RetroOOD: Understanding Out-of-Distribution Generalization in Retrosynthesis Prediction

Yemin Yu[1,5*], **Luotian Yuan**[2*], **Ying Wei**[3†], **Hanyu Gao**[4], **Fei Wu**[2,5], **Zhihua Wang**[5], **Xinhai Ye**[5]

[1]City University of Hong Kong
[2]Zhejiang University
[3]Nanyang Technological University
[4]Hong Kong University of Science and Technology
[5]Shanghai Institute for Advanced Study of Zhejiang University
yeminyu2-c@my.cityu.edu.hk, ying.wei@ntu.edu.sg

Abstract

Machine learning-assisted retrosynthesis prediction models have been gaining widespread adoption, though their performances oftentimes degrade significantly when deployed in real-world applications embracing out-of-distribution (OOD) molecules or reactions. Despite steady progress on standard benchmarks, our understanding of existing retrosynthesis prediction models under the premise of distribution shifts remains stagnant. To this end, we first formally sort out two types of distribution shifts in retrosynthesis prediction and construct two groups of benchmark datasets. Next, through comprehensive experiments, we systematically compare state-of-the-art retrosynthesis prediction models on the two groups of benchmarks, revealing the limitations of previous in-distribution evaluation and re-examining the advantages of each model. More remarkably, we are motivated by the above empirical insights to propose two model-agnostic techniques that can improve the OOD generalization of arbitrary off-the-shelf retrosynthesis prediction algorithms. Our preliminary experiments show their high potential with an average performance improvement of 4.6%, and the established benchmarks serve as a foothold for further retrosynthesis prediction research towards OOD generalization.

Introduction

Retrosynthesis is the fundamental step in the field of organic synthesis (Corey 1991), which involves the application of various strategies to break down a target molecule into simpler building-block molecules. One of the biggest challenges for the pharmaceutical industry is finding reliable and effective ways to make new compounds.

Recently, there has been growing interest in computer-aided synthesis planning due to its potential to reduce the effort required for manually designing retrosynthesis strategies with chemical knowledge. Numerous machine learning models have been developed to learn these strategies from a fixed training dataset and to generalize this knowledge to new molecules. The training process involves either learning an explicit set of hard-coded templates as fixed rules

in a template-based manner or learning an implicit high-dimensional mapping from the product to the precursors in a template-free approach. Under the standard independent and identically distributed (i.i.d.) train-test data split, both lines of approaches have yielded promising results.

However, all retrosynthesis models exhibit a performance discrepancy between an in-distribution (ID) and an out-of-distribution (OOD) test set, which is a common issue when deploying the retrosynthesis model to a real-world environment. In general, we conclude that this discrepancy can be essentially attributed to the two different types of distributional shifts: the **label shift** of the retrosynthesis strategies (retro-strategy) and the **covariate shift** of the target molecules. Our research is aimed to analyze this discrepancy caused by the distributional shifts under various retrosynthesis prediction baselines. By understanding the distributional shifts that occur, we hope to mitigate their effects and gain a deeper understanding of the behaviors of various types of retrosynthesis models under different distributional shifts. To the best of our knowledge, no study has been conducted to investigate this topic rigorously.

To systematically analyze these two distributional shifts, we create and conduct experiments on multi-dimensional OOD dataset splits using various types of single-step retrosynthesis prediction baseline models. For each type of distributional shift, we propose a model-agnostic approach to alleviate the performance degradation. Our paper contributes to this field in three ways:

- We systematically define the distributional shifts under the context of retrosynthesis prediction.
- We construct multi-dimensional out-of-distribution datasets benchmarks and analyze the degree of performance discrepancies on a comprehensive set of baselines.
- We propose model-agnostic invariant learning and concept enhancement techniques to reduce performance degradations and provide our insights.

Related Work

Single-step retrosynthesis prediction The role of machine learning in retrosynthesis prediction is becoming increasingly pronounced (Jiang et al. 2023), especially in the piv-

---

otal stage of single-step retrosynthesis prediction. Single-step retrosynthesis aims to predict the set of molecules that chemically react to form a given product, towards which existing approaches fall into three major categories, including template-based (TB), semi-template-based (semi-TB), and template-free (TF) ones. Templates (Szymkuć et al. 2016) encode the changes in atom connectivity during the reaction, thereby applicable in converting a product back into the corresponding precursors. TB approaches such as NeuralSym (Segler and Waller 2017), Retrosim (Coley et al. 2017), and GLN (Dai et al. 2019) learn to select a standard reaction template to apply to the specified product for deriving the resulting precursors with subgraph isomorphism. However, TB methods have been criticized for their poor generalization capability to reactions outside the underlying training template set (Schwaller et al. 2022; Segler and Waller 2017; Jin et al. 2017). Semi-TB models alleviate the generalization problem via either constructing a more flexible template database with subgraph extraction (Chen and Jung 2021; Yan et al. 2022) or decomposing retrosynthesis into two sub-tasks of i) center identification and ii) synthon completion (Yan et al. 2020; Somnath et al. 2021). On the other hand, TF approaches completely eliminate using reaction templates and instead learn chemical transformations implicitly. Using various molecule representations, existing TF solutions formulate retrosynthesis as a string (Liu et al. 2017; Schwaller et al. 2019a; Sun et al. 2021; Yu et al. 2022) or a graph (Shi et al. 2020; Sacha et al. 2021) translation problem.

OOD generalization for molecule-related tasks Notwithstanding extensive literature on the evaluation of ID generalization, some attempts have been made to explore the frequent distributional shifts in real-world molecule-related tasks, including retrosynthesis prediction. The works (Ji et al. 2022; Bender and Cortés-Ciriano 2021; Deng et al. 2022) systematically study the shift in molecular size and structure as well as labels, and present several OOD benchmark datasets. However, they set their sights on molecular property prediction for drug discovery, which substantially differs from retrosynthesis prediction as a molecular generation task. Molecular generation also introduces additional complexity in the definition of label space, which is more complicated than a single value in conventional regression and classification in property prediction. As a matter of fact, there have been some works devoted to investigating the factors that cause label shift in retrosynthesis prediction, including the change in template radius, size, and subgraph isomorphism (Heid et al. 2021; Tu et al. 2022; Schwaller et al. 2021) between training and testing reactions. Unfortunately, the influence of such label shift on the performance of existing single-step retrosynthesis prediction approaches remains largely unknown, though the two approaches in (Seidl et al. 2022) and (Su et al. 2022) as a TB and TF approach respectively attempt to evaluate the "zero-shot" reaction prediction performances. However, the definitions of "zero-shot" used are arbitrary and lack consistency, with (Seidl et al. 2022) considering new reaction templates as "zero-shot", while (Su et al. 2022) defines "zero-shot" samples as new reaction types.

Besides the lack of comprehensive performance evaluation under label shift, benchmark datasets that support such evaluation are also in urgent demand. Existing dataset splits for distribution shift in retrosynthesis prediction, either by reaction type bias (Kovács, McCorkindale, and Lee 2021) or by time period (Segler, Preuss, and Waller 2018), struggle to explicitly disentangle label shift from covariate shift. These early exploratory studies motivate a more rigorous and systematic analysis of the impact of distribution shift on retrosynthesis prediction, covering (1) the disentanglement of two types of shift, (2) benchmark datasets for each type of shift, (3) extensive empirical evaluation of state-of-the-art retrosynthesis prediction algorithms, and (4) two model-agnostic techniques to handle both shifts.

Preliminaries

In this section, we formally define the distributional shift in single-step retrosynthesis prediction and establish the notation used throughout the paper.

Out-Of-Distribution Retrosynthesis Prediction

Single-step retrosynthesis prediction is a task where the model receives a target molecule $m \in \mathcal{M}$ as input and predicts a set of precursor source precursors $r \in \mathcal{R}$ that can synthesize m. The model can use different molecular transformation rules to generate various precursors for target molecules. Depending on the definitions of these transformation rules, retrosynthesis models can be classified into two main categories: *template-based* and *template-free*. Template-based approaches utilize reaction templates to predict the precursors required for synthesizing a product. These templates encode the changes in atom connectivity during the reaction that represent a specific type of molecular transformation. On the other hand, template-free models use a generative model to generate the precursors for a given target directly. These models typically use the SMILES (Weininger 1988) string or a graph structure to represent molecules and implicitly learn high-dimensional transformation rules between the hidden representations of precursors and molecules. However, such transformation rules can always be mapped back to reaction templates after the reaction generation.

Without loss of generality, we denote the retro-strategy as $t \in \mathcal{T}$ to represent such transformation rules from target molecule $m \in \mathcal{M}$ to source precursors $r \in \mathcal{R}$, which are meant to be general and not specific to any particular model or approach. $\mathcal{T}$ represents the space of transformation rules applied to a target molecule to generate its precursors. It's crucial to note that our introduction of $\mathcal{T}$ is not merely restricted to a template-based interpretation. In essence, all retrosynthesis prediction models, in an end-to-end fashion, intake a target product and output a set of precursors. While the exact realization of these retro-strategies might differ among models, our evaluation still remains model-agnostic and is conducted solely on the exact matching of the output precursors.

Subsequently, the training and testing datasets for retrosynthesis are denoted as $\mathcal{D}^{tr} = \{(m_i, t_i)\}_{i=1}^{N}$ and

$\mathcal{D}^{tst} = \{(m_i^*, t_i^*)\}_{i=N+1}^{N+N^*}$. The out-of-distribution retrosynthesis prediction problem can be defined as follows:

Definition 1. Given the observational training reactions $\mathcal{D}^{tr} = \{(m_i, t_i) \sim P_{tr}(\mathcal{M}, \mathcal{T})\}_{i=1}^{N}$ and testing data $\mathcal{D}^{tst} = \{(m_i^*, t_i^*) \sim P_{tst}(\mathcal{M}, \mathcal{T})\}_{i=1}^{N^*}$ where $P_{tr}(\mathcal{M}, \mathcal{T}) \neq P_{tst}(\mathcal{M}, \mathcal{T})$ and N/N^* is the sample sizes of train/test data, the goal of out-of-distribution retrosynthesis prediction is to learn a model in training distribution $P_{tr}(\mathcal{M}, \mathcal{T})$ to generalize to the test distribution $P_{tst}(\mathcal{M}, \mathcal{T})$ accurately.

OOD Retrosynthesis Prediction Benchmarks

In this section, we rigorously define and investigate two types of distributional shifts in the context of retrosynthesis: **label shift** of retro-strategies, $P(\mathcal{T})$, and **covariate shift** of target molecules, $P(\mathcal{M})$. Subsequently, we create two out-of-distribution dataset splits for each shift on the benchmark retrosynthesis prediction dataset under different domain settings. These datasets are used in subsequent empirical studies to analyze performance gaps and evaluate the effectiveness of our proposed OOD generalization approaches.

Label Shift in Retro-strategy $P(\mathcal{T})$

In the context of retrosynthesis prediction, we define the label space as the set of retro-strategies, denoted as $\mathcal{T}$, that map from the space of target molecules, $\mathcal{M}$, to the space of precursors, $\mathcal{R}$. In general, the label shift refers to the change of distribution of retro-strategy $P_{tr}(\mathcal{T}) \neq P_{tst}(\mathcal{T})$. However, the definition of the retro-strategy can vary significantly among different types of retrosynthesis prediction models. For template-based models, the retro-strategy is a discrete set of reaction templates extracted from the training set during data pre-processing. On the other hand, for template-free models, the retro-strategy is learned inherently during training, which is a function space that maps $\mathcal{M}$ to $\mathcal{R}$ in the latent space. It is widely acknowledged in studies (Tu et al. 2022; Heid et al. 2021; Lin et al. 2020; Schwaller et al. 2019a) that template-free models can generalize to novel or unseen reaction templates, whereas template-based models are confined to the predefined set of extracted templates.

Nevertheless, our findings reveal that the claimed generalization ability of the template-free models highly depends on the granularity of templates. As shown in Fig. 1, we focus on two different granularity of templates: minimal-template (radius=0) and retro-template (radius=1+). The key difference is that a reaction can *only* be mapped to one distinct minimal template, while it is possible to be mapped to multiple retro-templates. Although almost all previous template-based methods used retro-template as the template definition, we discover that the nuisance in retro-strategy granularity will result in distinct performance differences in the OOD label shift. We provide a more detailed investigation of template granularity in the Appendix [1].

Covariate Shift in Target Molecule $P(\mathcal{M})$

In the context of retrosynthesis prediction, covariate shift refers to the change in the distribution of the target

---

[1] Please refer to the Arxiv version https://arxiv.org/pdf/2312.10900.pdf.

molecules as $P_{tr}(\mathcal{M}) \neq P_{tst}(\mathcal{M})$. This phenomenon is often studied in conjunction with the concept distribution $P(\mathcal{T}|\mathcal{M})$, as the fundamental assumption for accurately evaluating covariate shift is that the concept distribution remains constant, $P_{tr}(\mathcal{T}|\mathcal{M}) = P_{tst}(\mathcal{T}|\mathcal{M})$. Typically, previous works(Peters, Bühlmann, and Meinshausen 2016; Arjovsky et al. 2019) addressed the covariate shift by adopting a causal perspective and dividing the input into two separate parts: the invariant feature $\mathcal{M}\_inv$ and the variant (spurious) feature $\mathcal{M}\_var$. The invariance property holds that using the invariant feature $\mathcal{M}\_inv$ alone is sufficient to fully recover the concept, such that $P(\mathcal{T}|\mathcal{M}) = P(\mathcal{T}|\mathcal{M}\_inv)$. Therefore, a pure covariate shift dataset should be designed in such a way that all shifts in the distribution occur on the variant feature $\mathcal{M}_{var}$ when $P_{tr}(\mathcal{M}) \neq P_{tst}(\mathcal{M})$ to maintain the invariance properties.

Covariate shift in molecular structures is prevalent in molecular property prediction and material design tasks. Similarly, the invariance property assumes that specific patterns inside a molecule, such as functional groups or scaffold substructures, play a crucial role in predicting a specific property. Generally, the substructure invariance rules that govern these relationships are task-specific for each property and have been validated through extensive study and observation as prior knowledge. (Phanus-umporn et al. 2018; Klekota and Roth 2008; Zhu et al. 2022) Following the same concept, we assume that certain features or substructures $\mathcal{M}\_inv$ in the target molecule are crucial for the model to make an invariant prediction for different retro-strategies to maintain $P(\mathcal{T}|\mathcal{M}) = P(\mathcal{T}|\mathcal{M}\_inv)$. Naturally, the reaction center (radius=0) should always be included as part of the invariant feature; otherwise, applying the templates to the target molecule would result in automatic failure. In addition, other substructures not limited to the reaction center can simultaneously impact the applicability of a particular template in terms of chemo-, regio-, or stereo-selectivity, which are the features we aim to identify as additional parts of $\mathcal{M}_{inv}$.

OOD Benchmark Dataset Split We introduce the benchmark dataset construction process for label shift and covariate shift dataset split. The detailed construction process on the benchmark is elaborated in the Appendix.

Label shift benchmark dataset To systematically evaluate the generalizability of different models when facing label shift in retro-strategies, we generate two OOD dataset splits as `USPTO-50K_T`, on the benchmark USPTO50K dataset (Schneider, Stiefl, and Landrum 2016) using the different granularity of labels. As shown in Fig. 2, we extract the minimal-templates and retro-templates for each reaction and arrange them in descending order based on their occurrence frequency. We also deliberately ensured that the total template set does *not* intersect between ID and OOD subsets to investigate the ability of different models to generalize to novel retro-strategies.

Covariate shift benchmark dataset We adopt the similar definition of covariate split settings proposed in (Koh et al. 2021; Ji et al. 2022) by using the molecular size and scaffold as criteria to construct the covariate OOD dataset split
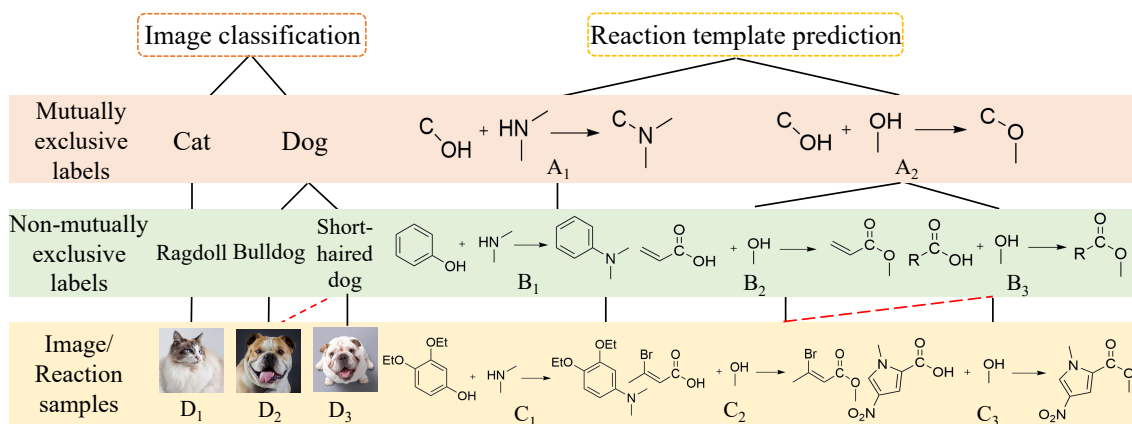
Figure 1: Minimal-templates and retro-templates. Left: In the image classification task, cats and dogs are typically regarded as mutually exclusive, while ragdolls, bulldogs, and short-haired dogs are not. Though image D_2 is only annotated as a bulldog, it has the potential label of a short-haired dog. Right: In retrosynthesis, similar to the image D_2, both labels B_2 and B_3 are viable options for generating the correct reaction C_2, but only one template B_2 is exposed as the positive label in the training dataset due to non-mutual exclusivity.
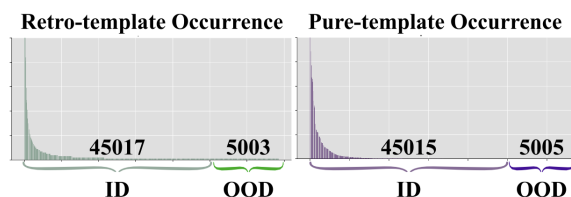

Figure 2: Retro-/minimal-template ID/OOD dataset split for label shift dataset USPTO50K_T.

USPTO-50k_M. The approach involves arranging the samples based on the molecular size or scaffold differences of the target molecule in ascending order and selecting larger or more complex target molecules as the OOD subset. Generally, a target molecule with a larger size or a more complex scaffold contains a larger proportion of variant features $\mathcal{M}\_var$ (Koh et al. 2021; Ji et al. 2022). As shown in Fig. 3, to eliminate the irrelevant influence of label shift, we conduct the data split independently for *each* minimal-template and then combine the results. This approach ensures that all covariate shifts occur on the variant feature $\mathcal{M}_{var}$ during $P_{tr}(\mathcal{M}) \neq P_{tst}(\mathcal{M})$, and guarantees that the ground-truth disconnection site stays consistent among all samples within a specific template class.

State-of-the-art Retrosynthesis Prediction Models under Distributional Shift

In this section, we introduce five representative retrosynthesis prediction models for empirical studies and analyze their baseline performance under the two distributional shifts mentioned above.

Baseline Methods

We select five representative models, namely *GLN* (Dai et al. 2019), *Molecular Transformer (MT)* (Schwaller et al.
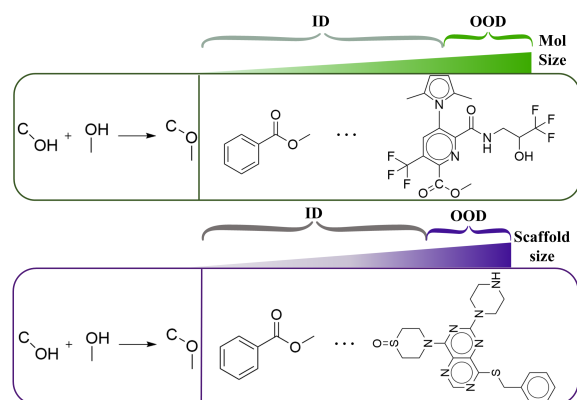

Figure 3: Molecular size and scaffold ID/OOD dataset split for covariate shift dataset USPTO50K_M.

2019a), *GraphRetro (G_Retro)* (Somnath et al. 2021), *Retro-Composer (R_Composer)* (Yan et al. 2022), and *MHN* (Seidl et al. 2022), as our baseline methods for empirical studies. These models comprehensively cover SMILES-based and graph-based representation under template-based, semi-template-based, and template-free categories. All baseline models are re-trained on each of the four OOD datasets separately for evaluation. We use the widely accepted top-k accuracy as the evaluation metric, which is still the most appropriate quantifiable metric for retrosynthesis prediction.

Baseline Results

The baseline results under covariate shift and label shift are listed in Tab. 1 and Tab. 2, respectively, with subscript $_{base}$.

Covariate shift In Tab. 1, we note a significant decline in performance, specifically a reduction of 30-40% in top-1 accuracy when comparing the ID test set to the OOD set. The present findings support our prior hypothesis that larger

Mol-Size	GLN_base(irm)	MT_base(irm)	G_Retro_base(irm)	R_Composer_base(irm)	MHN_base(irm)
ID Top-1	54.5%(54.9%)	52.5%(52.2%)	54.9%(**55.8%**)	55.1%(55.3%)	52.5%(52.2%)
ID Top-3	69.0%(70.5%)	75.7%(76.1%)	70.7%(71.1%)	79.7%(**80.5%**)	76.4%(76.6%)
ID Top-5	76.9%(77.6%)	80.1%(81.4%)	74.7%(75.4%)	86.0%(**87.7%**)	84.0%(84.2%)
ID Top-10	85.1%(85.7%)	83.6%(85.0%)	77.7%(78.2%)	90.0%(**90.9%**)	89.8%(90.2%)
OOD Top-1	37.6%(38.0%)	29.9%(30.3%)	38.5%(39.6%)	41.2%(**41.6%**)	34.0%(33.8%)
OOD Top-3	50.7%(51.2%)	46.7%(47.7%)	57.2%(58.4%)	67.3%(**68.2%**)	57.3%(57.5%)
OOD Top-5	58.9%(59.4%)	53.8%(55.0%)	64.5%(65.9%)	75.1%(**76.3%**)	67.5%(67.9%)
OOD Top-10	70.7%(71.5%)	58.1%(60.1%)	70.9%(72.8%)	83.1%(**84.9%**)	78.9%(79.2%)
Mol-Scaffold	GLN_base(irm)	MT_base(irm)	G_Retro_base(irm)	R_Composer_base(irm)	MHN_base(irm)
ID Top-1	55.7%(56.0%)	51.8%(51.4%)	56.2%(**56.2%**)	52.4%(52.8%)	52.3%(51.9%)
ID Top-3	69.8%(70.7%)	75.5%(76.0%)	70.2%(71.2%)	78.2%(**79.2%**)	76.6%(76.5%)
ID Top-5	77.2%(77.9%)	80.3%(81.6%)	73.8%(74.6%)	84.8%(**86.4%**)	84.0%(84.3%)
ID Top-10	85.5%(86.1%)	82.9%(84.7%)	76.6%(77.1%)	89.5%(**90.3%**)	90.1%(90.2%)
OOD Top-1	38.9%(39.5%)	37.9%(38.6%)	39.9%(40.1%)	40.7%(**41.2%**)	35.1%(34.8%)
OOD Top-3	53.3%(53.9%)	57.7%(59.0%)	57.8%(59.7%)	65.5%(**66.4%**)	60.3%(60.4%)
OOD Top-5	61.0%(61.7%)	64.0%(65.6%)	64.7%(66.5%)	75.1%(**76.2%**)	69.4%(69.7%)
OOD Top-10	72.4%(73.5%)	69.2%(70.2%)	71.2%(73.4%)	82.6%(**84.5%**)	79.9%(80.1%)

Table 1: The performance of five baselines and their IRM variants on covariate shift $P(\mathcal{M})$. The best IRM result is reported with *center-token masking IRM* for MT, *center prediction IRM* for GLN, *graph edit IRM* for GraphRetro, and *template composer IRM* for R_Composer, respectively.

molecules introduce more complexity in predicting the correct retro-strategy, regardless of the model employed, and these complexities are limited to specific feasible disconnection sites for a given target molecule. Among the five baseline models, MT exhibits the most significant decline in performance, since larger target molecules result in larger precursors with longer sequences of SMILES tokens as error accumulation, thereby intensifying the challenges associated with the covariate shift. RetroComposer outperforms most baselines in both splits, which can be attributed to its subgraph selection mechanism in discovering robust substructure invariance within the training samples.

Label shift In Tab. 2, the results are more varied between retro-template and minimal-template split. The average performance degradation is around 40-50% in the retro-template split and almost 100% in the minimal-template split. For retro-template split, we conclude that it's not rigorous to assume that template-based approaches cannot generalize to new templates without specifying the radius boundary, since our result shows that both GLN and MHN successfully generalize to a portion of unseen retro-templates due to their non-mutually exclusive nature. Additionally, we discover that when facing the same label shift in retro-templates, template-free models do not exhibit a clear advantage over template-based models in generalizability. On the other hand, the results in minimal-template split align with the previously held assumption that template-free models have only a limited ability to generalize to unseen templates when compared with template-based models.

Invariant Learning for Covariate Shift

In order to handle the covariate shift, our objective is to learn a robust parametric representation $\Phi(\cdot)$ that can accurately capture the full invariant features in the target molecule that satisfies the invariance property for predicting the retro-strategy. Specifically, we adopt Invariant Risk Minimization (IRM) to learn this invariant feature representation, which requires that the feature representation is simultaneously optimal across different domains. While the IRM regularizer is a known model-agnostic method, its precise application and optimization in a retrosynthesis model is a non-trivial problem and needs to be handled carefully in a model-specific way. We elaborate on the detailed IRM implementation for each baseline in the Appendix.

Performance Analysis

The best results are listed in Table 1 for the `USPTO50k_M`. Overall, we discover that applying IRM regularization to the specific reaction center identification stage improves the performance of GLN, GraphRetro, and RetroComposer, but the improvement is marginal for MT and MHN due to the nature of sequence-to-sequence generation and entanglement of center prediction and precursor generation, respectively. The overall insignificant improvement using IRM can be attributed to the uncontrollable concept drift on $P(\mathcal{T}|\mathcal{M}_{inv})$ presented within the dataset. The reason is that the collection of the reactions in USPTO50K is subject to the prior selection bias of different chemists during distinct wet-lab experiments under unobserved covariates (such as chemical conditions, etc.). Therefore, the ideal assumption for the invariance property is often violated, resulting in incoherency from the concept drift that hinders IRM from learning an optimal invariant predictor. In addition, the improvement resulting from the application of IRM could be more substantial if the distribution of training data were less biased towards specific retro-strategies.

Apart from the statistical result, we observe that using IRM regularization reduces spurious correlation on variant

Retro-template	GLN_base(irm)	MT_base(irm)	G_Retro_base(irm)	R_Composer_base(irm)	MHN_base(irm)
ID Top-1	50.9%(51.8%)	47.1%(49.0%)	53.2%(**53.9%**)	53.0%(53.3%)	51.9%(52.2%)
ID Top-3	68.2%(70.3%)	64.6%(67.0%)	68.6%(69.4%)	78.1%(**78.5%**)	75.1%(75.8%)
ID Top-5	76.2%(78.4%)	68.1%(72.9%)	72.3%(73.5%)	85.2%(**86.4%**)	82.7%(83.4%)
ID Top-10	85.2%(87.8%)	71.2%(77.1%)	74.8%(75.7%)	90.2%(**90.9%**)	89.9%(90.7%)
OOD Top-1	22.9%(24.5%)	23.8%(25.2%)	27.0%(**28.6%**)	25.4%(26.9%)	18.7%(20.4%)
OOD Top-3	31.8%(36.6%)	35.8%(41.2%)	40.3%(42.5%)	41.7%(**43.5%**)	33.1%(36.1%)
OOD Top-5	38.8%(43.4%)	39.8%(48.7%)	44.3%(46.6%)	47.6%(**49.8%**)	40.5%(42.8%)
OOD Top-10	46.6%(52.6%)	43.9%(**55.9%**)	47.4%(49.4%)	52.9%(55.4%)	49.6%(52.4%)

Minimal-template	GLN_base(irm)	MT_base(irm)	G_Retro_base(irm)	R_Composer_base(irm)	MHN_base(irm)
ID Top-1	51.9%(53.3%)	48.1%(49.5%)	53.6%(54.2%)	53.9%(**54.2%**)	52.9%(53.1%)
ID Top-3	68.9%(69.9%)	64.6%(67.2%)	68.3%(69.9%)	78.6%(**79.3%**)	74.2%(74.7%)
ID Top-5	76.5%(78.3%)	69.9%(73.3%)	74.8%(76.4%)	85.6%(**86.7%**)	83.2%(83.9%)
ID Top-10	86.6%(88.9%)	74.2%(80.2%)	76.4%(79.1%)	89.7%(90.5%)	90.6%(**91.3%**)
OOD Top-1	0%(0%)	**2.8%**(2.3%)	0%(0%)	0.1%(0.1%)	0.0%(0.1%)
OOD Top-3	0%(0%)	3.8%(**4%**)	0.1%(0.2%)	0.4%(0.4%)	0.1%(0.1%)
OOD Top-5	0%(0.2%)	4.2%(**4.7%**)	0.1%(0.3%)	0.7%(0.9%)	0.2%(0.2%)
OOD Top-10	0%(0.2%)	5.0%(**5.7%**)	0.1%(0.3%)	1.2%(1.2%)	0.3%(0.4%)

Table 2: The performance of five baselines and enhanced versions on label shift $P(T)$. The best-enhanced result is reported with $n = 5$ for MT,GLN, and MHN, and $n = 2$ for GraphRetro and Retrocomposer.

substructures $\mathcal{M}_{var}$ and an increased convergence towards the invariant substructures $\mathcal{M}_{inv}$ as shown in Fig. 7. We also comprehensively evaluated the results of applying IRM regularizer to different loss components as an ablation study presented in Tab. 3 in the Appendix.

Concept Enhancement for Label Shift

Besides covariate shift in molecular space, retrosynthesis prediction suffers significantly from the label shift $P(\mathcal{T})$. The reason behind the label shift is that the current benchmark dataset only includes reactions deemed most favorable by different chemists, thus manifesting a high precision. Still, it indiscriminately regards other unobserved potentially feasible reactions as equally infeasible, resulting in a low recall. Essentially, retrosynthesis is a many-to-many problem (Thakkar et al. 2022; Schwaller et al. 2019b), where the target molecule M can potentially be synthesized through various distinct retro-strategies T, and vice versa. To mitigate the low recall issue, we aim to enhance the concept of template applicability by transforming the binary criteria of the observed ground truth $P_{gt}(\mathcal{M}, \mathcal{T})$ into a continuous approximation using a probabilistic model. By utilizing a probabilistic model, we have greater flexibility to evaluate the boundary cases from the potentially feasible reactions, thereby constructing a more robust training set and improving recall without compromising precision.

However, modeling such probability is non-trivial since we need to perform counterfactual inference of unobserved reactions. One intuitive approach is to assume the distribution $P(\mathcal{M}, \mathcal{T})$ follows a Gaussian Process (GP) in order to construct a posterior predictive distribution for the feasibility of unobserved reactions. However, this assumption has limited expressiveness and may over-simplify the complex probabilistic structure of the selection bias among chemists.

Inspired by the recent advancements of Energy-based Model (EBM) (Grathwohl et al. 2019; Liu et al. 2020),

which offers greater flexibility and expressiveness compared to traditional probabilistic models, we adopt the EBM architecture to approximate $P(\mathcal{M}, \mathcal{T})$. EBM represents the likelihood of a probability distribution $p_D(\mathbf{x})$ for $x \in \mathbb{R}^D$ as $p_\theta(\mathbf{x}) = \frac{\exp(-F_\theta(\mathbf{x}))}{Z(\theta)}$, where the function $F_\theta(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ is known as the energy function, and $Z(\theta) = \int_{\mathbf{x}} \exp(-F_\theta(\mathbf{x}))$ is known as the partition function. Typically, directly evaluating $p_\theta(\mathbf{x})$ requires an intractable integration in partition function $Z(\theta)$ over all possible target-template tuples. Fortunately, the gradient for training the EBM, $\nabla_\theta \log p_\theta(\mathbf{x})$, can be expressed in the alternative form:

$$\nabla_\theta \log p_\theta(\mathbf{x}) = \mathbb{E}_{p_\theta(\mathbf{x}')} \left[\nabla_\theta F_\theta(\mathbf{x}') \right] - \nabla_\theta F_\theta(\mathbf{x}) \quad (1)$$

Thus, the question left for us is finding a surrogate for samples $\mathbf{x}'$ from the distribution $p_{(\theta)}(\mathbf{x}')$ to approximate the gradient of the training loss. In the next section, we elaborate on using a k-hop subgraph extraction algorithm on a bipartite graph to build the tractable EBM training loss.

Approach

The complete enhancement process is shown in Fig. 4. To begin, we model the set of ground-truth reactions as a target-template bipartite graph, where the ground-truth graph $G_{gt} = (M, T, E_{gt})$ contains all the target molecules and template as nodes and the observed ground-truth reactions as edges. The complete bipartite graph $G_{full} = (M, T, E_{full})$ can be obtained by connecting all template T nodes with the molecules node M. However, G_{full} contains a mixture of feasible, infeasible, and invalid reactions, which should be further denoised. Naturally, our problem is transformed into obtaining the best-enhanced graph $G_{enh} = (M, T, E_{enh})$ such that $E_{gt} \subset E_{enh} \subset E_{full}$.

Stage A: In the first stage, we use domain knowledge and a rule-based approach to filter out edges E_{fail} that generate
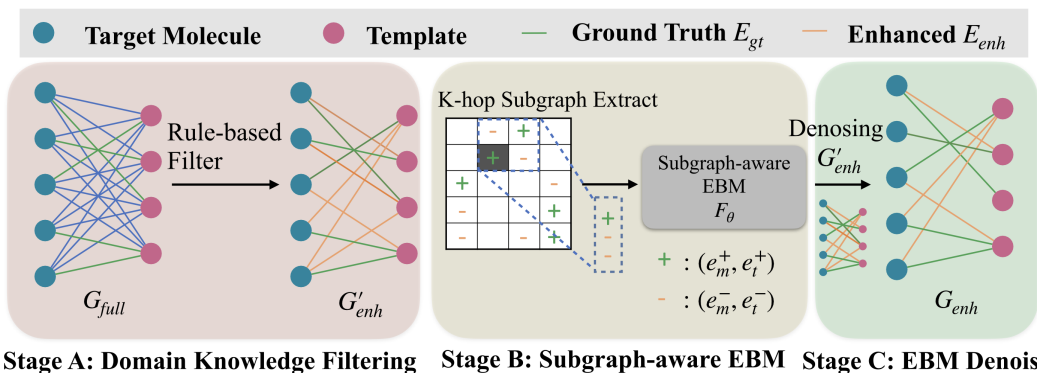
Figure 4: The process of acquiring the enhanced bipartite graph G_{enh} between target molecules and templates.

invalid reactions from target-template subgraph mismatch or syntactically illegal structures. We obtain a potentially enhanced graph $G'_{enh} = (M, T, E'_{enh})$, which contains all observed ground-truth reactions and unobserved potential reactions. However, G'_{enh} still contains edges that might produce chemically infeasible reactions, creating a trade-off between template diversity recall and observed selection bias precision. Naturally, this is the point where we can resort to EBM to evaluate the deviation from the observed ground-truth reaction to the unobserved enhanced reactions to obtain the best-enhanced graph. However, directly using the full set of the unobserved reaction (> 2 million) as the surrogate for $\mathbf{x}'$ in Eq. 1 is still computationally infeasible. To make the training realizable and reliable, we designed a tractable subgraph-aware EBM loss to realize the training.

Stage B: We design a subgraph-aware sampling method to select the most informative subsets to build EBM loss. Specifically, for each ground-truth reaction $e_{mt} \in E_{gt}$, we adopt a k-hop reaction subgraph extraction algorithm to acquire a subgraph $G^{m,t}_{sub} \in G'_{enh}$. This algorithm aims to extract subgraphs that contain sufficient neighborhood information in both the molecule and template dimensions to approximate the divergence of the unobserved potential reactions from the ground truth reactions. The complete algorithm is listed as Alg. 1 in the Appendix.

Within a selected subgraph, we can derive our EBM training objective. We omit the superscript ^{m,t} in the following notations for simplicity. Each subgraph G_{sub} can be further divided into two counterparts: G^+_{sub} with edges $E^+_{sub} = E_{sub} \cap E_{gt}$ and G^-_{sub} with edges $E^-_{sub} = E_{sub} \cap E_{enh}$. We define the tractable subgraph EBM loss in Eq. to push the energy score lower for the positive edges E^+_{sub} and higher for the negative edges E^-_{sub}. Specifically, for $e^+_{mt} \in E^+_{sub}$, we have:

$$L(\theta) = -\frac{1}{|E^+_{sub}|} \log \left(\frac{\exp(-F_\theta(e^+_m, e^+_t)/\tau)}{\sum_{e^-_{mt} \in E^-_{sub}} \exp(-F_\theta(e^-_m, e^-_t)/\tau)} \right)$$

Intuitively, each extracted subgraph G_{sub} contains a set of similar reactions that reflect a selection bias towards certain types of retrosynthesis strategies. Therefore, we apply $\frac{1}{|E^+_{sub}|}$ as an importance weighting coefficient to alleviate the

selection bias that exists in the original ground-truth distribution (Cortes et al. 2008).

Stage C: In the denoising stage, we similarly extract the k-hop reaction subgraph for each ground-truth reaction in E_{gt} and select the top-n reactions $E^n_{enh} \subset E^-_{sub}$ with the highest energy scores to form the enhanced set $E_{enh} = E_{gt} \bigcup E^n_{enh}$. Eventually, we can obtain the final enhanced graph E_{enh} used for pre-training the downstream baseline models. The complete architecture details of the EBM and the ablation study on the different settings of n are elaborated in the Appendix.

Performance Analysis

As shown in Table. 2, there is a significant performance improvement for conceptually-enhanced models over the baselines in both ID and OOD set for retro-templates and ID set for minimal-templates, which proves that concept enhancement is effective towards countering label shifts. Nevertheless, this approach has minimal effect on the minimal-template OOD set, as the algorithm can only use retro-templates for enhancement. Among the five baselines, we discover MT demonstrates the greatest improvements, which mainly due to its "template assembly" capability with the enhanced dataset to further derive novel implicit retro-strategies. Specifically, we discover that MT is capable of "assembling templates" from the training set to generate new templates (Appx. Fig.8). Therefore, we claim that MT can potentially learn to "invent" unseen minimal-templates from the training data in such a manner.

Conclusion

In this study, we examined the distributional shifts in retrosynthesis prediction and proposed two model-agnostic approaches, invariant learning, and concept enhancement, to address these shifts. Furthermore, we gained insights into the impact of covariate shift and label shift on multiple baseline performances through empirical analysis and evaluation of various baseline models. Future works can extend the coverage of the reactions in the benchmark dataset by exploiting a larger private or licensed dataset to obtain a more comprehensive outcome.

Acknowledgements

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization.

Bender, A.; and Cortés-Ciriano, I. 2021. Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, 26(2): 511–524.

Chen, S.; and Jung, Y. 2021. Deep Retrosynthetic Reaction Prediction using Local Reactivity and Global Attention. *JACS Au*, 1(10): 1612–1620.

Coley, C. W.; Green, W. H.; and Jensen, K. F. 2019. RD-Chiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *Journal of chemical information and modeling*, 59(6): 2529–2537.

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. 1991. *The logic of chemical synthesis*.

Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, 38–53. Springer.

Dai, H.; Li, C.; Coley, C.; Dai, B.; and Song, L. 2019. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32.

Deng, J.; Yang, Z.; Ojima, I.; Samaras, D.; and Wang, F. 2022. Artificial intelligence in drug discovery: applications and techniques. *Briefings in Bioinformatics*, 23(1).

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2019. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*.

Heid, E.; Liu, J.; Aude, A.; and Green, W. H. 2021. Influence of template size, canonicalization, and exclusivity for retrosynthesis and reaction prediction applications. *Journal of Chemical Information and Modeling*, 62(1): 16–26.

Ji, Y.; Zhang, L.; Wu, J.; Wu, B.; Huang, L.-K.; Xu, T.; Rong, Y.; Li, L.; Ren, J.; Xue, D.; Lai, H.; Xu, S.; Feng, J.; Liu, W.; Luo, P.; Zhou, S.; Huang, J.; Zhao, P.; and Bian, Y. 2022. DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery – A Focus on Affinity Prediction Problems with Noise Annotations.

Jiang, Y.; Yu, Y.; Kong, M.; Mei, Y.; Yuan, L.; Huang, Z.; Kuang, K.; Wang, Z.; Yao, H.; Zou, J.; Coley, C. W.; and Wei, Y. 2023. Artificial Intelligence for Retrosynthesis Prediction. *Engineering*, 25: 32–50.

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.

Klekota, J.; and Roth, F. 2008. Chemical Substructures that Enrich for Biological Activity. *Bioinformatics (Oxford, England)*, 24: 2518–25.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 5637–5664.

Kovács, D. P.; McCorkindale, W.; and Lee, A. A. 2021. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications*, 12(1): 1695.

Landrum, G.; et al. 2016. Rdkit: Open-source cheminformatics software.

Lin, K.; Xu, Y.; Pei, J.; and Lai, L. 2020. Automatic retrosynthetic route planning using template-free models. *Chemical science*, 11(12): 3355–3364.

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, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33: 21464–21475.

Peters, J.; Bühlmann, P.; and Meinshausen, N. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 947–1012.

Phanus-umporn, C.; Shoombuatong, W.; Prachayasittikul, V.; Anuwongcharoen, N.; and Nantasenamat, C. 2018. Privileged substructures for anti-sickling activity: Via cheminformatic analysis. *RSC Advances*, 8: 5920–5935.

Pope, P. E.; Kolouri, S.; Rostami, M.; Martin, C. E.; and Hoffmann, H. 2019. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10772–10781.

Ramsauer, H.; Schäfl, B.; Lehner, J.; Seidl, P.; Widrich, M.; Adler, T.; Gruber, L.; Holzleitner, M.; Pavlović, M.; Sandve, G. K.; et al. 2020. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Sacha, M.; Błaż, M.; Byrski, P.; Dąbrowski-Tumański, P.; Chromiński, M.; Loska, R.; Włodarczyk-Pruszyński, P.; and Jastrzębski, 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.; 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.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019a. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5.

Schwaller, P.; Nair, V. H.; Petraglia, R.; and Laino, T. 2019b. Evaluation Metrics for Single-Step Retrosynthetic Models. In *Second Workshop on Machine Learning and the Physical Sciences*. NeurIPS Vancouver, Canada.

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

Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; and Laino, T. 2022. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5): e1604.

Segler, M. H.; Preuss, M.; and Waller, M. P. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698): 604–610.

Segler, M. H. S.; 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.; Wegner, J. K.; Segler, M.; Hochreiter, S.; and Klambauer, G. 2022. Improving few-and zero-shot reaction template prediction using modern hopfield networks. *Journal of chemical information and modeling*, 62(9): 2111–2120.

Shi, C.; Xu, M.; Guo, H.; Zhang, M.; and Tang, J. 2020. A Graph to Graphs Framework for Retrosynthesis Prediction. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8818–8827. PMLR.

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.

Su, A.; Wang, X.; Wang, L.; Zhang, C.; Wu, Y.; Wu, X.; Zhao, Q.; and Duan, H. 2022. Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions. *Physical Chemistry Chemical Physics*, 24(17): 10280–10291.

Sun, R.; Dai, H.; Li, L.; Kearnes, S.; and Dai, B. 2021. Towards understanding retrosynthesis by energy-based models.

Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; and Grzybowski, B. A. 2016. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition*, 55(20): 5904–5937.

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.

Thakkar, A.; Vaucher, A.; Byekwaso, A.; Schwaller, P.; Toniato, A.; and Laino, T. 2022. Unbiasing Retrosynthesis Language Models with Disconnection Prompts.

Tu, H.; Shorewala, S.; Ma, T.; and Thost, V. 2022. Retrosynthesis Prediction Revisited. In *NeurIPS 2022 AI for Science: Progress and Promises*.

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.

Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; and Huang, J. 2020. RetroXpert: Decompose Retrosynthesis Prediction like a Chemist.

Yan, C.; Zhao, P.; Lu, C.; Yu, Y.; and Huang, J. 2022. RetroComposer: Composing Templates for Template-Based Retrosynthesis Prediction. *Biomolecules*, 12(9): 1325.

Yu, Y.; Wei, Y.; Kuang, K.; Huang, Z.; Yao, H.; and Wu, F. 2022. Grasp: Navigating retrosynthetic planning with goal-driven policy. *Advances in Neural Information Processing Systems*, 35: 10257–10268.

Zhang, W.; Feng, Y.; Meng, F.; You, D.; and Liu, Q. 2019. Bridging the Gap between Training and Inference for Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Zhu, J.; Liu, Y.; Wen, C.; and Wu, X. 2022. DGDFS: Dependence Guided Discriminative Feature Selection for Predicting Adverse Drug-Drug Interaction. *IEEE Transactions on Knowledge and Data Engineering*, 34(1): 271–285.