models to generate SMILES strings of reactants directly. MT (Schwaller et al. 2020, 2019) firstly uses Transformer in retrosynthesis prediction. SCROP (Zheng et al. 2019) designs a syntax correcter to improve the correctness of generated strings. DualTF (Sun et al. 2021) formulates retrosynthesis by energy-based models and adds an additional loss of forward prediction. Our method aims to overcome the particular challenges in template-free methods summarized in the Introduction by introducing more chemical information to the model.

Chemical Pre-training

Transformer (Vaswani et al. 2017) is wildly used in the NLP area and has achieved tremendous success combined with the paradigm of pre-training. SMILES representation of chemical molecules opens the door to pre-train molecular transformer-based models. ChemBerta (Chithrananda, Grand, and Ramsundar 2020) transfers Roberta (Liu et al. 2019) to the chemical area. X-MOL (Xue et al. 2021), SMILES-BERT (Wang et al. 2019) and MolBert (Fabian et al. 2020) introduce chemical features to the transformer-based pre-training model. DMP (Zhu et al. 2021) pre-trains a transformer-based model together with a graph-based model. However, these models only work with molecules instead of reactions, so that they cannot learn the chemical rules in reactions. Rxnfp (Schwaller et al. 2021b) attempts to map the space of chemical reactions with a transformer-based pre-training encoder. T5Chem (Lu and Zhang 2022) is the most related work, which directly adapts the T5 framework and fine-tunes on multiple tasks. MolR (Wang et al. 2022) pre-trains a GNN encoder for molecules with reactions. Different from the above works, our work focuses on reaction-level tasks including retrosynthesis with reactions as training data; it is the first to pre-train a sequence-to-sequence model for reactions with chemically-targeted pre-training tasks.

Single-step Retrosynthesis

Sequence-to-Sequence Based Single-Step Retrosynthesis Prediction

We use Simplified Molecular Input Line Entry System (SMILES) (Weininger 1988), which represents a three-dimensional molecule formula as a one-dimensional string. As a result, a chemical reaction is represented as two strings – a precursor string and a product string.

We denote $(x, y) \in (\mathcal{X}, \mathcal{Y})$ as a chemical reaction, where $x = (x_1, x_2, \cdots, x_m)$ is the target product represented by a m-token SMILES string and the $y = (y_1, y_2, \cdots, y_n)$ denotes the precursors of the reaction with n tokens. In retrosynthesis prediction, the products $\mathcal{X}$ is the source domain and the precursors $\mathcal{Y}$ is the target domain. The objective function of a retrosynthesis prediction model is

$$\mathcal{L}(\theta; (\mathcal{X}, \mathcal{Y})) = -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})} \log P(y|x;\theta). \qquad (1)$$

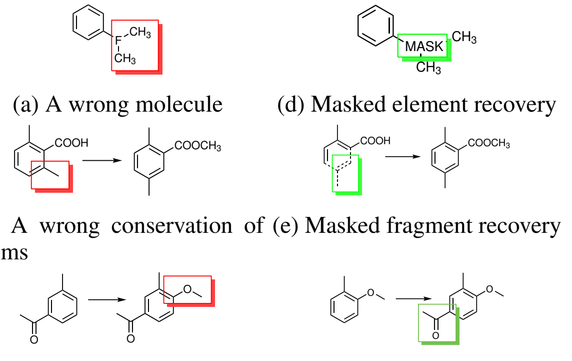
More specifically, the retrosynthesis prediction is more likely to be a conditional generation than a translation, because $\mathcal{X}$ and $\mathcal{Y}$ share the same vocabulary and all atoms in the product appear in the precursors in a reaction.

Challenges in Single-Step Retrosynthesis

We evaluate several single-step retrosynthesis models and thereupon pinpoint three major types of errors in single-step retrosynthesis, as shown in Figure 1.

Basic Chemical Rules of Molecules All molecules are expected to obey the basic valence bond theory. For example, a fluorine atom should never connect with three other atoms, as the invalid molecule shown in Figure 1(a). As a generative problem, it is necessary to learn chemical rules so that generated molecules are valid. However, these chemical rules are implicit and models do not produce valid molecules exactly according to the rules.

Conservation in Reactions We observe that many incorrect results generated by single-step retrosynthesis prediction models, especially by template-free models, do not allow *Law of Conservation of Atoms*. That is, the parts of the reactants other than the reaction center change in the reaction. For example, in Figure 1(b), the methyl group is mistakenly attached to the ortho position of the carboxyl group, which is completely different from the product.



(a) A wrong molecule

(b) A wrong conservation of atoms

(c) A wrong reaction center

(d) Masked element recovery

(e) Masked fragment recovery

(f) Reaction classification

Figure 1: Three kinds of errors in retrosynthesis and their solutions.

Selectivity of Reaction Centers The most challenging problem lies in selecting the correct reaction center which is a small region in a product. Given a target product, there usually exist several candidate reaction centers, but some of them being infeasible are against the mechanism of chemical reactions. On the benzene ring in Figure 1(c), a methoxy group cannot be attached to the para position of the acetyl group. Therefore, it is necessary for the model to also learn such chemical knowledge to avoid this kind of mistakes.

These challenges in single-step retrosynthesis prediction are all about chemical rules. In order to learn generalized and correct chemical rules, pre-training on a large dataset with chemically informed pre-training tasks serves as a viable solution. To this end, we propose a pre-trained model for single-step retrosynthesis.

Reagents Generation

Only Schwaller et al. (2020) attempted the concurrent prediction of reactants and reagents (*e.g.* solvents and catalysts). However, reagents are significant conditions of reactions. Both theoretically and practically, reagent prediction is necessary for reaction validation and automatic experiments. Meanwhile, it is more difficult to generate all precursors including reactants and reagents because of the variety of reagents.

Methodology

Solutions to the Challenges of Retrosynthesis

Generation of Valid Molecules The model that generates a wrong atom or a wrong bond in a certain position of a molecule is not desired. Thus, we design a masked element recovery task. Given a molecule, we mask some atoms and bonds of it and recover these masked elements by the model. This task helps the model learn the basic rules of molecules and avoid "illegal" atoms or bonds. For example, as shown in Figure 1(d), the model would not recover the masked element with a fluorine atom. In addition, more training data also improve the quality of generated molecules.

Conservation of Products and Precursors In retrosynthesis prediction, the most cases that break the conservation law are that functional groups in the product are dislocated in the generated precursors. Therefore, we design a masked fragment recovery task, in which the model recovers masked consecutive elements of precursors given the corresponding product. Each masked segment contains several atoms and even functional groups, and the model needs to place them in a correct order according to the given product. Different from the masked element recovery task, the masked fragment recovery task encourages the model to keep the consistency between the product and the precursors. For example, while the methyl can attach to either the ortho or the meta position of the carboxyl group in Figure 1(e), only the meta position is reasonable according to the product.

Besides, as for rare atoms, we propose to use the pointer-generator (See, Liu, and Manning 2017; Nishida et al. 2019) which offers opportunities to copy an element directly from the product. The copying mechanism helps to generate some rare atoms that appear in the product rather than miss them.

Selection of Correct Reaction Centers Empirical results of all previous works (Dai et al. 2019; Coley et al. 2017) witness a considerable performance improvement after adding the information of reaction types. It is because the reaction type categorized by patterns of reaction centers (Schneider et al. 2016) provides invaluable insight into reaction centers, and the performance of retrosynthesis prediction is largely dependent on the correctness of the identified reaction center of a reaction. Thus, a reaction classification task suffices to teach the model to focus more on reaction centers – only if the model focus on reaction centers, it can correctly predict the reaction type. In Figure 1(f), the most likely reaction type is the C-C bond formation so that the methoxy group is almost impossible to act as the reaction center.

Data Augmentation

Single-step retrosynthesis prediction usually uses the canonical SMILES strings (Schwaller et al. 2019) to limit the randomness of generation. However, Tetko et al. (2020) proposed that random representation of molecules could be an augmentation method for retrosynthesis. More training data also helps the model to learn chemical rules. Further, we find that canonicalized SMILES representation does not maintain the consistency of the same sub-graph in different molecules. In other words, canonicalized SMILES benefits the generation of unique results but harms the learning of structural information from SMILES strings. During pre-training, we thus provide different SMILES strings of the same structure, which helps the model to learn the equivalence between different strings. We follow the augmentation method proposed in Tetko et al. (2020) by including random SMILES representations of products, precursors, and reverse precursors and products.

An Overview of PMSR

As shown in Figure 2, PMSR is a sequence-to-sequence chemical reaction model with a transformer-based encoder and a transformer-based decoder. We argue that the retrosynthesis prediction task is more like conditional generation,

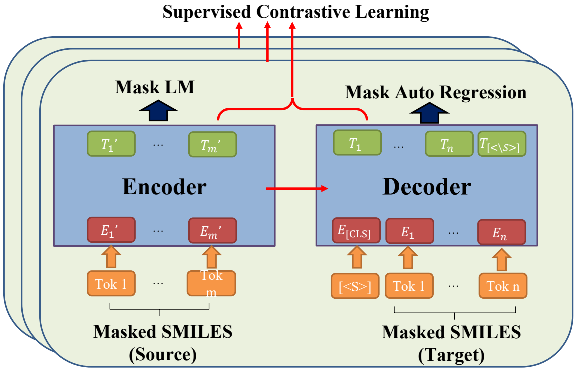


Figure 2: An overview of PMSR.

so that we adopt the pointer-generator (See, Liu, and Manning 2017; Nishida et al. 2019) in our model. The pointer-generator architecture allows the model to copy atoms from a product directly, improving the generation of rare atoms.

We design three pre-training tasks, i.e., molecule recovery (MR), auto-regression (AR) and contrastive classification (CC). **MR** and **AR** are pre-trained on both the encoder and decoder; **CC** is learned in a batch of data, which is like regularization of the model. These three pre-training tasks are optimized simultaneously.

Pre-training Tasks and Fine-tuning

Auto-regression Our pre-training data are purely reactions, so that we first design a supervised auto-regression task which is the same as our downstream tasks of retrosynthesis. In the auto-regression task, we have

$$\mathcal{L}_{AR} = -\log P(y|x) = -\sum_{t=1}^{n} \log P(y_t|y_{<t}, x), \qquad (2)$$

where x is the SMILES string of the product and y is the sequence of the precursors with n tokens.

Molecule Recovery We consider the masked element recovery and masked fragment recovery tasks in molecule recovery jointly. Concretely, we use the span-mask (Joshi et al. 2020) to generate masks in lengths of $[1, 10]$ for each molecule. Masks are later applied to SMILES strings of molecules. A one-token mask covers one atom or one bond, targeting masked element recovery. A multi-token mask covers a part of SMILES strings, which hides *at least* one fragment of a molecule. The decoder predicts the masked tokens from the given product and unmasked parts of precursors. In Figure 3, a 6-token mask can be recovered by the product and other parts of the benzene ring of the reactant. In this way, the decoder tends to generate precursors under the chemical rules. Together with auto-regression, the loss of the decoder is defined as

$$\mathcal{L}_{dec} = \mathcal{L}_{AR\&MR} = -\log P(y|x) = -\textstyle\sum_{t=1}^{n} \log P(y_t|\tilde{y}_{<t}, x), \qquad (3)$$

where $\tilde{y}$ indicates the masked precursors.

In order to increase the generalization of the encoder, we also add an MR task on the encoder. The encoder recovers

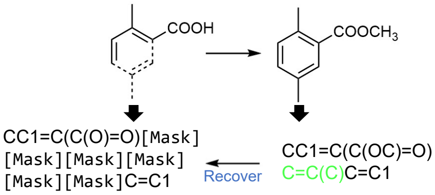


Figure 3: Molecule recovery for the decoder

the masked tokens by understanding the context and grammar of source sequences during pre-training. The loss function of the encoder is

$$\mathcal{L}_{enc} = \mathcal{L}_{MR} = -\log P(x|\tilde{x}) = -\sum_{p=1}^{s} \log P(x_{t_p}|\tilde{x}), \qquad (4)$$

where $\tilde{x}$ is the masked version of x with s masked tokens and t_p indicates the position of the p-th masked token.

Contrastive Classification Previous transformer-based methods, like MT (Schwaller et al. 2020), SCROP (Zheng et al. 2019) and DMP (Zhu et al. 2021), did not consider reaction centers which however are key to retrosynthesis prediction. As described before, we design a contrastive classification task to help the model learn the features of different reaction centers. We do not formulate a classification task for each reaction directly, considering that contrastive learning is more robust to corruptions; many reactions are labeled as unrecognized which mismatch all type templates of NameRXN (NextMoveSoftware 2021). The contrastive loss enforces all reactions in the same type to have similar embeddings, and the model to focus on reaction centers. In contrastive classification, features of the product are extracted by the encoder, i.e., $r_{src} = \mathrm{mean}(encoder([x_1, x_2, \cdots, x_m]))$. Similarly, features of precursors are extracted by the decoder, i.e., $r_{tgt} = \mathrm{mean}(decoder([y_1, y_2, \cdots, y_n]))$. Afterwards, we combine these two parts by concatenation, i.e., $r = \mathrm{concatenate}(r_{src}, r_{tgt})$. Following Khosla et al. (2020), we add a fully-connected layer as a projection layer by $z = \mathrm{FC}(r)$. The contrastive classification loss, therefore, is

$$\mathcal{L}_{CC} = \textstyle\sum_{i \in I} \mathcal{L}_{sup,i} = \sum_{i \in I} \frac{-1}{|C(i)|} \sum_{c \in C(i)} \log \frac{\exp(z_i \cdot z_c / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)}, \qquad (5)$$

where i is the i-th sample in the batch I, $A(i) = I \backslash i$, and $C(i) = \{c \in A(i) : type_c = type_i\}$ is the set of all other samples in batch I sharing the same reaction type with the i-th sample.

Fine-tuning on Downstream Tasks During fine-tuning, we share all parameters of the encoder and decoder and only keep the projection head of auto-regression. Molecule recovery and contrastive classification are removed in fine-tuning.

Experiments

PMSR Pre-training

Dataset We pre-train our model on Pistachio (Mayfield, Lowe, and Sayle 2017), which is automatically extracted from U.S., European and WIPO patents, including 13.3 million reactions. We remove invalid and redundant reactions and pick out all reactions which contain the same products as the test set of fine-tuning datasets to avoid data leaks. Then, we split the remaining data into a training set with 3.74M reactions and a validation set with 0.2M reactions. We augment training data 100 times. All input data are processed by RDKit toolkit (Landrum 2021). We do not remove reagents during pre-training so that our model has a chance of fine-tuning on retrosynthesis prediction with or without reagents. We use the super-class classified by NameRXN (NextMoveSoftware 2021) for the contrastive classification.

Model Architecture Our main transformer architecture consists of a 6-layer encoder and a 6-layer decoder with 768 embedding size, 2048 feed-forward filter size and 8 attention heads. We also evaluate an 8-layer PMSR on USPTO50K.

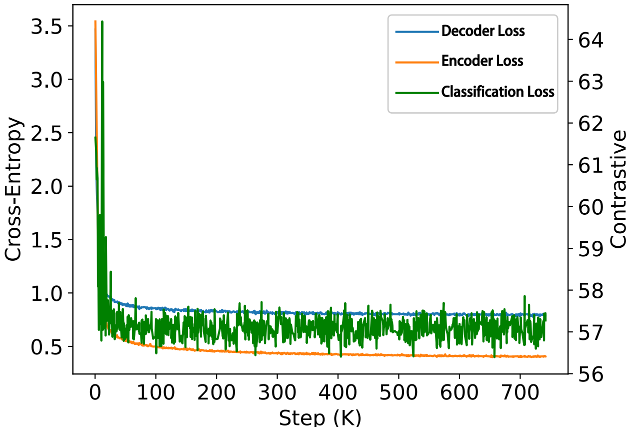


Figure 4: Loss curves during pre-training

Pre-training Details In the encoder, we use cross-entropy loss on masked tokens. Besides, we use label-smoothed cross-entropy loss with a label-smoothing factor of 0.1 in auto-regression and molecule recovery task of the decoder. In contrastive classification, the projection layer has 2048 hidden units, and we regularize the contrastive loss by a weight of 0.1. We use Adam optimizer (Kingma and Ba 2015) and vary the learning rate with Noam (Vaswani et al. 2017) schedule with 8000 warm-up steps. The pre-training process runs on 8 NVIDIA A100 GPU cards for 740K steps and the batch size is 14000 tokens. The loss curves of the pre-training are shown in Figure 4.

Single-Step Retrosynthesis Prediction without Reagents

Dataset We fine-tune our pre-trained model on single-step retrosynthesis prediction without reagents. We conduct experiments on USPTO-50K (Schneider, Stiefl, and Landrum 2016; Coley et al. 2017; Liu et al. 2017) and USPTO-full (Dai et al. 2019; Yan et al. 2020). We split the datasets in the same way as Dai et al. (2019). Additionally, we also evaluate our model on Pistachio (Mayfield, Lowe, and Sayle 2017) which contains more data than USPTO sets.

Evaluation Metrics We use top-k accuracy as our evaluation metrics. Following previous works (Coley et al. 2017; Dai et al. 2019; Shi et al. 2020; Seo et al. 2021), we compute top-k ($k = 1, 3, 5, 10$) accuracy by comparing whether one of the top-k generated results exactly match the ground-truth reactants in canonical format.

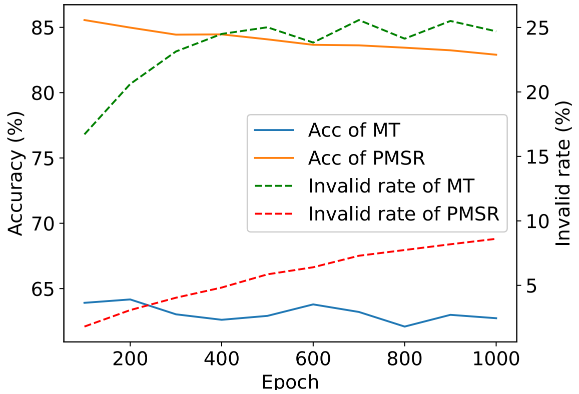


Figure 5: Changes of top-10 accuracy and invalid rate with the training on USPTO-50K

Results on USPTO-50K We fine-tune on USPTO-50K with reaction type known and unknown for 100 epochs with a learning rate of 5×10^{-4}. We use cross-entropy loss and decay the learning rate with Noam schedule (Devlin et al. 2019). We report our results in Table 1. All compared baselines are described in Related Work. Under the setting of unknown reaction type, PMSR not only outperforms all previous template-free methods, but also exceeds template-based methods and semi-template-based methods on Top-1 and Top-3 accuracy. We would highlight that PMSR has the smallest gap of given reaction class as a prior or not, which proves our model places more attention on reaction centers after pre-training. Even without the information on reaction type, our model still mines the hint of the reaction center from the input and predicts correctly. Compared with other pre-trained models (DMP and T5Chem), our pre-training tasks are more targeted to retrosynthesis prediction. In reaction class known cases, our method still far surpasses all template-free models on top-1 and top-10 accuracy. We should admit that the other two kinds of methods use the information of reaction class more directly during template retrieval or reaction center selection. However, real-world application scenarios are more in line with the former setting.

To show PMSR learns more chemical rules, we train 1000 epochs on USPTO-50K and trace the changes of the ratio of invalid generated top-10 reactants. As illustrated in Figure 5, both MT and PMSR overfit the training data as the top-10 accuracy on the test set is decreasing. However, the invalid rate of SMILES strings generated by MT is much higher than that generated by PMSR. Besides, we ask chemists to check 300 top-1 results generated by PMSR, MT and MEGAN which are different from the ground-truth. 99% results of PMSR keep the conservation, while the numbers of MT and MEGAN are 88% and 98%. 48% of re-

Model	Top-k Accuracy(%)							
	Reaction class unknown				Reaction class known			
$k =$	1	3	5	10	1	3	5	10
Template-based								
RetroSim	37.3	54.7	63.3	74.1	52.9	73.8	81.2	88.1
NeuralSym	44.4	65.3	72.4	78.9	55.3	76.0	81.4	85.1
GLN	52.5	74.6	80.5	86.9	64.2	79.1	85.2	90.0
MHN	50.5	73.9	81.0	87.9	–	–	–	–
LocalRetro	53.4	**77.5**	**85.9**	**92.4**	–	–	–	–
DualTB	**55.2**	74.6	80.5	86.9	**67.7**	**84.8**	**88.9**	**92.0**
Semi-template-based								
G2Gs	48.9	67.6	72.5	75.5	61.0	81.3	86.0	88.7
RetroXpert	50.4	61.1	62.3	63.4	62.1	75.8	78.5	80.9
GTA	51.1	67.6	74.8	81.6	–	–	–	–
GraphRetro	**53.7**	68.3	72.2	75.5	**63.9**	81.5	85.2	88.1
MEGAN	48.1	**70.7**	**78.4**	**86.1**	60.7	**82.0**	**87.5**	**91.6**
Template-free								
MT	42.3	61.9	67.5	72.9	54.2	73.6	78.3	81.3
SCROP	43.7	60.0	65.2	68.7	59.0	74.8	78.1	81.1
DMP	46.1	65.2	70.4	74.3	57.5	75.5	80.2	83.1
T5Chem	46.5	64.4	70.5	–	–	–	–	–
DualTF	53.6	70.7	74.6	77.0	65.7	81.9	84.7	85.9
PMSR (6-layer)	59.3	76.9	81.8	85.6	66.4	81.6	84.8	87.0
PMSR (8-layer)	**62.0**	**78.4**	**82.9**	**86.8**	**67.1**	**82.1**	**85.2**	**87.3**

Table 1: Top-k exact match accuracy on USPTO-50K

sults choose the same reaction center with the ground-truth, while only 28% and 36% answers generated by MT and MEGAN select the same reaction center. These results prove that PMSR learns more chemical rules and pre-training decreases wrong predictions caused by the challenges.

Results on USPTO-full We also evaluate PMSR on a larger dataset USPTO-full. We fine-tune the model with the same learning rate with USPTO-50K for 450 epochs. The results of baselines are copied from their corresponding paper. As shown in Table 2, the accuracy of template-based methods and semi-template-based methods decreases because of the larger number of templates and the higher error rate of AAM. Template-free methods, especially our method, exhibit excellent generalization. The only exception is GTA (Seo et al. 2021) as it only uses AAM mildly. However, PMSR also outperforms all template-free methods here.

Model	Top-k Accuracy(%)			
$k =$	1	3	5	10
Template-based				
RetroSim	32.8	–	–	56.1
NeuralSym	35.8	–	–	60.8
GLN	39.3	–	–	63.7
Semi-template-based				
MEGAN	33.6	–	–	63.9
GTA	**46.0**	–	–	**70.0**
Template-free				
MT	42.9	59.0	62.4	66.8
DMP	45.0	59.6	63.9	67.9
PMSR	**45.5**	**60.9**	**65.5**	**70.1**

Table 2: Top-k exact match accuracy on USPTO-full

Results on Pistachio Due to the poor quality of AAM in Pistachio, we only compare two baseline methods. The results on Pistachio show the strong performance of our model cannot be attributed solely to using more data. Training with the same number of data, PMSR still achieves state-of-the-art performance.

Single-Step Retrosynthesis Prediction with Reagents

Dataset We fine-tune on USPTO-50K (Schneider, Stiefl, and Landrum 2016) and USPTO-MIT (Jin et al. 2017) to evaluate the performance of single-step retrosynthesis prediction with reagents. We process USPTO-50K and USPTO-MIT following Dai et al. (2019) to resolve multi-product reactions into single-product reactions. We split USPTO-50K into a training set with 40012 reactions, a validation set with 5000 reactions and a test set with 4997 reactions randomly. The separation of USPTO-MIT is the same as Jin et al. (2017).

Model	Top-k Accuracy(%)			
$k =$	1	3	5	10
RetroXpert	39.2	48.8	50.9	52.5
MT	39.4	55.6	60.5	64.7
PMSR	**45.3**	**61.7**	**66.1**	**69.6**

Table 3: Top-k exact match accuracy on Pistachio

Dataset	Model	Top15(%)	RT(%)	Cov.(%)	ismi(%)
50K	MT	14.4	57.7	96.3	13.3
	PMSR	30.9	89.5	99.8	1.1
MIT	MT	30.0	80.4	99.7	1.5
	PMSR	31.5	90.8	99.7	0.3

Table 4: Experimental Results of Single-step Retrosynthesis Prediction with reagents

Evaluation Metrics Many reagents are replaceable, so we follow Schwaller et al. (2020)'s work to evaluate results with a forward prediction model. We train a forward prediction model on Pistachio (Mayfield, Lowe, and Sayle 2017) and generated precursors can be validated by back-translation. We compute round-trip accuracy (RT) and coverage (Cov.) of top-15 results as Schwaller et al. (2020). Round-trip accuracy measures the percentage of generated precursors that can convert to the target product. Coverage quantifies the ratio of target products for which the retrosynthesis model produces at least one valid candidate. Besides, we still report the ratio of invalid molecules in top-15 candidates generated by models.

We compare the baseline of the original transformer (MT) (Schwaller et al. 2020) with our PMSR model. As shown in Table 4, top-15 accuracy proves PMSR fits the distribution of the dataset better. Round-trip accuracy shows that PMSR generates more valid reactions, which means that PMSR understands *Law of Conservation of Atoms* deeper. As same as the experiment without reagents, almost all molecules generated by PMSR are valid, whereas MT generates 13.3% invalid molecules in top-15 candidates on USPTO-50K.

Ablation Study

We first evaluate the influence of each pre-training task. As shown in Table 5, after adding masks, top-10 accuracy increases by 2.7 points and after adding contrastive loss, top-10 accuracy is further improved by 2.9 points. The result shows the effectiveness of our pre-training tasks.

Additionally, we also study the necessity of the pre-training and fine-tuning scheme. We train a model from scratch with pre-training data and the training data of USPTO-50K and get poor performance. This shows the performance cannot be improved only by increasing training data. Our model can also adapt to different distributions of reactions.

At last, we would highlight that PMSR shows more power on retrosynthesis with more layers. An 8-layer PMSR outperforms all baselines on USPTO-50K as shown in Table 1.

Pre-training task	Top-k Accuracy(%)			
$k =$	1	3	5	10
AR	55.8	72.9	77.6	80.0
AR+MR	57.8	76.2	80.5	82.7
AR+MR+CC	**59.3**	**76.9**	**81.8**	**85.6**

Table 5: Ablation study of pre-training tasks on USPTO-50K

Different Data Distributions between Pre-training and Downstream Tasks

We have evaluated several downstream tasks to show the generalization of our pre-trained model. However, we still want to further consolidate the robustness of PMSR when the data distribution of pre-training is obviously different from the downstream task. We increase the gap between the pre-training data and the downstream task, resulting in (1) the small (our default setting) where we remove pre-training reactions that contain the same products as the test set of fine-tuning datasets, (2) the medium where we remove the whole USPTO-50K from our pre-training dataset Pistachio, and (3) the large where we remove all reactions in USPTO during pre-training. PMSR still achieves similar performance as shown in Table 6, which shows PMSR is robust to different data domains and not only dependent on more similar reactions in pre-training data.

Distribution Gap	Top-k Accuracy(%)			
$k =$	1	3	5	10
Large (without USPTO)	57.1	74.1	79.1	83.7
Medium (without 50K)	59.3	76.9	81.7	85.1
Small (full)	**59.3**	**76.9**	**81.8**	**85.6**

Table 6: Domain Adaptivity on USPTO-50K

Conclusion

In this paper, we first summarize three challenges of single-step retrosynthesis prediction. Then, we propose PMSR, a pre-trained model for single-step retrosynthesis. We formulate retrosynthesis prediction as a conditional generation problem and construct a transformer-based model with a pointer generator. After that, we design three pre-training tasks, auto-regression, molecule recovery and contrastive classification, customized to the challenges of retrosynthesis. We fine-tune PMSR on different datasets, and our method outperforms the baseline by a large margin on retrosynthesis with or without reagents. Further work can aim to extend more applications of our pre-training model, like reagent completion and reaction performance prediction.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (U20A20387, No. 62006207, No. 62037001), Project by Shanghai AI Laboratory (P22KS00111), Program of Zhejiang Province Science and Technology (2022C01044), the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), Fundamental Research Funds for the Central Universities (226-2022-00142, 226-2022-00051).

References

Chen, S.; and Jung, Y. 2021. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10): 1612–1620.

Chithrananda, S.; Grand, G.; and Ramsundar, B. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Coley, C. W.; Rogers, L.; Green, W. H.; and Jensen, K. F. 2017. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12): 1237–1245.

Corey, E. J.; and Wipke, W. T. 1969. Computer-assisted design of complex organic syntheses. *Science*, 166(3902): 178–192.

Dai, H.; Li, C.; Coley, C.; Dai, B.; and Song, L. 2019. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, 8872–8882.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186.

Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M. H. S.; Meyers, J.; Fiscato, M.; and Ahmed, M. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *CoRR*, abs/2011.13230.

Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; and Grzybowski, B. A. 2019. Automatic mapping of atoms across both simple and complex chemical reactions. *Nature communications*, 10(1): 1–11.

Jin, W.; Coley, C.; Barzilay, R.; and Jaakkola, T. 2017. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *Advances in Neural Information Processing Systems*, 30.

Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8: 64–77.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems*, 33.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Landrum, G. 2021. RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org/ Accessed June 24, 2021.

Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; and Pande, V. 2017. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10): 1103–1113.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; and Zhang, Y. 2022. Unified Deep Learning Model for Multitask Reaction Predictions with Explanation. *Journal of Chemical Information and Modeling*, 62(6): 1376–1387.

Mayfield, J.; Lowe, D.; and Sayle, R. 2017. Pistachio: Search and faceting of large reaction databases. In *Abstracts of Papers of the American Chemical Society*, volume 254. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA.

NextMoveSoftware. 2021. NameRxn. https://www.nextmovesoftware.com/namerxn.html. Accessed: 2022-01-04.

Nishida, K.; Saito, I.; Nishida, K.; Shinoda, K.; Otsuka, A.; Asano, H.; and Tomita, J. 2019. Multi-style Generative Reading Comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2273–2284. Florence, Italy: Association for Computational Linguistics.

Sacha, M.; Błaz, M.; Byrski, P.; Dabrowski-Tumanski, P.; Chrominski, M.; Loska, R.; Włodarczyk-Pruszynski, P.; and Jastrzebski, S. 2021. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7): 3273–3284.

Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; and Landrum, G. A. 2016. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *Journal of medicinal chemistry*, 59(9): 4385–4402.

Schneider, N.; Stiefl, N.; and Landrum, G. A. 2016. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12): 2336–2346.

Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; and Laino, T. 2021a. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15): eabe4166.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9): 1572–1583.

Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; and Laino, T. 2020. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12): 3316–3325.

Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; and Reymond, J.-L. 2021b. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2): 144–152.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1073–1083. Vancouver, Canada: Association for Computational Linguistics.

Segler, M. H.; and Waller, M. P. 2017. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25): 5966–5971.

Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Segler, M.; Wegner, J. K.; Hochreiter, S.; and Klambauer, G. 2021. Modern Hopfield Networks for Few-and Zero-Shot Reaction Template Prediction. *arXiv preprint arXiv:2104.03279*.

Seo, S.-W.; Song, Y. Y.; Yang, J. Y.; Bae, S.; Lee, H.; Shin, J.; Hwang, S. J.; and Yang, E. 2021. GTA: Graph Truncated Attention for Retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 531–539.

Shi, C.; Xu, M.; Guo, H.; Zhang, M.; and Tang, J. 2020. A Graph to Graphs Framework for Retrosynthesis Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, 8818–8827.

Somnath, V. R.; Bunne, C.; Coley, C.; Krause, A.; and Barzilay, R. 2021. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34.

Sun, R.; Dai, H.; Li, L.; Kearnes, S.; and Dai, B. 2021. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34.

Tetko, I. V.; Karpov, P.; Van Deursen, R.; and Godin, G. 2020. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1): 1–11.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, H.; Li, W.; Jin, X.; Cho, K.; Ji, H.; Han, J.; and Burke, M. 2022. Chemical-Reaction-Aware Molecule Representation Learning. In *International Conference on Learning Representations*.

Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; and Huang, J. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 429–436.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1): 31–36.

Xue, D.; Zhang, H.; Xiao, D.; Gong, Y.; Chuai, G.; Sun, Y.; Tian, H.; Wu, H.; Li, Y.; and Liu, Q. 2021. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv*, 2020–12.

Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; and Huang, J. 2020. RetroXpert: Decompose Retrosynthesis Prediction like A Chemist. In *Advances in Neural Information Processing Systems*, 8872–8882.

Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; and Yang, Y. 2019. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1): 47–55.

Zhu, J.; Xia, Y.; Qin, T.; Zhou, W.; Li, H.; and Liu, T.-Y. 2021. Dual-view Molecule Pre-training. *arXiv preprint arXiv:2106.10234*.