LEARNING SINGLE-STEP RETROSYNTHESIS WITH PSEUDO-REACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrosynthesis analysis aims to design reaction pathways and intermediates for a target compound. Emerging works have been developed to automate this process using artificial intelligence (AI). Although there are multiple synthesis pathways to synthesize one target product, most of the existing models were optimized to select only one of them. One of the potential reason is due to the absence of other potential reactions in the available reaction dataset, as one product usually only appears one time in the reaction dataset. In this work, we generate virtually validated pseudo-reactions using local reaction templates and reaction outcome prediction model to optimize the retrosynthesis model to predict multiple synthesis pathways. With the aid of newly generated pseudo-reactions, the top-10 exact match accuracy is increased from 93.1% to 94.2% and the top-10 round trip accuracy is increased from 81.3% to 87.6% with higher prediction confidence and diversity on a public reaction dataset.

1 INTRODUCTION

Retrosynthesis, referring to the process of decomposing a target molecule to its precursors, has been studied for decades since the first formalization by Corey (1991). Along with the advances of machine learning (ML)-based inverse molecular design, the number of ML based retrosynthesis models, including single-step and multi-step retrosynthesis, is increasing rapidly to bring the virtually designed molecules into real world (Coley et al. (2018); de Almeida et al. (2019)). While the total synthesis of most molecules requires multiple synthesis steps, the quality of multi-step retrosynthesis model is highly dependent to the quality of single-step retrosynthesis model since the multi-step model repeatedly applies single-step models to search the building blocks (Hassen et al. (2022)).

In general, single-step retrosynthesis models can be generally categorized into template-based (Coley et al. (2017); Segler & Waller (2017); Dai et al. (2019); Chen & Jung (2021); Seidl et al. (2022)) and template-free models (Liu et al. (2017); Tetko et al. (2020); Tu & Coley (2022); Zhong et al. (2022)) depending on using reaction templates or not. Template-based models classify which of the pre-defined reaction templates should be applied to a given product to obtain its precursors. For example, Dai et al. (2019) proposed Graph Logic Network (*GLN*) to maximize the likelihood of predicting the correct reaction template among a set of pre-defined reaction templates for a given product. Chen & Jung (2021) trained *LocalRetro* to learn the potential reaction center and the local reaction template to be applied to the given product. On the other hand, template-free models predict the precursors without pre-defining reaction templates. Most of the template-free models are based on sequence translation model, which "translate" the SMILES string of the given product to SMILES strings of its reactants (Liu et al. (2017); Tetko et al. (2020); Zhong et al. (2022)).

Although the existing models have shown promising results, most of the methods formulate the retrosynthesis problem as a one-to-one classification problem, which train the model to select one of the pre-defined templates (template-based) or tokens (template-free). Given the fact that one product has more than one potential synthesis pathway, retrosynthesis prediction is naturally a one-to-many classification problem. As only one of the reactant is given for a single product in most of the reactions in the dataset, training a retrosynthesis model by one-to-many classification is a challenging task due to the lack of other synthesis knowledge.

In this work, we generate virtually validated pseudo-reactions to fill the blank of potential retrosynthesis pathways in the reaction dataset using local reaction template (LRTs, Chen & Jung (2021)) and *LocalTransform* (Chen & Jung (2022)), and use these reactions to optimize the model to select multiple reaction templates at the predicted reaction center. By training LocalRetro (Chen & Jung (2021)) with pseudo-reactions, we show the potential direction of improving the diversity as well as the predictive performance of retrosynthetic models. Overall, the top-10 exact match accuracy is increased by 1.1% by masking the loss from pseudo-reactions, and the top-10 round trip accuracy is increased by 6.3% by imposing extra losses from pseudo-reactions.

2 METHODS

2.1 PSEUDO-REACTION GENERATION

While most of the product in the reaction dataset has only one or few records of synthesis pathway, it is challenging to train a retrosynthesis model with multiple reactions. In this section, we introduce a novel method to generate virtually feasible pseudo-reactions by enumerating local reaction templates (LRTs) at the known reaction center and validating the generated reactions by the reaction outcome prediction model *LocalTransform* (Chen & Jung (2022)).

For each chemical reaction in the dataset, an LRT can be extracted by comparing the changes before and after the reaction to represent the net changes of the chemical reaction. By extracting LRTs from all the reactions in a dataset, we build an LRT library and use this library to generate pseudo-reactions from existing reactions. Using this LRT library, we enumerate applicable LRTs found in the library at the same reaction center to generate pseudo-reactions. Considering the synthesizability of pseudo-reactants, we only generate pseudo-reactions using similar LRT, which yields at least one identical reactant in the original reaction.

For example, the C-C bond between two aromatic rings in Figure 1a is formed by Suzuki coupling using boronic acid and bromide in the reaction dataset. By searching the LRT library derived from reaction dataset, such C-C bond between two aromatic rings can also be formed by similar LRTs either changing the halide to chloride or iodine or changing the boronic acid to a magnesium ion (1+) or trimethylstannyl group from the original LRT. With these similar LRTs, we generate four pseudo-reactions that can potentially be used to synthesize the same product. To ensure these pseudo-reactions are chemically plausible, we use LocalTransform to filter out reactions if the predicted product is different from the original product. To avoid potential stereo hindrance, we limit the number of atoms in the generated pseudo-reaction no more than two atoms than the original reaction.

2.2 LOSS-MASKING AND PSEUDO-LABELING

In this paper, we use *LocalRetro* (Chen & Jung (2021)) to predict the reactants from a given product. Because the score generated from original LocalRetro is the probability of LRT applied on the predicted reaction center, the score does not have the physical meaning of reaction feasibility. Here, we binarize the prediction score of each reactant set R of a given product O to a binary probability $p(R|O)$, which is the product of the probability of selecting a reaction center for a given product $p(o|O)$ and probability of local reaction template predicted at the reaction center $p(T|o)$.

$$p(R|O) = p(T|o) * p(o|O) \tag{1}$$

The total training loss $Loss_{total}$ is the average of the loss of predicting reaction center $Loss_{RC}$ and the loss of predicting local reaction template $Loss_{LRT}$.

$$Loss_{total} = mean(Loss_{RC}, Loss_{LRT}) \tag{2}$$

$Loss_{RC}$ is the binary cross entropy between labeled reaction center y_n and predicted probability of each atom or bond being reaction center p_n,

$$Loss_{RC} = \sum_{n=1}^{N} -(y_n \log(p_n) + (1 - y_n) \log(1 - p_n)) \tag{3}$$

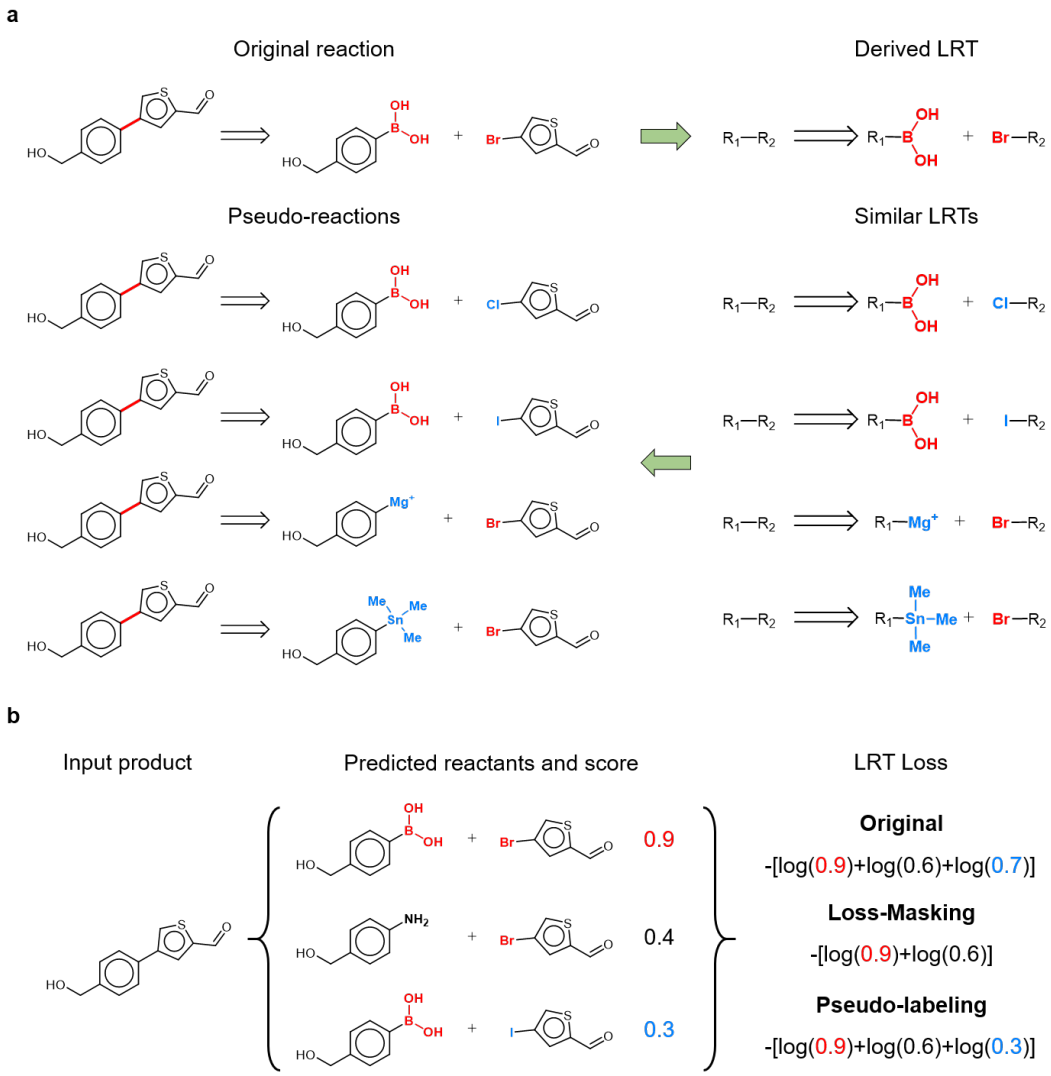

Figure 1: Pseudo-reactions generation and the training loss of three different training settings. (a) The pseudo-reactions are generated by enumerating similar LRTs at the known reaction center, yielding at least one identical reactant with the original reaction. (b) Different training losses are used to optimize the model prediction in different training settings.

and $Loss_{LRT}$ is the binary cross entropy between labeled LRT y_t and predicted probability of each LRT p_t applied at the known reaction center.

$$Loss_{LRT} = \sum_{t=1}^{T} -(y_t \log(p_t) + (1 - y_t) \log(1 - p_t)) \tag{4}$$

While there are potentially multiple reaction centers and LRTs can be applied to generate feasible reactants for given product molecule, only one reaction center and one LRT is known from the dataset. Therefore, labeling all unknown labels to negative data may also lead to a large number of false-negative reactions, which will discourage the prediction of potential synthesis pathways. To resolve this issue, we introduce the pseudo-reactions obtained from the previous section train LocalRetro using two different settings, namely loss-masking and pseudo-labeling.

In the loss-masking setting, we mask the loss of a LRT if it can be used to generate plausible pseudo-reactions. Thus, the $Loss_{LRT}$ of pseudo-reactions is zeroed to avoid imposing incorrect penalty when the model predicts pseudo-reactions. By masking the loss of potential LRTs, the model is not

discouraged to predict potential LRT even it does not yield the existing reactions in the dataset. This setting is conceptually similar to positive-unlabeled learning (PU learning, Elkan & Noto (2008); Jang et al. (2020)). In the pseudo-labeling setting, we turn the negative LRT labels into positive labels if the LRTs can be used to generate one of the pseudo-reactions. Therefore, the $Loss_{LRT}$ of pseudo-reactions is increased if the model predicts low score on pseudo-reactions. By enriching the positive labels with pseudo-reactions, the model is trained to be aware of both existing reaction and pseudo reactions. An example of different training losses under different training settings is shown in Figure 1b.

To investigate the effect of introducing different number of pseudo-reactions to train LocalRetro, we sample pseudo-reactions with probability 0.1 and 0.9. Since each model is trained with a different set of sampled pseudo-reactions, the final prediction scores are obtained by averaging the prediction scores generated from ensemble modeling. We used averaged the prediction scores from 6 models for all the experiments shown in this paper.

3 EXPERIMENTS

3.1 DATASET

We evaluate our methods on USPTO-50K dataset (Schneider et al. (2016)) with the same training/validation/test splits described in Coley et al. (2017). We extracted 658 LRTs from the 40K reactions in the training set of USPTO-50K dataset, and generated 49,707 pseudo-reactions from the training set. For the experiments focusing on rarely seen LRTs, we define LRTs having reaction popularity fewer than average (60.8) as rare LRTs and reactions using these LRTs as "rare reactions". A total number of 478 rare LRTs are found, which is 72.6% of the total number of LRTs but only cover 12% of training reactions and 11% of test reactions.

3.2 EVALUATION METRIC

In this paper, we use top-k exact match accuracy and round trip accuracy to evaluate the performance of LocalRetro and its variants. Top-k exact match accuracy calculates the whether the ground-truth reactant is predicted within top-k predictions by retrosynthetic models. In addition to top-k exact match accuracy, round trip accuracy (Schwaller et al. (2020)) is used to evaluate the feasibility of predicted reactants by comparing whether the product of predicted reactants predicted by reaction output prediction model matches the original product or not. We use LocalTransform to conduct reaction outcome prediction for round trip accuracy evaluation.

3.3 MAIN RESULTS

The top-k exact match accuracy of LocalRetro and its variants tested on USPTO-50K is shown in Table 1. Models trained with loss-masking with p=0.1 shows an 0.4 $\sim$1.1% improvement compared to the original model across all the k values. These improvements are more significant (0.6$\sim$4.0%) when focusing on rare reactions. Using a large portion of pseudo-reactions to train the model tends to decrease the top-k accuracy using both loss-masking and pseudo-labeling settings. Interestingly, training the models with pseudo-labeling decrease the top-k accuracy across all k values.

To evaluate the feasibility of the reactants predicted by LocalRetro and its variant models, we calculate the round trip accuracy and the number of predictions at different score thresholds within top-10 predictions in Table 2. In general, the round trip accuracy increases when using both loss-masking and pseudo-labeling models, and this improvement is larger when using the larger portion of pseudo-reactions during training. In particular, the model trained with pseudo-labeling (p=0.9) shows the best round trip accuracy. When predicting reactants for all products, LocalRetro with pseudo-labeling (p=0.9) shows a 6.3% higher overall round trip accuracy compared to original LocalRetro, and 27.6% of the predictions having prediction score greater than 0.5 shows a high 97.1% round trip accuracy. Among the predictions having score greater than 0.5, LocalRetro with pseudo-labeling (p=0.9) predicts 25 times more rare reactions and 10.3% higher round trip accuracy than original LocalRetro, which is a clear evidence of how pseudo-labeling can encourage the model to predict feasible rare reactions.

Table 1: Top-k exact match accuracy on USPTO-50K dataset. The highest accuracy is highlighted in bold font.

Reaction type	**Model**	**Top-k accuracy %**				
		k=1	3	5	10	50
All reactions	Original model	54.4	78.5	86.1	93.1	98.3
	Loss-masking (p=0.1)	**55.0**	**79.4**	**87.6**	**94.2**	**98.7**
	Loss-masking (p=0.9)	53.7	77.9	86.2	93.8	98.6
	Pseudo-labeling (p=0.1)	54.2	78.3	84.8	92.1	98.5
	Pseudo-labeling (p=0.9)	48.2	65.2	73.3	86.9	98.2
Rare reactions	Original model	25.0	44.9	54.6	72.9	95.0
	Loss-masking (p=0.1)	**26.9**	**48.1**	**58.6**	**75.7**	**95.6**
	Loss-masking (p=0.9)	26.1	44.2	55.5	73.4	**95.6**
	Pseudo-labeling (p=0.1)	25.2	45.1	53.5	68.8	94.7
	Pseudo-labeling (p=0.9)	22.5	34.3	44.5	60.9	92.6

Table 2: Top-10 round trip accuracy and prediction score on USPTO-50K dataset. The highest value is highlighted in bold font.

Reaction type	**Model**	**round trip accuracy % (ratio of predictions %)**		
		overall	score>0.1	score>0.5
All reactions	Original model	81.3 (100%)	89.0 (21.0%)	93.7 (5.3%)
	Loss-masking (p=0.1)	82.7 (100%)	89.3 (21.7%)	93.9 (5.6%)
	Loss-masking (p=0.9)	84.9 (100%)	91.4 (31.8%)	95.2 (9.1%)
	Pseudo-labeling (p=0.1)	85.6 (100%)	90.6 (26.6%)	94.5 (5.5%)
	Pseudo-labeling (p=0.9)	**87.6 (100%)**	**92.0 (67.6%)**	**97.1 (27.6%)**
Rare reactions	Original model	74.9 (23.8%)	78.1 (1.4%)	86.4 (0.2%)
	Loss-masking (p=0.1)	77.2 (23.8%)	78.2 (1.5%)	86.9 (0.2%)
	Loss-masking (p=0.9)	80.1 (21.9%)	84.3 (2.3%)	89.7 (0.3%)
	Pseudo-labeling (p=0.1)	80.8 (23.6%)	82.8 (1.8%)	87.1 (0.2%)
	Pseudo-labeling (p=0.9)	**83.5 (26.6%)**	**90.0 (15.1%)**	**96.7 (5.2%)**

The above results suggest that models training with small portion of loss-masking can boost both top-k exact match accuracy and round trip accuracy. On the other hand, although the models training with high portion of pseudo-labeling show decreased top-k exact match accuracy, training the models with large portion of pseudo-labeling is found to highly encourage the predictions toward rare predictions, generating confident and diverse, yet feasible synthesis pathways.

3.4 Prediction examples

To qualitatively understand the difference between LocalRetro and its variant models, we show an example of predicting the single-step retrosynthesis pathway of 2-methylbiphenyl in Figure 2. All of the LocalRetro models successfully predict the existed reactant in the dataset at rank 1 prediction, and the remaining predictions are predicted to yield the same target product validated by LocalTransform. Except model training with pseudo-labeling (p=0.9), the first 5 predictions are all same using the remaining four models. Notably, The prediction scores of the reactants predicted by original LocalRetro are all lower than 0.1, and training the models with pseudo-reactions is shown to increase the prediction score using both loss-masking and pseudo-labeling. Interestingly, while other models all suggest using Suzuki cross coupling to form the C-C bond between aromatic rings, model training with pseudo-labeling (p=0.9) predicts two reactants using cross electrophile coupling, which is rarely seen in the USPTO-50K dataset, to form the C-C bond with high prediction scores (0.391 and 0.353 for top-4 and top-5 prediction, respectively).

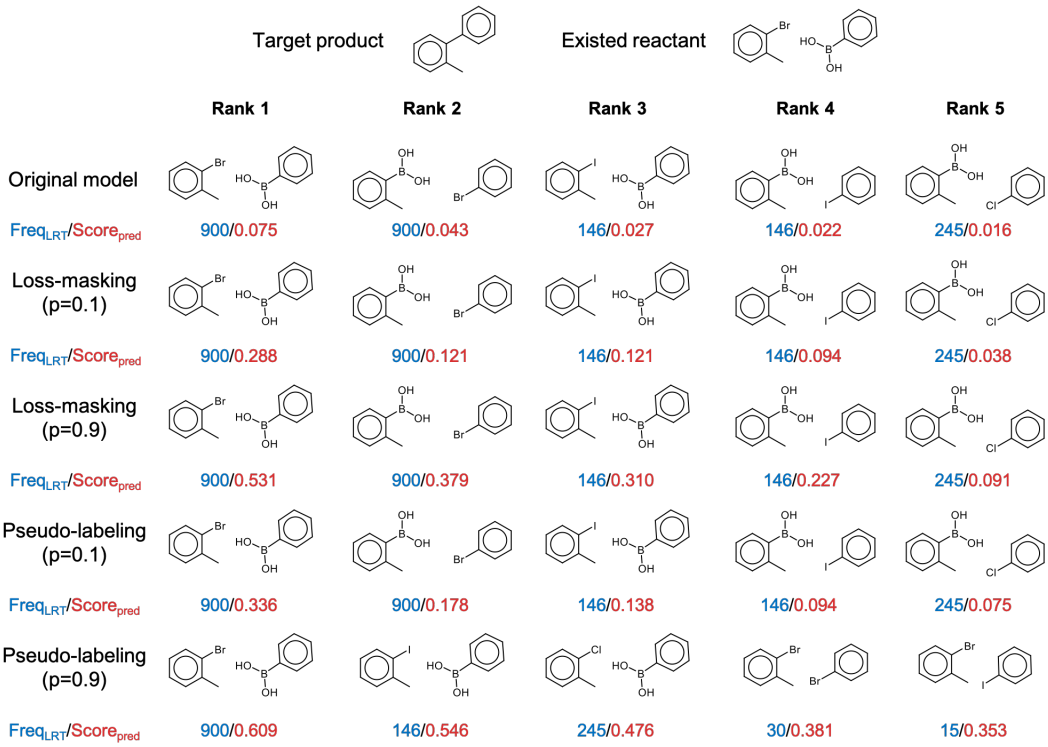

Figure 2: Example of predicting the single-step retrosynthetic pathway of 2-methylbiphenyl. The popularity of the predicted LRT in USPTO-50K train set ($Freq_{LRT}$) and the prediction scores ($Score_{pred}$) are display below the predicted reactant structures.

4 Conclusion

In this paper, we formulate the retrosynthertic prediction as a one-to-many classification problem and present two training settings, loss-masking and pseudo-labeling, to train LocalRetro models with pseudo-reactions. The top-k exact match accuracy, especially for the rare reactions, is increased at all k values when training the model by loss-masking. On the other hand, the model is highly encouraged to predict more diverse reactions with a strongly enhanced prediction score and round trip accuracy when the model is trained with pseudo-labeling. We suggest future retrosynthesis models may consider retrosynthesis prediction as a one-to-many classification problem and develop efficient methods to train the model with potential chemical synthesis data without actual experiments.

Author Contributions

S.C. and Y.J. conceived the project. S.C. designed the methods and performed the computational experiments and analyses. S.C. and Y.J. discussed the results and wrote the manuscript. Y.J. supervised the project.

Acknowledgments

We acknowledge the support from the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government under grant No. 2021-0-02068 (Artificial Intelligence Innovation Hub) and the National Science Foundation of Korea (NRF2019M3D3A1A01069099 and NRF-2019M3D1A1079303).

REFERENCES

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

Shuan Chen and Yousung Jung. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4(9):772–780, 2022.

Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

Connor W Coley, William H Green, and Klavs F Jensen. Machine learning in computer-aided synthesis planning. *Accounts of chemical research*, 51(5):1281–1289, 2018.

Elias James Corey. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (nobel lecture). *Angewandte Chemie International Edition in English*, 30(5):455–465, 1991.

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

A Filipa de Almeida, Rui Moreira, and Tiago Rodrigues. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry*, 3(10):589–604, 2019.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 213–220, 2008.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Alan Kai Hassen, Paula Torren-Peraire, Samuel Genheden, Jonas Verhoeven, Mike Preuss, and Igor Tetko. Mind the retrosynthesis gap: Bridging the divide between single-step and multi-step retrosynthesis prediction. *arXiv preprint arXiv:2212.11809*, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jidon Jang, Geun Ho Gu, Juhwan Noh, Juhwan Kim, and Yousung Jung. Structure-based synthesizability prediction of crystals using partially supervised learning. *Journal of the American Chemical Society*, 142(44):18836–18843, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Mufei Li, Jinjing Zhou, Jiajing Hu, Wenxuan Fan, Yangkang Zhang, Yaxin Gu, and George Karypis. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega*, 2021.

Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113, 2017.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.

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

Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jörg K Wegner, Marwin Segler, Sepp Hochreiter, and Günter Klambauer. Improving few-and zero-shot reaction template prediction using modern hopfield networks. *Journal of chemical information and modeling*, 62(9):2111–2120, 2022.

Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1): 1–11, 2020.

Zhengkai Tu and Connor W Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling*, 62 (15):3503–3513, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022.

A Model Details

A.1 LocalRetro

The LocalRetro model used in this paper is modified from the GitHub page of Chen & Jung (2021), which is a graph-based model using message passing neural network (MPNN, Gilmer et al. (2017) for local reactivity and attention mechanism (Vaswani et al. (2017)) for global reactivity.

First, all the ReLU activation functions (Nair & Hinton (2010)) are changed to GELU activation functions (Hendrycks & Gimpel (2016)), where $\Phi(x)$ is the cumulative distribution function of Gaussian distribution.

$$GELU(x) = x\Phi(x) \tag{5}$$

Second, the updated bond features is concatenated with its own original bond features to avoid the vanishing of bond feature over MPNN, where v'_a and v'_b are the updated atom features after MPNN, e_{ab} is the original features of chemical bond.

$$e'_{ab} = FullyConnected(v'_a||v'_b||e_{ab}) \tag{6}$$

Third, a binary classifier of classifying whether an atom or bond can be the reaction center o of a given product molecule O, where v'' is the atom or bond features after a global reactivity attention block.

$$p(o|O) = FullyConnected(v''_o) \tag{7}$$

A.2 LocalTransform

We train the LocalTransform model described in Chen & Jung (2022) using reagent-free reactions in USPTO-50K training dataset. Because the reagent is not specified during training and testing, one reactant set may yield different products with different reaction conditions. The prediction results of LocalTransform tested on the test set of USPTO-50K is summerzied in Table 3. We consider the reactant is feasible to synthesize a target product if the prediction score of the target product is greater than 0.5 to achieve the balance between accuracy and coverage.

Table 3: Exact match accuracy and coverage on USPTO-50K dataset of LocalTransform.

Reaction type	**Exact match accuracy %** **(coverage of test set %)**		
	score>0.1	score>0.5	score>0.9
All reactions	45.0 (97.2)	79.3 (91.7)	97.0 (74.3)
Rare reactions	51.1 (91.2)	74.1 (82.8)	94.9 (58.0)

B EXPERIMENTAL DETAILS

B.1 IMPLEMENTATION

We use the open-source Python library RDKit (Landrum (2013)) for all the molecular pre- and post-processing. The machine learning models are highly dependent on dgllife (Li et al. (2021)) and Pytorch (Paszke et al. (2019)) deep learning libraries. All experiments were conducted under a machine equipped with an Intel Core i9-12900K @3.20 GHz, 128 GB of RAM, and NVIDIA GeForce RTX 3090 GPU.

B.2 TRAINING CONDITIONS

We train our models for 30 epochs with a batch size 16 by using the Adam optimizer (Kingma & Ba (2014)) with a weight decay of 10^{-6} and set the initial learning rate to 10^{-4}. The learning rate is reduced by a factor of 0.1 after each 8 epochs and the training is terminated if the validation loss does not decrease for three successive epochs. Model gradients are clipped at a maximum norm of 20.