# Supporting Information

## Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes

Yun Zhang, ‡[a] Ling Wang, ‡[a] Xinqiao Wang, ‡[a] Chengyun Zhang, [a] Jiamin Ge, [a] Jing Tang, [a] An Su[*b] and Hongliang Duan[*a]

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

[b]College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, China.

*E-mail: hduan@zjut.edu.cn and ansu0912@outlook.com

## Table of content

## Section S1 Detailed information about the transformer model

### S1.1 SMILES tokenization

The SMILES tokenization is a specific language that explicitly describes molecular structures in strings and the input tokens and output tokens are converted to vectors in transformer model.[1-2] For the vocabulary files automatically generated by the model contain input tokens and output tokens of our model can be found in *https://github.com/hongliangduan/Transformer-model-for-prediction-in-low-chemical-data-regimes.*

In the course of experiment, the reactions are translated to SMILES and inputted to the transformer model. And the outputting tokens from the model are also a sequence of SMILES.

### S1.2 Hyperparameters of the models

Before using the transformer model to predict the target of Baeyer-Villiger reaction, we first debugged and adjusted the hyperparameters of the transformer model based on the previous work of our laboratory which solve reaction predictions task. [3]

In the pretraining step, the transformer model is trained on a general chemical reaction dataset in which containing 380k data to obtain the basic chemical information. With the model being pretrained to a certain degree, the model can be applied to training dataset of Baeyer-Villiger for capturing the feature of Baeyer-Villiger reaction. Finally, the training process is stopped when the reaching a steady state.

Here are hyperparameters selections of the transformer models:

```
"activation_dtype": "float32"
"add_relative_to_values": false
"attention_dropout": 0.1
"attention_dropout_broadcast_dims": ""
"attention_key_channels": 0
"attention_value_channels": 0
"attention_variables_3d": false
"batch_size": 6144
"causal_decoder_self_attention": true
"clip_grad_norm": 0.0
"compress_steps": 0
"conv_first_kernel": 3
"daisy_chain_variables": true
"data_dir":"./t2t_data"
"dropout": 0.2
"eval_drop_long_sequences": false
"eval_run_autoregressive": false
"eval_steps": 100
"factored_logits": false
"ffn_layer": "dense_relu_dense"
"filter_size": 2048
"force_full_predict": false
"grad_noise_scale": 0.0
"heads_share_relative_embedding": false
```

```
"hidden_size": 256
"initializer": "uniform_unit_scaling"
"initializer_gain": 1.0
"input_modalities": "default"
"kernel_height": 3
"kernel_width": 1
"label_smoothing": 0.1
"layer_postprocess_sequence": "da"
"layer_prepostprocess_dropout": 0.3
"layer_prepostprocess_dropout_broadcast_dims": ""
"layer_preprocess_sequence": "n"
"learning_rate": 0.2
"learning_rate_constant": 2.0
"learning_rate_cosine_cycle_steps": 250000
"learning_rate_decay_rate": 1.0
"learning_rate_decay_scheme": "noam"
"learning_rate_decay_staircase": false
"learning_rate_decay_steps": 5000
"learning_rate_minimum": null
"learning_rate_schedule":"constant*linear_warmup*rsqrt_decay*rsqrt_hiden_size"
"learning_rate_warmup_steps": 16000
"length_bucket_step": 1.1
"max_input_seq_length": 0
"max_length": 256
"max_relative_position": 0
"max_target_seq_length": 0
"min_length": 0
"min_length_bucket": 8
"model_dir":"./t2t_train/translate_retro_syn/transformer-
transformer_base_single_gpu"
"moe_hidden_sizes": "2048"
"moe_k": 2
"moe_loss_coef": 0.001
"moe_num_experts": 16
"moe_overhead_eval": 2.0
"moe_overhead_train": 1.0
"multiply_embedding_mode": "sqrt_depth"
"multiproblem_class_loss_multiplier": 0.0
"multiproblem_label_weight": 0.5
"multiproblem_mixing_schedule": "constant"
"multiproblem_reweight_label_loss": false
"multiproblem_schedule_max_examples": 10000000.0
"multiproblem_schedule_threshold": 0.5
"nbr_decoder_problems": 1
```

```
"no_data_parallelism": false
"norm_epsilon": 1e-06
"norm_type": "layer"
"num_decoder_layers": 0
"num_encoder_layers": 0
"num_heads": 8
"num_hidden_layers": 6
"optimizer": "Adam"
"optimizer_adafactor_beta1": 0.0
"optimizer_adafactor_beta2": 0.999
"optimizer_adafactor_clipping_threshold": 1.0
"optimizer_adafactor_decay_type": "pow"
"optimizer_adafactor_factored": true
"optimizer_adafactor_memory_exponent": 0.8
"optimizer_adafactor_multiply_by_parameter_scale": true
"optimizer_adam_beta1": 0.9
"optimizer_adam_beta2": 0.997
"optimizer_adam_epsilon": 1e-09
"optimizer_momentum_momentum": 0.9
"optimizer_momentum_nesterov": false
"optimizer_multistep_accumulate_steps": null
"parameter_attention_key_channels": 0
"parameter_attention_value_channels": 0
"pos": "timing"
"prepend_mode": "none"
"pretrained_model_dir": ""
"proximity_bias": false
"relu_dropout": 0.1
"relu_dropout_broadcast_dims": ""
"sampling_method": "argmax"
"sampling_temp": 1.0
"schedule": "continuous_train_and_eval"
"scheduled_sampling_gold_mixin_prob": 0.5
"scheduled_sampling_prob": 0.0
"scheduled_sampling_warmup_steps": 50000
"self_attention_type": "dot_product"
"shared_embedding": false
"shared_embedding_and_softmax_weights": true
"split_to_length": 0
"summarize_grads": false
"summarize_vars": false
"symbol_dropout": 0.0
"symbol_modality_num_shards": 16
"symbol_modality_skip_top": false
```

```
"target_modality": "default"
"train_steps": 2000000
"use_fixed_batch_size": false
"use_pad_remover": true
"use_target_space_embedding": true
"video_num_input_frames": 1
"video_num_target_frames": 1
"vocab_divisor": 1
"warm_start_from": null
"weight_decay": 0.0
"weight_dtype": "float32"
"weight_noise": 0.0
```

## Section S2 Preparation of Baeyer-Villiger reaction

The Baeyer-Villiger reaction dataset we filtered out from the Reaxys database is splatted into three parts: training, validation and test dataset. We make further analysis of the Baeyer-Villiger reaction dataset' classification to confirm the effectiveness of the transformer-baseline, transformer-transfer learning and transformer-transfer learning with onefold SMILES augmentation models. According to the classification of functional groups containing in the reactants of the Baeyer-Villiger reaction, the reactions can be divided into two types: the one is reaction with aldehyde compound as reactant and another one is reactions with ketone compounds as reactant. Table S1 shows the detailed distributions of reactions in the three dataset we used to pretrain, valid and test the performance of the transformer-baseline, transformer-transfer learning and transformer-transfer learning with onefold SMILES augmentation models. In the limited dataset composed of 2254 Baeyer-Villiger reaction, there are 392 reactions of aldehyde compound as reactants and 1862 reactions are ketone compound as reactants. The number of reactions in which the aldehyde compounds are referred to as reactants accounts for 77.0% of the total training dataset, and this kind of reactions correspond correspondingly accounts for 11.5% in the validation and test dataset, respectively. As for the Baeyer-Villiger reaction with the ketone compounds as reactants, they account for 80.6% in the training dataset and 19.4% of reactions of ketone reactants are equally divided into Validation and test dataset. In other words, the distributions of reactions with splitting into three datasets is in accord with the scaffold splitting condition.

**Table S1.** The detailed classifications of Baeyer-Villiger reaction by reactants' type on training, validation and test dataset.

| Dataset | Reactant type | | Total |
|---|---|---|---|
| | aldehyde compound | ketone compounds | |
| Training dataset | 302 | 1501 | 1803 |
| Validation dataset | 45 | 181 | 226 |

| | | | |
|---|---|---|---|
| Test dataset | 45 | 180 | 225 |
| Total | 392 | 1862 | 2254 |

## Section S3 Analysis of cross-validations experiments

### S3.1 Cross-validations of transformer models on Baeyer-Villiger reaction dataset

In order to avoid the contingency of models' performance caused by the data splitting, such as prediction results depend too much on favourable or adverse data splitting procedure, we randomly split the Baeyer-Villiger reaction dataset for ten times and do experiments, respectively. The detailed top-n accuracies of transformer-baseline (trained and tested on Baeyer-Villiger reaction dataset), transformer-transfer learning (pretrained and trained on general chemical reaction and Baeyer-Villiger reaction datasets respectively, and tested on Baeyer-Villiger reaction dataset), transformer-transfer learning with different levels SMILES augmentation models are described in Table S2, Table S3, Table S4, Table S5 and Table S6. Furthermore, we list the average top-n accuracies of these models in Table S7. All of the average top-1 accuracies of transformer-baseline, transformer-transfer learning, transformer-transfer learning with data augmentation models demonstrate that these models could be applied into reaction predictions. In addition, the transformer-transfer learning model achieves around 25% improvement and transformer-transfer learning with data augmentations further improves 3.8% in finishing this task. To some extent, with the introduction of pretraining knowledge obtained from transfer learning and SMILES augmentation strategy, the transformer-baseline model expresses a better performance on addressing the limitation of small data in chemistry field. It is clear that transfer learning and data augmentation play a critical role in promoting the transformer model's ability of predicting reaction, and the transformer model does not achieve good results by pretrained only on big data (general chemical reaction dataset) rather than trained on specific Baeyer-Villiger reaction dataset.

**Table S2.** The top-n accuracies of transformer-baseline models.

| Entry | Transformer-baseline model | | | |
|---|---|---|---|---|
| | Top-1 (%) | Top-2 (%) | Top-3 (%) | Top-5 (%) |
| 1 | 58.4 | 66.7 | 68.4 | 71.1 |
| 2 | 58.4 | 67.7 | 70.8 | 71.2 |
| 3 | 55.7 | 66.1 | 68.3 | 69.1 |
| 4 | 53.3 | 62.2 | 65.8 | 67.5 |
| 5 | 56.4 | 64.8 | 69.4 | 68.5 |
| 6 | 54.5 | 63.5 | 65.3 | 67.4 |
| 7 | 56.7 | 67.2 | 68.2 | 70.3 |
| 8 | 58.2 | 67.8 | 70.1 | 71.6 |
| 9 | 59.2 | 66.3 | 65.7 | 67.4 |

|  |  |  |  |  |
|---|---|---|---|---|
| 10 | 53.6 | 62.9 | 65.2 | 67.2 |
| average | 56.4 | 65.5 | 67.7 | 69.1 |

**Table S3.** The top-n accuracies of transformer-transfer learning model.

| Entry | Transformer-transfer learning model |  |  |  |
|---|---|---|---|---|
|  | Top-1 (%) | Top-2 (%) | Top-3 (%) | Top-5 (%) |
| 1 | 81.8 | 86.2 | 89.3 | 90.7 |
| 2 | 81.4 | 88.9 | 91.5 | 94.2 |
| 3 | 81.0 | 88.4 | 90.7 | 93.1 |
| 4 | 81.3 | 88.9 | 90.2 | 92.0 |
| 5 | 84.0 | 88.0 | 90.2 | 92.7 |
| 6 | 82.2 | 89.1 | 90.5 | 92.4 |
| 7 | 81.7 | 88.5 | 90.6 | 93.5 |
| 8 | 82.5 | 88.7 | 90.4 | 92.9 |
| 9 | 80.7 | 89.0 | 90.7 | 93.1 |
| 10 | 81.6 | 88.3 | 91.2 | 92.8 |
| average | 81.8 | 88.4 | 90.5 | 92.7 |

**Table S4.** The top-n accuracies of transformer-transfer learning model with onefold augmentation.

| Entry | Transformer-transfer learning with onefold augmentation |  |  |  |
|---|---|---|---|---|
|  | Top-1 (%) | Top-2 (%) | Top-3 (%) | Top-5 (%) |
| 1 | 86.7 | 92.4 | 94.2 | 94.2 |
| 2 | 84.0 | 91.5 | 91.6 | 93.6 |
| 3 | 85.7 | 91.1 | 91.6 | 92.9 |
| 4 | 86.2 | 91.6 | 93.7 | 94.2 |
| 5 | 85.1 | 90.5 | 92.8 | 92.8 |
| 6 | 85.2 | 90.3 | 92.1 | 93.9 |
| 7 | 84.7 | 89.9 | 91.7 | 92.4 |
| 8 | 85.6 | 91.3 | 92.5 | 92.7 |
| 9 | 85.4 | 91.9 | 94.1 | 94.2 |
| 10 | 86.3 | 92.2 | 93.7 | 93.7 |
| average | 85.5 | 91.3 | 92.8 | 93.5 |

**Table S5.** The top-n accuracies of transformer-transfer learning model with twofold augmentation model.

| Entry | Transformer-transfer learning with twofold augmentation |
|---|---|

|  | Top-1 (%) | Top-2 (%) | Top-3 (%) | Top-5 (%) |
|---|---|---|---|---|
| 1 | 84.0 | 92.4 | 94.2 | 94.2 |
| 2 | 82.8 | 90.1 | 92.6 | 92.5 |
| 3 | 81.3 | 89.5 | 93.5 | 93.9 |
| 4 | 81.9 | 89.5 | 91.6 | 92.7 |
| 5 | 82.4 | 90.3 | 92.4 | 93.6 |
| 6 | 83.5 | 91.4 | 92.5 | 92.7 |
| 7 | 81.6 | 90.2 | 93.1 | 93.0 |
| 8 | 82.9 | 91.7 | 94.0 | 94.5 |
| 9 | 83.5 | 91.7 | 92.7 | 93.2 |
| 10 | 83.2 | 92.0 | 93.5 | 93.7 |
| average | 82.7 | 90.9 | 93.0 | 93.4 |

**Table S6.** The top-n accuracies of transformer-transfer learning model with fourfold augmentation model.

| Entry | Transformer-transfer learning with fourfold augmentation |  |  |  |
|---|---|---|---|---|
|  | Top-1 (%) | Top-2 (%) | Top-3 (%) | Top-5 (%) |
| 1 | 82.7 | 90.2 | 93.3 | 94.2 |
| 2 | 82.8 | 90.9 | 93.6 | 94.5 |
| 3 | 81.3 | 89.8 | 92.4 | 93.1 |
| 4 | 82.4 | 90.3 | 92.8 | 93.8 |
| 5 | 81.9 | 89.4 | 91.5 | 92.9 |
| 6 | 82.3 | 89.9 | 90.7 | 92.3 |
| 7 | 82.1 | 89.7 | 92.1 | 94.1 |
| 8 | 81.5 | 90.3 | 93.3 | 94.7 |
| 9 | 81.7 | 88.5 | 92.1 | 93.2 |
| 10 | 83.0 | 91.2 | 92.4 | 94.0 |
| average | 82.2 | 90.0 | 92.4 | 93.7 |

**Table S7**. The average accuracies of top-n on transformer-baseline, transformer-transfer learning, transformer-transfer learning with three level SMILES augmentations.

| Model | The average accuracies of top-n | | | |
|---|---|---|---|---|
| | Top-1 | Top-2 | Top-3 | Top-5 |
| Transformer-baseline model | 56.4 | 65.5 | 67.7 | 89.1 |
| Transformer-transfer learning model | 81.8 | 88.4 | 90.5 | 92.9 |
| Transformer-transfer learning model with data augmentation ×1 | 85.5 | 91.3 | 82. | 93.5 |
| Transformer-transfer learning model with data augmentation ×2 | 82.7 | 90.8 | 93.0 | 93.4 |
| Transformer-transfer learning model with data augmentation ×4 | 82.2 | 90.0 | 92.4 | 93.7 |

## S3.2 Analysis of recurrent neural network baseline model

To make the results more convincing, we perform this experiment on a recurrent neural network model. The recurrent neural network model is a neural machine translation (NMT) model to reaction prediction, which has been used by Liu *et al.* to perform retrosynthesis reaction prediction.[4] Duan *et al.* also used this model in their work to accomplish the task of reaction prediction.[5] Table S8 list the results of recurrent neural network baseline model based on the corresponding task with transformer-transfer learning model integrated with onefold augmentation. For example, the average top-1 accuracy of recurrent neural network-transfer learning model with onefold augmentation is 75.4%, and average accuracies of top-2, top-3, top-5 are 79.9%, 81.6%,86.1%. respectively, which are 6%~9% lower than the transform-er-transfer learning with onefold augmentation. Furthermore, our previous work has demonstrated the transformer model is more superiority than the recurrent neural network baseline model.[6]

**Table S8.** The top-n accuracies of cross-validation results by recurrent neural network baseline model.

| Entry | Recurrent Neural Network Baseline Model[a] | | | |
|---|---|---|---|---|
| | Top-1 (%) | Top-2 (%) | Top-3 (%) | Top-5 (%) |
| 1 | 75.1 | 80.9 | 83.1 | 90.7 |
| 2 | 75.7 | 78.8 | 80.8 | 84.0 |
| 3 | 73.8 | 76.4 | 79.1 | 91.5 |
| 4 | 72.4 | 77.7 | 78.2 | 78.6 |
| 5 | 78.6 | 83.5 | 85.8 | 86.7 |
| 6 | 76.1 | 81.8 | 83.5 | 85.3 |
| 7 | 73.3 | 77.9 | 79.6 | 84.0 |
| 8 | 75.5 | 80.0 | 81.9 | 86.7 |
| 9 | 76.7 | 81.3 | 82.7 | 88.4 |
| 10 | 76.4 | 80.4 | 81.4 | 84.9 |
| average | 75.4 | 79.9 | 81.6 | 86.1 |

[a]The baseline model refers to that the sequence-to-sequence model with transfer learning and onefold data augmentation.

## Section S4 Analysis of top-2 predictions

The top-n represents that once the rank1 to rank n predictions are found, the prediction results of the model scan will stop. The Baeyer-Villiger reaction is an oxygen insertion reaction, which has two potential reaction sites for the reaction with ketone compounds as reactants. Therefore, we concern more about its' top-2 predictions.

For transformer-transfer learning model, the top-2 accuracy is 86.2%. We make further analysis based on the correctly top-2 predictions. In the rank1 predictions, the correctly predictions accounts for 89.7%, and the rest of right products appear in the rank2. Furthermore, the predictions which the rank 1 predicts correctly and rank2 meets the regioselectivity of Baeyer-Villiger reaction account for 28.3%. On the other hand, for these correct reactions with rank1 fitting the regioselectivity account for 8.2% in the top-2 prediction of transformer-transfer learning model.

For transformer-transfer learning model with onefold augmentation, the top-2 accuracy is 92.4%. The result of rank1 prediction indicate that the correctly predictions accounts for 93.2%. Moreover, for the predictions that either rank1 or rank2 meets regioselectivity account for 22.4%.

## Section S5 References

1 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.* 1988, 28, 31-36.

2 D. Weininger, A. Weininger and J. L. Weininger, SMILES. 2.Algorithm for generation of unique SMILES notation, *J. Chem. Inf. Model.* 1989, 29, 97−101.

3 L. Wang, C. Zhang, R. Bai, J. Li and L. H. Duan, Heck reaction prediction using a transformer model based on a transfer learning strategy, *Chem. Commun.* 2020, 56, 9368-9371.

4 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, Retrosynthetic reaction prediction using neural Sequence-to-Sequence models, *ACS Cent. Sci.* 2017, 3, 1103-1113.

5 H. Duan, L. Wang, C. Zhang, L. Guo and J. Li, Retrosynthesis with attention-based NMT model and chemical analysis of "wrong" predictions, *RSC Adv.* 2020, 10, 1371-1378.

6 R. Bai, C. Zhang, L. Wang, C. Yao, J. Ge and H. Duan, Transfer learning: making retrosynthetic predictions based on a small chemical reaction dataset scale to a new level. *Molecules* 2020, 25, 2357.