Retrosynthesis prediction enhanced by in-silico reaction data augmentation

Xu Zhang[1], Yiming Mo[2], Wenguan Wang[1*] and Yi Yang[1*]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, 310058, Zhejiang, China.
[2]College of Chemical and Biological Engineering, Zhejiang University, Hangzhou, 310058, Zhejiang, China.

*Corresponding author(s). E-mail(s): wenguanwang@zju.edu.cn; yangyics@zju.edu.cn;
Contributing authors: xu.zhang@zju.edu.cn; yimingmo@zju.edu.cn;

Supplementary Information

Contents

S1 Details of baseline

In this work, we adopt the vanilla transformer (Vaswani et al, 2017) as the network architecture. A typical transformer model consists of two major parts called encoder and decoder. There are several identical layers of transformer encoder and each has three separate blocks, named as "Layer Norm", "Multi-head Self Attention (MSA)", and "Feedforward Network (FFN)". Among them, the attention mechanism is the most critical part of transformer, where three different vectors Keys(K), Queries(Q) and Values(V) of dimension d are employed for each input token. For computing the self attention metric, the dot product of Queries and all the Keys are calculated and scaled by $1/\sqrt{d}$ in order to prevent the dot products from generating very large numbers. This matrix is then converted into a probability matrix through the $softmax$ function and is multiplied to the Values to produce the attention metric as follows:

$$Attention = softmax(\frac{QK^T}{\sqrt{d}})V. \tag{1}$$

Besides, the baseline utilizes Root-aligned SMILES (R-SMILES) (Zhong et al, 2022) as our SMILES augmentation strategy. R-SMILES has a better augmentation effect because it adopts a tightly aligned one-to-one mapping between the product and the reactant to predict retrosynthesis more effectively. Specifically, it adopts the same atom as the root (*i.e.*, the starting atom) of the SMILES strings for both the products and the reactants, which successfully resolves the one-to-many problem in random augmentation and enriches the SMILES representation compared to using canonical SMILES.

S2 Iterative training

RetroWISE is a self-boosting framework and could benefit from iterative training. Specifically, a better base model will result in better in-silico reactions, leading to improved predictions for retrosynthesis. If we can build a better base model with the in-silico reactions, then we can continue repeating this process: utilizing the base model to generate in-silico reactions, and building an even better base model with these reactions to generate higher-quality reactions for training. In other words, the key idea is to build a better base model with previous in-silico reactions for iteratively augmenting real paired data. Table S1 suggests that adding one more iteration enhances the prediction performance of the retrosynthesis model (*e.g.*, +1.1% top-1 accuracy). However, iterative training also has several drawbacks, such as significantly increasing

Table S1 Iterative training yields high-quality in-silico reactions and accurate prediction.

Method	$k=1$	3	5	10	20	50
Baseline	63.8	83.0	87.6	91.7	94.1	95.1
RetroWISE +Iterative	**64.9**	**83.8**	**88.0**	**91.9**	**94.3**	**96.1**

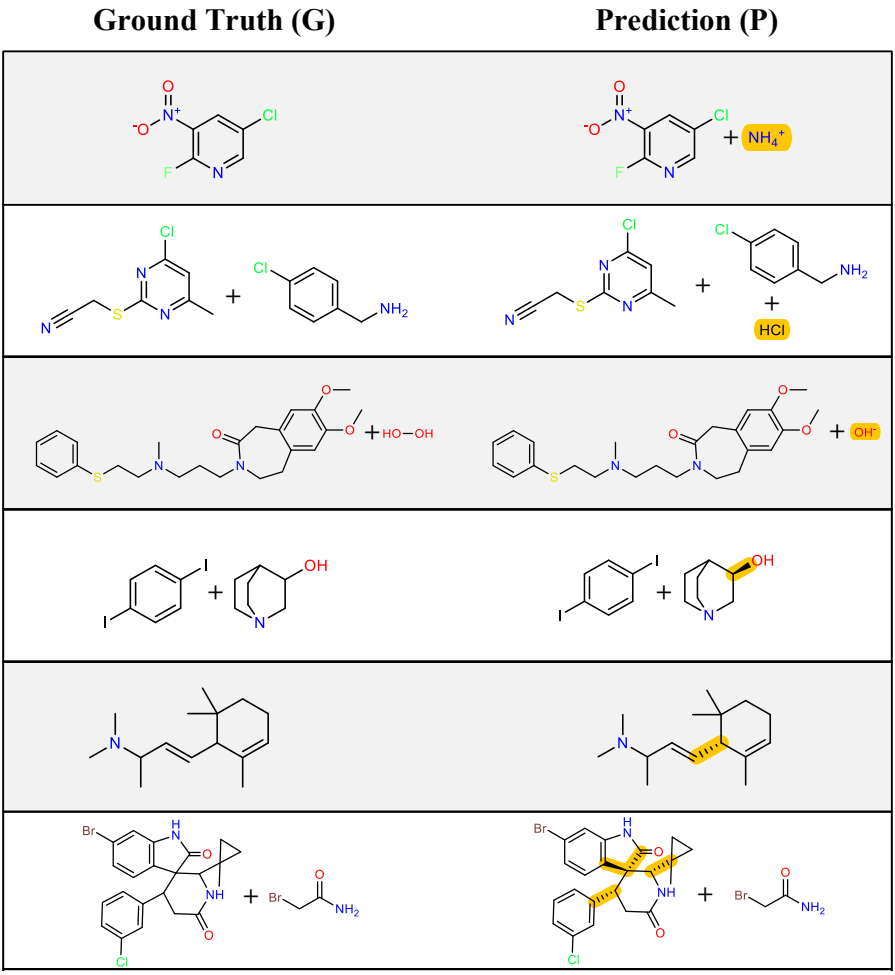

Fig. S1 Predictions with very high molecule similarity ($Tc \geq 0.95$) but inconsistent with the ground truth. We highlight differences between the ground truth and the prediction.

the training and generation time with too many iterations, and introducing more biases during iterative training.

S3 Discussions of highly scored inaccurate predictions

Two chemical structures are typically considered similar if the Tanimoto coefficient (T_c) is above 0.85 (Maggiora et al, 2014). We previously presented an inaccurate prediction with a high similarity ($T_c = 0.91$) in the main manuscript for a more proper evaluation, which demonstrates the prediction diversity of RetroWISE. As illustrated in Fig. S1, we provide more examples from the USPTO-50K test set with higher similarity ($T_c \geq 0.95$), highlighting some

challenges faced by machine learning (ML)-based methods: (1) the tendency of ML-based models to generate unnecessary reagents like NH_4^+, HCL, and OH^- due to learning bias; (2) the failure of models to accurately represent molecular information of stereochemistry, such as using incorrect symbols (/ or \) to denote directional single bonds adjacent to a double bond or creating accurate molecules but with incorrect chirality (*e.g.*, C@H v.s. C@@H).

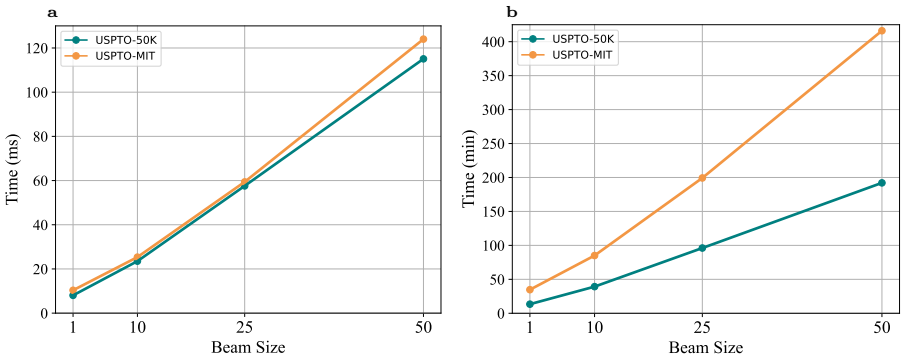

Fig. S2 Inference time of RetroWISE with different beam sizes. The inference time is measured with **a**, "Time per product" and **b**, "Total time". "Time per product" measures the time required to generate one reactant(s) from a given product using RetroWISE. "Total time" is the time to decode the whole test set into prediction results.

S4 Computational and memory efficiency

The proposed RetroWISE framework prioritizes memory and computational efficiency during inference, which enables further applications, such as multi-step retrosynthesis planning. RetroWISE employs a transformer architecture with approximately 44.5M parameters as the sequence-based model for USPTO-50K and USPTO-MIT. Compared with previous transformer-based method such as RetroPrime (Wang et al, 2021) having 75.4M parameters, RetroWISE is more lightweight and easy to deploy. Fig. S2 illustrates the inference speed on different datasets, measured with a single GPU (GeForce RTX 4090). The time per product varies with different beam sizes. On USPTO-50K, it is between 8.03 ms and 115.07 ms, while on USPTO-MIT, which has longer sequences, it is between 10.35 ms and 123.98 ms. The total time also depends on the beam size and the dataset. For the USPTO-50K test set with $10,014$ products, it varies from 13.41 min to 192.06 min. For the USPTO-MIT test set with $201,325$ products, the range is from 34.73 min to 416.11 min. These results highlight the computational and memory efficiency of RetroWISE.

References

Maggiora G, Vogt M, Stumpfe D, et al (2014) Molecular similarity in medicinal chemistry: miniperspective. Journal of medicinal chemistry 57(8):3186–3204

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems

Wang X, Li Y, Qiu J, et al (2021) Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. Chemical Engineering Journal 420:129,845

Zhong Z, Song J, Feng Z, et al (2022) Root-aligned smiles: a tight representation for chemical reaction prediction. Chemical Science 13(31):9023–9034