Electronic Supplementary Material (ESI) for RSC Advances.
This journal is © The Royal Society of Chemistry 2022

Electronic Supplementary Material (ESI) for RSC Advances.
This journal is © The Royal Society of Chemistry 2022

Supporting Information

Transformer-based Multi-task Learning for Reaction Prediction under Low-resource Circumstance

Haoran Qiao, [a] Yejian Wu , [b] Yun Zhang, [b] Chengyun Zhang, [b] XinYi Wu, [b] Zhipeng Wu, [b] Qingjie Zhao, [c] Xinqiao Wang, [b] Huiyu Li, *[b] Hongliang Duan*[bd]

[a] College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China.

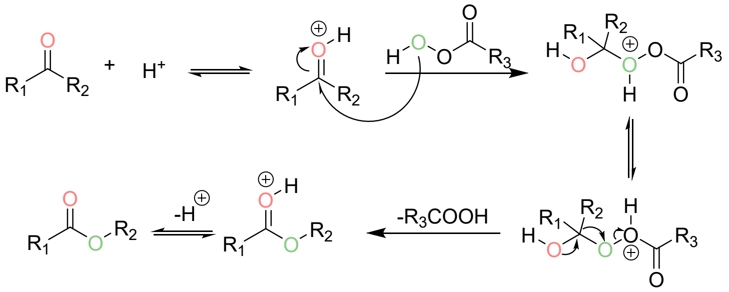
[b] Artificial Intelligence Aided Drug Discovery Institute, College of Pharmaceutical Sciences Zhejiang University of Technology, Hangzhou 310014, China.

[c] Innovation Research Institute of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China.

*Email: huiyuli@shiep.edu.cn; hduan@zjut.edu.cn

[d] Shanghai Institute of Materia Medica (SIMM), Chinese Academy of Sciences, Shanghai 201203, China.

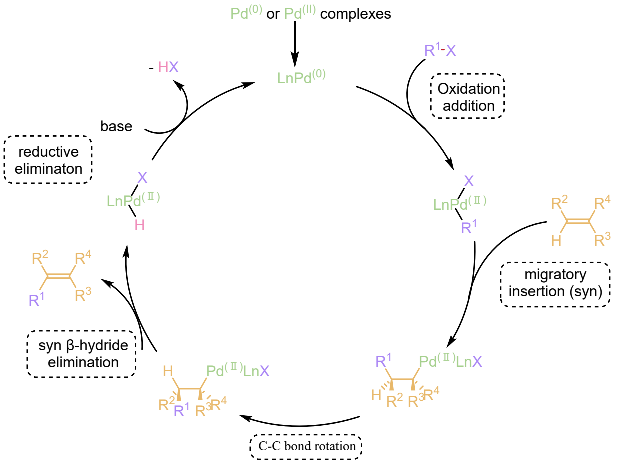
Section S1. Detailed information about Baeyer-Villiger reaction



Supplementary Fig. 1 The general mechanism of Baeyer-Villiger reaction

An example of a typical small-scale reaction is the Baeyer-Villiger reaction[1]. An ester can be formed from a ketone or an aldehyde using a peroxyacid or a peroxide. This reaction, as a rearrangement reaction, is crucially characterized by the fact that the regional chemistry is dependent on the ability of the group to migrate. Typically, groups are ranked in terms of their ability to migrate as follows: tertiary alkyl > secondary al-kyl > aryl > methyl. There is a general mechanism for this reaction that can be found in Supplementary Figure 1. First, a proton activates the carbonyl group of the reactant, al-lowing it to be attacked more readily by the peroxyacid. Secondly, carbonyl groups are attacked by peroxyacids to form Criegee intermediates. The carboxylic acid then leaves the intermediate, leaving an electron-deficient oxygen cation. As the electron-deficient oxygen is unstable, the hydrocarbon group will rapidly move to the electron-deficient oxygen of the hydrogen peroxide group. And a protonated ester is obtained, which is rapidly deprotonated to form the final product.

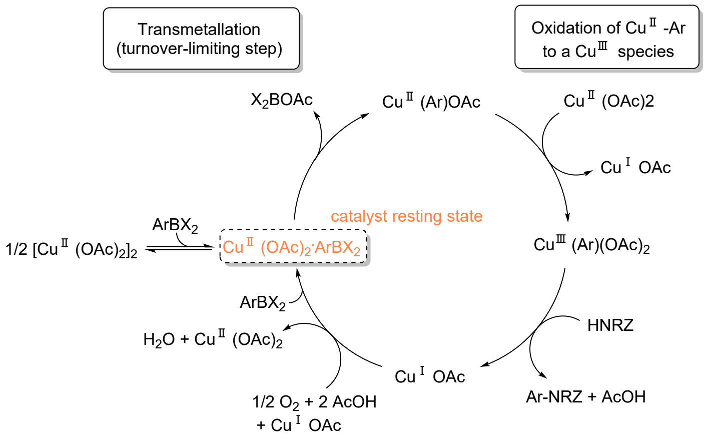
Section S2. Detailed information about Heck reaction



The other example of a small data set reaction is the Heck reaction, which forms a new olefin through the coupling of an olefin with an organohalide or trifluoride[2]. Heck reactions can be divided into two categories: intermolecular and intramolecular Heck reactions. Generally speaking, alkenes with more substituent groups of reactants will react again more slowly. Therefore, reaction rates are roughly in the following order CH2=CH2 > CH2=CHOAc > CH2=CHMe > CH2=CHPh > CH2=C(Me)Ph in the Heck reaction. The catalytic cycle process is a currently generally accepted mechanism of Heck reaction

(Supplementary Fig. 2). There are four stages in this catalytic cycle, firstly, catalyst precursors of Pd(II) are activated to produce low coordination number Pd(0), followed by oxidative addition of activated palladium and haloaromatic hydrocarbons in a second stage. Translocational insertion of the alkene is the third stage of the Heck reaction and determines the regioselectivity of the overall reaction. During the final stage, the product is produced in the process of eliminating the β-hydrogen. In this step, it should be noted that the hydridopalladium complex is presented. However, after elimination by alkali reduction this complex will be regenerated.

Section S3. Detailed information about Chan-Lam coupling reaction



Supplementary Fig. 3 The general mechanism of Chan-Lam coupling reaction

The other example of a small data set reaction we used is the Chan-Lam coupling reactions[3]. These reactions are aromatic, alkenyl and alkylation reactions in which substrates containing NH/OH/SH groups are oxidatively cross-coupled with organo-boronic acid compounds in air under weak base conditions, catalyzed by copper acetate. The coupling process is a currently generally accepted mechanism of Chan-Lam coupling reaction (Supplementary Fig. 3). First, complexation of the aryl boronic acid and the divalent copper complex occurs to form an aryl divalent copper intermediate and a boronic acid. Then, the divalent copper intermediate is oxidized to a trivalent copper intermediate in the presence of oxygen. This intermediate undergoes reductive elimination with amine to produce the final coupling product and a monovalent copper complex. In addition, the generated monovalent copper complex is oxidized by oxygen to a divalent copper complex, completing the catalyst cycle.

Section S4. Detailed information about top-n accuracy of RFRPT model

Supplementary Table 1 Comparison of the Baseline and RFRPT model's performance.

Dataset	Task	Top-N accuracy (%)			
		Top-1	Top-2	Top-5	Top-10
Baeyer-Villiger	Forward reaction prediction	69.0	77.9	80.5	81.4
	Retrosynthesis	77.9	82.7	85.4	85.8
Heck	Forward reaction prediction	73.3	77.5	79.4	79.7

	Retrosynthesis	37.6	52.0	62.9	65.9
Chan-Lam	Forward reaction prediction	65.2	71.0	72.9	74.1
	Retrosynthesis	57.4	63.8	69.1	70.6
Baeyer-Villiger	Forward reaction prediction	75.7	80.5	81.9	82.3
	Retrosynthesis	81.9	77.9	80.5	81.4
Heck	Forward reaction prediction	81.0	84.4	86.4	87.0
	Retrosynthesis	55.2	65.3	70.2	74.9
Chan-Lam	Forward reaction prediction	83.0	86.0	86.7	87.5
	Retrosynthesis	66.5	70.8	74.6	76.3

In this paper, we train 3 RFRPT models and 6 baseline models on our datasets, with results as shown in Supplementary Table 1.

Section S5. Detailed information about beam search

The model is autoregressive and in the last layer of the model is softmax, its output is considered as a vector of all words in the lexicon that should be predicted as the probability of the current output. We logicized each element and summed it with the result of the previous time step as the new result, found the candidate with the highest number N of the current result summed, and then re-entered the N results into the model to get the new vector and filtered it again until all the words became the ending identifier <end> or reached the truncation length we set (300).

Section S6. Detailed information about uncertainty metric

To assess the level of confidence of the model in the prediction, we evaluated the models separately using multiple evaluation metrics to validate the models following the method of Schwaller et al.[4]

We treated the predictions that were identical to the products reported in the patent with a confidence score above the threshold as true-positives (TPs), the predictions that were not identical to the reported products and were below the threshold as true-negatives (TNs), the predictions that were identical to the reported products but were below the threshold as false-negatives (FNs), and finally, the predictions that were not identical to the reported products but were above the threshold as false-positives (FPs). In addition, In addition, the fomulations of other evaluation indicators are as follows.

Accuracy: Percentage of correct samples to predicted samples:

$$Accuracy=\frac{TP+TN}{TP+TF+TN+FN}$$

Precision: Percentage of correct positive sample predictions among those predicted to be positive:

$$Precsion=\frac{TP}{TP+FP}$$

Specificity: Percentage of true negative samples among actual negative samples:

$$Specificity=\frac{TN}{FP+TN}$$

True Positive Rate (TPR): Percentage of true positive samples among actual positive samples:

$$TPR=\frac{TP}{TP+FN}$$

False Positive Rate(FPR): Percentage of false positive samples among actual negative samples:

$$FPR = \frac{FP}{FP + TN}$$

Matthews correlation coefficient(MCC): Measuring the similarity of the true distribution and predicted outcomes:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Recipient operating characteristic curve (ROC) The horizontal coordinate of the recipient operating characteristic curve is TPR and the vertical coordinate is FPR, AUC (area under curve) is defined as the area enclosed with the coordinate axis under the ROC curve.

Section S7. Detailed information about hyperparameter search

We performed a hyperparametric comparison in the baseline model using the Baeyer-Villiger dataset, where a drop out can reduce the complex co-adaptive relationships between neurons so that the weight update no longer depends on the joint action of implicit nodes with fixed relationships. It can avoid overfitting in the training process; we perform a set of experiments with the drop out varying from 0.1 to 0.7 . The experimental results are shown in the Supplementary Table 2, the model works best when the drop out is 0.3. The learning rate affects whether and when the objective function converges to a local minimum. We performed a set of experiments with the learning rate varying from 1e-2 to 1e-6 . The model works best when the leaning rate is 1e-3 as shown in the Supplementary Table 3. We chose the parameters of dropout 0.3 and lr 1e-3, and we used the parameters previously optimized in our lab as the rest of the parameters.[5]

Supplementary Table 2 Hyperparametric search for drop out on Baeyer-Villiger dataset

Drop out	Top-1 Accuracy
0.1	0.665
0.2	0.664
0.3	0.712
0.4	0.681
0.5	0.664
0.6	0.044
0.7	0.004

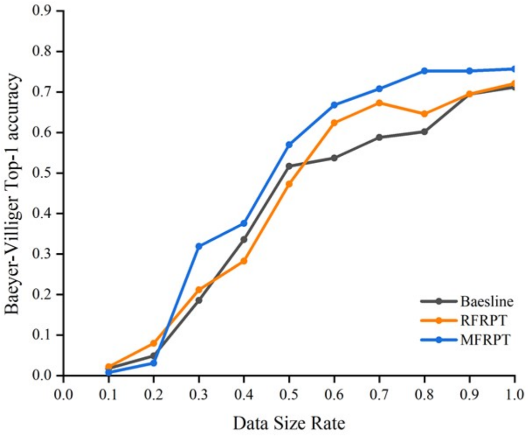
Supplementary Table 3 Hyperparametric search for learning rate on Baeyer-Villiger dataset

Learning rate	Top-1 Accuracy

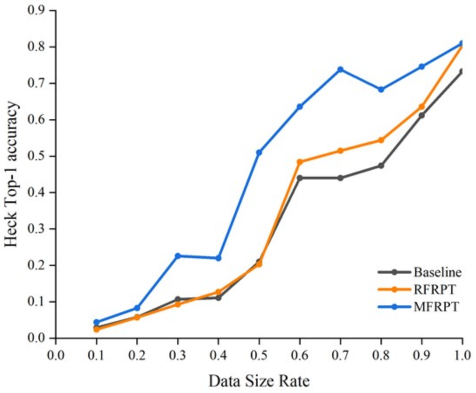
1e-1	0
1e-2	0
1e-3	0.712
1e-4	0.442
1e-5	0.106
1e-6	0.008
1e-7	0

Section S8. Detailed information about Impact of different data sizes on the model

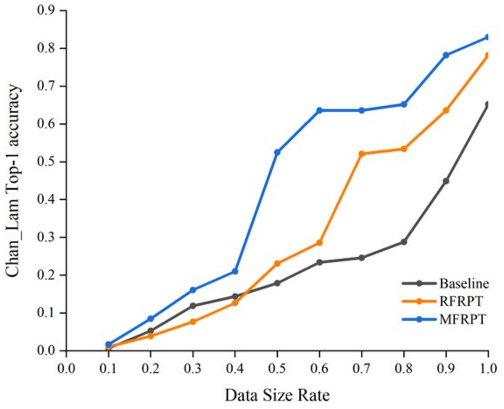
As shown in Supplementary Fig 4 and Supplementary Table 4, we trained 10 baseline models, RFRPT and MFRPT based on different data points with Baeyer-Villiger dataset. The difference between the accuracy of RFRPT and that of the baseline is highest at 60% data rate, while the difference between MFRPT and that of the baseline is highest at 80% data rate as shown in Supplementary Fig 5 and Supplementary Table 5, under the Heck reaction dataset. The model exhibits lower accuracy when the data rate is small, which is due to the underfitting phenomenon caused by the small data rate. When the data rate reaches 70%, the difference between RFRPT and baseline model accuracy is highest, and MFRPT reaches the largest difference with baseline model accuracy at 50% data rate. As shown in Supplementary Fig 6 and Supplementary Table 6, under the Chan-lam reaction dataset, the same as before, the models show underfitting when the data rate is small. When the data rate is 70% the difference between RFRPT and the baseline model accuracy is the highest, while MFRPT reaches the highest difference with the baseline model accuracy at 60%. In the case of a fixed data set, RFRPT generalizes better than the baseline without underfitting, but these differences generally decrease as the dataset increases, depending on the type of reaction data. MFRPT is generally out of overfitting in smaller data points. This is because MFRPT has three data inputs at each data point; for the encoder the input data comes from three, and the generalization ability of MFRPT is stronger than that of the baseline model. Again, these gaps decrease as the input dataset increases. The above results show that MFRPT and RFRPT generalize better than the baseline model when the data set is small, but when the data set is large enough, the generalization of the baseline model is closer to the effect of our model.



Supplementary Fig. 4 Top-1 accuracy in baseline, RFRPT, and MFRPT models with Baeyer-villiger dataset for different data size rates.



Supplementary Fig. 5 Top-1 accuracy in baseline, RFRPT, and MFRPT models with Heck dataset for different data size rates.



Supplementary Fig. 6 Top-1 accuracy in baseline, RFRPT, and MFRPT models with Chan-lam dataset for different data size rates.

Supplementary Table 4 Increment of top1 accuracy of RFRPT, MFRPT at different data size rate with baseline on Baeyer-villiger dataset

Data Size Rate	RFRPT increment	MFRPT increment
0.1	0.004	-0.010
0.2	0.031	-0.018
0.3	0.026	0.133
0.4	-0.053	0.040
0.5	-0.044	0.053
0.6	0.087	0.131
0.7	0.085	0.120
0.8	0.074	0.150
0.9	0	0.057
1.0	0.009	0.045

Supplementary Table 6 Increment of top1 accuracy of RFRPT, MFRPT at different data size rate with baseline on Heck dataset

Data Size Rate	RFRPT increment	MFRPT increment
0.1	-0.005	0.015
0.2	-0.001	0.025
0.3	-0.014	0.119
0.4	0.016	0.109
0.5	-0.007	0.300
0.6	0.044	0.196
0.7	0.075	0.268
0.8	0.070	0.259
0.9	0.024	0.134
1.0	0.071	0.077

Supplementary Table 7 Increment of top1 accuracy of RFRPT, MFRPT at different data size rate with baseline on Chan-Lam dataset

Data Size Rate	RFRPT increment	MFRPT increment
0.1	0.004	0.010
0.2	-0.014	0.032
0.3	-0.042	0.042
0.4	-0.018	0.066
0.5	0.052	0.346
0.6	0.052	0.402
0.7	0.275	0.390
0.8	0.246	0.364
0.9	0.187	0.333
1.0	0.130	0.178

Section S9. Detailed information about split data method

We used python 3.9 and sklearn module [6] to split all the data randomly, in the ratio of 8:1:1, for training set: validation set: testing set. The number of data sets is shown in Supplementary Table 8. All the splitting results are in https://github.com/qiaohaoran/MFRPT-and-RFRPT/tree/main/rawdata . During training, the training set was used to train the model, the validation set was used to observe the convergence of the model during training and to apply an early stopping strategy without taking part in the updating of the model parameters. After training, we input the test set into the trained model to obtain our results to evaluate the generalization performance of the model.

Supplementary Table 8 The split data set size in the different datasets

Dataset	Train	Valid	Test	Total
Chan-lam	4220	527	527	5274
Heck	7967	996	996	9959
Baeyer-villiger	1808	226	226	2260

References

1. G.-J. Ten Brink, I. Arends and R. Sheldon, *Chemical Reviews*, 2004, **104**, 4105-4124.
2. R. F. Heck, *Journal of the American Chemical Society*, 1968, **90**, 5518-5526.

3. P. Y. Lam, C. G. Clark, S. Saubern, J. Adams, M. P. Winters, D. M. Chan and A. Combs, *Tetrahedron Letters*, 1998, **39**, 2941-2944.

4 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, ACS Cent. Sci., 2019, 5, 1572–1583.

5 Duan, H., Wang, L., Zhang, C., Guo, L., and Li, J.RSC Adv., 2020, **10**, 1371-1378

6 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.