AgentRetro: Agent-Enhanced Molecular Spectral Domain Generalization Framework for Retrosynthesis Prediction under Mixed OOD Shifts

Jiahang Shen[1,2] , **Hongxin Xiang**[1,2,*] , **Yiping Liu**[1,2,*]

[1]State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

[2]Yuelushan Center for Industrial Innovation, Changsha, China

shenjiahang@hnu.edu.cn, xianghx@hnu.edu.cn, yiping0liu@gmail.com

Abstract

Retrosynthesis prediction is a cornerstone of drug discovery, enabling the synthesis of novel therapeutic candidates. However, current deep learning models falter when navigating the unexplored chemical space essential for innovation. In these realistic scenarios, models face a challenging mixed out-of-distribution (OOD) shift: they must handle both novel molecular scaffolds (covariate shift) and unseen reaction templates (label shift), leading to severe performance degradation. To address this, we propose AgentRetro, a framework integrating spectral domain generalization with Large Language Model (LLM) agents. It features a Spectral Domain Enhancer (SDE) that partitions chemical space via spectral clustering to capture diverse domain structures, and a Multi-Knowledge Agent (MKA) that enriches data with LLM-generated OOD reactions and textual rationales. These components are unified by a dual training strategy: meta-learning for template adaptability and adversarial learning for structural invariance. Crucially, our analysis reveals that explicitly decoupling structural invariance from mechanistic adaptability is essential for overcoming the interference between conflicting distribution shifts. Experiments show AgentRetro consistently improves state-of-the-art baselines, notably boosting EditRetro's OOD Top-10 accuracy from 41.7% to 50.1%. The source code is available at https://github.com/jiahangshen/IJCAI26-AgentRetro.

1 Introduction

Retrosynthesis is a foundational challenge in organic chemistry, aiming to decompose complex target molecules into simpler, commercially available precursors [Corey, 1991]. The advent of deep learning has yielded a new generation of retrosynthesis models capable of suggesting plausible synthetic routes with high accuracy [Segler and Waller, 2017; Schwaller *et al.*, 2019; Yan *et al.*, 2022]. These approaches, spanning template-based, template-free, and semi-template paradigms, consistently demonstrate impressive performance on standard in-distribution (ID) benchmarks [Jiang *et al.*, 2023]. However, their practical utility is severely hampered by a shared, critical flaw: a sharp performance decline when encountering the out-of-distribution (OOD) data common in real-world chemical exploration.

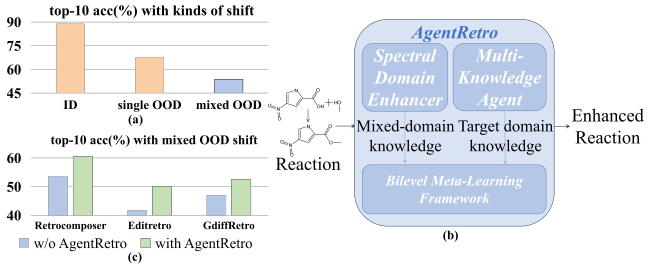

Figure 1: **The AgentRetro framework and the mixed-shift challenge.** **(a)** The mixed-shift setting, combining covariate and label shifts, causes a catastrophic performance drop compared to single shifts alone, as demonstrated on the RetroComposer baseline. **(b)** AgentRetro's architecture: An internal Spectral Domain Enhancer (SDE) and an external Multi-Knowledge Agent (MKA) provide knowledge to a Bilevel Meta-Learning Framework that trains a robust prediction model. **(c)** Empirical results showing AgentRetro consistently and significantly boosts OOD accuracy across multiple backbone models.

Foundational work has characterized this OOD challenge as a combination of two distinct phenomena: covariate shift, caused by novel molecular scaffolds not seen in training, and label shift, stemming from unseen reaction rules or templates [Yu *et al.*, 2023; Ye *et al.*, 2022]. The prevailing wisdom has been to treat these as separate issues, developing isolated strategies for each. This isolated perspective, while analytically useful, ignores a critical aspect of practical synthesis planning. In real-world scenarios, these shifts rarely occur in isolation. Instead, we argue that models are consistently confronted by what we formally define as a mixed distribution shift, where novel molecular structures and unseen reaction rules appear concurrently. This is not merely an additive challenge; as graphically demonstrated in Figure 1(a), this compound effect leads to a drastically more severe performance degradation than either individual shift alone. However, the field has lacked a formal recognition of this mixed-shift problem, and no framework has been specifically engineered to address it.

In this paper, we directly confront this gap. To the best of our knowledge, AgentRetro is the first framework specifically engineered to resolve the fundamental Invariance-Adaptability Paradox inherent in retrosynthesis under mixed shifts. In Retrosynthesis, models face competing objectives: they must disregard irrelevant scaffold variations (*structural noise*) while remaining highly sensitive to subtle changes in reaction rules (*mechanistic signal*). Standard domain generalization methods fail here because enforcing invariance (e.g., via DANN) often suppresses necessary mechanistic details, whereas meta-learning can become over-sensitive to scaffold noise [Ye *et al.*, 2022].

AgentRetro addresses this by structurally decoupling the chemical space, as conceptually illustrated in Figure 1(b). It achieves this through two novel, cooperative components: (i) A **Spectral Domain Enhancer (SDE)** that constructs a composite manifold based on molecular scaffolds and reaction mechanisms. Unlike generic clustering, SDE effectively isolates structural domains from mechanistic ones, identifying the underlying sources of mixed shift. (ii) A **Multi-Knowledge Agent (MKA)**, built upon a Large Language Model, which actively explores the unseen chemical space. The MKA generates targeted "pseudo-OOD" data and distills high-level chemical rationales, providing the mechanistic signal that cannot be inferred from the training data alone.

These components are unified by a principled bilevel meta-learning strategy designed to apply distinct optimization pressures: the inner loop learns to ignore scaffold variance (via adversarial training on SDE-defined structural domains), while the outer loop optimizes adaptation to new reaction types (via meta-learning on a mix of real and MKA-generated data). This dual objective directly resolves the two-pronged nature of the mixed-shift challenge.

In summary, our paper contributes to this field in three ways:

- We formally define the mixed-shift problem in retrosynthesis and propose AgentRetro, a chemistry-grounded bilevel learning framework engineered to resolve the invariance-adaptability paradox.
- We introduce a synergistic knowledge integration method that combines an internal Spectral Domain Enhancer (SDE) for structural disentanglement with an external, LLM-driven Multi-Knowledge Agent (MKA).
- We establish new state-of-the-art results on our proposed mixed-shift benchmark, boosting the Top-10 accuracy of a strong baseline, EditRetro, from 41.7% to 50.1% (+8.4%).

2 Defining and Benchmarking the Mixed-Shift Challenge

To rigorously evaluate model performance under realistic out-of-distribution (OOD) scenarios, we construct a new benchmark designed to feature mixed distribution shifts, where shifts in molecular structure and reaction strategy occur simultaneously. A mixed shift, denoted as $P(\mathcal{MT})$, arises from the combination of a covariate shift in target molecules ($P(\mathcal{M})$), where $P_{tr}(\mathcal{M}) \neq P_{tst}(\mathcal{M})$, and a label shift in retrosynthesis strategies ($P(\mathcal{T})$), where $P_{tr}(\mathcal{T}) \neq P_{tst}(\mathcal{T})$ [Yu *et al.*, 2023]. The core challenge lies in handling their co-occurrence, as real-world chemical discovery often involves both novel scaffolds and new reaction types.

Our benchmark construction process is designed to systematically isolate reactions representative of such mixed shifts. This is achieved by applying a dual-criterion selection process to the standard USPTO-50k dataset, focusing on both molecular and reaction-level properties.

2.1 Inducing Covariate Shift via Product Properties

To create a covariate shift, we select target molecules based on properties known to challenge generalization. Following established practices in OOD benchmarks [Koh *et al.*, 2021; Ji *et al.*, 2022], we rank molecules in ascending order of molecular size and scaffold complexity. Molecules with larger sizes or more complex scaffolds, which naturally contain a higher proportion of variant features ($\mathcal{M}_{var}$) compared to essential invariant features ($\mathcal{M}_{inv}$) like reaction centers [Peters *et al.*, 2015], are designated for the OOD set.

2.2 Inducing Label Shift via Template Frequency

To introduce a label shift, we focus on the rarity of the underlying reaction strategy, which we encode using reaction templates-subgraphs representing the core structural transformation. For each reaction, we identify its minimal-template (radius=0) and retro-template (radius=1+) and rank them by their occurrence frequency in the training set, similar to [Yu *et al.*, 2023]. Reactions governed by low-frequency templates are selected for the OOD set, as they represent less common or novel synthetic pathways that test a model's adaptability beyond memorized patterns.

By integrating these two criteria, our final mixed-shift OOD set comprises reactions that are simultaneously characterized by low-frequency reaction templates and high-complexity molecular structures. We ensure no template overlap between the in-distribution (ID) and OOD subsets for the selected rare templates. This construction creates a challenging and realistic benchmark that directly tests a model's ability to navigate the complex interplay of structural and strategic novelty inherent in retrosynthesis.

3 Method: AgentRetro

To address the mixed-shift challenge, we propose **AgentRetro**, a framework that synergistically leverages intrinsic data structures and extrinsic knowledge from an LLM agent. The overall architecture is depicted in Figure 2, and its core components are detailed in the following sections.

3.1 Spectral Domain Enhancer

The primary objective of the SDE is to automatically partition the training data into chemically meaningful domains, which is a prerequisite for our domain generalization strategy. This process is designed to capture the mixed-shift nature of retrosynthesis by considering both molecular and reaction-template similarities.

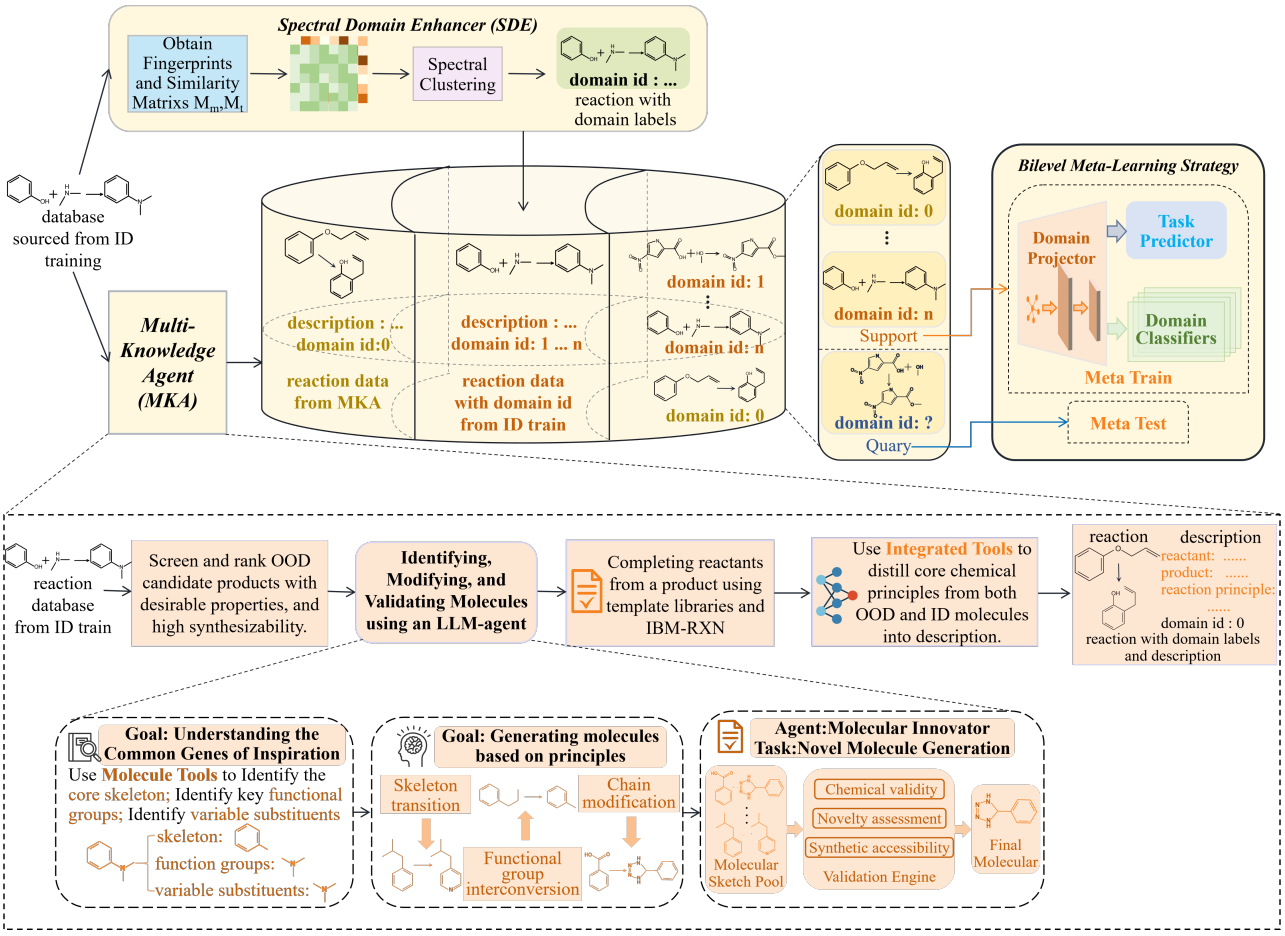

Figure 2: The AgentRetro Framework. Our framework consists of three core components. (A) Spectral Domain Enhancer (SDE): The training data is partitioned into fine-grained domains by performing spectral clustering on both molecular similarity ($\mathbf{M}_{\mathcal{M}}$) and reaction template similarity ($\mathbf{M}_{\mathcal{T}}$). (B) Multi-Knowledge Agent (MKA): An LLM agent generates targeted "pseudo-OOD" reactions and distills textual chemical rationales to augment the training set with external knowledge. (C) Bilevel Meta-Learning Strategy: A bilevel process uses the SDE-defined domains for adversarial training (to learn scaffold-invariant features) and the MKA-augmented data for meta-learning (to learn adaptable reaction principles), training a robust retrosynthesis model.

Feature and Similarity Matrix Construction. Let a reaction r_i in our dataset be represented by a product molecule m_i and a set of reactants R_i. The underlying transformation is captured by a reaction template t_i. To quantify the similarity between any two reactions r_i and r_j, we compute two metrics:

- **Molecular Similarity:** We first compute the Extended Connectivity Fingerprints (ECFP) for each product molecule, denoted as $\mathrm{ECFP}(m_i)$. The structural similarity between two products is then measured by the Tanimoto coefficient, yielding a molecular similarity matrix $\mathbf{M}_{\mathcal{M}}$ where each element $[\mathbf{M}_{\mathcal{M}}]_{ij} = \mathrm{Tanimoto}(\mathrm{ECFP}(m_i), \mathrm{ECFP}(m_j))$.
- **Reaction Template Similarity:** To capture reaction mechanisms with richer and more distinct information, we select a radius-1 retrosynthesis template (t_i) for each reaction. We then compute its "difference fingerprint" representing the transformation from product to reactants by taking the XOR operation between the ECFP of the product and reactants. This results in a template fingerprint, denoted as $\mathrm{TemplateFP}(r_i)$. Specifically, we use ECFP with a radius of 2 and 2048 bits. For reactions with multiple reactants, their individual ECFPs are summed before the XOR operation with the product's ECFP. The template similarity matrix $\mathbf{M}_{\mathcal{T}}$ is subsequently constructed, with its elements defined as $[\mathbf{M}_{\mathcal{T}}]_{ij} = \mathrm{Tanimoto}(\mathrm{TemplateFP}(r_i), \mathrm{TemplateFP}(r_j))$. This matrix quantifies the mechanistic relationships across all reactions, which is crucial for characterizing the label shift component of our mixed-shift problem.

Domain Partitioning via Composite Spectral Clustering. To explicitly capture the interplay between structural and mechanistic shifts, we perform spectral clustering on each similarity dimension independently, leveraging its strength in discovering manifold structures within the non-Euclidean chemical similarity space. We construct an affinity matrix using a k-nearest neighbors (k=10) graph and compute the nor-

malized symmetric graph Laplacian ($\mathbf{L}_{sym}$). The first $K_{\mathcal{M}}$ and $K_{\mathcal{T}}$ eigenvectors are then used as input to a standard k-means algorithm to obtain the final cluster assignments.

First, we apply spectral clustering to the molecular similarity matrix $\mathbf{M}_{\mathcal{M}}$, partitioning the dataset into $K_{\mathcal{M}}$ domains based on product structure and assigning each reaction r_i a molecular domain label $d_i^{(\mathcal{M})} \in \{1, \ldots, K_{\mathcal{M}}\}$. Concurrently, we apply the same process to the reaction template similarity matrix $\mathbf{M}_{\mathcal{T}}$, partitioning the data into $K_{\mathcal{T}}$ domains based on reaction mechanism and assigning a template domain label $d_i^{(\mathcal{T})} \in \{1, \ldots, K_{\mathcal{T}}\}$. A final, composite domain label d_i is then created for each reaction by taking the Cartesian product of its individual cluster assignments:

$$d_i = (d_i^{(\mathcal{M})}, d_i^{(\mathcal{T})}) \tag{1}$$

This process yields a rich set of up to $K_{\mathcal{M}} \times K_{\mathcal{T}}$ unique domains, which are subsequently mapped to $K_{\mathcal{M}} \times K_{\mathcal{T}}$ single categorical labels (i.e., $d_i \in \{1, \ldots, 12\}$) for utilization in the domain classifier. Each domain represents a specific combination of molecular structure and reaction mechanism. The result is a domain-annotated dataset $\mathcal{D}_{\text{ID}} = \{(r_i, d_i)\}_{i=1}^{N_{\text{ID}}}$, whose fine-grained partitioning is crucial for our bilevel meta-learning strategy.

Unlike heuristic approaches (e.g., scaffold splitting) that impose rigid, single-dimensional boundaries, our SDE captures the complex, non-Euclidean manifold of chemical reactivity. As empirically validated in our ablation studies (see Section 4.4), heuristic partitioning fails to identify latent chemical sub-populations that SDE uncovers, proving that modeling this intrinsic chemical structure is essential for robust generalization.

To visualize the effectiveness of this composite partitioning, we projected the spectral embedding of the training data into a 2D space using t-SNE, colored by the final 12 domain IDs. As shown in Figure 3, the domains form coherent and well-separated clusters, confirming that our SDE method successfully uncovers the latent structure of the reaction space.

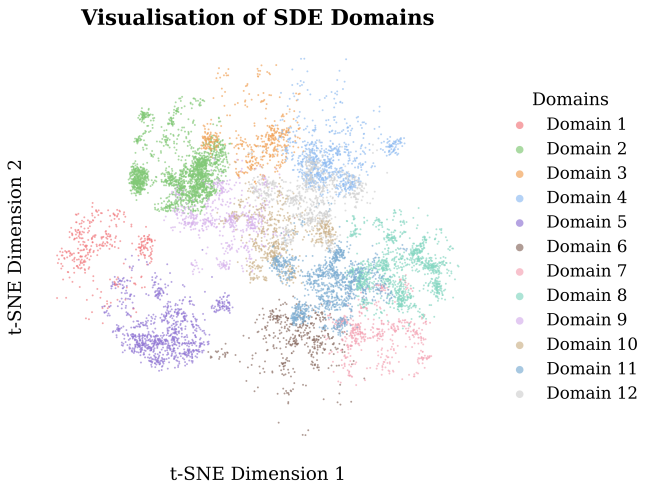

Figure 3: t-SNE visualization of the 12 chemical domains generated by the SDE on the training set. Each color represents a unique composite domain ID.

3.2 Multi-Knowledge Agent (MKA)

To address both covariate and label shifts, we introduce the MKA, an intelligent agent designed to augment the training data with targeted OOD knowledge. The MKA utilizes an LLM to perform two key functions: generating challenging "pseudo-OOD" reaction data and distilling high-level chemical principles into rich textual descriptions. These descriptions are then encoded and used as auxiliary input features during training, providing the model with rich semantic context. At inference time, when no textual description is available, a zero-vector is used as a placeholder to maintain architectural consistency. Crucially, to prevent data leakage and unfairly inflating OOD performance, all 'pseudo-OOD' reactions generated by the MKA were strictly filtered. We ensured that their underlying reaction templates (via SMARTS matching) and Bemis-Murcko scaffolds did not appear in the held-out OOD test set. As shown in Figure 2, this process follows a multi-step workflow.

Candidate Screening. The process begins with a strategic screening phase. The agent first interrogates external databases (e.g., PubChem) to identify molecules structurally dissimilar from the ID set (e.g., Tanimoto similarity $<$0.7). This initial dissimilarity filter is critical to ensure we are targeting the true OOD space. Subsequently, these candidates are ranked based on key metrics like drug-likeness and synthetic accessibility. This second filtering step is crucial to guide the LLM towards generating molecules that are not only novel but also chemically plausible and relevant to real-world synthesis challenges, thus avoiding the generation of exotic but impractical structures.

LLM-driven Molecule Generation and Validation. This core step involves a structured, principle-driven process for generating novel molecules. It begins with *principle distillation*, where the agent uses molecular tools to analyze selected seed molecules, deconstructing them into core components: the foundational skeleton, key functional groups, and variable substituents. This analysis establishes a set of design principles. Guided by these principles, the LLM agent then performs *principle-based generation*, creating a "Molecular Sketch Pool" of new structures via various modification strategies (e.g., skeleton transition, functional group interconversion, chain modification). Finally, a "Molecular Innovator" validation engine performs *multi-faceted validation*, rigorously assessing each sketch against three key criteria: (1) chemical validity (using RDKit), (2) novelty (by checking against existing databases), and (3) synthetic accessibility. Only molecules that pass this comprehensive validation are considered final, high-quality OOD products.

Reactant Completion and Knowledge Distillation. For each validated OOD product, established tools (e.g., IBM-RXN) or local template libraries are used to predict the corresponding reactants, completing the novel reaction. In the final and crucial step, the MKA prompts the LLM to perform knowledge distillation. The agent provides the LLM with the new reaction alongside related ID reactions and prompts it to generate a rich, natural-language 'rationale' explaining the re-

action's principles, product structure, and reactant properties. This textual knowledge, once embedded, provides critical semantic context that a purely structural model cannot access.

The output of this pipeline is a supplementary OOD dataset $\mathcal{D}_{\text{OOD}} = \{(r_j, d_j, \text{desc}_j)\}_{j=1}^{N_{\text{OOD}}}$, where each reaction is paired with a unified "pseudo-OOD" domain label ($d_j = 0$) and a detailed textual description. This multi-modal dataset provides a richer source of knowledge for the subsequent training stage.

3.3 Bilevel Meta-Learning Strategy

To operationalize the knowledge from SDE and MKA, we propose a bilevel training strategy that unifies structural invariance with mechanistic adaptability. Leveraging the domain-annotated source dataset $\mathcal{D}_{\text{src}} = \mathcal{D}_{\text{ID}} \cup \mathcal{D}_{\text{OOD}}$, we tackle the mixed-shift problem by integrating two complementary objectives:

Inner-Loop: Invariance via Adversarial Training. To handle covariate shift (structural noise), we adopt the domain-adversarial objective of DANN [Ganin *et al.*, 2016]. A Gradient Reversal Layer (GRL) is inserted to maximize the loss of a domain classifier $\mathcal{L}_{\text{adv}}$, forcing the feature extractor to learn scaffold-invariant representations that confuse the domain discriminator.

Outer-Loop: Adaptability via Meta-Learning. To handle label shift (novel templates), we employ the MLDG framework [Li *et al.*, 2017] to simulate domain shifts during training. In each step, we partition domains into meta-train and meta-test sets. We first compute a *hypothetical update* on the meta-train set, which simultaneously minimizes the task loss $\mathcal{L}_{\text{task}}$ and the adversarial loss:

$$\phi' = \phi - \beta \nabla_\phi \big(\mathcal{L}_{\text{task}}(\mathcal{D}_{\text{meta-train}}) - \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(\mathcal{D}_{\text{meta-train}}) \big) \tag{2}$$

where β is the inner-loop learning rate. The meta-objective $\mathcal{L}_{\text{meta}}$ is then evaluated using these updated parameters ϕ' on the held-out meta-test domain. Finally, the model parameters are updated to minimize the total objective, ensuring robustness to both shifts:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{meta}} \mathcal{L}_{\text{meta}}(\phi'; \mathcal{D}_{\text{meta-test}}) \tag{3}$$

This bilevel optimization provides a principled architecture where the inner loop enforces structural invariance and the outer loop explicitly optimizes for mechanistic adaptability. Algorithm 1 summarizes the full procedure.

Algorithm 1 AgentRetro Bilevel Training Loop

Require: Source data $\mathcal{D}_{\text{src}}$, hypers $\lambda_{\text{adv}}, \lambda_{\text{meta}}$, rates α, β.
Initialize parameters $\phi, \theta, \omega, \psi$.
while not converged **do**
  Sample domains; split into $\mathcal{D}_{\text{meta-train}}, \mathcal{D}_{\text{meta-test}}$.
  `#Inner Loop Update (on meta-train)`
  Update domain classifier to minimize its loss:
  $\psi \leftarrow \psi - \alpha \nabla_\psi \mathcal{L}_{\text{adv}}(\mathcal{D}_{\text{meta-train}})$
  Compute hypothetical parameters for the feature extractor, applying GRL to the adversarial loss component:
  $\phi' \leftarrow \phi - \beta \nabla_\phi (\mathcal{L}_{\text{task}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}})$
  `#Outer Loop Update (on meta-test)`
  Compute outer loss on the meta-test set using ϕ':
  $\mathcal{L}_{\text{outer}} \leftarrow \mathcal{L}_{\text{task}}(\phi', \theta; \mathcal{D}_{\text{meta-test}})$
  `#Global Parameter Update`
  Update the original parameters using the meta-objective:
  $\phi \leftarrow \phi - \alpha \nabla_\phi (\mathcal{L}_{\text{task}} + \lambda_{\text{meta}} \mathcal{L}_{\text{outer}})$
  $\theta \leftarrow \theta - \alpha \nabla_\theta (\mathcal{L}_{\text{task}} + \lambda_{\text{meta}} \mathcal{L}_{\text{outer}})$
end while
return trained model parameters ϕ, θ.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. All experiments are conducted on the widely used USPTO-50k dataset. To specifically evaluate OOD performance, we utilize the mixed-shift OOD benchmark constructed in Section 3. Following our procedure, the full dataset is partitioned into training, validation, in-distribution (ID) test, and out-of-distribution (OOD) test sets with a 7:1:1:1 ratio. The validity and difficulty of this benchmark are analyzed in Section 4.5.

Baselines. To demonstrate the broad applicability of our framework, we apply AgentRetro to several state-of-the-art retrosynthesis models, whose original, unmodified versions serve as our primary baselines. We apply AgentRetro to three state-of-the-art baselines representing different paradigms: **RetroComposer** [Yan *et al.*, 2022], **EditRetro** [Han *et al.*, 2024] and **GDiffRetro** [Sun *et al.*, 2025].

For all baselines, we use their official open-source implementations and follow their recommended hyperparameter configurations to ensure a fair comparison.

Implementation Details. Our framework is implemented in PyTorch, and all models were trained on NVIDIA 4090D GPUs. For AgentRetro, the balancing hyperparameters λ_{adv} and λ_{meta} were tuned via a grid search on the validation set. Based on this tuning, the number of molecular domains $K_{\mathcal{M}}$ and template domains $K_{\mathcal{T}}$ were set to 4 and 3, respectively.

Evaluation Metrics. Following standard practice in single-step retrosynthesis prediction, we report Top-k accuracy. This metric measures whether the ground-truth reactants are present within the top k predictions generated by the model. We report results for k=1, 3, 5 and 10.

4.2 Main Results on OOD Generalization

To answer our first research question, we compare the performance of the baseline models against their AgentRetro-enhanced counterparts on both the ID and OOD test sets. Table 1 shows that AgentRetro significantly boosts performance across all baselines, particularly on the OOD test set.

This validates our primary hypothesis that AgentRetro is a versatile and effective solution for mitigating performance degradation under mixed-shift conditions. The most dramatic improvement is observed with EditRetro, where AgentRetro increases the Top-10 OOD accuracy from 41.7% to 50.1%, a substantial absolute gain of **+8.4%**. Notably, our method also

Backbone	Method	Top-1		Top-3		Top-5		Top-10	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD
RetroComposer	Baseline	56.7±0.2	17.3±0.4	79.6±0.4	34.7±0.5	85.1±0.2	43.5±0.3	89.0±0.2	53.7±0.6
	+ AgentRetro	**57.6±0.4**	**23.0±0.5**	**80.8±0.3**	**40.5±0.4**	**85.8±0.3**	**48.3±0.5**	**89.8±0.3**	**60.5±0.4**
	Δ	*+0.9*	*+5.7*	*+1.2*	*+5.8*	*+0.7*	*+4.8*	*+0.8*	*+6.8*
EditRetro	Baseline	58.0±0.3	14.9±0.4	77.8±0.4	28.6±0.5	83.2±0.3	34.6±0.4	87.1±0.4	41.7±0.4
	+ AgentRetro	**59.8±0.2**	**20.6±0.3**	**80.2±0.2**	**35.3±0.3**	**86.0±0.2**	**41.9±0.4**	**89.7±0.2**	**50.1±0.3**
	Δ	*+1.8*	*+5.7*	*+2.4*	*+6.7*	*+2.8*	*+7.3*	*+2.6*	*+8.4*
GDiffRetro	Baseline	58.9±0.5	15.3±0.6	79.1±0.4	31.5±0.5	81.9±0.4	39.9±0.6	83.8±0.3	47.0±0.5
	+ AgentRetro	**59.6±0.4**	**19.5±0.5**	**80.0±0.3**	**36.2±0.4**	**82.5±0.5**	**43.5±0.4**	**84.5±0.4**	**52.6±0.4**
	Δ	*+0.7*	*+4.2*	*+0.9*	*+4.7*	*+0.6*	*+3.6*	*+0.7*	*+5.6*

Table 1: Performance comparison of baseline models and their AgentRetro-enhanced counterparts on the mixed-shift benchmark. Results are Top-k accuracies (%) on In-Distribution (ID) and Out-of-Distribution (OOD) sets, reported as mean ± std dev over 3 runs. The Δ row shows the absolute OOD improvement.

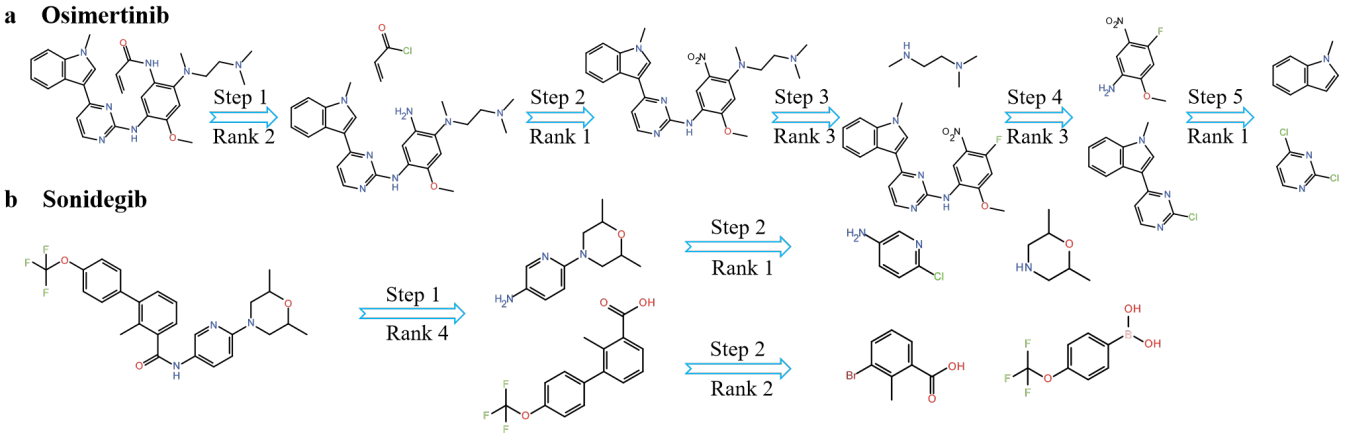

Figure 4: Multi-step retrosynthesis predictions by AgentRetro for Osimertinib and Sonidegib. Both pathways align well with literature-reported routes, demonstrating the model's practical utility in complex synthesis planning.

yields mild improvements on the ID test set, suggesting that the acquired generalization capabilities do not come at the cost of in-distribution performance, but rather reflect a more fundamentally robust understanding of chemistry.

4.3 Case Study

To evaluate the practical utility of our single-step prediction method in synthesis planning, we applied AgentRetro (with EditRetro) to generate multi-step pathways for two pharmaceutically significant compounds absent from our training set: the EGFR inhibitor **Osimertinib** and the Hedgehog pathway inhibitor **Sonidegib**. We used an iterative workflow where plausible, high-ranked precursors were selected for subsequent disconnection steps.

In both examples, as shown in Figure 4, AgentRetro produces retrosynthetic pathways that closely align with literature-reported strategies, with most key steps ranked high among the predictions. For the first case, Osimertinib (Figure 4a), our model successfully identifies a critical and chemically selective disconnection on the indole core, overcoming a challenge where pattern-based baselines like EditRetro typically falter. For the second case, Sonidegib (Figure 4b), the model's top-ranked (Rank 1) prediction is the strategic disconnection for a Suzuki coupling, correctly prioritizing a powerful C-C bond formation. These results demonstrate the strong practical potential of our method in guiding real-world retrosynthesis planning.

4.4 Ablation Studies

To understand the synergy within AgentRetro, we performed a multi-stage ablation study (Table 2) analyzing: (1) standard DG baselines, (2) our core architectural components, and (3) our specific training objectives. This progressive analysis reveals how each element is essential to the framework's overall effectiveness.

Comparing against standard DG methods contextualizes AgentRetro's performance. While both **DANN-only** (+4.2% Top-10) and **MLDG-only** (+3.6%) significantly improve upon the baseline, they fall well short of the full AgentRetro model. This highlights a key finding: neither domain invariance nor meta-adaptation alone is sufficient to fully tackle the mixed-shift challenge. Ablating our core components re-

Method	Top-1	Top-3	Top-5	Top-10
EditRetro (Base)	14.9±0.4	28.6±0.5	34.6±0.4	41.7±0.4
+ DANN-only	17.5±0.5	33.0±0.4	38.9±0.5	45.9±0.5
+ MLDG-only	17.2±0.4	32.8±0.6	38.5±0.4	45.3±0.4
w/o MKA (Full)	16.1±0.3	31.0±0.5	37.5±0.6	45.1±0.5
- w/o Pseudo-OOD	18.4±0.6	32.7±0.2	38.8±0.3	45.9±0.4
- w/o Text Desc.	18.8±0.3	33.0±0.1	39.5±0.2	46.8±0.5
w/o SDE	18.0±0.4	33.7±0.4	39.2±0.3	46.8±0.3
w/o Adv Loss	18.5±0.5	33.8±0.5	39.9±0.4	46.7±0.6
w/o MLDG	18.2±0.3	34.1±0.4	39.5±0.5	46.5±0.4
AgentRetro	**20.6±0.3**	**35.3±0.3**	**41.9±0.4**	**50.1±0.3**

Table 2: Ablation study of AgentRetro components. We include fine-grained ablations for the MKA module (*italics*) to isolate the contributions of pseudo-OOD data and textual descriptions. All values are Top-k accuracies (%) on the OOD test set.

veals their distinct and critical roles. Removing the **Multi-Knowledge Agent (w/o MKA)** causes the largest performance drop (-5.0% in Top-10), demonstrating the indispensable need for the MKA's external knowledge. To further dissect this gain, we performed fine-grained ablations shown in Table 2. Removing only the generated molecules (**w/o Pseudo-OOD**) leads to a 4.2% drop in Top-10 accuracy, confirming that targeted exploration of the chemical space is crucial. Furthermore, removing only the textual rationales (**w/o Text Desc.**) also results in a significant degradation (-3.3%), validating that textual modalities provide critical semantic context (e.g., reaction conditions and types) that graph encoders often miss. This aligns with findings in prior multimodal chemical learning works [Edwards *et al.*, 2021]. Removing the **Spectral Domain Enhancer (w/o SDE)** also results in a substantial drop (-3.3%), proving that our generalization algorithms are ineffective without the meaningful chemical domains that SDE provides as a structural foundation. Finally, ablating the training objectives confirms their specific roles. Removing the **Adversarial Loss*** impairs scaffold-invariance (-3.4% Top-10), while removing the **meta-learning component (w/o MLDG)** harms adaptability to new reaction types (-3.6%), confirming their necessity for tackling covariate and label shifts, respectively.

These results demonstrate a clear synergistic effect. The gains from individual DG baselines (+4.2% and +3.6%) are significantly smaller than the +8.4% improvement from the full AgentRetro framework. This super-additive gain confirms that AgentRetro is not a mere collection of parts, but a truly synergistic system: SDE provides the necessary structure, MKA provides the critical external knowledge, and the dual training objectives effectively leverage both.

4.5 Validation of the Mixed-Shift Benchmark

To rigorously validate the difficulty of our benchmark, we quantified the distribution shift between the **training** and test sets using three distinct theoretical metrics:

- **Maximum Mean Discrepancy (MMD^2)** [Muandet *et al.*, 2017]: Measures the distance between mean embeddings in a Reproducing Kernel Hilbert Space (RKHS) using a Tanimoto kernel.

Metric	vs. ID Test	vs. OOD Test
Product Space (Covariate Shift)		
MMD^2	0.0000	**0.0052**
SWD	0.0026	**0.0192**
EMD	0.6269	**0.6741**
Template Space (Label Shift)		
MMD^2	0.0000	**0.0078**
SWD	0.0026	**0.0235**
EMD	0.5602	**0.7461**
Combined Space (Mixed Shift)		
MMD^2	0.0000	**0.0064**
SWD	0.0026	**0.0223**
EMD	0.6547	**0.7109**

Table 3: Quantitative analysis of distribution shift using MMD^2, SWD, and EMD.

- **Earth Mover's Distance (EMD)** [Rubner *et al.*, 2000]: Intuitively measures the minimum "work" required to transform one distribution into another, computed using Jaccard distance.
- **Sliced-Wasserstein Distance (SWD)** [Lee *et al.*, 2019]: A computationally efficient approximation of EMD that compares distributions along random one-dimensional projections.

The quantitative results, detailed in Table 3, are unequivocal. Across all three metrics and feature spaces, the discrepancy scores for the OOD test set are consistently an order of magnitude larger than the negligible scores for the ID test set. This provides strong numerical evidence that the observed performance degradation is rooted in a genuine and substantial distribution shift.

5 Conclusion

In this paper, we formally defined and addressed the critical challenge of mixed shifts in retrosynthesis, where models face concurrent covariate and label shifts. We proposed AgentRetro, a novel framework whose core insight is the synergistic combination of internal domain discovery (via SDE) with external, LLM-driven knowledge augmentation (via MKA), unified under a principled bilevel learning strategy. Our extensive experiments demonstrate the effectiveness of this approach: AgentRetro significantly boosts OOD performance across multiple backbones, most notably improving EditRetro's Top-10 accuracy by a substantial +8.4%. This work showcases a promising path toward building more robust and generalizable models for scientific discovery by integrating structured, data-driven insights with the vast knowledge encoded in large language models. Future work could focus on multi-step synthesis planning and enhancing model interpretability.

Acknowledgements

Hongxin Xiang and Yiping Liu are corresponding authors. This work was supported by the National Science and Tech-

nology Major Project (Grant No. 2023ZD0120902), the National Natural Science Foundation of China (Grant Nos. 62522110 and 62472152), the Hunan Provincial Natural Science Foundation of China (Grant No. 2024JJ4015), and the Project of Yuelushan Center for Industrial Innovation (Grant No. 2025YCII0214).

References

[Corey, 1991] Elias James Corey. *The logic of chemical synthesis*. 1991.

[Edwards *et al.*, 2021] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks, May 2016. arXiv:1505.07818 [stat].

[Han *et al.*, 2024] Yuqiang Han, Xiaoyang Xu, Chang-Yu Hsieh, Keyan Ding, Hongxia Xu, Renjun Xu, Tingjun Hou, Qiang Zhang, and Huajun Chen. Retrosynthesis prediction with an iterative string editing model. *Nature Communications*, 15(1):6404, July 2024. Publisher: Nature Publishing Group.

[Ji *et al.*, 2022] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, Houtim Lai, Shaoyong Xu, Jing Feng, Wei Liu, Ping Luo, Shuigeng Zhou, Junzhou Huang, Peilin Zhao, and Yatao Bian. DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery – A Focus on Affinity Prediction Problems with Noise Annotations, January 2022. arXiv:2201.09637 [cs, q-bio].

[Jiang *et al.*, 2023] Yinjie Jiang, Yemin Yu, Ming Kong, Yu Mei, Luotian Yuan, Zhengxing Huang, Kun Kuang, Zhihua Wang, Huaxiu Yao, James Zou, Connor W. Coley, and Ying Wei. Artificial Intelligence for Retrosynthesis Prediction. *Engineering*, 25:32–50, June 2023.

[Koh *et al.*, 2021] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts, July 2021. arXiv:2012.07421 [cs].

[Lee *et al.*, 2019] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10277–10287, 2019.

[Li *et al.*, 2017] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to Generalize: Meta-Learning for Domain Generalization, October 2017. arXiv:1710.03463 [cs].

[Muandet *et al.*, 2017] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1–2):1–141, 2017.

[Peters *et al.*, 2015] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, November 2015. arXiv:1501.01332 [cs, stat].

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[Schwaller *et al.*, 2019] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, September 2019. Publisher: American Chemical Society.

[Segler and Waller, 2017] Marwin H. S. Segler and Mark P. Waller. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chemistry – A European Journal*, 23(25):6118–6128, 2017.

[Sun *et al.*, 2025] Shengyin Sun, Wenhao Yu, Yuxiang Ren, Weitao Du, Liwei Liu, Xuecang Zhang, Ying Hu, and Chen Ma. GDiffRetro: Retrosynthesis Prediction with Dual Graph Enhanced Molecular Representation and Diffusion Generation, January 2025. arXiv:2501.08001 [cs].

[Yan *et al.*, 2022] Chaochao Yan, Peilin Zhao, Chan Lu, Yang Yu, and Junzhou Huang. RetroComposer: Composing Templates for Template-Based Retrosynthesis Prediction. *Biomolecules*, 12(9):1325, September 2022. arXiv:2112.11225 [physics, q-bio].

[Ye *et al.*, 2022] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. OoD-Bench: Quantifying and Understanding Two Dimensions of Out-of-Distribution Generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7937–7948, June 2022. ISSN: 2575-7075.

[Yu *et al.*, 2023] Yemin Yu, Luotian Yuan, Ying Wei, Hanyu Gao, Xinhai Ye, Zhihua Wang, and Fei Wu. RetroOOD: Understanding Out-of-Distribution Generalization in Retrosynthesis Prediction, December 2023. arXiv:2312.10900 [cs, q-bio].