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

Yemin Yu\textsuperscript{1,5*}, Luotian Yuan\textsuperscript{2*}, Ying Wei\textsuperscript{3†}, Hanyu Gao\textsuperscript{4}, Fei Wu\textsuperscript{2,5}, Zhihua Wang\textsuperscript{5}, Xinhai Ye\textsuperscript{5}

\textsuperscript{1}City University of Hong Kong \\
\textsuperscript{2}Zhejiang University \\
\textsuperscript{3}Nanyang Technological University \\
\textsuperscript{4}Hong Kong University of Science and Technology \\
\textsuperscript{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

\begin{itemize}
  \item \textsuperscript{1}These authors contributed equally.
  \item \textsuperscript{2}Corresponding author.
  \item Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
\end{itemize}
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 (Heidt 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 notion 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, take 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
The out-of-distribution retrosynthesis prediction problem can be defined as follows:

**Definition 1.** Given the observational training reactions \( \mathcal{D}^{tr} = \{(m_i, t_i)\}_{i=1}^{N} \) and testing data \( \mathcal{D}^{tst} = \{(m_i', t_i')\}_{i=1}^{N'} \), the goal of out-of-distribution retrosynthesis prediction is to learn a model in training distribution \( P_{tr}(\mathcal{M}, T) \) to generalize to the test distribution \( P_{tst}(\mathcal{M}, T) \) 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, \( P(T) \), 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 \( P(T) \)

In the context of retrosynthesis prediction, we define the label space as the set of retro-strategies, denoted as \( T \), 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 \( P_{tr}(T) \neq P_{tst}(T) \). 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 molecules as \( P_{tr}(\mathcal{M}) \neq P_{tst}(\mathcal{M}) \). This phenomenon is often studied in conjunction with the concept distribution \( P(T|\mathcal{M}) \), as the fundamental assumption for accurately evaluating covariate shift is that the concept distribution remains constant, \( P_{tr}(T|\mathcal{M}) = P_{tst}(T|\mathcal{M}) \). 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 \( P(T|\mathcal{M}) = P(T|\mathcal{M}_{inv}) \). 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 a priori knowledge. (Phanusumporn 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 \( P(T|\mathcal{M}) = P(T|\mathcal{M}_{inv}) \). 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 $M_{var}$ (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 $M_{var}$ during $P_{tr}(M) \neq P_{tst}(M)$, 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. 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.
Table 1: The performance of five baselines and their IRM variants on covariate shift $P(M)$. 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.

<table>
<thead>
<tr>
<th>Mol-Size</th>
<th>GLN_base(irm)</th>
<th>MT_base(irm)</th>
<th>G_Retro_base(irm)</th>
<th>R_Composer_base(irm)</th>
<th>MHN_base(irm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Top-1</td>
<td>54.5%(54.9%)</td>
<td>52.5%(52.2%)</td>
<td>54.9%(55.8%)</td>
<td>55.1%(55.3%)</td>
<td>52.5%(52.2%)</td>
</tr>
<tr>
<td>ID Top-3</td>
<td>69.0%(70.5%)</td>
<td>75.7%(76.1%)</td>
<td>70.7%(71.1%)</td>
<td>79.7%(80.5%)</td>
<td>76.4%(76.6%)</td>
</tr>
<tr>
<td>ID Top-5</td>
<td>76.9%(77.6%)</td>
<td>80.1%(81.4%)</td>
<td>74.7%(75.4%)</td>
<td>86.0%(87.7%)</td>
<td>84.0%(84.2%)</td>
</tr>
<tr>
<td>ID Top-10</td>
<td>85.1%(85.7%)</td>
<td>83.6%(85.0%)</td>
<td>77.7%(78.2%)</td>
<td>90.0%(90.9%)</td>
<td>89.8%(90.2%)</td>
</tr>
<tr>
<td>OOD Top-1</td>
<td>37.6%(38.0%)</td>
<td>29.9%(30.3%)</td>
<td>38.5%(39.6%)</td>
<td>41.2%(41.6%)</td>
<td>34.0%(33.8%)</td>
</tr>
<tr>
<td>OOD Top-3</td>
<td>50.7%(51.2%)</td>
<td>46.7%(47.7%)</td>
<td>57.2%(58.4%)</td>
<td>67.3%(68.2%)</td>
<td>57.3%(57.5%)</td>
</tr>
<tr>
<td>OOD Top-5</td>
<td>58.9%(59.4%)</td>
<td>53.8%(55.0%)</td>
<td>64.5%(65.9%)</td>
<td>75.1%(76.3%)</td>
<td>67.5%(67.9%)</td>
</tr>
<tr>
<td>OOD Top-10</td>
<td>70.7%(71.5%)</td>
<td>58.1%(60.1%)</td>
<td>70.9%(72.8%)</td>
<td>83.1%(84.9%)</td>
<td>78.9%(79.2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mol-Scaffold</th>
<th>GLN_base(irm)</th>
<th>MT_base(irm)</th>
<th>G_Retro_base(irm)</th>
<th>R_Composer_base(irm)</th>
<th>MHN_base(irm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Top-1</td>
<td>55.7%(56.0%)</td>
<td>51.8%(51.4%)</td>
<td>56.2%(56.2%)</td>
<td>52.4%(52.8%)</td>
<td>52.3%(51.9%)</td>
</tr>
<tr>
<td>ID Top-3</td>
<td>69.8%(70.7%)</td>
<td>75.5%(76.0%)</td>
<td>70.2%(71.2%)</td>
<td>78.2%(79.2%)</td>
<td>76.6%(76.5%)</td>
</tr>
<tr>
<td>ID Top-5</td>
<td>77.2%(77.9%)</td>
<td>80.3%(81.6%)</td>
<td>73.8%(74.6%)</td>
<td>84.8%(86.4%)</td>
<td>84.0%(84.3%)</td>
</tr>
<tr>
<td>ID Top-10</td>
<td>85.5%(86.1%)</td>
<td>82.9%(84.7%)</td>
<td>76.6%(77.1%)</td>
<td>89.5%(90.3%)</td>
<td>90.1%(90.2%)</td>
</tr>
<tr>
<td>OOD Top-1</td>
<td>38.9%(39.5%)</td>
<td>37.9%(38.6%)</td>
<td>39.9%(40.1%)</td>
<td>40.7%(41.2%)</td>
<td>35.1%(34.8%)</td>
</tr>
<tr>
<td>OOD Top-3</td>
<td>53.3%(53.9%)</td>
<td>57.7%(59.0%)</td>
<td>57.8%(59.7%)</td>
<td>65.5%(66.4%)</td>
<td>60.3%(60.4%)</td>
</tr>
<tr>
<td>OOD Top-5</td>
<td>61.0%(61.7%)</td>
<td>64.0%(65.6%)</td>
<td>64.7%(66.5%)</td>
<td>75.1%(76.2%)</td>
<td>69.4%(69.7%)</td>
</tr>
<tr>
<td>OOD Top-10</td>
<td>72.4%(73.5%)</td>
<td>69.2%(70.2%)</td>
<td>71.2%(73.4%)</td>
<td>82.6%(84.5%)</td>
<td>79.9%(80.1%)</td>
</tr>
</tbody>
</table>

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. 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 $P(T|M_{inv})$ 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
substructures $\mathcal{M}_{\text{var}}$ and an increased convergence towards the invariant substructures $\mathcal{M}_{\text{inv}}$ 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 $P(T)$. 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 $P_{gt}(M,T)$ 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 $P(M,T)$ 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 $P(M,T)$. EBM represents the likelihood of a probability distribution $p_D(x)$ for $x \in \mathbb{R}^D$ as $p_D(x) = \exp(-F_D(x))/Z(\theta)$, where the function $F_D(x) : \mathbb{R}^D \to \mathbb{R}$ is known as the energy function, and $Z(\theta) = \int \exp(-F_D(x)) dx$ is known as the partition function. Typically, directly evaluating $p_D(x)$ requires an intractable integration in partition function $Z(\theta)$ over all possible target-template tuples. Fortunately, the gradient for training the EBM, $\nabla_\theta \log p_\theta(x)$, can be expressed in the alternative form:

$$\nabla_\theta \log p_\theta(x) = E_{p_\theta(x')} [\nabla_\theta F_\theta(x')] - \nabla_\theta F_\theta(x) \quad (1)$$

Thus, the question left for us is finding a surrogate for samples $x'$ from the distribution $p_\theta(x')$ 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

<table>
<thead>
<tr>
<th>Retro-template</th>
<th>GLN_base(irm)</th>
<th>MT_base(irm)</th>
<th>G_Retro_base(irm)</th>
<th>R_Composer_base(irm)</th>
<th>MHN_base(irm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Top-1</td>
<td>50.9%(51.8%)</td>
<td>47.1%(49.0%)</td>
<td>53.2%(53.9%)</td>
<td>53.0%(53.3%)</td>
<td>51.9%(52.2%)</td>
</tr>
<tr>
<td>ID Top-3</td>
<td>68.2%(70.3%)</td>
<td>64.6%(67.0%)</td>
<td>68.6%(69.4%)</td>
<td>78.1%(78.5%)</td>
<td>75.1%(78.5%)</td>
</tr>
<tr>
<td>ID Top-5</td>
<td>76.2%(78.4%)</td>
<td>68.1%(72.9%)</td>
<td>72.3%(73.5%)</td>
<td>85.2%(86.4%)</td>
<td>82.7%(83.4%)</td>
</tr>
<tr>
<td>ID Top-10</td>
<td>85.2%(87.8%)</td>
<td>71.2%(77.1%)</td>
<td>74.8%(75.7%)</td>
<td>90.2%(90.9%)</td>
<td>89.9%(90.7%)</td>
</tr>
</tbody>
</table>

| 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%) |

<table>
<thead>
<tr>
<th>Minimal-template</th>
<th>GLN_base(irm)</th>
<th>MT_base(irm)</th>
<th>G_Retro_base(irm)</th>
<th>R_Composer_base(irm)</th>
<th>MHN_base(irm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID Top-1</td>
<td>51.9%(53.3%)</td>
<td>48.1%(49.5%)</td>
<td>53.6%(54.2%)</td>
<td>53.9%(54.2%)</td>
<td>52.9%(53.1%)</td>
</tr>
<tr>
<td>ID Top-3</td>
<td>68.9%(69.9%)</td>
<td>64.6%(67.2%)</td>
<td>68.3%(69.9%)</td>
<td>78.6%(79.3%)</td>
<td>74.2%(74.7%)</td>
</tr>
<tr>
<td>ID Top-5</td>
<td>76.5%(78.3%)</td>
<td>69.9%(73.3%)</td>
<td>74.8%(76.4%)</td>
<td>85.6%(86.7%)</td>
<td>83.2%(83.9%)</td>
</tr>
<tr>
<td>ID Top-10</td>
<td>86.6%(88.9%)</td>
<td>74.2%(80.2%)</td>
<td>76.4%(79.1%)</td>
<td>89.7%(90.5%)</td>
<td>90.6%(91.3%)</td>
</tr>
</tbody>
</table>

| 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%(4.4%)  | 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.
invalid reactions from target-template subgraph mismatch or syntactically illegal structures. We obtain a potentially enhanced graph $G_{enh} = (M, T, E', E_{enh})$, 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 $x'$ 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 $G_{sub} \subset G_{enh}$. 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: $E_{sub}^+ = E_{sub} \cap E_{gt}$ 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 $1/|E_{sub}|$ 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 $E_{enh}^+_n \subset E_{sub}$ with the highest energy scores to form the enhanced set $E_{enh} = E_{gt} \cup E_{enh}^+$. Eventually, we can obtain the final enhanced graph $G_{enh}$ 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 retrosynthesis. 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

This work is sponsored by the Starry Night Science Fund at Shanghai Institute for Advanced Study (Zhejiang University) and Shanghai AI Laboratory.

References


