

CMAS: Enhancing Data Augmentation for Molecular Graph Classification via a Collaborative Multi-Agent System

Xuanting Fan^{1*}, Chenyu Wang^{1*}, Xin Liu^{1*}, Weiyi Ren¹, Yifan Wang^{1†}, Xiao Luo²

¹University of International Business and Economics ²University of Wisconsin–Madison

Correspondence: yifanwang@uibe.edu.cn

Abstract

Data augmentation is widely adopted in molecular graph learning due to its practical benefits in enhancing model performance. However, conventional augmentation like edge deletion often destroys critical chemical semantics, as molecular properties are highly sensitive to structural changes. To address this, we introduce CMAS, a novel Collaborative Multi-Agent System for enhancing data augmentation in molecular graph classification. Specifically, for each input molecule, we first use a BRICS-based decomposition strategy with prompt-guided LLM generation to produce structurally diverse candidates while preserving the core scaffold. To ensure that the generated candidates preserve the physicochemical properties of the original molecule, we retrieve domain-specific chemical knowledge from a knowledge graph and prompt an LLM to assess property consistency, effectively filtering out semantically divergent samples. Finally, we perform rule-based sanitization with iterative feedback to ensure chemical validity and preserve distributional diversity, resulting in a high-quality augmented molecular set. Extensive experiments on MoleculeNet benchmarks validate the superiority of CMAS against several state-of-the-art baselines. Code is available at <https://github.com/Catfromuibe/CMAS>.

1 Introduction

Molecular property prediction is a fundamental task in cheminformatics, drug discovery, and computational biology (Gilmer et al., 2017). Graph Neural Networks (GNNs) have been widely used to learn expressive representations from molecular graphs. However, molecular graph classification often suffers from limited labeled data and severe class imbalance, where practically important molecules,

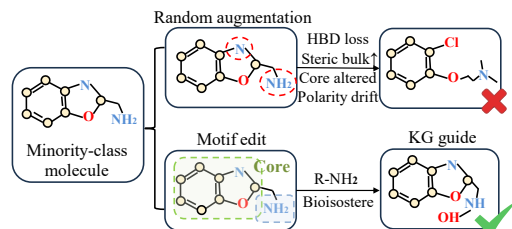


Figure 1: Random molecular augmentation may damage label-related motifs and cause property drift, whereas KG-guided motif-level editing aims to generate label-preserving and chemically reasonable molecules.

such as biologically active or toxic compounds, are fewer than dominant classes (Wu et al., 2018). For example, the dataset HIV has an imbalance factor (IF) of about 27.6. This imbalance biases models toward majority classes and weakens their generalization to minority-class molecules.

Data augmentation provides a promising solution. Existing graph augmentation methods perturb graph structures through node dropping, edge perturbation, subgraph masking, and attribute masking (You et al., 2020; Li et al., 2025a). Although effective for general graph learning, molecules are constrained by strict chemical rules. Uncontrolled perturbations may violate chemical validity, damage key functional groups, or alter intrinsic molecular properties, thereby introducing noisy samples into training. Recent molecular data augmentation methods have attempted to incorporate chemical heuristics or substructure-level information (Wang et al., 2022; Fang et al., 2024), and motif-level alignment has also been explored in graph-text pre-training (Li et al., 2025b).

However, we argue that molecular augmentation for imbalanced graph classification remains challenging due to the following two key issues: (i) *Limited control over motif-level editing positions*. Label-related semantics are often associated with key motifs. As shown in Figure 1, without controlling which motifs can be edited, augmentation may

*Equal contribution.

†Corresponding author.

disrupt critical substructures and introduce noisy samples for downstream classification. (ii) *Limited chemical guidance for modification directions*. Even when editable motifs are identified, reasonable modification still requires chemical knowledge. Inappropriate substitutions may significantly alter key physicochemical properties, causing property drift and potential label inconsistency.

Towards this end, we propose CMAS, a multi-agent molecular graph augmentation framework for imbalanced molecular graph classification. CMAS augments minority-class molecules through motif-level editing and knowledge graph-guided refinement. Given a minority-class SMILES, we first decompose it into BRICS fragments as motif-level editing units. A motif-level agent is prompted to identify editable motifs and perform controlled replacement or deletion while preserving core motifs. To improve chemical reasonableness, a knowledge-level agent retrieves top- k relevant triples from a chemical knowledge graph using the original SMILES and motif editing information as the query. The retrieved triples are organized into a knowledge prompt to guide the refinement agent in checking property consistency and revising unreasonable substitutions. The validated molecules are then added to the minority-class augmentation pool for downstream GNN training.

In summary, we briefly outline the main contributions of this paper as follows:

- ❶ **Conceptual:** We highlight the importance of motif controllability and chemical knowledge guidance in molecular data augmentation.
- ❷ **Methodological:** We propose CMAS, a novel framework that combines controllable motif-level editing with chemical knowledge graph-guided refinement.
- ❸ **Experimental:** We conduct comprehensive experiments on public datasets to evaluate the effectiveness of CMAS. The results outperform existing state-of-the-art methods.

2 Related Work

2.1 Molecular Graph Learning

Since molecules can be represented as graph structures, molecular graph learning has become a central paradigm for molecular property prediction. Early studies learned task-specific molecular features directly from molecular structures (Duvenaud et al., 2015). Subsequently, several studies designed more expressive molecular graph en-

coders (Yang et al., 2019). GROVER (Rong et al., 2020) combines message passing architecture with the attention mechanism for molecular representation learning via millions of unlabeled molecules. Beyond molecular topology, recent methods further incorporate external knowledge. Fang et al. (Fang et al., 2022) constructed a chemical element knowledge graph and leveraged it to enhance molecular contrastive learning. Despite these advances, Jiang et al. (Jiang et al., 2025) pointed out that learnable molecular representations remain severely constrained under data-scarce and highly imbalanced classification scenarios.

2.2 Data Augmentation

Data augmentation (Chen et al., 2026) has been regarded as an effective strategy for improving the generalization ability of GNNs (Fu et al., 2025, 2026). In general graph learning, GraphCL (You et al., 2020) constructs different graph views through node dropping, edge perturbation, attribute masking, and subgraph sampling. However, molecular graphs are different from generic graphs. Random perturbations on molecular graphs may disrupt key atoms. (Lee et al., 2025a). To adapt augmentation to molecular graph learning, MolCLR (Wang et al., 2022) constructs augmented molecular views through structure-aware perturbations. SMILES-based augmentation methods (Kimber et al., 2021) generate multiple equivalent SMILES representations, thereby increasing input diversity. In addition, MGSSL (Zhang et al., 2021) points out that functional groups and molecular motifs often contain indicative information related to molecular properties. Nevertheless, existing augmentation methods still rely on random or heuristic perturbations and lack explicit constraints on substructure semantics and chemical properties.

3 Model

In this section, we first formalize the problem setting and then introduce our proposed multi-agent framework for molecular data augmentation. To generate diverse and chemically plausible candidates, the Motif-Level Agent first edits peripheral motifs while preserving the core scaffold. The Knowledge-Level Agent then examines each candidate against the original molecule by querying a knowledge graph, computing an alignment score that reflects physicochemical consistency. At the final stage, the Validation Agent imposes validity,

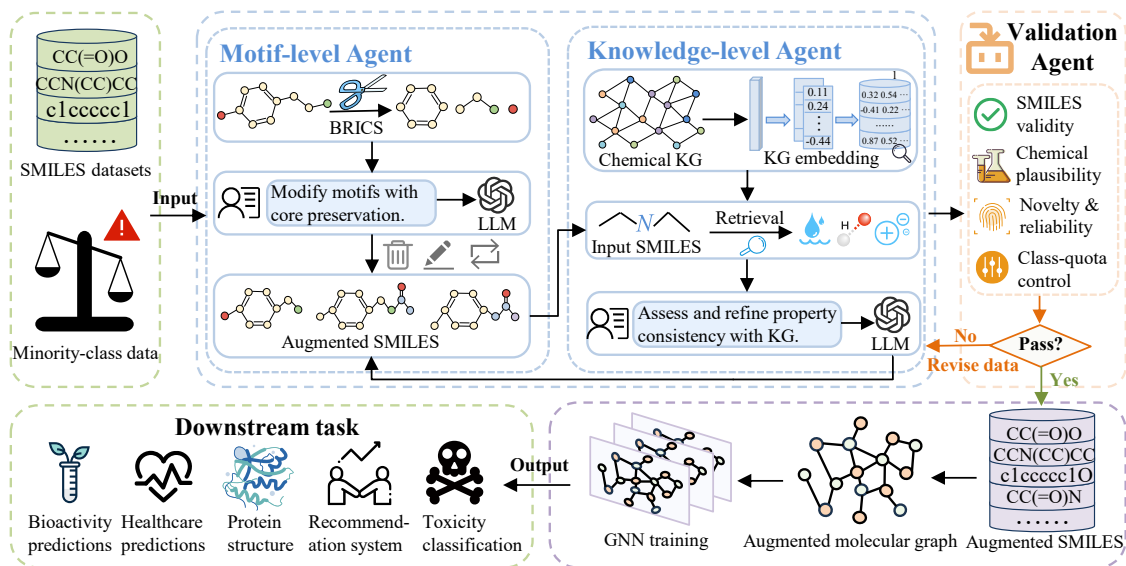


Figure 2: An overview of the proposed CMAS for molecular graph classification, which augments minority-class molecules via a Motif-Level Agent that performs core-preserving motif editing, a Validation Agent that enforces validity, chemical plausibility, novelty and class-quota control with iterative revision, and a Knowledge-Level Agent that conducts KG-guided property-consistent refinement.

diversity, and class-quota filters, and any rejected candidate is sent back to the generation agent for iterative refinement. Algorithm 1 summarises the entire augmentation procedure, and Figure 2 illustrates the overall pipeline, with agent details elaborated in the following subsections.

3.1 Problem Definition.

Considering a molecular dataset $\mathcal{D} = \{(S_i, y_i)\}_{i=1}^N$, in which S_i represents the SMILES string of the i -th molecule and $y_i \in \mathcal{Y}$ represents its ground-truth property label. The dataset is partitioned into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{valid}}$, and $\mathcal{D}_{\text{test}}$. Our goal is to address the molecular data imbalance in molecular graph classification by developing a multi-agent system for data augmentation. Specifically, we apply the proposed augmentation algorithm $\mathcal{F}(\cdot)$ to the training set $\mathcal{D}_{\text{train}}$ to synthesize high-quality virtual molecules, yielding an augmented dataset $\mathcal{D}_{\text{aug}} = \{(S'_j, y'_j)\}_{j=1}^M$. The final training corpus is then constructed as $\tilde{\mathcal{D}}_{\text{train}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{aug}}$. This enrichment of training data diversity facilitates model learning and boosts classifier performance.

3.2 Motif-Level Generation Agent: Scaffold-Constrained Molecular Editing.

The goal of the Motif-Level Agent is to generate diverse molecules by replacing or deleting functional motifs while preserving the structural integrity of the original molecule (Li et al., 2025b). The in-

put molecule SMILES S is first converted into a molecular graph $G = (\mathcal{V}, \mathcal{E})$. To obtain a set of functional groups, we decompose G into motifs using the BRICS algorithm (Zhang et al., 2021). BRICS defines 16 retrosynthesis-based cleavage rules, which identify a set of breakable bonds:

$$E_{\text{break}} = \{(u, v) \in E \mid \text{Rule}(u, v) = \text{True}\}, \quad (1)$$

where $\text{Rule}(u, v)$ indicates whether the bond between atoms u and v matches any of the predefined reaction types (Details can be seen in the appendix). We then remove these breakable bonds from the original edge set to obtain the residual edge set:

$$\mathcal{E}' = \mathcal{E} \setminus E_{\text{break}}. \quad (2)$$

The resulting residual graph is $G' = (\mathcal{V}, \mathcal{E}')$, which is generally disconnected. Each connected component of G' corresponds to a chemically meaningful fragment. We define the motif set $\mathcal{M}$ as the collection of all such connected components:

$$\mathcal{M} = \{m_1, m_2, \dots, m_K\}, \quad (3)$$

where each motif $m_i = (\mathcal{V}_i, \mathcal{E}_i)$ is a maximal connected subgraph of G' . Specifically, $\mathcal{V}_i \subseteq \mathcal{V}$ denotes the vertex set of the i -th connected component, and $\mathcal{E}_i \subseteq \mathcal{E}$ contains all edges whose two endpoints belong to $\mathcal{V}_i$. Each vertex set $\mathcal{V}_i$ satisfies the maximal property, meaning that no extra vertex from $\mathcal{V}$ can be incorporated into the subgraph

without destroying its internal connectivity in G' . Dummy atoms are attached at each cleavage site to mark connection points, thereby providing an explicit structural scaffold that prevents the LLM from generating chemically incoherent sequences.

Based on the original molecule S and its motif set $\mathcal{M}$, we construct a structured system prompt:

$$\mathcal{P}(S, \mathcal{M}) = \text{Concat}(\mathcal{R}_c, \mathcal{C}_c, \mathcal{I}_f, \mathcal{I}_d, \mathcal{C}_s), \quad (4)$$

where $\mathcal{R}_c$ assigns the LLM the role of a domain expert, $\mathcal{C}_c$ supplies the original SMILES S and the motif set $\mathcal{M}$, $\mathcal{I}_f$ enforces that core pharmacophores and ring systems remain unchanged, $\mathcal{I}_d$ specifies perturbation rules for peripheral motifs, and $\mathcal{C}_s$ requires the output to be a single syntactically complete SMILES string.

Finally, the prompt $\mathcal{P}(S, \mathcal{M})$ is fed into a large language model, which acts as a conditional sampler to generate N candidate molecules:

$$c^{(n)} = \text{LLM}(\mathcal{P}(S, \mathcal{M})), \quad n = 1, 2, \dots, N, \quad (5)$$

where each $c^{(n)}$ is a candidate molecule expected to preserve the semantic property label of the original molecule S .

3.3 Knowledge-Level Agent: KG-Guided Property Alignment.

The Knowledge-Level Agent evaluates whether a generated candidate molecule preserves the physicochemical characteristics of the original molecule. Given a molecule represented by a SMILES string, we first parse it into a molecular graph using RDKit and extract its atom set $\mathcal{A} = \{a_1, a_2, \dots, a_n\}$. Each atom a is mapped to its corresponding entity $\phi(a)$ in ElementKG (Fang et al., 2022), from which we retrieve the associated knowledge triples:

$$\mathcal{G}(a) = \{(\phi(a), r, v) \mid (\phi(a), r, v) \in \text{KG}\}, \quad (6)$$

where r and v denote the relation type and attribute value, respectively.

To reduce noisy relations, we manually assign a predefined priority score to each relation type according to its chemical relevance (details are provided in the Appendix). The top- k highest-priority triples are selected:

$$\mathcal{R}_k(a) = \text{Top-}k_{\text{priority}}(\mathcal{G}(a)). \quad (7)$$

The molecular knowledge summary is then constructed by aggregating the selected relations as:

$$\mathcal{K}(S) = \bigcup_{a \in \mathcal{A}} \{(a, \mathcal{R}_k(a))\}. \quad (8)$$

Algorithm 1: The proposed CMAS for molecular graph augmentation.

Input: D_{train} ; minority labels; R ; τ ; L_{max} ;
 τ_{kg} , δ_{div} ; LLMs; KG

Output: D_{aug} ; $\tilde{D}_{\text{train}} = D_{\text{train}} \cup D_{\text{aug}}$

```

1  $D_{\text{aug}} \leftarrow \emptyset$ ;
2 foreach minority  $S$  with label  $y$  do
3   Extract BRICS motifs  $M$  from  $S$ ;
4    $\text{count} \leftarrow 0$ ;
5   while  $\text{count} < R$  do
6      $\text{retry} \leftarrow 0$ ,  $\text{accepted} \leftarrow \text{False}$ ;
7     while  $\neg \text{accepted}$  and
8        $\text{retry} < L_{\text{max}}$  do
9        $c \leftarrow \text{LLM}_{\text{motif}}(\mathcal{P}(S, M))$ ;
10       $p_{\text{align}} \leftarrow \text{LLM}_{\text{kg}}(\mathcal{P}_{\text{kg}}(K(S), K(c)))$ ;
11      if  $\text{RDKitValid}(c)$  and
12         $p_{\text{align}} \geq \tau_{\text{kg}}$  then
13        if  $\text{Sim}_{\text{max}}(c, D_{\text{train}} \cup D_{\text{aug}}) \leq \delta_{\text{div}}$  then
14          Add  $c$  with label  $y$  to  $D_{\text{aug}}$ ;
15           $\text{count} \leftarrow \text{count} + 1$ ;
16           $\text{accepted} \leftarrow \text{True}$ ;
17         $\text{retry} \leftarrow \text{retry} + 1$ ;
18      if  $\neg \text{accepted}$  then
19        break;
20 return  $\tilde{D}_{\text{train}} = D_{\text{train}} \cup D_{\text{aug}}$ ;

```

Similarly, we construct $\mathcal{K}(c)$ for a generated candidate molecule c . Based on the two summaries, we formulate an LLM judging prompt:

$$\mathcal{P}_{\text{kg}} = \mathcal{P}_{\text{kg}}(\mathcal{K}(S), \mathcal{K}(c)). \quad (9)$$

The LLM then automatically performs a comprehensive, fine-grained semantic consistency assessment between the original molecule and the generated candidate, efficiently producing a reliable normalized alignment score:

$$p_{\text{align}} = \text{LLM}(\mathcal{P}_{\text{kg}}) \in [0, 1]. \quad (10)$$

The alignment score serves as a lightweight proxy for preserving molecular property consistency. A candidate is accepted if $p_{\text{align}} \geq \tau_{\text{kg}}$, where τ_{kg} is a predefined threshold. Otherwise, the candidate SMILES c together with the rejection reason is returned to the Motif-Level Agent for regeneration until a qualified candidate is obtained or the retry limit is reached.

3.4 Validation Agent: Chemical Constraint Verification and Diversity Regularization.

The Validation Agent performs final candidate filtering by enforcing chemical validity, semantic alignment, and structural diversity constraints. Given a generated molecule c , we first apply RDKit sanitization to verify valence correctness, aromaticity consistency, and bond feasibility. Molecules containing invalid valence states, disconnected structures, or chemically infeasible aromatic systems are directly rejected.

To avoid generating near-duplicate molecules, we further compute the Tanimoto similarity between c and molecules in the training set based on molecular fingerprints. Candidates with average similarity scores exceeding the predefined threshold δ_{div} are filtered out to preserve augmentation diversity and reduce distributional collapse. In addition, only candidates whose alignment scores satisfy $p_{\text{align}} \geq \tau_{\text{kg}}$ are retained to maintain physicochemical consistency with the original molecule.

Rejected candidates, together with their corresponding rejection reasons, are returned to the Motif-Level Agent for regeneration until a qualified molecule is obtained or the retry limit is reached.

4 Experiment

In this section, we aim to answer the following research questions: (RQ1) What are the advantages of CMAS compared with other baseline methods? (RQ2) How do different parameter settings influence the performance of our method? (RQ3) What are the standalone and joint effects of the three agents? (RQ4) How do different LLM backbones affect the performance of our model? (RQ5) What advantages do the augmented data generated by our method have over those produced by other data augmentation approaches?

4.1 Experiment Setup

Datasets. We adopt eight widely used molecular datasets from MoleculeNet (Wu et al., 2018), including SIDER, BACE, BBBP, ClinTox, Tox21, HIV, MUV, and ToxCast. In terms of IF, BACE (about 1.1) and BBBP (about 2) are relatively low, while the others generally have IF around or above 10, indicating a clear imbalance. Detailed statistics are provided in Table 4.

Evaluation Metrics. We evaluate our method using standard metrics tailored to each task. For all classification tasks, we report both accuracy

and ROC-AUC to assess the discriminative performance of baselines and our method.

Baseline Methods. We compare CMAS with seven baselines, including one topology-based augmentation method, DropEdge (Rong et al., 2019); four mixup-based methods, M-Evolve (Zhou et al., 2020), Graph Transplant (Park et al., 2022), G-Mixup (Han et al., 2022), and Mix-key (Jiang et al., 2024); one diffusion-based method, GenMol (Lee et al., 2025b); and one LLM-based augmentation method, DemoGraph (Feng et al., 2025). These methods collectively represent the mainstream approaches to graph data augmentation.

Implementation Details. In our experiment, we employ four GNN backbones: GCN (Kipf and Welling, 2016), GIN (Xu et al., 2018), GAT (Veličković et al., 2017), and GraphSAGE (Hamilton et al., 2017), each consisting of 3 layers with a hidden dimension of 128, followed by global mean pooling and dropout of 0.3 after each layer. To better assess the model’s generalization ability, we employ random scaffold splitting with an 8:1:1 train/validation/test ratio for all datasets. During training, we apply the Adam optimizer with a fixed learning rate of 0.001, a batch size of 32, and 150 epochs. For our method, we adopt Qwen 3.6 as the LLM for both the motif-level agent and knowledge-level agent, retrieve the 5 closest knowledge samples, fix the motif replacement rate $\tau = 0.2$, and set the knowledge-alignment threshold $\tau_{\text{kg}} = 0.3$ and diversity threshold $\delta_{\text{div}} = 0.4$ as optimal values. For other baselines, we use their paper’s code with needed modifications.

4.2 The Outperformance of CMAS (RQ1)

To address this question, we conduct comprehensive experiments across eight benchmark datasets using different GNN backbones with 7 different augmentation-based methods, with results reported in Table 1. The following observations are obtained from a detailed analysis:

- CMAS consistently outperforms augmentation baselines across all datasets and different GNN backbones. This stems from its multi-agent synergy: the Motif-level Agent enables perturbations via motif deletion and replacement, thereby preserving structural similarity and conformational constraints. Thus, the knowledge-level Agent retrieves relevant chemical knowledge to construct guided filters, which screen the molecules generated by the former and select those sharing physico-

Table 1: Performance comparison of data augmentation methods on molecular property prediction. Results are reported as Accuracy (%), mean $\pm$ standard deviation). The best and second-best results in each column are highlighted with blue and yellow backgrounds, respectively. For the full set of results, please refer to Table 5.

Encoder	Method	SIDER	BACE	BBBP	ClinTox	ToxCast	HIV	MUV	Tox21
GCN	DropEdge	72.44 $\pm$ 1.02	69.60 $\pm$ 3.20	80.60 $\pm$ 1.40	91.07 $\pm$ 3.11	83.06 $\pm$ 0.83	92.32 $\pm$ 0.76	97.36 $\pm$ 0.25	85.17 $\pm$ 1.58
	M-Evolve	75.24 $\pm$ 1.60	75.79 $\pm$ 3.44	86.04 $\pm$ 1.95	92.21 $\pm$ 1.89	76.21 $\pm$ 8.53	96.45 $\pm$ 0.76	98.71 $\pm$ 0.17	89.72 $\pm$ 0.69
	Graph Transplant	61.90 $\pm$ 2.70	65.00 $\pm$ 6.65	82.34 $\pm$ 5.11	90.81 $\pm$ 3.34	83.01 $\pm$ 1.03	96.48 $\pm$ 0.25	98.99 $\pm$ 0.08	90.82 $\pm$ 3.61
	G-Mixup	76.13 $\pm$ 1.16	74.74 $\pm$ 4.72	85.56 $\pm$ 5.00	91.34 $\pm$ 2.91	82.65 $\pm$ 1.12	96.27 $\pm$ 0.38	99.79 $\pm$ 0.03	92.32 $\pm$ 3.70
	Mix-Key	61.70 $\pm$ 4.10	73.70 $\pm$ 3.10	83.20 $\pm$ 1.90	90.74 $\pm$ 2.95	85.91 $\pm$ 5.70	89.48 $\pm$ 0.66	97.15 $\pm$ 0.30	86.68 $\pm$ 1.70
	Genmol	73.50 $\pm$ 3.90	75.20 $\pm$ 2.90	85.40 $\pm$ 1.80	91.38 $\pm$ 4.20	88.32 $\pm$ 5.50	94.45 $\pm$ 0.50	98.57 $\pm$ 0.16	90.54 $\pm$ 1.60
	DemoGraph	75.09 $\pm$ 1.87	75.35 $\pm$ 4.29	84.96 $\pm$ 2.15	91.38 $\pm$ 2.57	87.58 $\pm$ 3.02	94.72 $\pm$ 0.42	97.84 $\pm$ 0.28	90.01 $\pm$ 4.03
	Ours	77.24 $\pm$ 1.40	78.16 $\pm$ 4.02	87.64 $\pm$ 2.40	94.30 $\pm$ 2.24	90.12 $\pm$ 2.85	96.83 $\pm$ 0.30	99.89 $\pm$ 0.03	92.58 $\pm$ 0.45
GIN	DropEdge	59.30 $\pm$ 1.70	76.84 $\pm$ 6.01	85.56 $\pm$ 5.74	92.15 $\pm$ 2.87	83.33 $\pm$ 0.92	96.64 $\pm$ 0.65	97.15 $\pm$ 1.13	86.03 $\pm$ 0.87
	M-Evolve	73.95 $\pm$ 0.69	75.39 $\pm$ 3.40	86.07 $\pm$ 4.29	93.55 $\pm$ 1.54	74.18 $\pm$ 8.67	95.81 $\pm$ 0.61	98.37 $\pm$ 0.33	92.54 $\pm$ 0.68
	Graph Transplant	63.70 $\pm$ 2.80	79.32 $\pm$ 1.20	83.32 $\pm$ 6.99	91.14 $\pm$ 2.74	83.36 $\pm$ 1.02	95.49 $\pm$ 1.53	98.69 $\pm$ 0.16	92.89 $\pm$ 3.39
	G-Mixup	62.80 $\pm$ 0.80	78.03 $\pm$ 2.93	85.85 $\pm$ 6.35	90.60 $\pm$ 3.26	59.94 $\pm$ 3.46	96.44 $\pm$ 1.02	98.79 $\pm$ 0.76	92.14 $\pm$ 4.12
	Mix-Key	63.10 $\pm$ 2.40	79.10 $\pm$ 1.50	84.30 $\pm$ 2.00	83.87 $\pm$ 5.10	86.91 $\pm$ 4.90	94.72 $\pm$ 0.90	98.07 $\pm$ 0.11	86.69 $\pm$ 1.70
	Genmol	73.50 $\pm$ 2.30	81.30 $\pm$ 1.40	86.00 $\pm$ 1.90	92.77 $\pm$ 4.90	88.91 $\pm$ 4.70	96.32 $\pm$ 0.80	98.37 $\pm$ 0.09	91.79 $\pm$ 1.50
	DemoGraph	73.99 $\pm$ 1.22	82.38 $\pm$ 3.05	84.94 $\pm$ 0.91	84.95 $\pm$ 3.42	88.44 $\pm$ 2.69	95.07 $\pm$ 2.17	97.20 $\pm$ 0.24	89.95 $\pm$ 0.41
	Ours	76.14 $\pm$ 1.09	84.79 $\pm$ 3.23	87.02 $\pm$ 0.83	93.48 $\pm$ 3.69	90.61 $\pm$ 2.84	97.45 $\pm$ 1.09	99.23 $\pm$ 0.19	93.17 $\pm$ 0.46
GAT	DropEdge	57.30 $\pm$ 2.60	65.90 $\pm$ 2.90	80.60 $\pm$ 4.00	92.35 $\pm$ 2.73	82.17 $\pm$ 0.63	96.23 $\pm$ 0.37	98.53 $\pm$ 0.08	85.21 $\pm$ 0.73
	M-Evolve	74.67 $\pm$ 1.20	74.87 $\pm$ 3.59	86.27 $\pm$ 2.50	92.61 $\pm$ 1.79	78.62 $\pm$ 11.71	96.52 $\pm$ 0.83	99.52 $\pm$ 0.02	92.08 $\pm$ 1.02
	Graph Transplant	61.00 $\pm$ 2.40	71.50 $\pm$ 4.40	83.61 $\pm$ 5.99	90.47 $\pm$ 3.10	82.91 $\pm$ 1.14	96.47 $\pm$ 0.33	99.65 $\pm$ 0.01	92.44 $\pm$ 3.84
	G-Mixup	61.50 $\pm$ 1.50	71.97 $\pm$ 4.79	84.88 $\pm$ 6.62	91.28 $\pm$ 2.95	82.83 $\pm$ 1.20	96.51 $\pm$ 0.78	99.79 $\pm$ 0.03	91.76 $\pm$ 3.72
	Mix-Key	61.30 $\pm$ 2.00	70.60 $\pm$ 3.90	82.40 $\pm$ 3.00	88.37 $\pm$ 8.50	85.32 $\pm$ 7.10	92.05 $\pm$ 1.00	97.62 $\pm$ 0.93	85.18 $\pm$ 2.40
	Genmol	73.00 $\pm$ 1.90	74.50 $\pm$ 4.20	85.90 $\pm$ 2.80	91.77 $\pm$ 6.82	87.12 $\pm$ 6.90	95.65 $\pm$ 0.90	98.92 $\pm$ 0.14	90.38 $\pm$ 2.20
	DemoGraph	75.52 $\pm$ 1.86	73.53 $\pm$ 0.98	84.39 $\pm$ 2.31	92.00 $\pm$ 2.30	88.00 $\pm$ 2.94	94.57 $\pm$ 0.22	98.66 $\pm$ 0.03	90.11 $\pm$ 0.48
	Ours	77.34 $\pm$ 1.74	75.66 $\pm$ 0.93	86.34 $\pm$ 2.22	94.41 $\pm$ 2.23	90.07 $\pm$ 2.88	96.89 $\pm$ 0.19	99.91 $\pm$ 0.02	92.73 $\pm$ 0.45
GraphSAGE	DropEdge	57.60 $\pm$ 3.70	62.80 $\pm$ 2.60	78.80 $\pm$ 2.10	79.70 $\pm$ 4.90	78.99 $\pm$ 0.93	79.88 $\pm$ 1.47	97.37 $\pm$ 0.15	88.52 $\pm$ 3.27
	M-Evolve	74.34 $\pm$ 1.24	63.48 $\pm$ 1.35	80.46 $\pm$ 3.26	86.41 $\pm$ 1.87	76.09 $\pm$ 12.03	88.31 $\pm$ 2.31	98.01 $\pm$ 0.18	91.68 $\pm$ 0.79
	Graph Transplant	60.20 $\pm$ 3.90	66.20 $\pm$ 4.40	81.60 $\pm$ 1.20	90.81 $\pm$ 3.34	83.41 $\pm$ 0.98	95.01 $\pm$ 1.54	98.39 $\pm$ 0.16	92.79 $\pm$ 3.46
	G-Mixup	60.00 $\pm$ 0.80	75.13 $\pm$ 4.27	84.39 $\pm$ 6.82	90.94 $\pm$ 3.18	82.70 $\pm$ 1.06	87.69 $\pm$ 2.35	99.79 $\pm$ 0.03	91.80 $\pm$ 3.89
	Mix-Key	59.29 $\pm$ 2.70	74.57 $\pm$ 5.40	80.12 $\pm$ 6.30	87.23 $\pm$ 2.10	84.69 $\pm$ 0.80	80.09 $\pm$ 2.90	95.61 $\pm$ 0.90	75.09 $\pm$ 3.10
	Genmol	71.33 $\pm$ 2.80	74.66 $\pm$ 2.30	84.15 $\pm$ 2.70	90.36 $\pm$ 2.50	87.43 $\pm$ 5.28	92.77 $\pm$ 2.80	99.32 $\pm$ 0.10	90.64 $\pm$ 3.20
	DemoGraph	74.76 $\pm$ 1.69	72.92 $\pm$ 0.97	83.72 $\pm$ 2.15	92.17 $\pm$ 2.09	87.89 $\pm$ 2.76	94.03 $\pm$ 0.21	96.97 $\pm$ 0.15	80.09 $\pm$ 0.41
	Ours	76.82 $\pm$ 1.87	77.41 $\pm$ 2.24	85.37 $\pm$ 2.11	93.29 $\pm$ 2.90	90.11 $\pm$ 2.85	96.60 $\pm$ 0.19	99.89 $\pm$ 0.10	92.91 $\pm$ 0.55

ochemical similarity with the original while retaining functional analogy.

- We observe that CMAS exhibits notably lower performance variance compared to other baselines such as DropEdge and M-Evolve. This enhanced stability is directly attributed to the Validation Agent, which rigorously filters chemically invalid or semantically misaligned samples through pre-defined rule sets, thereby preventing distributional noise and ensuring high-quality augmentations. Consequently, this stability not only boosts average performance but also guarantees more consistent and reproducible results across different random seeds and data splits, a critical advantage for reliable drug discovery where prediction certainty is paramount.
- In contrast, existing methods such as DropEdge, Graph Transplant, and G-Mixup often inject noise due to random edge dropping, topology-agnostic interpolation, or unconstrained subgraph mixing. CMAS circumvents these issues through its multi-agent synergy, which preserves structural integrity, enforces physicochemical similarity, and incorpo-

rates rigorous sample validation, thereby generating cleaner, more task-relevant augmentations with superior performance, ensuring faithful structural and chemical augmentations.

4.3 Parameter Study (RQ2)

To address this question, we conduct a sensitivity analysis on the motif replacement rate τ , the knowledge-alignment threshold τ_{kg} , and the diversity threshold δ_{div} . Specifically, τ and τ_{kg} are chosen from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, while δ_{div} is varied from 0.1 to 0.9 with a step size of 0.1, covering a wide and practical range.

- **Effect of the motif replacement rate τ .** Figure 3 shows that across all evaluated datasets, the accuracy first rises and then gradually falls as τ increases, with its peak occurring at $\tau \approx 0.2$. This is because moderate motif editing effectively enriches structural diversity, whereas excessive replacement disrupts molecular scaffolds and thereby substantially weakens the intrinsic relationship between structure and property.
- **Effect of τ_{kg} and δ_{div} .** Figure 4 shows that mod-

Table 2: Ablation study on agent components (values are in %). Results are reported as Accuracy and AUC (%), mean $\pm$ standard deviation). The best results in each column are highlighted with blue backgrounds.

Module			SIDER		BACE		BBBP		Tox21	
Motif	Knowledge	Validation	acc	auc	acc	auc	acc	auc	acc	auc
✗	✗	✗	59.40 $\pm$ 2.80	55.10 $\pm$ 1.70	64.80 $\pm$ 2.90	71.70 $\pm$ 4.10	80.30 $\pm$ 4.10	77.50 $\pm$ 4.70	90.94 $\pm$ 2.84	77.49 $\pm$ 5.31
✓	✗	✗	65.23 $\pm$ 2.48	60.14 $\pm$ 1.87	68.47 $\pm$ 3.12	74.83 $\pm$ 3.47	82.13 $\pm$ 2.76	80.18 $\pm$ 3.24	91.82 $\pm$ 2.67	78.47 $\pm$ 4.98
✗	✓	✗	70.52 $\pm$ 2.17	63.48 $\pm$ 1.82	72.84 $\pm$ 2.68	79.47 $\pm$ 2.93	84.26 $\pm$ 2.13	83.14 $\pm$ 2.47	90.84 $\pm$ 2.43	79.76 $\pm$ 4.52
✓	✗	✓	69.83 $\pm$ 1.86	62.82 $\pm$ 1.58	72.51 $\pm$ 2.43	79.53 $\pm$ 2.57	83.47 $\pm$ 1.78	83.23 $\pm$ 2.08	90.51 $\pm$ 2.31	79.24 $\pm$ 4.23
✗	✓	✓	74.84 $\pm$ 1.67	65.53 $\pm$ 1.47	76.23 $\pm$ 2.08	84.18 $\pm$ 2.17	86.54 $\pm$ 1.52	87.76 $\pm$ 1.76	91.47 $\pm$ 2.12	80.53 $\pm$ 3.84
✓	✓	✗	75.48 $\pm$ 1.48	66.24 $\pm$ 1.37	77.53 $\pm$ 1.87	86.47 $\pm$ 1.68	87.13 $\pm$ 4.02	90.47 $\pm$ 1.38	89.82 $\pm$ 1.97	80.87 $\pm$ 3.52
✓	✓	✓	77.24$\pm$1.40	66.83$\pm$4.03	78.16$\pm$4.02	87.38$\pm$1.14	87.64$\pm$2.40	91.41$\pm$1.12	92.58$\pm$0.45	81.75$\pm$0.86

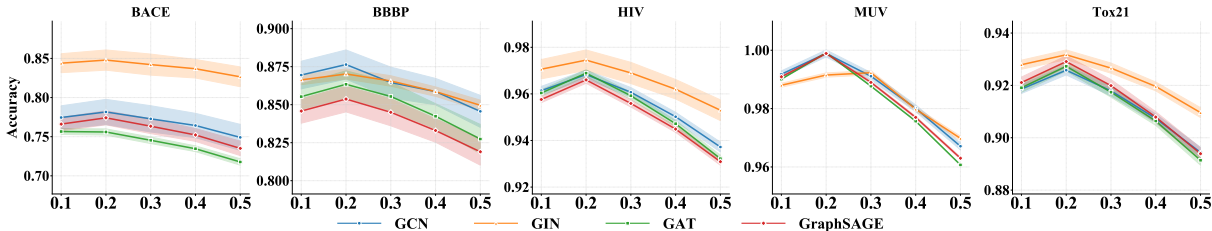


Figure 3: Effect of the motif replacement rate τ across five different datasets.

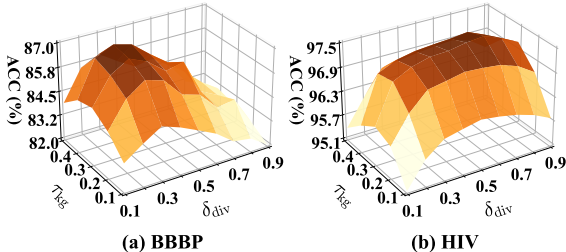


Figure 4: Sensitivity analysis of the knowledge-alignment threshold τ_{kg} and diversity threshold δ_{div} on the GAT encoder.

erate values of τ_{kg} and δ_{div} yield the best results. Increasing τ_{kg} improves semantic consistency by filtering property-inconsistent molecules, while a suitable δ_{div} balances structural novelty and distributional stability. In contrast, thresholds that are too strict or too loose reduce performance due to insufficient diversity or excessive semantic deviation. Moreover, BBBP exhibits a sharper performance peak, indicating higher sensitivity to parameter changes, whereas HIV shows a smoother performance surface and remains stable across a wider parameter range. This difference suggests that larger and more structurally diverse datasets are more tolerant to augmentation perturbations.

Therefore, we fix $\tau = 0.2$ and adopt moderate values of τ_{kg} and δ_{div} in all main experiments.

4.4 Ablation Study (RQ3)

In the preceding sections, we have shown that our method achieves strong performance on the molecular property prediction task. To further dissect

the contribution of each component in our framework, we conduct ablation experiments on four benchmark datasets using GCN as the backbone encoder. As shown in Table 2, we evaluate seven combinations of the three agents: Motif (motif-level agent), Knowledge (knowledge-level agent), and the validation agent. The variant with none of the agents serves as the baseline, which applies no augmentation at all. For the Motif agent specifically, when it is disabled (denoted as ✗), we replace it with random edge deletion as a naive augmentation strategy. For the other agents, disabling simply removes them. Based on the experimental results, we make the following key observations.

- We first evaluate the Motif and Knowledge agents in isolation. Unlike random edge deletion that disrupts molecular topology, the Motif agent preserves structural integrity by applying local perturbations at the functional group level, which avoids breaking the original backbone connections. This allows it to outperform the traditional edge-dropping baseline. The Knowledge agent surpasses the baseline, as it enforces physicochemical similarity instead of random feature masking.
- With all three agents integrated into a full model, this complete architecture achieves the best results across all datasets, outperforming every two-agent combination and all ablated variants. The synergistic integration clearly yields the most discriminative molecular representations, thus validating the effectiveness of our proposed design. This confirms that the three agents are complementary

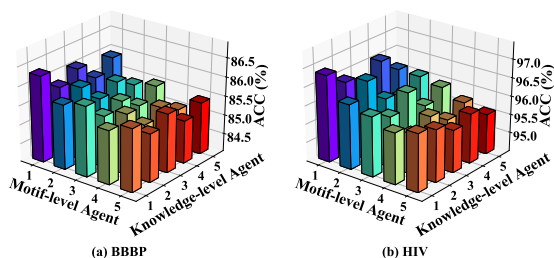


Figure 5: Analysis of LLM backbone. Both axes enumerate the same five backbones in fixed order: Qwen3.6, DeepSeek-v4, Qwen3:8b, Llama3.1:8b, and GLM5.1. (The axes label them as 1–5 for brevity.)

Table 3: Distribution learning performance comparison on BBBP dataset.

Model	Validity $\uparrow$	Diversity $\uparrow$	Similarity $\uparrow$	FCD $\downarrow$	Synth. $\downarrow$
DropEdge	72.84	0.9211	0.5634	28.441	1.874
Mix-Key	86.72	0.9415	0.8422	13.227	1.462
DemoGraph	89.31	0.9037	0.9014	8.964	1.218
Ours	95.84	0.9116	0.9485	5.771	0.986

rather than redundant, jointly addressing structural diversity, chemical semantics, and sample validity. The consistent gains across all benchmarks further demonstrate that their collaboration is essential for robust molecular augmentation.

4.5 Different LLM Backbones Empirical Study (RQ4)

In light of the importance of LLM backbones in our knowledge-guided augmentation framework, we further investigate the effects of using different LLMs for the Motif-Level Agent and Knowledge-Level Agent. We conduct experiments with several representative open-source and proprietary LLMs accessed through APIs, as illustrated in Figure 5. The results show that different backbone combinations lead to only relatively small performance fluctuations, with the overall ACC remaining within a narrow range. Among all configurations, the combination of Qwen3.6 for both the Motif-Level Agent and the Knowledge-Level Agent achieves the best performance in HIV and BBBP. At the same time, most other combinations also maintain competitive results. This phenomenon suggests that the proposed framework is not overly dependent on a specific LLM backbone. Instead, the collaborative augmentation pipeline itself contributes the major performance gains by jointly ensuring structural diversity and semantic consistency. This stability validates the robustness of our framework, demonstrating that the augmentation pipeline does not over-rely on a single proprietary model to achieve reliable molecular generation.

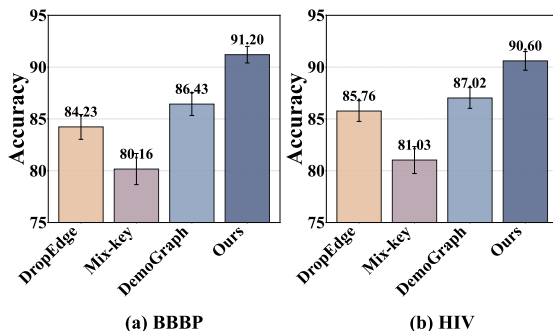


Figure 6: Analysis of data faithfulness

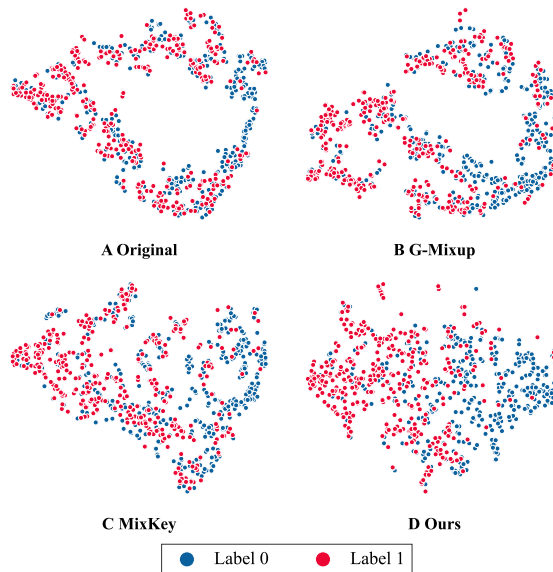


Figure 7: T-SNE visualization of molecular representations on the BACE dataset.

4.6 Data Quality Evaluation Study (RQ5)

The previous experiments have demonstrated the exceptional performance of the data generated by our CMAS for model training. In this section, we further construct a comprehensive experiment to evaluate the intrinsic quality of augmented molecular graphs generated by our method and baseline approaches from two core perspectives: distribution learning and data faithfulness. (Long et al., 2024).

Distribution Learning. We evaluate the alignment between generated molecules and the original data distribution using five metrics: Validity, Diversity (Bajusz et al., 2015), Similarity, FCD (Preuer et al., 2018), and Synth (Qiao et al., 2025). As illustrated in Table 3, CMAS achieves good performance across the majority of evaluation metrics, with particularly strong results in validity, similarity, and synthesizability, while maintaining competitive diversity. Specifically, our method achieves higher chemical validity and better synthesizability while maintaining a favorable trade-off between structural diversity and molecular similarity. It also

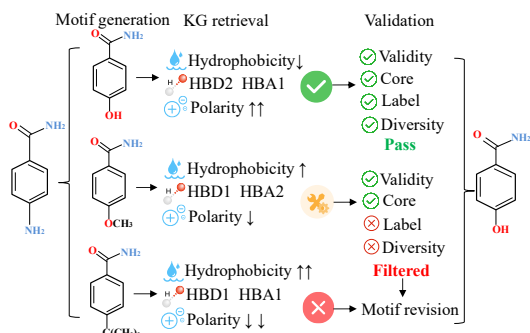


Figure 8: Visualization of CMAS augmentation process.

obtains lower FCD values, demonstrating that our augmented molecules better conform to chemical rules, align with real-world distributions, and provide more stable controllability than baselines.

Data Faithfulness. We combine a pre-trained GNN with human inspection to evaluate the label faithfulness of augmented samples: an augmented molecule is considered faithful if its assigned label is consistent with the model prediction and is judged reasonable by humans. Accuracy (the proportion of faithful samples among all augmented data) is adopted as the evaluation metric, which provides a direct quantitative measure of label preservation. As shown in Figure 6, CMAS consistently outperforms all baselines, achieving higher faithfulness across the board and clearly demonstrating its ability to generate more reliably labeled molecules.

4.7 Visualization.

Visualization of molecular embeddings. We apply t-SNE (Van der Maaten and Hinton, 2008) to map the learned molecular embeddings to a 2D space, with different colors indicating distinct property labels. Figure 7 shows the results on the BACE dataset. Our method achieves the clearest separation between classes, whereas baseline methods still have substantial feature overlap. This indicates that our multi-agent framework helps GNNs learn more discriminative molecular representations.

Visualization of CMAS Pipeline. We present the overall pipeline of our molecular augmentation framework in Figure 8. Starting from one molecular SMILES, we generate various candidate molecules through systematic substituent modification. The knowledge-guided module further analyzes key physicochemical properties to guide the generation process, while the validation module filters molecules based on validity, structural integrity, label consistency, and diversity, thereby retaining high-quality samples for downstream tasks.

5 Conclusion

In this study, we propose CMAS, a multi-agent data augmentation framework for molecular graph classification on imbalanced datasets. The core functions include generating structurally diverse molecular candidates while ensuring chemical validity and physicochemical consistency via knowledge retrieval and rule-based validation. By coordinating three specialized agents, it addresses key challenges in existing augmentation methods, where uncontrolled perturbations may violate chemical validity, damage key functional groups, or alter intrinsic properties, introducing noisy samples into training. Experiments demonstrate that CMAS consistently outperforms state-of-the-art baselines on MoleculeNet, advancing the reliability of molecular graph augmentation and model capability.

Limitations

While CMAS has achieved performance improvements on molecular property prediction tasks, it still presents notable limitations. The model’s rigorous generation-and-verification mechanism can lead to increased token usage and computational overhead when the Validation Agent repeatedly rejects outputs from the Motif-Level Agent, which affects overall efficiency. Moreover, the ElementKG knowledge base currently contains only attributes and properties of individual elements, lacking broader information about functional groups and bond types. This incomplete knowledge may prevent the LLM from making fully informed decisions, thereby constraining overall performance. In future work, we plan to introduce ElementKG 2.0, an enhanced knowledge graph that incorporates functional groups, bond types, and other higher-level chemical properties, enabling the Knowledge-Level Agent to perform more comprehensive and accurate consistency checks.

Acknowledgments

This paper is partially supported by the Fundamental Research Funds for the Central Universities in UIBE (Grant No. 23QN02), the Humanities and Social Sciences Research Fund of the Ministry of Education of China under Grant 25YJCZH275 and Open Research Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems.

References

- Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20.
- Chunchun Chen, Xing Wei, Jiayi Yang, Chenrun Wang, Yiwei Fu, Yuxing Zhang, Xin Sun, Rui Fan, and Wei Ye. 2026. Edge self-adversarial augmentation enhances graph contrastive learning against neighborhood inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 40, pages 20005–20013.
- Jorg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. 2008. On the art of compiling and using ‘drug-like’ chemical fragment spaces. *ChemMedChem*, 3(10):1503.
- David K Duvenaud, Dougal Maclaurin, Jorge Iparaguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28.
- Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *International Conference on Learning Representations*, volume 2024, pages 48221–48251.
- Yin Fang, Qiang Zhang, Haihong Yang, Xiang Zhuang, Shumin Deng, Wen Zhang, Ming Qin, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2022. Molecular contrastive learning with chemical element knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3968–3976.
- Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553.
- Yushi Feng, Tsai Hor Chan, Guosheng Yin, and Lequan Yu. 2025. Democratizing large language model-based graph data augmentation via latent knowledge graphs. *Neural Networks*, page 107777.
- Yiwei Fu, Chunchun Chen, Yuxing Zhang, Weizhi Zhang, Guanjie Zheng, Wei Ye, Xiao Luo, Philip S Yu, and Minghua Deng. 2026. Mask: Minority-aware semantic alignment via knowledge enhancement for long-tailed graph clustering. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management*.
- Yiwei Fu, Yuxing Zhang, Chunchun Chen, Jianwen Ma, Quan Yuan, Rong-Cheng Tu, Xinli Huang, Wei Ye, Xiao Luo, and Minghua Deng. 2025. Mark: Multi-agent collaboration with ranking guidance for text-attributed graph clustering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6057–6072.
- Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. 2016. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301.
- Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. Pmlr.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.
- Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. 2022. G-mixup: Graph data augmentation for graph classification. In *International conference on machine learning*, pages 8230–8248. PMLR.
- Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simonov. 2016. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3:85.
- Jian Jiang, Chunhuan Zhang, Lu Ke, Nicole Hayes, Yueying Zhu, Huahai Qiu, Bengong Zhang, Tianshou Zhou, and Guo-Wei Wei. 2025. A review of machine learning methods for imbalanced data challenges in chemistry. *Chemical Science*, 16(18):7637–7658.
- Tianyi Jiang, Zeyu Wang, Wenchao Yu, Jinhuan Wang, Shanqing Yu, Xiaoze Bao, Bin Wei, and Qi Xuan. 2024. Mix-key: graph mixup with key structures for molecular property prediction. *Briefings in bioinformatics*, 25(3):bbae165.
- Talia B Kimber, Maxime Gagnebin, and Andrea Volkamer. 2021. Maxsmi: maximizing molecular property prediction performance with confidence estimation using smiles augmentation and deep learning. *Artificial Intelligence in the Life Sciences*, 1:100014.
- Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079.
- O-Joun Lee and 1 others. 2025a. Pre-training graph neural networks on molecules by using subgraph-conditioned graph information bottleneck. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17204–17213.
- Seul Lee, Karsten Kreis, Srividya Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saeed Paliwal, Weili Nie, and Arash Vahdat. 2025b. Genmol: A drug discovery generalist with discrete diffusion. *arXiv preprint arXiv:2501.06158*.

- Ruochang Li, Xiao Luo, Zhiping Xiao, Wei Ju, and Ming Zhang. 2025a. HEAL: Hybrid enhancement with LLM-based agents for text-attributed hypergraph self-supervised representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6815–6829, Suzhou, China. Association for Computational Linguistics.
- Yibo Li, Yuan Fang, Mengmei Zhang, and Chuan Shi. 2025b. Advancing molecular graph-text pre-training via fine-grained alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 1589–1599.
- Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey, 2024. URL <https://arxiv.org/abs/2406.15126>, 2406.
- Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. 2012. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697.
- Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. 2016. Deeptox: toxicity prediction using deep learning. *Frontiers in environmental science*, 3:167215.
- Joonhyung Park, Hajin Shim, and Eunho Yang. 2022. Graph transplant: Node saliency-guided graph mixup with local structure preservation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7966–7974.
- Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. 2018. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741.
- Anjie Qiao, Zhen Wang, Chuan Chen, DeFu Lian, and Enhong Chen. 2025. Composable score-based graph diffusion model for multi-conditional molecular generation. *arXiv preprint arXiv:2509.09451*.
- Ann M Richard, Richard S Judson, Keith A Houck, Christopher M Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T Martin, John F Wambaugh, and 1 others. 2016. Toxcast chemical landscape: paving the road to 21st century toxicology. *Chemical research in toxicology*, 29(8):1225–1251.
- Sebastian G Rohrer and Knut Baumann. 2009. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *Journal of chemical information and modeling*, 49(2):169–184.
- Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571.
- Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. 2019. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*.
- Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational modeling of β -secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.
- Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2018. Deep graph infomax. *arXiv preprint arXiv:1809.10341*.
- Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.
- Zeyu Wang, Tianyi Jiang, Jinhuan Wang, Jiafei Shao, Bin Wei, Qi Xuan, and Hong Wang. 2025. Molecular connectivity index-based data augmentation for molecular property prediction. *IEEE Transactions on Computational Biology and Bioinformatics*.
- Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.
- Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z Li. 2022. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM web conference 2022*, pages 1070–1079.
- Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. 2023. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

- Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, and 1 others. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388.
- Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. 2021. Graph contrastive learning automated. In *International conference on machine learning*, pages 12121–12132. PMLR.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823.
- Xinyi Zhang, Yanni Xu, Changzhi Jiang, Lian Shen, and Xiangrong Liu. 2024. Molemcl: a multi-level contrastive learning framework for molecular pre-training. *Bioinformatics*, 40(4):btac164.
- Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. 2021. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems*, 34:15870–15882.
- Jiajun Zhou, Jie Shen, Shanqing Yu, Guanrong Chen, and Qi Xuan. 2020. M-evolve: structural-mapping-based data augmentation for graph classification. *IEEE Transactions on Network Science and Engineering*, 8(1):190–200.

A Description of Datasets

We evaluate on eight classification benchmarks from MoleculeNet (Wu et al., 2018). Each molecule is converted from SMILES to an undirected graph: nodes are heavy atoms and edges are chemical bonds. Following common practice (Wang et al., 2025), we adopt **random scaffold split** with ratio 8:1:1 for train/validation/test to reduce scaffold leakage. Table 4 summarizes dataset statistics (graph counts and average sizes are computed after RDKit parsing). For class-imbalanced sets (e.g., BACE, BBBP), we optionally upsample minority-class training molecules via our augmentation pipeline before GNN training.

Benchmark descriptions. The eight sets are briefly introduced below; all task definitions and labels follow the MoleculeNet benchmark protocol (Wu et al., 2018). (i) **SIDER** (Side Effect Resource) links marketed drugs to MedDRA system-organ-class side effects and is widely used for multi-label adverse-event prediction (Kuhn et al., 2016); we follow the standard binary subset packaged in MoleculeNet. (ii) **BACE** collects β -secretase 1 (BACE-1) inhibitors with qualitative binding labels and is a canonical target-specific benchmark from medicinal chemistry screening (Subramanian et al., 2016; Wu et al., 2018). (iii) **BBBP** (blood-brain barrier penetration) distinguishes compounds that permeate the CNS from those that do not, originating from in silico BBB models and experimental permeability annotations (Martins et al., 2012). (iv) **ClinTox** contrasts FDA-approved drugs against drugs removed from the market due to toxicity, encoding clinical failure risk at the molecular level as in MoleculeNet (Wu et al., 2018; Gayvert et al., 2016). (v) **ToxCast** aggregates thousands of high-throughput screening assays from the U.S. EPA ToxCast program; MoleculeNet provides a large collection of assay-wise binary endpoints (Richard et al., 2016; Mayr et al., 2016; Wu et al., 2018). (vi) **HIV** derives from the NCI anti-HIV screening set and labels whether a compound inhibits HIV replication, a standard virtual-screening task in chemoinformatics (Wu et al., 2018). (vii) **MUV** (maximum unbiased validation) was designed with near-zero structural redundancy between training and test splits to stress-test model generalization (Rohrer and Baumann, 2009). (viii) **Tox21** comes from the NIH Tox21 collaborative initiative and includes nuclear-receptor and stress-response pathway assays; we

use the 12-assay binary suite distributed in MoleculeNet (Huang et al., 2016; Mayr et al., 2016; Wu et al., 2018). Together, these benchmarks span dissimilar label semantics (single-target, clinical, and high-throughput toxicity), sizes, and average molecular graphs, as reflected in Table 4.

B Performance Comparison

Table 5 reports the complete comparison on all eight datasets. We pair each augmentation method with four graph encoders (GCN, GIN, GAT, and GraphSAGE) and report **Accuracy** and **AUC** (% , mean $\pm$ standard deviation over multiple runs). Baselines include structure-level augmentations (DropEdge, Graph Transplant, G-Mixup), evolution- and mixup-based methods (M-Evolve, Mix-Key), and recent graph-molecular augmentations (Genmol, DemoGraph). The best and second-best entries in each column are highlighted in **bold** and underlined, respectively. Our method achieves the best or tied-best results on most dataset-encoder pairs, with especially clear gains on SIDER, BACE, BBBP, ClinTox, and ToxCast where baselines are more spread out.

Comparison with contrastive learning-based methods As contrastive learning has become prevalent in data augmentation, we further evaluate the competitiveness of our augmentation strategy by comparing it with 8 state-of-the-art methods, using GraphSAGE as the graph encoder. The performance on 6 benchmark datasets is summarized in Table 6. Two key observations can be drawn:

- CMAS surpasses all existing contrastive learning-based augmentation methods across six benchmarks. With the GraphSAGE encoder, our method achieves the top ROC-AUC on every dataset. Our approach also excels on ToxCast, Sider, and HIV, confirming its robustness and broad applicability. Unlike prior methods that show large performance fluctuations across different datasets, ours delivers consistently strong results with low variance, underscoring its reliability and transferability.
- We find that conventional contrastive methods often fail to capture local molecular information. GraphCL and SimGRACE lack chemical guidance; GraphMVP and JOAO focus on global views and ignore local functional groups, leading to inferior performance. In contrast, our method uses two agents: one performs local motif perturbations for structural similarity, and the other selects samples with physicochemical similarity via chemical

Table 4: Dataset statistics used in our experiments.

Dataset	# Graphs	Avg. # Nodes	Avg. # Edges	Node Dim	# Tasks	IF	Split
SIDER	1427	~ 33.6	~ 70.7	9	27	~ 9	Scaffold (8:1:1)
BACE	1513	~ 34.1	~ 73.7	9	1	~ 1.1	Scaffold (8:1:1)
BBBP	2050	~ 23.9	~ 51.6	9	1	~ 2	Scaffold (8:1:1)
ClinTox	1484	~ 26.1	~ 55.5	9	2	~ 10	Scaffold (8:1:1)
HIV	41127	~ 25.5	~ 54.9	9	1	~ 30	Scaffold (8:1:1)
MUV	93087	~ 24.2	~ 52.6	9	17	~ 175	Scaffold (8:1:1)
Tox21	7831	~ 18.6	~ 38.6	9	12	~ 18	Scaffold (8:1:1)
ToxCast	8576	~ 18.8	~ 19.3	9	617	~ 167	Scaffold (8:1:1)

knowledge. This yields better augmentations and more accurate predictions. As shown, our method achieves the highest ROC-AUC with low variance across all datasets, demonstrating its effectiveness and stability.

C Visual Summary Across Encoders

Visualization of Prompts. We provide an example of the prompt in Figure 9. This example illustrates the template designed for the Motif Agent, including the motif decomposition results and the associated validation constraints. The prompt templates for the Knowledge Agent, which involve retrieving domain-specific chemical knowledge, are provided in the Appendix due to space limitations.

D Efficiency Analysis

As summarized in Table 7, the token consumption, running time, and molecular statistics of all evaluated datasets are presented. Overall, the computational cost of our framework scales consistently with dataset size. Smaller datasets, such as BBBP and BACE, typically require relatively low token usage and running time, while larger datasets, such as HIV and MUV, naturally incur higher costs due to repeated LLM-based generation and verification processes. Despite the increased overhead introduced by the multi-agent augmentation pipeline, the overall efficiency remains fully practical.

Time and Token Overhead Comparison To further dissect the computational efficiency at a granular level, we benchmark the per-molecule running time and token consumption of CMAS against existing heuristic-based and LLM-driven baseline methods on the BBBP dataset.

As presented in Table 8, non-LLM baselines that rely on random edge masking or subgraph permutations exhibit lower per-molecule latency, but they lack chemical semantic awareness. Compared to the state-of-the-art LLM-based method DemoGraph, CMAS achieves a **31.5% reduction**

in latency (174s vs. 254s per molecule) and a **28.6% decrease in token consumption** (~2,000 vs. ~2,800 tokens). This demonstrates that our knowledge-guided agent coordinates generation more concisely, avoiding redundant prompting iterations.

Scalability on Large-Scale Screening To evaluate whether CMAS can scale effectively to real-world high-throughput screening applications, we conduct experiments on the massive-scale ogbg-molpcba benchmark, which comprises 437,929 molecules evaluated across 128 assay tasks.

As reported in Table 9, although the sequential theoretical runtime across the entire molecular library is substantial, the embarrassingly parallel nature of our augmentation pipeline significantly compresses the empirical wall-clock time down to ~56.0 hours. Under this setting, CMAS outperforms DemoGraph by **+3.6 percentage points** while cutting the actual processing time nearly in half. This confirms that our offline data augmentation pipeline is practically viable for large-scale pharmaceutical screening.

Data Privacy and Local Model Deployment

While our primary experiments utilize API-based LLMs to ensure direct comparability with prior work (e.g., DemoGraph), enterprise and pharmaceutical drug discovery pipelines often involve proprietary molecular structures subject to strict data confidentiality constraints.

To assess offline deployment feasibility without external network transmission, we deploy an open-weight model (Qwen3: 8B) locally on private hardware. Table 10 compares the downstream accuracy of CMAS driven by API access versus the locally hosted model.

The results indicate that the locally deployed variant incurs a negligible performance difference (< 0.2% across all benchmarks) while guaranteeing complete data isolation and elim-

Table 5: Performance comparison of different methods on molecular property prediction. Results are reported as Accuracy/AUC (%), mean $\pm$ standard deviation). The best and second-best results in each column are highlighted in **bold** and underlined, respectively.

Encoder	Method	SIDER		BACE		BBBP		ClinTox		ToxCast		HIV		MUV		Tox21	
		acc	auc	acc	auc	acc	auc	acc	auc	acc	auc	acc	auc	acc	auc	acc	auc
GCN	DropEdge	72.44 $\pm$ 1.02	60.40 $\pm$ 0.20	69.60 $\pm$ 3.20	75.26 $\pm$ 4.19	80.60 $\pm$ 1.40	87.45 $\pm$ 3.92	91.07 $\pm$ 3.11	87.10 $\pm$ 5.97	83.06 $\pm$ 0.83	63.58 $\pm$ 5.72	92.32 $\pm$ 0.76	75.32 $\pm$ 3.65	97.36 $\pm$ 0.25	66.43 $\pm$ 2.37	85.17 $\pm$ 1.58	72.04 $\pm$ 0.36
	M-Evolve	75.24 $\pm$ 1.60	62.95 $\pm$ 2.67	75.79 $\pm$ 3.44	83.30 $\pm$ 4.07	86.04 $\pm$ 1.95	87.24 $\pm$ 4.60	92.21 $\pm$ 1.89	81.10 $\pm$ 4.83	76.21 $\pm$ 8.53	73.67 $\pm$ 10.21	96.45 $\pm$ 0.76	74.05 $\pm$ 4.53	98.71 $\pm$ 0.17	67.05 $\pm$ 2.21	89.72 $\pm$ 0.69	77.58 $\pm$ 1.34
	Graph Transplant	61.90 $\pm$ 2.70	57.20 $\pm$ 1.30	65.00 $\pm$ 6.65	70.38 $\pm$ 5.83	82.34 $\pm$ 5.11	84.63 $\pm$ 4.93	90.81 $\pm$ 3.34	80.74 $\pm$ 7.96	83.01 $\pm$ 1.03	61.83 $\pm$ 6.06	96.48 $\pm$ 0.25	73.63 $\pm$ 4.79	98.99 $\pm$ 0.08	67.21 $\pm$ 1.66	90.82 $\pm$ 3.61	63.53 $\pm$ 1.18
	G-Mixup	76.13 $\pm$ 1.16	62.40 $\pm$ 2.30	74.74 $\pm$ 4.72	80.72 $\pm$ 5.02	85.56 $\pm$ 5.00	83.32 $\pm$ 4.93	91.34 $\pm$ 2.91	64.20 $\pm$ 2.27	82.65 $\pm$ 1.12	60.80 $\pm$ 2.47	96.27 $\pm$ 0.38	70.94 $\pm$ 3.21	99.79 $\pm$ 0.03	62.38 $\pm$ 5.11	92.32 $\pm$ 3.70	60.24 $\pm$ 7.31
	Mix-Key	61.70 $\pm$ 4.10	58.57 $\pm$ 1.60	73.70 $\pm$ 3.10	78.30 $\pm$ 2.10	83.20 $\pm$ 1.90	84.42 $\pm$ 4.00	90.74 $\pm$ 2.95	86.71 $\pm$ 5.89	85.91 $\pm$ 5.70	63.45 $\pm$ 1.80	89.48 $\pm$ 0.66	77.39 $\pm$ 4.00	97.15 $\pm$ 0.30	66.47 $\pm$ 7.90	86.68 $\pm$ 1.70	78.35 $\pm$ 1.90
	Gemmol	73.50 $\pm$ 3.90	63.47 $\pm$ 2.13	75.20 $\pm$ 2.90	81.23 $\pm$ 3.45	85.40 $\pm$ 1.80	86.57 $\pm$ 1.86	91.38 $\pm$ 4.20	87.42 $\pm$ 3.97	88.32 $\pm$ 5.50	66.83 $\pm$ 5.12	94.45 $\pm$ 0.50	76.94 $\pm$ 2.31	98.57 $\pm$ 0.16	64.78 $\pm$ 4.52	90.54 $\pm$ 1.60	77.87 $\pm$ 1.63
	DemoGraph	75.09 $\pm$ 1.87	64.36 $\pm$ 3.52	75.35 $\pm$ 4.29	85.15 $\pm$ 1.36	84.96 $\pm$ 2.15	89.32 $\pm$ 1.08	91.38 $\pm$ 2.57	89.23 $\pm$ 4.41	87.58 $\pm$ 3.02	65.37 $\pm$ 1.69	94.72 $\pm$ 0.42	78.15 $\pm$ 1.24	97.84 $\pm$ 0.28	69.30 $\pm$ 3.17	90.01 $\pm$ 4.03	79.32 $\pm$ 0.91
GIN	Ours	77.24 $\pm$ 1.40	66.83 $\pm$ 1.03	78.16 $\pm$ 4.02	87.38 $\pm$ 1.14	87.64 $\pm$ 2.42	91.41 $\pm$ 1.12	94.30 $\pm$ 2.24	91.58 $\pm$ 4.70	90.12 $\pm$ 2.85	75.14 $\pm$ 1.82	96.83 $\pm$ 0.30	81.01 $\pm$ 1.18	99.89 $\pm$ 0.03	71.92 $\pm$ 2.94	92.58 $\pm$ 0.45	81.75 $\pm$ 0.86
	DropEdge	59.30 $\pm$ 1.70	58.86 $\pm$ 3.09	76.84 $\pm$ 6.01	84.12 $\pm$ 5.39	85.56 $\pm$ 5.74	88.48 $\pm$ 5.40	92.15 $\pm$ 2.87	85.26 $\pm$ 7.86	83.33 $\pm$ 0.92	65.86 $\pm$ 4.85	96.64 $\pm$ 0.65	79.45 $\pm$ 4.28	97.15 $\pm$ 1.13	67.12 $\pm$ 2.19	86.03 $\pm$ 0.87	77.35 $\pm$ 0.34
	M-Evolve	73.95 $\pm$ 0.69	60.59 $\pm$ 2.83	75.39 $\pm$ 3.40	82.13 $\pm$ 3.14	86.07 $\pm$ 4.29	85.92 $\pm$ 5.71	93.55 $\pm$ 1.54	77.92 $\pm$ 14.71	74.18 $\pm$ 8.67	63.36 $\pm$ 11.18	95.81 $\pm$ 0.61	74.35 $\pm$ 2.63	98.37 $\pm$ 0.33	68.75 $\pm$ 3.91	92.54 $\pm$ 0.68	77.09 $\pm$ 1.38
	Graph Transplant	63.70 $\pm$ 2.80	55.83 $\pm$ 3.24	79.32 $\pm$ 1.20	83.42 $\pm$ 4.25	83.32 $\pm$ 6.99	85.52 $\pm$ 5.65	91.14 $\pm$ 2.74	84.64 $\pm$ 8.48	83.36 $\pm$ 1.02	63.49 $\pm$ 4.85	95.49 $\pm$ 1.53	72.69 $\pm$ 2.08	98.69 $\pm$ 0.16	69.30 $\pm$ 3.87	92.89 $\pm$ 3.39	68.85 $\pm$ 1.45
	G-Mixup	62.80 $\pm$ 0.80	57.33 $\pm$ 2.32	78.03 $\pm$ 2.93	83.96 $\pm$ 3.82	85.85 $\pm$ 6.35	85.23 $\pm$ 7.11	90.60 $\pm$ 3.26	65.26 $\pm$ 2.86	59.94 $\pm$ 3.46	60.94 $\pm$ 1.28	96.44 $\pm$ 1.02	74.03 $\pm$ 4.87	98.79 $\pm$ 0.76	69.51 $\pm$ 3.07	92.14 $\pm$ 4.12	60.51 $\pm$ 6.18
	Mix-Key	63.10 $\pm$ 2.40	68.93 $\pm$ 1.90	79.10 $\pm$ 1.50	81.69 $\pm$ 2.60	84.30 $\pm$ 2.00	85.41 $\pm$ 3.50	93.87 $\pm$ 5.10	87.90 $\pm$ 6.90	86.91 $\pm$ 4.90	64.31 $\pm$ 1.00	94.72 $\pm$ 0.90	78.33 $\pm$ 17.20	98.07 $\pm$ 0.11	67.61 $\pm$ 4.00	86.69 $\pm$ 1.70	78.15 $\pm$ 0.70
	Gemmol	73.50 $\pm$ 2.30	64.82 $\pm$ 2.45	81.30 $\pm$ 1.40	83.56 $\pm$ 2.78	86.00 $\pm$ 1.90	87.13 $\pm$ 1.92	92.77 $\pm$ 3.90	89.34 $\pm$ 4.12	88.91 $\pm$ 4.70	66.47 $\pm$ 4.78	96.32 $\pm$ 0.80	78.46 $\pm$ 2.67	98.37 $\pm$ 0.09	67.23 $\pm$ 3.89	91.79 $\pm$ 1.50	78.52 $\pm$ 1.74
GAT	DemoGraph	73.99 $\pm$ 1.22	74.34 $\pm$ 1.03	82.38 $\pm$ 3.05	84.14 $\pm$ 1.96	84.94 $\pm$ 0.91	87.96 $\pm$ 1.78	84.95 $\pm$ 3.42	91.43 $\pm$ 1.95	88.44 $\pm$ 2.69	65.36 $\pm$ 0.82	95.07 $\pm$ 2.17	84.05 $\pm$ 3.58	97.20 $\pm$ 0.24	68.52 $\pm$ 1.31	89.95 $\pm$ 0.41	79.01 $\pm$ 1.52
	Ours	76.14 $\pm$ 1.09	76.61 $\pm$ 0.95	84.79 $\pm$ 3.23	86.46 $\pm$ 1.84	87.02 $\pm$ 0.83	90.25 $\pm$ 1.87	93.48 $\pm$ 3.69	94.07 $\pm$ 1.82	90.61 $\pm$ 2.84	67.61 $\pm$ 0.71	97.45 $\pm$ 1.09	86.51 $\pm$ 3.70	99.23 $\pm$ 0.19	71.09 $\pm$ 1.19	93.17 $\pm$ 0.46	81.36 $\pm$ 1.43
	DropEdge	57.30 $\pm$ 2.60	56.91 $\pm$ 2.48	65.90 $\pm$ 2.90	80.07 $\pm$ 4.49	80.60 $\pm$ 4.00	88.83 $\pm$ 3.45	92.35 $\pm$ 2.73	87.71 $\pm$ 6.48	82.17 $\pm$ 0.63	60.82 $\pm$ 1.64	96.23 $\pm$ 0.37	71.14 $\pm$ 3.42	98.52 $\pm$ 0.08	64.17 $\pm$ 4.26	85.21 $\pm$ 0.73	71.25 $\pm$ 0.97
	M-Evolve	74.67 $\pm$ 1.20	61.47 $\pm$ 2.33	74.87 $\pm$ 3.59	81.66 $\pm$ 3.83	86.27 $\pm$ 2.50	82.10 $\pm$ 7.64	92.61 $\pm$ 1.79	78.91 $\pm$ 7.84	78.62 $\pm$ 11.71	64.20 $\pm$ 7.48	96.52 $\pm$ 0.83	74.56 $\pm$ 3.46	99.52 $\pm$ 0.02	69.54 $\pm$ 4.07	92.08 $\pm$ 1.02	73.20 $\pm$ 2.46
	Graph Transplant	61.00 $\pm$ 3.40	55.74 $\pm$ 1.49	71.50 $\pm$ 4.40	77.90 $\pm$ 3.80	83.61 $\pm$ 5.99	85.76 $\pm$ 4.34	90.47 $\pm$ 3.10	81.88 $\pm$ 5.93	82.91 $\pm$ 1.14	61.56 $\pm$ 5.50	96.47 $\pm$ 0.33	74.98 $\pm$ 3.47	99.65 $\pm$ 0.01	70.18 $\pm$ 4.05	92.44 $\pm$ 3.84	63.19 $\pm$ 12.62
	G-Mixup	61.50 $\pm$ 1.50	54.89 $\pm$ 2.10	71.97 $\pm$ 4.79	79.52 $\pm$ 4.48	84.88 $\pm$ 6.62	81.95 $\pm$ 10.17	91.28 $\pm$ 2.95	60.88 $\pm$ 7.13	82.83 $\pm$ 1.20	59.03 $\pm$ 3.36	96.51 $\pm$ 0.78	70.87 $\pm$ 6.22	99.79 $\pm$ 0.03	61.97 $\pm$ 5.73	91.76 $\pm$ 3.72	60.14 $\pm$ 6.53
	Mix-Key	61.30 $\pm$ 2.00	58.09 $\pm$ 1.40	70.60 $\pm$ 3.90	81.61 $\pm$ 2.70	82.40 $\pm$ 3.00	84.41 $\pm$ 3.80	88.37 $\pm$ 8.50	85.00 $\pm$ 1.20	85.32 $\pm$ 7.10	60.12 $\pm$ 1.72	92.05 $\pm$ 1.00	76.45 $\pm$ 4.00	97.62 $\pm$ 0.93	66.83 $\pm$ 7.20	85.18 $\pm$ 2.40	78.75 $\pm$ 0.70
GraphSAGE	Gemmol	73.00 $\pm$ 1.90	62.38 $\pm$ 2.67	74.50 $\pm$ 4.20	80.47 $\pm$ 3.82	85.90 $\pm$ 2.80	86.28 $\pm$ 2.14	91.77 $\pm$ 6.82	88.56 $\pm$ 4.56	87.12 $\pm$ 6.90	62.83 $\pm$ 2.51	95.65 $\pm$ 0.90	76.18 $\pm$ 2.89	98.92 $\pm$ 0.14	66.84 $\pm$ 4.17	90.38 $\pm$ 2.20	76.13 $\pm$ 1.92
	DemoGraph	75.52 $\pm$ 1.86	62.47 $\pm$ 1.28	73.53 $\pm$ 0.98	82.82 $\pm$ 1.19	84.39 $\pm$ 2.31	87.43 $\pm$ 1.72	92.00 $\pm$ 2.30	88.56 $\pm$ 4.25	88.00 $\pm$ 2.94	63.76 $\pm$ 1.51	94.57 $\pm$ 0.22	78.43 $\pm$ 7.70	98.66 $\pm$ 0.03	69.21 $\pm$ 0.79	90.11 $\pm$ 0.48	79.88 $\pm$ 0.82
	Ours	77.34 $\pm$ 1.74	64.04 $\pm$ 1.20	75.66 $\pm$ 0.93	84.60 $\pm$ 1.12	86.34 $\pm$ 2.22	89.67 $\pm$ 1.66	94.41 $\pm$ 2.23	90.24 $\pm$ 4.18	90.07 $\pm$ 2.88	65.19 $\pm$ 1.45	96.89 $\pm$ 0.19	80.32 $\pm$ 7.90	99.91 $\pm$ 0.02	70.97 $\pm$ 0.73	92.73 $\pm$ 0.45	81.42 $\pm$ 0.78
	DropEdge	57.60 $\pm$ 3.70	59.65 $\pm$ 2.70	62.80 $\pm$ 2.60	80.52 $\pm$ 6.69	78.80 $\pm$ 2.10	88.76 $\pm$ 3.83	79.70 $\pm$ 4.90	70.22 $\pm$ 6.54	78.99 $\pm$ 0.93	58.65 $\pm$ 6.21	79.88 $\pm$ 1.47	77.63 $\pm$ 1.96	97.37 $\pm$ 0.15	65.96 $\pm$ 3.66	88.52 $\pm$ 3.27	74.74 $\pm$ 11.11
	M-Evolve	74.34 $\pm$ 1.24	61.28 $\pm$ 2.97	63.48 $\pm$ 1.35	83.70 $\pm$ 2.90	80.46 $\pm$ 3.26	86.20 $\pm$ 5.83	86.41 $\pm$ 1.87	87.89 $\pm$ 5.25	76.09 $\pm$ 12.03	65.92 $\pm$ 8.97	88.31 $\pm$ 2.31	75.54 $\pm$ 2.02	98.01 $\pm$ 0.18	68.31 $\pm$ 2.81	91.68 $\pm$ 0.79	78.85 $\pm$ 1.17
	Graph Transplant	60.20 $\pm$ 3.90	58.10 $\pm$ 0.70	66.20 $\pm$ 4.40	78.40 $\pm$ 3.10	81.60 $\pm$ 1.20	81.80 $\pm$ 3.70	90.81 $\pm$ 3.34	80.06 $\pm$ 6.90	83.41 $\pm$ 0.98	61.67 $\pm$ 6.54	95.01 $\pm$ 1.54	73.06 $\pm$ 2.07	98.39 $\pm$ 0.16	69.58 $\pm$ 3.85	92.79 $\pm$ 3.46	66.04 $\pm$ 1.71
	G-Mixup	60.00 $\pm$ 0.80	55.91 $\pm$ 0.07	75.13 $\pm$ 4.27	81.94 $\pm$ 4.81	84.39 $\pm$ 6.82	83.11 $\pm$ 8.51	90.94 $\pm$ 3.18	65.88 $\pm$ 5.62	82.70 $\pm$ 1.06	60.33 $\pm$ 3.02	87.69 $\pm$ 2.35	72.70 $\pm$ 5.40	99.79 $\pm$ 0.03	65.20 $\pm$ 5.74	91.80 $\pm$ 3.89	60.23 $\pm$ 4.42
GraphSAGE	Mix-Key	59.29 $\pm$ 2.70	56.46 $\pm$ 1.20	74.57 $\pm$ 5.40	79.81 $\pm$ 2.90	80.12 $\pm$ 6.30	85.54 $\pm$ 2.60	90.23 $\pm$ 3.10	80.81 $\pm$ 5.00	84.69 $\pm$ 0.80	61.42 $\pm$ 15.90	80.09 $\pm$ 2.90	78.24 $\pm$ 1.40	95.61 $\pm$ 0.90	68.91 $\pm$ 2.90	75.09 $\pm$ 3.10	81.65 $\pm$ 0.90
	Gemmol	71.33 $\pm$ 2.80	61.47 $\pm$ 1.20	74.66 $\pm$ 2.30	81.85 $\pm$ 3.21	84.15 $\pm$ 2.70	85.94 $\pm$ 2.66	90.36 $\pm$ 2.50	84.73 $\pm$ 4.78	84.52 $\pm$ 5.67	64.52 $\pm$ 5.67	92.77 $\pm$ 2.80	75.62 $\pm$ 4.45	99.32 $\pm$ 0.10	65.37 $\pm$ 4.63	90.64 $\pm$ 3.20	78.46 $\pm$ 1.62
	DemoGraph	74.76 $\pm$ 1.69	61.91 $\pm$ 1.14	72.92 $\pm$ 0.97	82.11 $\pm$ 1.08	83.72 $\pm$ 2.15	87.60 $\pm$ 1.58	92.17 $\pm$ 2.09	84.39 $\pm$ 3.96	87.89 $\pm$ 2.76	62.48 $\pm$ 1.38	94.03 $\pm$ 0.21	78.07 $\pm$ 8.12	96.97 $\pm$ 0.15	68.41 $\pm$ 0.69	80.09 $\pm$ 0.41	78.94 $\pm$ 0.72
	Ours	76.82 $\pm$ 1.87	62.59 $\pm$ 0.91	77.41 $\pm$ 2.24	87.97 $\pm$ 1.45	85.37 $\pm$ 2.11	89.11 $\pm$ 1.94	93.29 $\pm$ 2.90	88.99 $\pm$ 2.25	90.11 $\pm$ 2.85	67.22 $\pm$ 1.81	96.60 $\pm$ 0.19	81.87 $\pm$ 1.79	99.89 $\pm$ 0.10	72.04 $\pm$ 1.24	92.91 $\pm$ 0.55	82.31 $\pm$ 0.78

Table 6: Performance comparison of contrastive learning methods on molecular property prediction. Results are reported as AUC (% , mean $\pm$ standard deviation). The best and second-best results in each column are highlighted with light blue and brown-yellow backgrounds, respectively.

Method	Tox21	ToxCast	Sider	clintox	HIV	Bace
InfoMax (Veličković et al., 2018)	75.10 $\pm$ 0.50	62.70 $\pm$ 0.40	58.00 $\pm$ 0.80	69.30 $\pm$ 2.30	76.10 $\pm$ 0.70	75.50 $\pm$ 1.70
GraphCL (You et al., 2020)	74.20 $\pm$ 0.50	62.80 $\pm$ 0.40	59.50 $\pm$ 0.60	70.10 $\pm$ 7.80	77.30 $\pm$ 1.10	74.40 $\pm$ 2.50
GraphMVP (Liu et al., 2021)	75.40 $\pm$ 0.50	63.50 $\pm$ 0.50	59.30 $\pm$ 1.10	76.10 $\pm$ 4.80	76.10 $\pm$ 1.70	78.60 $\pm$ 3.50
JOAO (You et al., 2020)	74.70 $\pm$ 0.20	63.60 $\pm$ 0.30	60.60 $\pm$ 1.10	78.10 $\pm$ 3.90	76.60 $\pm$ 1.00	77.30 $\pm$ 1.30
JOAOv2 (You et al., 2021)	74.90 $\pm$ 0.70	63.50 $\pm$ 0.30	61.10 $\pm$ 0.70	78.60 $\pm$ 3.30	76.90 $\pm$ 0.80	75.30 $\pm$ 3.70
SimGRACE(Xia et al., 2022)	73.80 $\pm$ 0.40	61.70 $\pm$ 0.30	59.20 $\pm$ 0.60	63.40 $\pm$ 2.30	74.30 $\pm$ 0.90	75.20 $\pm$ 0.60
Mole-Bert(Xia et al., 2023)	73.60 $\pm$ 0.50	63.50 $\pm$ 0.40	59.10 $\pm$ 1.00	72.60 $\pm$ 2.70	76.50 $\pm$ 1.40	77.10 $\pm$ 2.40
MoleMCL(Zhang et al., 2024)	76.20 $\pm$ 0.60	63.70 $\pm$ 0.20	59.00 $\pm$ 0.80	80.20 $\pm$ 1.60	79.60 $\pm$ 0.50	80.30 $\pm$ 0.80
Ours	82.31 $\pm$ 1.95	67.22 $\pm$ 1.81	62.59 $\pm$ 0.91	88.99 $\pm$ 2.25	81.87 $\pm$ 1.79	87.97 $\pm$ 1.45

Table 7: Statistics of the evaluated datasets. Total tokens and total time are reported in millions (M) and million seconds (M s), respectively.

Dataset	Tokens (M)	Time (M s)	Mean #Atoms	Mean #Motifs
SIDER	6.39	0.30	33.64	16.49
BACE	6.64	0.31	34.09	15.39
BBBP	4.85	0.36	24.06	8.98
ClinTox	5.98	0.28	26.16	11.37
ToxCast	31.35	1.44	18.78	7.51
HIV	157.79	7.28	25.51	9.35
MUV	355.51	16.34	24.23	9.38
Tox21	28.30	1.30	18.57	7.44
ALL	596.81	27.61	24.18	9.31

Table 8: Augmentation latency and token consumption per molecule on BBBP.

Method	Time (s/mol)	Tokens / mol
DropEdge	0.01	—
G-Mixup / Mix-Key	0.04	—
Graph Transplant	0.08	—
M-Evolve	0.17	—
GenMol	1.70	—
DemoGraph	254	$\sim$ 2,800
CMAS (Ours)	174	$\sim$2,000

inating API call overhead. This confirms that CMAS can be readily adapted into fully offline, privacy-preserving industrial workflows.

E BRICS Predefined Fragmentation Rules

The Motif Agent decomposes each training molecule into editable fragments using **BRICS** (Breaking of Retrosynthetically Interesting Chemical Substructures) (Degen et al., 2008). BRICS does not learn cut rules from data; it uses a fixed set of **16 bond-disconnection types** that approximate retrosynthetically plausible breaks. We implement fragmentation with RD-Kit’s Chem.BRICS.FindBRICSBonds and a post-processing routine that merges atom cliques and

Table 9: Scalability and efficiency evaluation on the large-scale ogbg-molpcba benchmark.

Method	Per-Mol Time (s)	Theoretical Time (h)	Actual Time (h)	Test AP (%)
DemoGraph	257	$\sim$ 31,256	$\sim$ 102.8	28.5
CMAS (Ours)	177	$\sim$ 21,535	$\sim$ 56.0	32.1

Table 10: Downstream performance comparison between API-based and local deployment.

Deployment Mode	SIDER	BACE	BBBP	ClinTox
CMAS (API-based)	77.24	78.16	87.64	94.30
CMAS (Local Qwen3:8B)	77.05	78.02	87.51	94.22

extracts per-motif SMILES for prompting.

Role in our method. BRICS defines where local edits may occur, not how replacements are chosen. Each fragment is listed in the Motif Agent prompt; the LLM may delete or replace motifs with probability τ while preserving the core scaffold and task label when possible. Candidates that fail RDKit sanitization or physicochemical filters are discarded by the Validation Agent before augmented SMILES are converted to graphs for GNN training.

Fragmentation procedure. Given an input SMILES, we (i) parse and sanitize the molecule, (ii) mark bonds from FindBRICSBonds, (iii) split bond endpoints into separate cliques, (iv) retain acyclic side chains when a clique crosses a ring/non-ring boundary, (v) separate acyclic branch atoms (degree > 2), and (vi) merge overlapping cliques. Each final clique is turned into a motif SMILES via MolFragmentToSmiles and passed to the Motif Agent together with the original SMILES (see the case study appendix). If no BRICS bond is detected, the entire molecule is treated as one motif.

Practical notes. Duplicate motif SMILES after canonicalization are allowed in prompts but may be

deduplicated when saving augmented data. BRICS fragmentation is independent of KG-RAG: knowledge retrieval and accept/reject judgment use full-molecule SMILES only *after* motif editing.

F Relation Priority Definition in ElementKG

To reduce noisy or weakly informative knowledge triples, we assign a predefined priority weight w_r to each relation type r in ElementKG according to its estimated contribution to physicochemical property preservation. Relations closely associated with molecular stability, bonding behavior, and electronic characteristics are assigned higher weights, while generic descriptive attributes receive lower weights.

Specifically, relations including `hasValence`, `hasElectronegativity`, `hasOxidationState`, `hasBondPreference`, `hasElectronAffinity`, and `participatesInAromaticity` are treated as high-priority relations with $w_r = 1.0$, since these attributes strongly influence molecular property consistency and functional group behavior. Relations including `hasAtomicRadius`, `hasIonizationEnergy`, `hasCovalentRadius`, `hasElectronConfiguration`, and `belongsToPeriod` are assigned medium priority with $w_r = 0.7$, as they provide supportive physicochemical context. Generic metadata relations such as `belongsToGroup`, `hasDiscoveryYear`, `hasAppearance`, and textual annotation relations are assigned lower priority with $w_r = 0.4$.

For each atom entity a , all associated triples are ranked according to their relation weights w_r , and the top- k highest-priority triples are selected:

$$\mathcal{R}_k(a) = \text{Top-}k_{w_r}(\mathcal{G}(a)). \quad (11)$$

In our implementation, we set $k = 5$ to balance knowledge coverage and prompt efficiency. This priority-based filtering mechanism reduces irrelevant knowledge injection and improves the stability of LLM-based semantic alignment assessment.

G ElementKG and Knowledge Retrieval in Our Pipeline

ElementKG is an element-centric chemical knowledge graph introduced for knowledge-enhanced molecular representation learning (Fang et al., 2023). It is distributed as an OWL ontology in which each chemical element is an entity, annotated with numeric attributes (e.g., electronegativ-

ity, density, ionization energy) and linked to other elements via typed relations (e.g., comparative boiling/melting points, hardness, or family/period associations). Functional groups are also represented and can be aligned with SMARTS patterns for substructure matching. We do *not* modify the ontology in experiments; we load it as a fixed knowledge source for retrieval-augmented augmentation.

How we query ElementKG. For each SMILES (original and candidate in the KG-RAG Judge), we (i) parse the structure and list element symbols plus functional groups matched by SMARTS patterns; (ii) query ElementKG for per-element properties and pairwise element relations involving atoms in the molecule; (iii) compute RD-Kit descriptors (molecular weight, LogP, H-bond donors/acceptors); (iv) encode the structure with a CMPN-based encoder (KANO(Fang et al., 2023)) to obtain atom- and functional-group embeddings; (v) retrieve nearest element and functional-group neighbors from precomputed embedding banks via FAISS L_2 search; and (vi) format the results into prompt text (elements, functional groups, descriptors, ElementKG properties/relations, and embedding neighbors). Steps (i)–(vi) are applied to both the source and candidate structures in dual-molecule KG-RAG judgment (see the case study appendix).

Role in our pipeline. ElementKG does not edit molecules; it supplies structured evidence for accept/reject decisions after motif editing (Appendix E). Invalid candidates are still filtered by RDKit validation in the main pipeline. Only symbols and relations appearing in the input SMILES are retrieved, so context size scales with molecular size rather than the full periodic table.

Resources bundled with our code. The implementation relies on four knowledge assets shipped with the project: (i) the ElementKG OWL file, (ii) a functional-group pattern list, (iii) pretrained element and functional-group embedding tables used to build FAISS indices, and (iv) CMPN weights for on-the-fly atom/FG embeddings. Only symbols and relations *present in the input SMILES* are queried, so context size scales with molecular size rather than the full periodic table.

Knowledge Granularity Analysis. We investigate how the granularity of retrieved chemical knowledge influences the reasoning capability of

Table 11: Ablation study on knowledge retrieval granularity and computational cost.

Knowledge Setting	Simple Edit (%)	Complex Motif (%)	BBBP ACC (%)	HIV ACC (%)	Time (h)	Tokens (M)
No Retrieval	84.9	76.8	85.96 $\pm$ 1.71	96.02 $\pm$ 0.52	0.00	0.00
Element + Functional Group	89.7	82.4	86.71 $\pm$ 1.55	96.38 $\pm$ 0.41	0.36	7.85
Element-only (Default)	91.3	89.2	87.64 $\pm$ 2.40	96.83 $\pm$ 0.30	0.18	3.92

Table 12: Downstream accuracy under varying retrieval parameter k .

Top- k	BBBP ACC (%)	HIV ACC (%)
1	85.81 $\pm$ 1.23	96.12 $\pm$ 0.45
3	86.92 $\pm$ 0.89	96.51 $\pm$ 0.67
5	87.64 $\pm$ 2.40	96.83 $\pm$ 0.30
7	87.23 $\pm$ 1.56	96.62 $\pm$ 0.92
10	86.53 $\pm$ 1.02	96.21 $\pm$ 0.58

the agent. Specifically, we compare three configurations: (1) no knowledge retrieval, (2) element-level combined with functional-group knowledge, and (3) element-only knowledge.

As summarized in Table 11, the **element-only** configuration yields the highest semantic editing consistency and downstream classification accuracy on both BBBP and HIV datasets. Incorporating explicit functional-group knowledge directly into the prompt context introduces redundant constraints that complicate LLM reasoning while nearly doubling token usage and runtime latency. Consequently, we maintain element-only retrieval as our default design choice.

Sensitivity of Top- k Knowledge Relations. We further assess the impact of the retrieval parameter $k \in \{1, 3, 5, 7, 10\}$, which controls the number of top-ranked relational triples retrieved per entity.

As shown in Table 12, model performance exhibits a unimodal response peaking at $k = 5$. When k is too small ($k < 5$), the prompt lacks sufficient relational context for the agent to establish reliable physical-chemical alignment; when k becomes overly large ($k > 5$), low-priority relations dilute the prompt with extraneous context and induce noise. We therefore adopt $k = 5$ as the optimal default across all experiments.

H Distribution Learning Metrics

Let $\mathcal{G}$ denote the set of generated molecules and $\mathcal{R}$ the reference set (training molecules on the target benchmark). Let $\mathcal{G}_v \subseteq \mathcal{G}$ be the valid subset after RDKit parsing and sanitization. All SMILES are canonicalized before evaluation; structural similarity uses Morgan fingerprints (radius = 2, 2048 bits) and the Tanimoto coefficient (Bajusz et al., 2015).

Validity. Validity is the percentage of generated molecules that are chemically well-defined: $\text{Validity} = |\mathcal{G}_v|/|\mathcal{G}| \times 100\%$. Higher values are preferred.

Diversity. Diversity is the mean pairwise structural dissimilarity within $\mathcal{G}_v$: $\text{Diversity} = \frac{2}{|\mathcal{G}_v|(|\mathcal{G}_v|-1)} \sum_{g_i < g_j} (1 - \text{Tanimoto}(g_i, g_j))$. Higher values indicate broader coverage of chemical space.

Similarity. Similarity is the mean, over $g \in \mathcal{G}_v$, of the maximum Tanimoto similarity between g and any molecule in $\mathcal{R}$: $\text{Similarity} = \frac{1}{|\mathcal{G}_v|} \sum_{g \in \mathcal{G}_v} \max_{r \in \mathcal{R}} \text{Tanimoto}(g, r)$. Higher values indicate closer alignment with reference structures.

FCD. Fréchet ChemNet Distance (FCD) (Preuer et al., 2018) compares the distributions of ChemNet penultimate-layer embeddings for $\mathcal{G}_v$ and $\mathcal{R}$ via the Fréchet distance between fitted Gaussians. Lower FCD indicates better global distributional match in chemically and biologically relevant feature space.

Accuracy. A GNN classifier is trained on the reference training split and applied to $\mathcal{G}_v$, where each molecule inherits the label $y(g)$ of its parent. Accuracy is the fraction of generated molecules with correct predictions: $\text{Accuracy} = \frac{1}{|\mathcal{G}_v|} \sum_{g \in \mathcal{G}_v} \mathbb{I}[\hat{y}(g) = y(g)]$. Higher values indicate better preservation of task-relevant labels.

Synth. Synth (Qiao et al., 2025) is the mean synthetic-accessibility score over $\mathcal{G}_v$ (e.g., SC-Score). Lower values indicate easier-to-synthesize structures.

I Case Study

Figure 9 illustrates one accepted augmentation for a minority-class BACE inhibitor (label = 1). The example follows the same few-shot instruction format used at inference: background (dataset and sample), editing rules, retrieved knowledge, and structured model output.

Motif editing. The Motif Agent applies BRICS decomposition and edits motifs with probability $\tau=0.2$. In this case, an aromatic ether (methoxy)

Case: BACE Inhibitor Augmentation with Motif Editing and Dual-Molecule KG-RAG Judgment

Few-shot Instruction

[Instruction] You are a medicinal-chemistry expert. For an imbalanced BACE training fold, a minority-class molecule with its label is selected from the training set. The Motif Agent first builds a candidate by editing BRICS motifs: each motif is independently deleted or replaced with probability tau. The KG-RAG Agent then retrieves structured knowledge for both the original SMILES and the candidate using Molecular Knowledge Extractor (ElementKG, KANO/CMPN embeddings, and FAISS with top-k neighbors). Decide whether the candidate should be accepted into the augmented training set with the same label as the original.

Accept only if: (1) the candidate passes RDKit validation (Smiles Validation Agent); (2) the edit is a local motif change, not a full scaffold change; and (3) knowledge for both molecules does not contradict keeping the original BACE inhibitor label.

[Dataset and sample] Benchmark: BACE (SMILES and Class columns). This example is a minority-class inhibitor (label = 1).

[Settings] Motif edit probability tau = 0.2; Knowledge retrieval top-k = 3; LLM temperature rho = 0.5.

[Original SMILES] S1(=O)(=O)N(c2cc(cc3c2n(cc3CC)CC1)C(=O)NC@HCc1cccc1)C

[Motif Agent] BRICS motifs come from brics_decomp and fragment SMILES. The agent is told: for each BRICS motif, delete or replace with probability tau; output one valid SMILES only. Example motifs (shortened): sulfonamide-linked fragment, benzene ring, amide linker, aromatic ether fragment, and others. At rho = 0.5, the aromatic ether motif is chosen for replace: the methoxy group (OCH3) on the terminal ring becomes hydroxyl (OH). Other motifs are left unchanged.

[Candidate SMILES] S1(=O)(=O)N(c2cc(cc3c2n(cc3CC)CC1)C(=O)NC@HCc1cc(O)cc1)C

[KG-RAG context] Knowledge is extracted separately for the original and the candidate.

Original: elements C, H, N, O, S; functional groups include amide, amine, ether, sulfonamide; MW 591.7, LogP 3.8, 4 H-bond donors, 8 acceptors; ElementKG element properties and relations; top-3 nearest functional groups include ether and amide.

Candidate: same elements C, H, N, O, S; functional groups include amide, amine, phenol, sulfonamide (ether removed, phenol added); MW 577.6, LogP 3.5, 5 H-bond donors, 8 acceptors; top-3 neighbors include phenol, then ether and amide.

These fields match Molecular Knowledge Extractor (elements, functional groups from SMARTS, MW/LogP/HBD/HBA, ElementKG text, embedding neighbors). Here they support accept/reject judgment, not generating multiple analogs in one shot.

[Judge task] Return JSON: decision = accept or reject, and a short reason.

Model Output

Step 1 (Motif): Only the terminal aromatic motif changes: methoxy to hydroxyl after a sampled replace at rho. The sulfonamide, amide, core heterocycle, and charged chain stay the same. This is a local motif edit, not a new scaffold.

Step 2 (KG-RAG): Both molecules share the same elements and main inhibitor-related groups. Ether becomes phenol; LogP drops slightly and H-bond donors increase by one, which is reasonable. Top-k retrieval for the candidate highlights phenol. Nothing in the retrieved knowledge argues against keeping the inhibitor label.

Step 3 (Validation): SMILES parses; sanitize passes; atom count and molecular weight pass the configured filters.

Step 4 (Label): The edit is a small peripheral change. Label 1 is kept for the augmented pair.

Decision: Accept. Add the candidate with label 1 to the augmented training set, rebalance classes, convert SMILES to graphs, train GraphSAGE with early stopping on the validation set, and predict on the test set.

Figure 9: Case study on BACE: motif editing (methoxy→hydroxyl), dual-molecule KG-RAG context, validation, and **accept** decision for a minority-class inhibitor. Settings: $\rho=0.5$, $\text{top-}k=5$, $\tau=0.2$.

on the terminal ring is replaced by a hydroxyl group while the sulfonamide, amide, and core heterocycle remain unchanged, yielding a local peripheral edit rather than a scaffold change.

Dual-molecule KG-RAG judgment. The KG-RAG Agent retrieves ElementKG-style context, functional groups, and physicochemical descriptors for *both* the original and candidate SMILES (top- $k=5$ neighbors). The original molecule shows amide, amine, ether, and sulfonamide groups with MW ≈ 591.7 and LogP ≈ 3.8 . The candidate replaces ether with phenol (MW ≈ 577.6 , LogP ≈ 3.5 , one additional H-bond donor). Retrieved neighbors (phenol-, ether-, and amide-related entries) remain consistent with β -secretase inhibitor chemistry, and no retrieved fact contradicts retaining label = 1.

Validation and decision. RDKit parsing and sanitization succeed; atom-count and molecular-weight filters are satisfied. The LLM judge returns accept with rationale that the edit is a small, label-preserving modification. The accepted pair is added to the balanced training set and later converted to a graph for GraphSAGE training, as described in the main text.