GLA: Grounding Large Language Models in Molecular Hierarchy for Chemical Understanding

Anonymous ACL submission

Abstract

Conventional Euclidean geometries lead to structural distortion and entangle core pharmacophoric identities with peripheral groups. Existing molecule-language models, relying on linear or uniform encodings, often obscure the hierarchical organization of chemical semantics. To address this, we propose **G**eometric-**L**anguage **A**lignment (GLA), a framework integrating intrinsic molecular topology into large language models. GLA employs a mixed-curvature encoder that maps scaffolds to hyperbolic space to capture hierarchy, while encoding side-chains in Euclidean space. These representations are aligned with text via a dual-view contrastive objective and injected into a frozen language model. Experiments on cross-modal retrieval, captioning, and property prediction benchmarks show GLA consistently improves performance over baselines, suggesting that modeling geometric heterogeneity enhances the grounding between molecular structure and chemical language.

1 Introduction

Large Language Models (LLMs) drive advancements in chemistry, ranging from property prediction to *de novo* molecule generation (Edwards et al., 2022; Wang et al., 2019; Fang et al., 2024b). Typically, these models use linearized notations like SMILES within sequence modeling frameworks (Weininger, 1988). Although innovations like MM-Deacon (Guo et al., 2022) broaden semantic scope using multilingual notations (e.g., SMILES and IUPAC), relying on 1D sequences obscures the inherent topological structure. This creates a semantic gap, preventing models from grounding chemical concepts in accurate structural representations.

To bridge this gap, recent work enhances LLM reasoning using external knowledge. For instance, MolRAG (Xian et al., 2025) improves property prediction by retrieving structurally analogous molecules, while MolTC (Fang et al., 2024a) integrates graph information for interaction modeling. Despite these strides, a challenge remains: flattening a molecule or treating it as a monolithic graph blurs the distinction between scaffolds and functional groups. As noted by the Molecular Structural Reasoning (MSR) framework (Jang et al., 2025), LLMs consequently struggle to reason about properties dependent on specific functional groups.

The root of this issue lies in the geometric encoding limitations of standard encoders. Molecules are structurally heterogeneous, combining hierarchical scaffolds with dense functional groups. Standard graph neural networks compress these diverse topologies into uniform Euclidean space, entangling global hierarchy with local detail. This "geometric homogenization" erodes semantic separability, confusing pharmacophoric cores with peripheral modifiers. Capturing these nuances is vital, as local substructures often dictate functionality (Wu et al., 2023). Existing frameworks compound this by prioritizing coarse-grained global alignment. Models such as MoMu (Su et al., 2022), MoleculeSTM (Liu et al., 2023a), and MolCA (Liu et al., 2023c) map whole-graph representations to complete texts. While effective for retrieval, this overlooks granular relationships between specific motifs and text phrases.

To address these limitations, we propose **G**eometric-**L**anguage **A**lignment (GLA). This framework aligns hierarchical molecular structures with structured semantic representations. Our contributions are summarized as follows:

- We introduce a **Mixed-Curvature Geometric Encoder** that disentangles molecular topology. By mapping hierarchical scaffolds to hyperbolic space and local side-chains to Euclidean space, we effectively preserve intrinsic hierarchical structures and minimize geometric distortion.

- We propose **Dual-View Contrastive Learning** for fine-grained semantic grounding. This aligns structural motifs with specific linguistic roles, resolving the ambiguity of coarse-grained global alignment through precise substructure-text synchronization.

- Extensive experiments on PCDes (Zeng et al., 2022), ChEBI-20 (Edwards et al., 2021), and MoleculeNet benchmark (Wu et al., 2018) demonstrate state-of-the-art performance. Our results validate that explicitly modeling geometric hierarchy is essential on diverse tasks.

2 Related Work

2.1 Geometric Representations for Molecular Encoding

The inductive bias of an encoder dictates its semantic capacity. Standard Euclidean Message Passing Neural Networks (MPNNs), such as GIN and GCN (Hu et al., 2020; Wang et al., 2022), capture local atomic interactions but struggle to represent the hierarchical latent space of molecular scaffolds due to polynomial volume growth. While Hyperbolic Graph Neural Networks (HGNNs) (Liu et al., 2019) naturally accommodate tree-like hierarchies, existing approaches typically impose a *monolithic* curvature. This is suboptimal for heterogeneous chemical structures, where rigid pharmacophoric trees coexist with locally dense functional groups. Forcing such diverse topologies into a single manifold introduces geometric distortion. Unlike general mixed-curvature graph approaches (Zhu et al., 2020), GLA explicitly synchronizes these geometric distinctions with linguistic granularity to resolve this bottleneck.

2.2 Molecule–Text Multimodal Alignment

Molecule–text alignment has evolved from linear sequence modeling (Liu et al., 2023b; Jiang et al., 2024) to graph-based contrastive learning (Su et al., 2022; Liu et al., 2023a; Luo et al., 2023). Recent works have further expanded this to reaction contexts (Liu et al., 2024b). However, these paradigms remain predominantly *coarse-grained* and *isotropic*, mapping entire molecular graphs to complete textual descriptions. This global pooling obscures the compositional nature of chemical language, where specific functional groups modulate a central scaffold. While recent efforts attempt local alignment via optimal transport or clustering (Min et al., 2024; Zhang et al., 2025), they lack explicit geometric grounding. GLA addresses this by employing dual-view contrastive learning to align specific structural regions with their corresponding semantic roles.

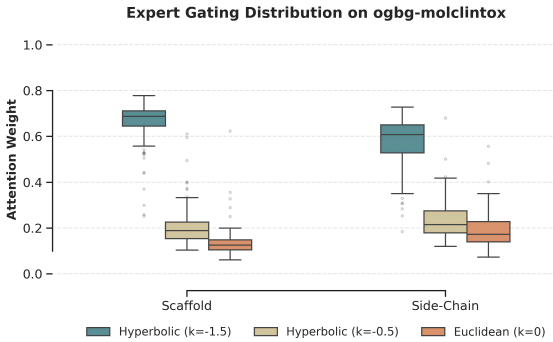


Figure 1: **Expert Gating Distribution on ogbg-molclintox.** The boxplot visualizes the attention weights across different curvature experts for scaffolds and side-chains. The model predominantly routes hierarchical scaffolds to the high-curvature hyperbolic expert ($k = -1.5$), while showing an increased preference for Euclidean geometry ($k = 0$) in local side-chain components.

2.3 Structural Injection into Scientific Language Models

Recent advancements utilize instruction tuning to adapt LLMs for scientific optimization (Dey et al., 2025; Pei et al., 2024). However, architectures like InstructMol (Cao et al., 2025) act as *passive feature bridges*, projecting monolithic embeddings into the LLM. This obscures structural hierarchies, preventing differential attention to scaffolds versus side-chains. In contrast, GLA injects *geometrically disentangled* representations. By **dynamically routing** features into hyperbolic or Euclidean spaces, our framework provides **latent structural priors** and distinct inductive biases, enabling fine-grained alignment between hierarchical motifs and chemical text.

3 Methodology

To bridge the semantic gap between linearized chemical language and hierarchical molecular topology, we propose **GLA**. Unlike prior approaches that compress heterogeneous structures into a monolithic embedding, GLA explicitly models the distinct geometric natures of pharmacophoric scaffolds and functional side-chains. As illustrated in Figure 2, our framework operates through a four-stage component: (1) decomposing the graph to isolate structural roles (Section 3.1);

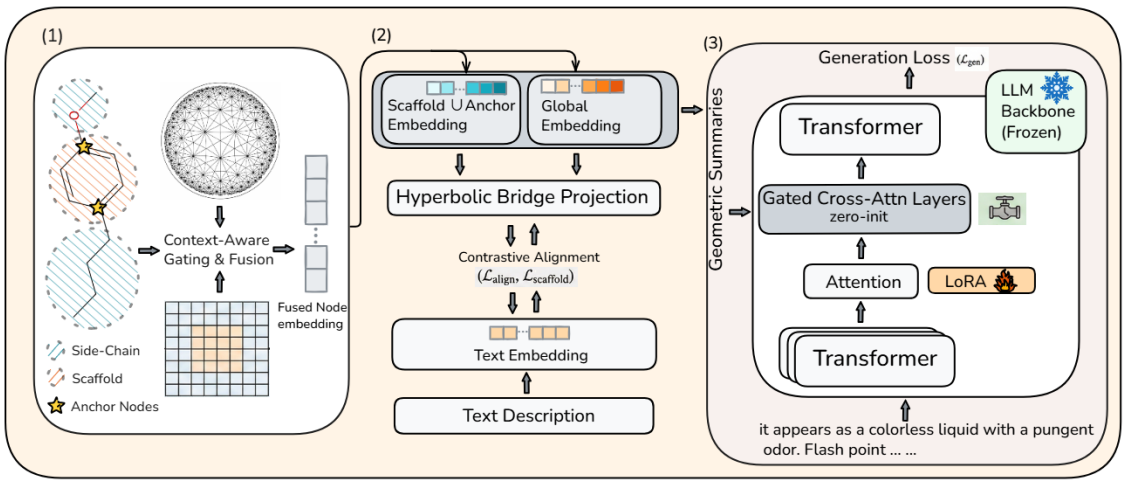


Figure 2: **The overall architecture of GLA.** The framework consists of three key phases: **(1) Refined Structural Disentanglement & Mixed-Curvature Encoding (Left):** Molecular graphs are decomposed into scaffolds and side-chains, encoded by hyperbolic and Euclidean experts respectively, and fused via a context-aware gating network. **(2) Dual-View Contrastive Alignment (Middle):** A hyperbolic bridge aligns geometric summaries (Global & Scaffold views) with textual semantics. **(3) Geometry-Conditioned Generative Injection (Right):** The aligned representations serve as soft prompts injected into the frozen LLM backbone (Qwen-2.5-3B) via zero-initialized Gated Cross-Attention layers, enabling structure-aware text generation.

(2) encoding these substructures in their native curvatures (hyperbolic vs. Euclidean) via a Mixed-Curvature MoE (Section 3.2); (3) synchronizing these geometric summaries with textual semantics via dual-view alignment (Section 3.3); and (4) injecting this grounded understanding into a frozen LLM for controllable generation (Section 3.4).

3.1 Refined Structural Disentanglement

Resolving the ambiguity between a molecule's core identity and its peripheral modifiers requires explicitly separating their structural representations. We address this by decomposing the molecular graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based on medicinal chemistry principles, rather than treating all atoms uniformly.

We partition the node set $\mathcal{V}$ into a rigid *scaffold set* $\mathcal{V}_S$ and a flexible *side-chain set* $\mathcal{V}_C$. To preserve the critical semantic connectivity between these regions (e.g., where a functional group modifies the core), we identify a set of **Anchor Nodes**:

$$\mathcal{V}_A = \{v \in \mathcal{V}_S \mid \exists u \in \mathcal{V}_C, (v, u) \in \mathcal{E}\} \quad (1)$$

Anchor nodes act as the structural interface, ensuring that the information flow between the Pharmacophoric Scaffold Graph ($\mathcal{G}_S$) and the Side-Chain Graph ($\mathcal{G}_C$) remains intact. This separation allows the subsequent encoder to apply distinct geometric inductive biases to the stable core and the flexible substituents.

3.2 Mixed-Curvature Hyperbolic Mixture-of-Experts

Mixture of Experts (MoE) (Jacobs et al., 1991; Shazeer et al., 2017) aims at training multiple experts with distinct skills, a paradigm widely employed to enhance model capacity. To accommodate the complex structural heterogeneity of molecules, we leverage this framework to design a **Mixed-Curvature Mixture-of-Experts** encoder. Instead of constraining the representation to a single geometry, we instantiate a diverse bank of experts $\{\mathcal{E}_k\}_{k=1}^{K}$ that span a spectrum of curvatures. This pool implicitly includes both Euclidean ($\kappa_k = 0$) and hyperbolic ($\kappa_k < 0$) components, thereby endowing the model with the sufficient geometric capacity to encapuslate disparate topological properties—ranging from locally dense substructures to hierarchically expansive scaffolds—within a unified latent space.

Manifold-Specific Encoding. Given node features $\mathbf{x}_i \in \mathbb{R}^d$, we map them to the corresponding manifold via the exponential map $\exp_{\mathbf{0}}^{\kappa_k}(\cdot)$. A **gating mechanism** then **dynamically routes** nodes to specific experts for curvature-aware message passing. To fuse these disparate geometries, hyperbolic embeddings are projected back to a shared tangent space at the origin using the logarithmic map:

$$\mathbf{z}_i^{(k)} = \log_{\mathbf{0}}^{\kappa_k}(\mathrm{GNN}_{\kappa_k}(\mathcal{G}, \mathbf{x}_i)) \in \mathbb{R}^d \quad (2)$$

where $\mathbf{z}_i^{(k)}$ denotes the tangent-space representation derived from the k-th expert.

Context-Aware Expert Fusion. To determine the intrinsic geometric preference of each atom, a gating network computes soft routing weights conditioned on the tangent-space features via a learnable context vector $\mathbf{w}_k$:

$$\alpha_i^{(k)} = \frac{\exp\left(\mathbf{w}_k^\top \mathbf{z}_i^{(k)}\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{w}_j^\top \mathbf{z}_i^{(j)}\right)} \tag{3}$$

$$\mathbf{h}_i = \sum_{k=1}^{K} \alpha_i^{(k)} \mathbf{z}_i^{(k)} \tag{4}$$

Crucially, this mechanism allows the model to *autonomously* learn geometric affinities. Rather than strictly regulating which geometry encodes which substructure, the gating network enables an adaptive alignment where nodes naturally gravitate towards the expert manifold (Euclidean or Hyperbolic) that best minimizes geometric distortion for their specific structural context.

3.3 Dual-View Contrastive Alignment

Merely obtaining disentangled embeddings is insufficient; they must be grounded in the linguistic compositionality of chemical descriptions. We propose a **Dual-View Contrastive Alignment** strategy to synchronize holistic molecular semantics with core structural semantics.

Dual-View Geometric Projections. We construct two complementary graph-level views. The *global embedding* $\mathbf{g}_{\text{global}}$ pools all nodes to capture the complete molecule, while the *scaffold embedding* $\mathbf{g}_{\text{scaffold}}$ pools only $\mathcal{V}_S \cup \mathcal{V}_A$ to isolate the pharmacophoric core. These are projected into the LLM's semantic space via an **adaptive multiscale Hyperbolic Bridge** $f_{\text{bridge}}(\cdot)$:

$$\mathbf{z}_{\text{geo}} = f_{\text{bridge}}(\mathbf{g}_{\text{global}}), \tag{5}$$

$$\mathbf{z}_{\text{scaffold}} = f_{\text{bridge}}(\mathbf{g}_{\text{scaffold}}) \tag{6}$$

In-Batch Contrastive Learning. To align these views with text, we extract textual representations $\mathbf{z}_{\text{text}}$ from the [EOS] token of the frozen LLM encoder. We optimize an InfoNCE-based objective that forces the geometric embeddings to be closer to their paired text description than to others in the batch B:

$$s_{ij} = \frac{\text{sim}\left(\mathbf{z}_{\text{geo}}^{(i)}, \mathbf{z}_{\text{text}}^{(j)}\right)}{\tau}. \tag{7}$$

$$\mathcal{L}_{\text{align}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})}. \tag{8}$$

An auxiliary loss $\mathcal{L}_{\text{scaffold}}$ is defined analogously for the scaffold view, ensuring the model explicitly learns to map core structures to their specific linguistic identifiers.

3.4 Geometry-Conditioned Generative Injection

Finally, to transfer this structural understanding to downstream tasks like captioning and property prediction, we must inject the aligned representations into the LLM without disrupting its pre-trained knowledge. Instead of using long, token-heavy atom sequences, we treat our compact global and scaffold embeddings as *Geometric Summaries* $\mathbf{Z} = [\mathbf{z}_{\text{geo}}, \mathbf{z}_{\text{scaffold}}]$

These summaries are integrated via **Gated Cross-Attention** layers inserted into the frozen LLM. For hidden states $\mathbf{H}_{\text{text}}^{(l)}$ at layer l, the injection is defined as:

$$\mathbf{A} = \text{Softmax}\left(\frac{(\mathbf{H}_{\text{text}}^{(l)}\mathbf{W}_Q)(\mathbf{Z}\mathbf{W}_K)^\top}{\sqrt{d}}\right) \tag{9}$$

$$\mathbf{O}^{(l)} = \mathbf{H}_{\text{text}}^{(l)} + \tanh(\gamma) \cdot \mathbf{A}(\mathbf{Z}\mathbf{W}_V) \tag{10}$$

where γ is initialized to zero to ensure training starts from a stable pre-trained state. The total training objective combines generation with **learnable gating** and **geometric consistency** alignment:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{scaffold}} \tag{11}$$

This formulation ensures the LLM's generation is guided by both the overall molecular properties and the specific geometry of the scaffold.

4 Experiments

Datasets and Tasks. To comprehensively evaluate GLA, we conduct experiments on three standard benchmarks: **PCDes** (Zeng et al., 2022) for **cross-modal** molecule-text retrieval, **ChEBI-20** (Edwards et al., 2021) for molecule captioning, and **MoleculeNet** (Wu et al., 2018) for **multitask classification** property prediction.

Research Questions Our evaluation addresses three pivotal questions:

- **(RQ1)** Can disentangled geometric representations surpass isotropic baselines in fine-grained cross-modal retrieval?

Methods	Text→Molecule				Molecule→Text			
	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
Euclidean Graph Alignment Baselines								
MoMu	4.90	14.48	20.69	10.33	5.08	12.82	18.93	9.89
MolCA	35.09	62.14	69.77	47.33	37.95	**66.81**	74.48	50.80
MolFM	16.14	30.67	39.54	23.63	13.90	28.69	36.21	21.42
Linearized Sequence Baselines								
MoleculeSTM	35.80	–	–	–	39.50	–	–	–
LLM-Centric Alignment Baselines								
Atomas-Base	39.08	59.72	66.56	47.33	37.88	59.22	65.56	47.81
Atomas-Large	**49.08**	68.32	73.16	57.79	**46.22**	66.02	72.32	55.52
GLA (Ours)	42.15	**70.59**	**76.52**	**58.14**	44.28	65.51	**75.89**	**56.94**

Table 1: Molecule–text retrieval performance on the PCDes test set. **Bold** and underlined indicate the best and second-best results, respectively. Baseline results are taken from Zhang et al. (2025).

Method	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
MoMu-large	0.599	0.515	-	-	0.593
InstructMol-GS	0.475	0.371	0.566	0.394	0.502
MolCA, Galac1.3B	0.620	0.531	0.681	0.537	0.618
GIT-Mol-GS	0.352	0.263	0.575	0.485	0.560
MolFM-base	0.585	0.498	0.653	0.508	0.594
MolT5-large	0.594	0.508	0.654	0.510	0.594
Text+Chem T5-augm	0.625	0.542	0.682	0.543	0.622
MolXPT	0.594	0.505	0.660	0.511	0.597
MolReGPT (GPT-4-0314)	0.607	0.525	0.634	0.476	0.562
Atomas-Base	0.632	0.549	0.685	0.545	0.626
GLA (Ours)	**0.641**	**0.558**	**0.692**	**0.554**	**0.635**

Table 2: Results of molecule captioning task on ChEBI-20 test set. **Bold** and underlined indicate the best and second-best results, respectively. Baseline results are adapted from Zhang et al. (2025).

- **(RQ2)** Does injecting mixed-curvature summaries into frozen LLMs enhance the generation of structurally faithful captions?
- **(RQ3)** Do these geometry-aware representations transfer effectively to discriminative property prediction tasks?

Baselines To answer these, we benchmark GLA against state-of-the-art methods across three tasks. Detailed baseline configurations and specific implementation settings are provided in Appendix A and Appendix B.

Molecule-Text Retrieval: We compare graph-based alignment models (**MoMu** (Su et al., 2022), **MolFM** (Luo et al., 2023), **MolCA** (Liu et al., 2023c)) and sequence-based methods (**MoleculeSTM** (Liu et al., 2023a)).

Molecule Captioning: Baselines include T5-based sequence models (**MolT5** (Edwards et al., 2022), **Text+Chem T5** (Christofidellis et al., 2023)), GPT-style architectures (**MolXPT** (Liu et al., 2023b), **MolReGPT** (Li et al., 2023)), and multi-modal LLM adapters (**GIT-Mol** (**Liu et al., 2024a**), **InstructMol** (Cao et al., 2023)).

Property Prediction: We evaluate against sequence transformers (**MoleculeSTM**), graph networks (**MoMu**, **MolFM**), and LLM-centric injectors (**MolCA**, **Atomas** (Zhang et al., 2025)).

4.1 Main Results

Molecule-Text Retrieval To validate whether resolving the "geometric homogenization" of flat encoders enhances semantic grounding, we evaluate GLA on the PCDes benchmark (Table 1). By explicitly modeling the hierarchical topology of molecules, GLA achieves **SOTA on comprehensive ranking metrics**, demonstrating superior robustness in the overall ranking space compared to the 7B-parameter baseline, Atomas-Large (e.g., **+1.42% in Molecule-to-Text MRR**). This performance gain indicates that while massive parameter scales may help identify top-1 candidates, our mixed-curvature approach effectively places the correct semantic targets within the local neighborhood of the query, even when the exact match is am-

Method	BBBP	Tox21	ToxCast	ClinTox	MUV	HIV	BACE	SIDER	Avg.
MoleculeSTM-SMILES	70.6	75.7	65.2	86.6	65.7	77.0	82.0	63.7	73.3
MolFM	72.9	77.2	64.4	79.7	76.0	78.8	83.9	64.2	74.6
MoMu	70.5	75.6	63.4	79.9	70.6	75.9	76.7	60.5	71.6
MolCA-SMILES	70.8	76.0	56.2	89.0	-	-	79.3	61.1	-
Atomas	73.7	77.9	66.9	93.2	**76.3**	80.6	83.1	64.4	77.0
GLA	**75.6**	**79.7**	**69.6**	**94.5**	75.6	**82.2**	**85.2**	**67.9**	**78.8**

Table 3: Results for molecular property prediction tasks (ROC-AUC) on MoleculeNet benchmark. **Bold** and underlined indicate the best and second-best results. Baseline results are adapted from Zhang et al. (2025).

biguous. Furthermore, the substantial improvement over Euclidean graph baselines like MolCA (e.g., **+6.33% in M2T R@1**) confirms that breaking the bottleneck of uniform geometric spaces is critical for aligning fine-grained biochemical descriptions with their corresponding structural motifs.

Molecule Captioning We further investigate whether grounding language models in explicit geometric motifs can bridge the semantic gap often observed in 1D sequence models. On the **ChEBI-20** captioning task (Table 5), GLA consistently outperforms the strongest baseline, Atomas-Base, across all generation metrics (e.g., +0.9 BLEU-4, +0.7 ROUGE-1). Unlike MolT5 and Text+Chem T5, which rely solely on linearized SMILES and struggle to maintain structural consistency, GLA's superior performance stems from its ability to attend differentially to the scaffold and side-chains. This geometric guidance ensures that the generated descriptions are not only linguistically fluent but also **structurally faithful**, accurately reflecting the molecule's functional groups rather than hallucinating generic chemical traits.

Molecular Property Prediction We evaluate if geometric disentanglement enhances discriminative reasoning, particularly under topological distribution shifts. As shown in Table 3, GLA consistently outperforms GNN and LLM baselines on MoleculeNet, with significant gains on scaffold-split datasets like HIV and BACE. This confirms that our **curvature-guided routing** provides stronger **geometric inductive priors** than standard Euclidean embeddings, mitigating overfitting when test topologies diverge. Notably, GLA achieves a 2.4% improvement over **Atomas** on the BBBP task. This gain validates that **disentangled modeling of lipophilic side-chains** allows the LLM to reason more effectively about property-determining substructures than models using monolithic encodings.

4.2 Ablation Study

To deconstruct the contribution of each component, we perform an ablation study on the PCDes test set (Table 4).

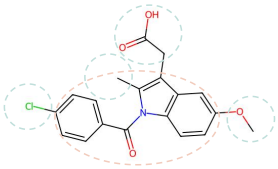
Effect of Hyperbolic Geometry. Replacing the hyperbolic MoE encoder with a standard Euclidean GCN leads to a significant performance drop of **3.13%** in R@1. This degradation confirms that the hierarchical complexity of molecular scaffolds is poorly modeled by flat Euclidean geometry, resulting in embeddings that lack the necessary topological separation for precise retrieval.

Impact of Structural Disentanglement. Removing the explicit scaffold/side-chain decomposition (Row 3) causes a decline of **2.23%**. This indicates that treating the molecule as a homogeneous graph dilutes the distinct semantic signals of the core scaffold and functional groups. By disentangling them, GLA allows the model to align specific textual phrases to their corresponding structural regions more effectively.

Necessity of Gated Injection. The most substantial drop (**3.58%**) is observed when replacing the gated cross-attention with a static projection. This finding validates that the alignment between chemical structure and natural language is not static; the model requires a dynamic gating mechanism to modulate its attention between the scaffold (for core classification) and side-chains (for property modification) depending on the textual context.

5 Analysis

The quantitative superiority of GeoGLA suggests that explicitly modeling geometric hierarchy benefits semantic alignment. In this section, we probe the internal mechanisms of the model to validate two central hypotheses: (1) that the mixed-curvature latent space successfully resolves the geometric collapse of Euclidean encoders by disentangling structural roles based on specificity, and



Indomethacin

Model	Text Alignment Visualization (on Ground Truth)
PubChem	Indometacin is a member of the class of indole-3-acetic acids that is indole-3-acetic acid in which the indole ring is substituted at positions 1, 2 and 5 by p-chlorobenzoyl, methyl, and methoxy groups, respectively. A non-steroidal anti-inflammatory drug, it is used in the treatment of musculoskeletal and joint disorders including osteoarthritis, rheumatoid arthritis, gout, bursitis and tendinitis. It is a N-acylindole, a member of monochlorobenzenes, an aromatic ether and a member of indole-3-acetic acids.
GLA (Ours)	Indometacin is a member of the class of indole-3-acetic acids that is indole-3-acetic acid in which the indole ring is substituted at positions 1, 2 and 5 by p-chlorobenzoyl, methyl, and methoxy groups, respectively. A non-steroidal anti-inflammatory drug... It is a N-acylindole, a member of monochlorobenzenes, an aromatic ether and a member of indole-3-acetic acids.
GLA (w/o Dual)	Indometacin is a member of the class of indole-3-acetic acidsthat is indole-3-acetic acid in which the indole ring is substituted at positions 1, 2 and 5 by p-chlorobenzoyl, methyl, and methoxy groups, respectively. A non-steroidal anti-inflammatory drug... It is a N-acylindole, a member of monochlorobenzenes, an aromatic ether and a member of indole-3-acetic acids.

Figure 3: Qualitative comparison using full PubChem description. **GLA (Ours)** accurately aligns the structural core (Orange) and specific substituents (Blue) within the long text.

Variant	M2T R@1	Δ
GLA (Full)	**44.28**	-
Geometric Components		
- w/o Hyperbolic Enc.	41.15	-3.13
- w/o Decomposition	42.05	-2.23
Interaction Mechanisms		
- w/o Gated Attention	40.70	-3.58
- w/o Dual-View Align	42.50	-1.78

Table 4: Ablation study on the PCDes test set. Removing hyperbolic geometry and gated injection causes significant performance drops.

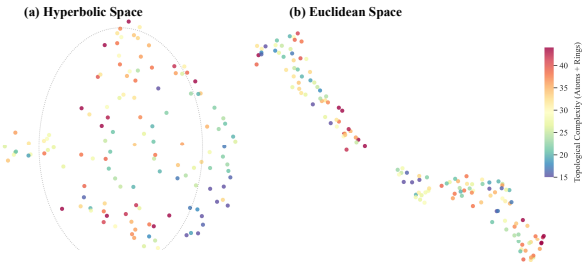


Figure 4: **Learned Manifold Visualization (UMAP).** Projections of graph embeddings on `ogbg-molclintox`, colored by topological complexity. While the Euclidean baseline (Right) suffers from structural collapse, GLA (Left) exhibits a clear radial hierarchy, effectively distributing complex structures towards the hyperbolic boundary.

(2) that the Dual-View Contrastive Alignment is indispensable for the LLM to ground fine-grained structural motifs without hallucinating local details.

5.1 Manifold Visualization and Geometric Fidelity

To empirically validate whether GeoGLA mitigates the representation collapse typical of flat graph encoders, we visualize the learned embedding manifolds using UMAP in Figure 4. Each point represents a molecular graph, color-coded by its topological complexity.

As shown in the Euclidean (Right), the manifold exhibits a pathological "filamentous collapse", where molecules of varying complexities are compressed into narrow, overlapping ribbon-like regions. This confirms that Euclidean geometry lacks the **intrinsic dimensionality** and **representation capacity** to accommodate the exponential growth of the molecular state space, leading to crowding artifacts that obscure structural distinctions.

In contrast, GeoGLA (Left) demonstrates a clear radial organization characteristic of hyperbolic embeddings. We observe a distinct radial trend correlating with structural complexity: generic scaffolds (cool hues) cluster near the semantic origin, while highly complex molecules (warm hues) are naturally distributed towards the Poincaré ball boundary. This result suggests that our mixed-curvature encoder effectively utilizes the hyperbolic radius to represent structural specificity and hierarchy, naturally distributing complex molecules towards the expansive boundary while keeping generic scaffolds at the core.

5.2 Correlating Hyperbolic Norm with Semantic Specificity

To quantify the efficacy of curvature-guided encoding, we analyze the relationship between the hyperbolic norm $||z||_2$ of the learned embeddings and the structural role of components. Guided by **dynamic gating**, standard hyperbolic embeddings typically place root nodes near the origin and leaf nodes near the boundary to accommodate the expo-

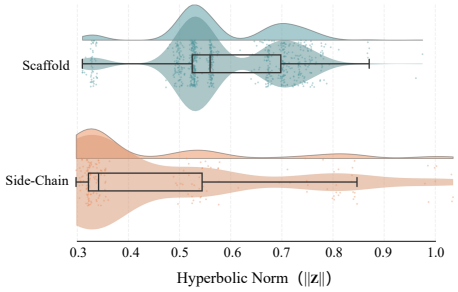


Figure 5: **Distribution of Hyperbolic Embedding Norms ($||z||_2$).** Pharmacophoric scaffolds (Top) are mapped towards the manifold boundary (high norm), while local side-chains (Bottom) cluster near the origin. This confirms that the model successfully disentangles structural roles based on geometric magnitude.

nential growth of the **structural hierarchy** within tree-like graphs.

However, as illustrated in Figure 5, our model exhibits a distinct distribution driven by **semantic specificity** rather than simple topological depth. For a more granular visualization of these distributions, explicitly contrasting the density and spread of scaffolds versus side-chains, please refer to Figure 6 in Appendix C.

Specifically, pharmacophoric scaffolds are mapped towards the boundary of the Poincaré ball (High Norm). This reflects their role as the primary determinants of biological activity—a highly specific semantic class that requires the expansive volume of the boundary to resolve fine-grained structural variations (e.g., distinguishing a benzodiazepine core from a similar fused ring). Conversely, functional side-chains cluster closer to the origin (Low Norm). Semantically, these substructures (e.g., methyl, hydroxyl groups) act as common, reusable modifiers that appear across diverse molecular families. Their proximity to the origin reflects their high generality and lower semantic entropy compared to the unique scaffolds. This geometric organization confirms that GeoGLA effectively translates the *chemical hierarchy*—from generic building blocks to specific drug cores—into a structured geometric manifold.

5.3 Qualitative Analysis

To intuitively demonstrate how **Dual-View Contrastive Alignment** supports fine-grained grounding, we present a qualitative case study in Figure 3. We compare captions generated by the full GLA model with those from a variant without dual-view alignment (w/o Dual) for the molecule *Indomethacin*.

The ablated model correctly identifies the global category (member of indole-3-acetic acids") but hallucinates or omits key local details, failing to specify substituents on the indole ring. In contrast, GeoGLA accurately captures functional groups, correctly generating fine-grained descriptions such as p-chlorobenzoyl" and "methoxy groups" at positions 1 and 5. This comparison highlights that while the global view encodes scaffold identity, alignment of local geometric summaries is crucial for grounding complex chemical modifications.

To examine whether the Mixed-Curvature MoE learns to route substructures to appropriate manifolds, we visualize the gating network's attention distribution on the ogbg-molclintox dataset (Figure 1). The model exhibits distinct routing behaviors for scaffolds and side-chains, supporting our geometric disentanglement hypothesis.

Pharmacophoric scaffolds are predominantly routed to the high-curvature hyperbolic expert ($\kappa = -1.5$), with median attention weights substantially higher than those of other experts, suggesting that the model leverages hyperbolic space to encode hierarchical molecular cores. In contrast, **side-chains** show a more diffuse distribution, with increased affinity for Euclidean ($\kappa = 0$) and low-curvature ($\kappa = -0.5$) experts. This indicates that functional groups are treated as local, relatively flat structures, effectively alleviating the "geometric homogenization" limitation of standard encoders.

6 Conclusion

In this work, we address the limitations of standard encoders in capturing chemical compositionality. We propose GLA, a framework that aligns LLMs with molecular hierarchy via mixed-curvature representation learning. Instead of relying on rigid rules, GLA learns to route molecular components to their optimal geometries (hyperbolic or Euclidean), effectively resolving the embedding collapse common in flat spaces. Experiments across retrieval, captioning, and property prediction demonstrate that this approach significantly enhances the grounding of textual concepts in structural motifs. Broadly, our findings suggest that incorporating geometric inductive biases is essential for moving scientific LLMs from surface pattern matching to structure-aware reasoning. Future work may explore hyperbolic attention mechanisms to further mitigate linearization constraints.

Limitations

Despite GLA's strong performance, we identify three critical limitations. First, Riemannian optimization and manifold-constrained operations introduce computational overhead during the backward pass, posing significant throughput challenges when scaling to billion-parameter pre-training or processing massive chemical libraries. Second, our curvature-guided decomposition provides a powerful inductive bias for complex, hierarchically-organized drug-like molecules, but its benefits may diminish for structurally homogeneous or simple linear compounds where the scaffold-sidechain distinction is semantically negligible. Finally, while GLA aligns global and motif-level features, it does not yet account for the dynamic conformational flexibility of molecules; capturing the ensemble of 3D spatial conformers remains an open challenge for achieving a truly holistic geometric-language alignment in future iterations.

References

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2025. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 354–379. Association for Computational Linguistics.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning*, pages 6140–6157. PMLR.

Vishal Dey, Xiao Hu, and Xia Ning. 2025. GeLLM³O: Generalizing large language models for multi-property molecule optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25192–25221, Vienna, Austria. Association for Computational Linguistics.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 375–413. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Junfeng Fang, Shuai Zhang, Chang Wu, Zhengyi Yang, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, and Xiang Wang. 2024a. MolTC: Towards molecular relational modeling in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1943–1958, Bangkok, Thailand. Association for Computational Linguistics.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024b. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhihui Guo, Pramod Sharma, Andy Martinez, Liang Du, and Robin Abraham. 2022. Multilingual molecular representation learning via contrastive pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3441–3453, Dublin, Ireland. Association for Computational Linguistics.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. 2020. Strategies for pre-training graph neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2025. Structural reasoning improves molecular understanding of LLM. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21016–21036, Vienna, Austria. Association for Computational Linguistics.

Yinuo Jiang, Xiang Zhuang, Keyan Ding, Qiang Zhang, and Huajun Chen. 2024. Enhancing cross text-molecule learning by self-augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9551–9565, Bangkok, Thailand. Association for Computational Linguistics.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2023. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *arXiv preprint arXiv:2306.06615*.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024a. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073.

Qi Liu, Maximilian Nickel, and Douwe Kiela. 2019. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8228–8239.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023a. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nat. Mac. Intell.*, 5(12):1447–1457.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023b. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1606–1616, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638.

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024b. ReactXT: Understanding molecular "reaction-ship" via reaction-contextualized molecule-text pretraining. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5353–5377, Bangkok, Thailand. Association for Computational Linguistics.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023. Molfm: A multimodal molecular foundation model. *CoRR*, abs/2307.09484.

Zijun Min, Bingshuai Liu, Liang Zhang, Jia Song, Jinsong Su, Song He, and Xiaochen Bo. 2024. Exploring optimal transport-based multi-grained alignments for text-molecule retrieval. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024, Lisbon, Portugal, December 3-6, 2024*, pages 2317–2324. IEEE.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. Biot5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1216–1240. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *CoRR*, abs/2209.05481.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2019, Niagara Falls, NY, USA, September 7-10, 2019*, pages 429–436. ACM.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.*, 4(3):279–287.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36.

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.

Zhenqin Wu, Jiahui Wang, Haonan Du, and 1 others. 2023. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14:2585.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. MolRAG: Unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531, Vienna, Austria. Association for Computational Linguistics.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862.

Yikun Zhang, Geyan Ye, Chaohao Yuan, Bo Han, Long-Kai Huang, Jianhua Yao, Wei Liu, and Yu Rong. 2025. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep graph contrastive representation learning. *CoRR*, abs/2006.04131.

A Baseline Details

To evaluate the effectiveness of GeoGLA, we compare it against a diverse set of baselines across three tasks. Below, we provide detailed descriptions of these methods.

A.1 Molecule-Text Retrieval Baselines

We compare our framework against several leading multi-modal pre-training models. **MoMu** (Su et al., 2022) is a pioneering framework that utilizes contrastive learning to align molecular graphs with natural language descriptions. **MoleculeSTM** (Liu et al., 2023a) bridges chemical structures and textual descriptions through a multi-modal contrastive learning strategy that explicitly incorporates chemical reactivity knowledge. **MolFM** (Luo et al., 2023) learns joint representations by integrating three separate unimodal encoders to process molecular structures, biomedical texts, and knowledge graphs respectively. Finally, **MolCA** (Liu et al., 2023c) employs a cross-modal projector to connect a graph encoder with a Large Language Model (LLM).

A.2 Molecule Captioning Baselines

For the captioning task, we benchmark against 9 state-of-the-art methods. **MolT5** (Edwards et al., 2022) is a T5-based encoder-decoder model pre-trained on molecule-text pairs. Similarly, **Text+Chem T5** extends the T5 architecture to process both natural language and SMILES within a unified framework. **MolXPT** adopts a GPT-style autoregressive architecture based on GPT-2$_{medium}$. **GIT-Mol** (Liu et al., 2024a) utilizes a GIT-Former to map molecular graphs, images, and text into a shared latent space. **MolReGPT** leverages in-context learning with frozen LLMs (e.g., GPT-3.5-turbo) to generate descriptions through retrieval-augmented prompting. **InstructMol** (Cao et al., 2023) aligns molecular encoders with LLMs using instruction tuning; we specifically compare against the **InstructMol+GS** variant. Additionally, we evaluate the generative capabilities of **MoMu**, **MolFM**, and **MolCA**.

A.3 Molecular Property Prediction Baselines

For molecular property prediction, we utilize the MoleculeNet benchmark (Wu et al., 2018). We compare against sequence-based models like **MoleculeSTM** and **MolT5**, which treat molecules as linear SMILES strings. We also include graph-based baselines such as **MolFM** and **MoMu**, which utilize GNNs to capture topological information. Furthermore, we benchmark against LLM-centric approaches like **MolCA** and **Atomas** (Zhang et al., 2025), which inject molecular features into frozen LLMs for downstream classification.

B Implementation Details

The framework is instantiated with **Qwen-2.5-3B** as the frozen LLM backbone. We apply Low-Rank Adaptation (LoRA) to the query and value projections of the attention layers with rank $r = 16$. The molecular encoder utilizes a 5-layer hyperbolic GIN initialized with curvature $K \in \{-1.5, -0.5, 0.0\}$.

Training is conducted using the **RiemannianAdam** optimizer for geometric parameters and AdamW for Euclidean parameters. We set the learning rate to 1e-4 with a linear warmup over the first 10% of steps. To enhance robustness, we incorporate Free Large-batch Adversarial Generation (FLAG) with a perturbation magnitude of 8e-3. All experiments are conducted on NVIDIA L20 GPU.

C Additional Geometric Visualizations

To further substantiate the geometric disentanglement discussed in Section 5.2, we visualize the distribution of hyperbolic embedding norms using a Raincloud plot.

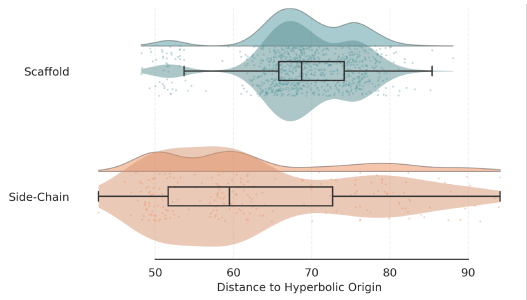
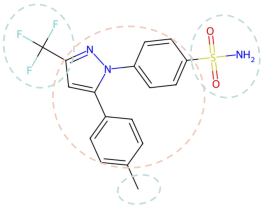


Figure 6: **Visualizing Hierarchical Disentanglement in Hyperbolic Space (Indomethacin).** In sharp contrast to the Euclidean embeddings in Figure **??**, where structural roles are often entangled, the hyperbolic geometry effectively disentangles the molecular hierarchy. The rigid scaffold nodes (purple) are pushed towards the manifold boundary, while flexible side-chain nodes (yellow) cluster near the origin, clearly separating their distinct semantic roles.

Method	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
MoMu-small	0.532	0.445	-	-	0.564
MoMu-base	0.549	0.462	-	-	0.575
MoMu-large	0.599	0.515	-	-	0.593
InstructMol-GS	0.475	0.371	0.566	0.394	0.502
MolCA, Galac1.3B	0.620	0.531	0.681	0.537	0.618
GIT-Mol-GS	0.352	0.263	0.575	0.485	0.560
MolFM-small	0.542	0.452	0.623	0.469	0.562
MolFM-base	0.585	0.498	0.653	0.508	0.594
MolT5-small	0.519	0.436	0.620	0.469	0.563
MolT5-base	0.540	0.457	0.634	0.485	0.578
MolT5-large	0.594	0.508	0.654	0.510	0.594
Text+Chem T5-augm	0.625	0.542	0.682	0.543	0.622
MolXPT	0.594	0.505	0.660	0.511	0.597
MolReGPT (GPT-3.5-turbo)	0.565	0.482	0.450	0.543	0.585
MolReGPT (GPT-4-0314)	0.607	0.525	0.634	0.476	0.562
Atomas-Base	0.632	0.549	0.685	0.545	0.626
GLA (Ours)	**0.641**	**0.558**	**0.692**	**0.554**	**0.635**

Table 5: Results of molecule captioning task on ChEBI-20 test set. **Bold** and underlined indicate the best and second-best results, respectively. Baseline results are adapted from Zhang et al. (2025).



Celecoxib

Model	**Text Alignment Visualization (on Ground Truth)**
PubChem	Celecoxib is a member of the class of pyrazoles that is 1H-pyrazole which is substituted at positions 1, 3 and 5 by 4-sulfamoylphenyl, trifluoromethyl and p-tolyl groups, respectively. A cyclooxygenase-2 inhibitor... It is a member of toluenes, a sulfonamide, a member of pyrazoles and an organofluorine compound.
GLA (Ours)	Celecoxib is a member of the class of pyrazoles that is 1H-pyrazole which is substituted at positions 1, 3 and 5 by 4-sulfamoylphenyl, trifluoromethyl and p-tolyl groups, respectively. A cyclooxygenase-2 inhibitor... It is a member of toluenes, a sulfonamide, a member of pyrazoles and an organofluorine compound.
GLA (w/o Dual)	Celecoxib is a member of the class of pyrazoles that is 1H-pyrazole which is substituted at positions 1, 3 and 5 by 4-sulfamoylphenyl, trifluoromethyl and p-tolyl groups, respectively. A cyclooxygenase-2 inhibitor... It is a member of toluenes, a sulfonamide, a member of pyrazoles and an organofluorine compound.

Figure 7: Visualization of Geometric-Language Alignment on Celecoxib. The text content is identical (Ground Truth) across rows. Colored words indicate successful alignment by the model. **GLA (Ours)** correctly aligns both the scaffold (orange) and side-chain (blue) concepts. In contrast, the **w/o Dual** variant only captures the coarse scaffold information, failing to align the specific side-chain descriptions (shown in plain text).