Improving Molecular Understanding of Large Language Model via Substructure-aware Instruction Tuning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown strong performance in molecular tasks, yet they often fail to capture fine-grained molecular information, particularly the presence of substructures and how they behave across diverse chemical contexts. Most existing approaches rely on surface-level cues, treating substructures as isolated markers rather than modeling their functional behaviors. We introduce `SubMol-Instructions`, a substructure-aware instruction tuning dataset that explicitly links molecular substructures to their task-specific functional behaviors. We further propose `StructMol`, a molecule LLM that builds a robust global molecular representation from multiple structural views and is trained to incorporate the fine-grained substructure-level supervision introduced by `SubMol-Instructions`. Experimental results on reaction prediction, property prediction, and molecule translation tasks show that our approach consistently outperforms powerful baselines, highlighting the importance of explicitly defining and learning substructural behaviors for improving molecular understanding of LLM.[1]

1 Introduction

LLMs have demonstrated remarkable versatility across a wide range of domains, serving as pivotal catalysts for scientific discovery. In drug development, Molecule LLMs have emerged as powerful tools to accelerate traditional, labor-intensive wet-lab pipelines (M. Bran et al., 2024; Cao et al., 2025). By adapting general-purpose language models via molecular multi-task instruction tuning (Fang et al., 2023), these systems have demonstrated the potential to perform challenging chemical tasks, such as reaction prediction (Wei et al., 2010), property prediction (Wu et al., 2018), and molecule-text translation (Edwards et al., 2022).

Despite these achievements, current Molecule LLMs still exhibit fundamental limitations in achieving deep molecular understanding. Mostly, these models primarily rely on a linear 1D representation, such as SMILES (Weininger, 1988) or SELFIES (Krenn et al., 2020). However, the distinct distribution of these representations diverges substantially from the natural language corpora used for pretraining, causing models to struggle with capturing precise structural details, particularly functional groups (Chen et al., 2025; Kim et al., 2025). Furthermore, the sparsity of molecular data restricts the effective assimilation of new chemical knowledge, thereby hindering the acquisition of fine-grained representations (Park et al., 2025). As a result, they often fail to capture fine-grained structural differences that play a crucial role in shaping a molecule's core functionalities and properties (Wu et al., 2023).

While a few studies have attempted to enhance fine-grained understanding by leveraging molecular substructures (Yang et al., 2025; Lin et al., 2025), they remain limited by their reliance on superficial cues, such as the mere presence of substructures or associated textual keywords. However, the inherent properties and functions of a molecule are deeply embedded in its structure, composition, and interactions (Edwards et al., 2025). Specifically, molecular tasks contain implicit structural cues, such as identifying which bond connectivity is severed during a chemical reaction (Coley et al., 2019) or determining whether specific motifs disproportionately govern a target property (Xiong et al., 2019). Expecting LLMs to implicitly internalize these complex insights solely from surface-level data may be insufficient, given their lack of topological understanding. This motivates us to explicitly model the substructure-level functional cues.

To address the aforementioned limitation, we introduce **`SubMol-Instructions`**, a dataset that links substructures to their functional behaviors

[1] Our data and code are available at `https://anonymous.4open.science/r/SubMol-Instructions`.

across diverse chemical tasks (Wei et al., 2010; Wu et al., 2018; Edwards et al., 2022). Using annotated molecules, we first isolate substructures and extract their relationships with task outcomes. These relationships comprise structural inclusion between substructures and their parent molecules or textual descriptions, substructure roles in chemical reactions, and substructure contributions to molecular properties. We then filter noisy instances using a noisy sample estimation method to ensure data quality (Arazo et al., 2019). The resulting dataset facilitates fine-grained molecular understanding by explicitly supervising substructure-level behaviors.

To fully harness this explicit substructure-level supervision from our curated dataset and deepen molecular understanding, we further introduce **`StructMol`**. Recognizing that local substructure behaviors are intrinsically governed by the global molecular context, `StructMol` constructs a coherent global representation through a dual-view strategy. This approach jointly generates SMILES and SELFIES to capture complementary structural insights: the explicit bond connectivity of SMILES aids in delineating substructural boundaries, while the robustness of SELFIES ensures chemically consistent generation (Leon et al., 2024). This design provides a coherent global context in which fine-grained substructure-level supervision can be effectively applied.

Empirical evaluations across a diverse range of tasks, including reaction prediction, property prediction, and molecule–text translation, demonstrate that our model consistently outperforms strong baselines despite its compact size. Remarkably, even when relying solely on 1D molecular representations, it often surpasses state-of-the-art methods that incorporate 2D graph information. These results underscore that advancing molecular understanding in LLMs hinges not merely on scaling model size or adding modalities, but on effectively leveraging intrinsic molecular structure. Ultimately, this validates that the fine-grained substructural learning facilitated by our dataset is key to enabling robust generalization across diverse chemical contexts. Our main contributions are as follows:

- We introduce `SubMol-Instructions`, a dataset explicitly bridging molecular substructures with their functional behaviors.
- We propose `StructMol`, a dual-view framework that synergizes SMILES and SELFIES to capture complementary structural details.
- Our method outperforms existing methods across diverse tasks, underscoring the importance of explicit substructural learning.

2 Related Work

2.1 Molecule Language Modeling

Researchers in natural language processing have long explored graph-structured data, motivating the interpretation of molecules as atom-level graphs and the application of NLP techniques to chemical domains (Wu et al., 2018; Chithrananda et al., 2020; Ross et al., 2022). These approaches have gained attention for enabling more efficient drug candidate screening than traditional wet-lab pipelines. More recently, molecule instruction-tuning methods (Fang et al., 2023; Yu et al., 2024a) have sought to leverage the adaptability of LLMs for biomolecular tasks. However, LLMs still exhibit limited structural understanding of molecules, motivating continued efforts to address this challenge (Park et al., 2024a; Hu et al., 2025). For example, InstructMol integrates a 2D graph encoder with an LLM (Cao et al., 2025), while subsequent work (Pei et al., 2025) further incorporates 3D molecular topology to improve generation reliability. Nevertheless, existing approaches continue to struggle with fine-grained molecular understanding, often underperforming small-sized specialist models (<0.3B parameters) (Park et al., 2025). In this work, we aim to bridge this gap by explicitly modeling substructure-level functional behaviors, enabling more precise molecular reasoning.

2.2 Fine-grained Molecular Understanding

Efforts to improve molecular understanding in language models have explored learning from fine-grained molecular decompositions. For property prediction, prior work has proposed fine-grained alignment across multiple molecular modalities (Feng et al., 2023; Yu et al., 2024b; Park et al., 2024b), while molecule–caption alignment methods associate molecular fragments with textual phrases to capture structure–language correspondence (Zhang et al., 2025; Park et al., 2025). To further enhance structural perception, hierarchical tokenization schemes encode atom-, motif-, and molecule-level information (Chen et al., 2025), and datasets with multi-level annotations have been introduced to better link molecular structures to textual descriptions (Yang et al., 2025). In parallel, large-scale synthesis and decomposition pre-

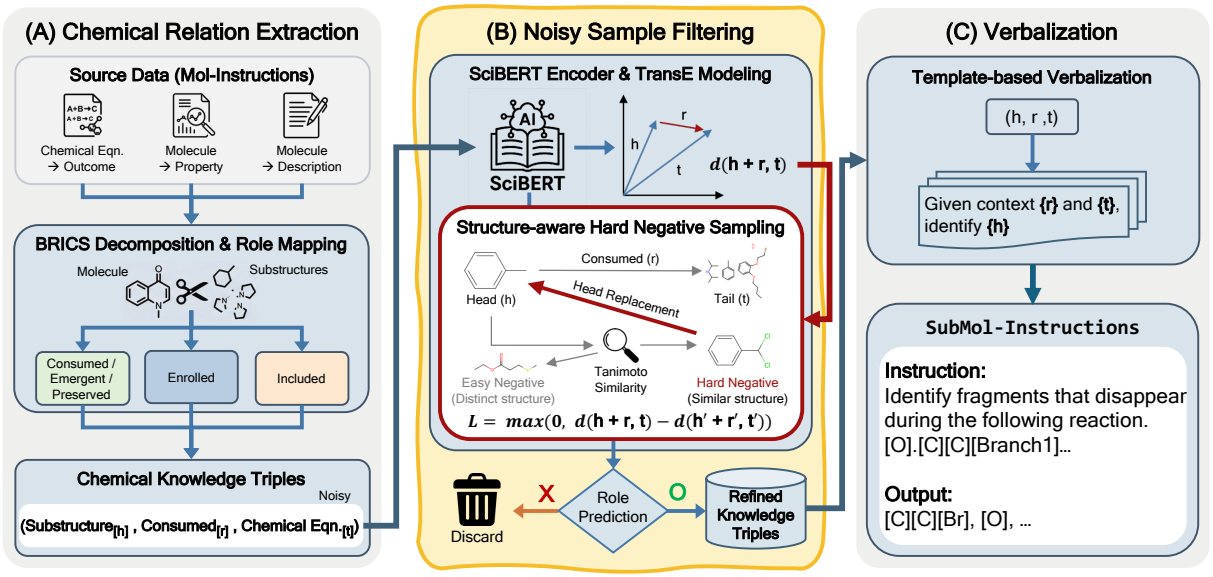


Figure 1: Data generation process of SubMol-Instructions. Data statistics are provided in Table 6 in the Appendix.

training has been shown to benefit reaction prediction tasks (Lin et al., 2025). However, these approaches primarily rely on surface-level signals, such as structural presence or alignment, without explicitly modeling how substructures influence task outcomes. In contrast, our work directly models substructure-level functional behaviors across diverse chemical tasks, enabling more explicit and task-oriented molecular reasoning.

3 Method

In this section, we introduce our proposed dataset and model. SubMol-Instructions provides explicit supervision on substructural cues, enabling models to capture fine-grained structural signals across diverse chemical tasks (§3.1). Using this dataset, we develop StructMol, a structure-aware molecule LLM that improves molecular understanding by learning our new dataset (§3.2).

3.1 SubMol-Instructions Construction

3.1.1 Chemical Relation Extraction

To effectively predict how molecules behave in tasks such as reaction prediction, property prediction, and molecule-text translation, it is essential to identify substructural cues. In this work, we define and extract meaningful substructures based on three key relationships, enabling the model to learn fine-grained molecular representations.[2] An overview of the data generation process is shown in Figure 1.

[2]More details are provided in Appendix B.

Structure-Reactivity Relation

In the context of chemical reactions, local structural features, such as the degree of connectivity and bond order, are critical determinants of reactivity (Coley et al., 2019). To enable the model to learn the causal link between structure and reactivity, we distinguish and extract substructures that are **consumed**, **emergent**, or **preserved** during reactions.

Structure-Property Relation

In the prediction of molecular properties, a critical premise is that molecules sharing similar substructures exhibit similar macroscopic properties (Le et al., 2012; Park et al., 2024b). To facilitate the learning of this mapping between structure and property, we extract all **enrolled** substructures that contribute to the manifestation of a specific property.

Structure-Description Relation

Textual descriptions of molecules are composed of keywords derived from specific substructures (Park et al., 2025). To capture the subtle semantic alignment between molecular structure and natural language descriptions, we extract substructures that are **included** in the description of the parent molecule.

We implement this schema by processing the training sets of Mol-Instructions (Fang et al., 2023),

covering forward reaction prediction, retrosynthesis (Lu and Zhang, 2022), property prediction (Wu et al., 2018), molecule captioning and generation (Kim et al., 2021).[3] For each molecule–output pair, we decompose the molecule using the BRICS algorithm (Degen et al., 2008) and map the resulting substructures to task entities according to the categories above. Task entities may take the form of an input chemical equation, a scalar property value, or a textual description. This process converts each instance into a structured chemical knowledge triple of the form

$$(\texttt{substructure}, \texttt{relation}, \texttt{entity}),$$

where both `substructure` and molecular `entity` are expressed in SELFIES format.

3.1.2 Learning-based Noisy Sample Filtering

While the relation extraction procedure provides broad coverage, the heuristic nature of rule-based mapping inevitably introduces noise. To improve dataset reliability, we propose a filtering mechanism grounded in the early-learning phenomenon, wherein deep neural networks prioritize clean, consistent patterns before memorizing noisy labels (Arazo et al., 2019). This filtering step serves to stabilize role supervision, ensuring that `SubMol-Instructions` reflects consistent functionality of substructures.[4]

We first train a knowledge graph embedding model using chemical knowledge triples curated in the above section. To facilitate the encoding of scientific data, we utilize SciBERT (Beltagy et al., 2019) as an entity encoder, which maps the `substructure` (h) and the `entity` (t) into a shared embedding space. These embeddings are optimized using the *TransE* scoring function (Bordes et al., 2013) with a margin-based ranking loss:

$$\mathcal{L} = \sum_{(h,r,t)\in\mathcal{S}} \sum_{(h',r',t')\in\mathcal{S}'} \Big[\gamma + d(\mathbf{h}+\mathbf{r},\mathbf{t}) - d(\mathbf{h}'+\mathbf{r}',\mathbf{t}')\Big]_+ \quad (1)$$

where $\mathcal{S}$ denotes the set of original triples, $\mathcal{S}'$ the set of negative triples, γ the margin, and $d(\cdot)$ the distance metric.

[3]Reagent prediction data were excluded due to the lack of high-quality annotations (Andronov et al., 2023; Yu et al., 2024a).

[4]Further analysis of the impact of our filtering method, along with filtered examples, is in Appendix E.

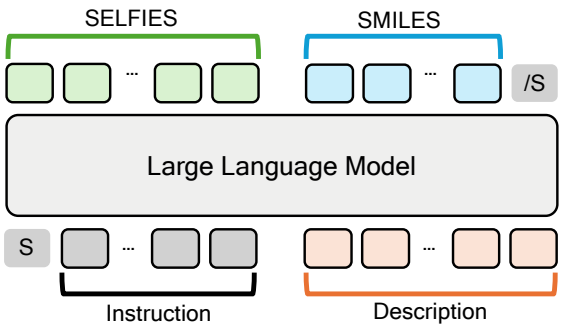


Figure 2: Illustration of Dual-view Molecule Alignment. Given alignment pairs of SMILES, SELFIES, and Description, LLM takes one of them to generate the others.

To construct negative triples, we corrupt original triples by replacing substructures with structurally similar molecules rather than using random sampling, guiding the model to distinguish fine-grained structural differences. During training, each original triple is corrupted by substituting either the head or the tail with one of its precomputed nearest neighbors based on structural similarity. We compute Morgan fingerprints (Rogers and Hahn, 2010a) and construct a similarity map using Tanimoto similarity.[5]

After training, the embedding model serves as a filter. For each triple, we rank candidate relations according to embedding distance. Triples whose ground-truth relation r does not appear within the model-predicted relations are considered inconsistent and are removed.

3.1.3 Verbalization

To bridge the gap between structured chemical knowledge and text generation objectives, we transform refined triples into natural language instructions using the templates listed in Table 9 in Appendix. By guiding the model to generate the target (h) conditioned on its functional context defined by the (r) and (t), we ensure that the model explicitly learns the correlations between structural patterns and their functional behaviors.

3.2 `StructMol`

To better learn local information of molecules, the model should first understand global structural information, since the functional behaviors of substructures are inextricably linked to the global molecular context. To construct this comprehensive structural representation, we propose a dual-view

[5]All molecular processing is performed using RDKit.

learning strategy that simultaneously predicts both SMILES and SELFIES. By combining the explicit connectivity of SMILES with the robust validity constraints of SELFIES (Leon et al., 2024), this approach captures complementary structural details that contribute to a coherent global molecular representation.

3.2.1 Step 1: Dual-view Molecule Alignment

Following prior work (Cao et al., 2025; Hu et al., 2025), we perform molecule–text pre-training using PubChem data (Kim et al., 2021) to integrate molecular knowledge into the internal representations of LLMs. While existing molecule LLMs primarily rely on SELFIES for generation, we integrate SMILES to compensate for the lack of explicit connectivity in SELFIES.

Specifically, we collect 430K molecule–text pairs and represent each molecule using both SMILES and SELFIES, forming triplets of (SMILES, SELFIES, Description). During pre-training, the model is conditioned on one component along with a task instruction and trained to generate the remaining components, as illustrated in Figure 2. This dual-view pre-training objective exposes the model to complementary structural signals and encourages consistent alignment across textual and molecular representations, yielding a more robust alignment as analyzed in Appendix C.

3.2.2 Step 2: Dual-view Instruction Tuning

Building upon the dual-view alignment established in Step 1, we further train the model using `SubMol-Instructions` to capture local information. In this stage, we adopt the same dual-view generation strategy. Specifically, when generating substructures, the model is encouraged to produce both SMILES and SELFIES representations simultaneously, allowing it to learn connected and structurally valid molecular fragments in a consistent manner. The same dual-view training strategy is also maintained during downstream molecular generation tasks, where the model jointly generates SMILES and SELFIES. By doing so, StructMol actively leverages the structural knowledge acquired during pre-training and instruction tuning, ensuring consistency between global molecular context and local substructure reasoning.

4 Experiments

In this section, we verify the efficacy of our new dataset, `SubMol-Instructions`, through extensive experiments and analyses using `StructMol` to answer the following questions:

- Can `StructMol` exhibit better molecule understanding? (§4.2, §4.4)
- Can the learned knowledge of `StructMol` be generalized to unseen tasks? (§4.3)
- Which components of `StructMol` are crucial for improving its effectiveness? (§4.5, §E)

4.1 Experimental Settings

Dataset. We use the training data from Mol-Instructions (Fang et al., 2023) as a seed to construct `SubMol-Instructions`. From the molecules in this dataset, we extract a total of 510,526 substructures, which are used to generate approximately 1.8 million substructure-related question–answer pairs. To decompose molecules into substructures, we impose a maximum atom count of 100 due to computational constraints. The molecule–text pairs used for pre-training `StructMol` are collected following the approach described in Park et al. (2025). For evaluation, we conduct molecule generation tasks, including forward reaction prediction, retrosynthesis, and description-guided molecule design, using Mol-Instructions. We additionally evaluate molecule captioning on the ChEBI-20 dataset (Edwards et al., 2022) and molecule property prediction tasks using the MoleculeNet benchmark (Wu et al., 2018).

Baselines. We compare `StructMol`, our model trained with `SubMol-Instructions`, against a diverse set of competitive generalist Molecule LLMs, including the model from Mol-Instructions (Fang et al., 2023), InstructMol (Cao et al., 2025), PRESTO (Cao et al., 2024), Omni-Mol (Hu et al., 2025), KnowMol (Yang et al., 2025), and UniMoT (Guo et al., 2025). In particular, we focus on comparisons with Omni-Mol, which shares the same backbone language model as our approach (Llama-3.2-1B-Instruct (Dubey et al., 2024)).[6]

Metrics. For molecule generation tasks, we employ Exact Match (EM), BLEU (Papineni et al., 2002), and Levenshtein distance (Lev) (Miller et al., 2009) to measure string level similarity, along with Validity to assess the grammatical correctness of generated molecules. We also adopt molecular fingerprint based similarity metrics, including MACCS FTS (MAC) (Durant et al., 2002), RDK

[6]Implementation details are in Appendix A.

Method	# Params	Modality	Exact↑	BLEU↑	Lev↓	RDK↑	MAC↑	Mor↑	Validity↑
Forward Reaction Prediction									
Mol-Instructions (Fang et al., 2023)	8B	1D	0.503	0.883	13.410	0.756	0.863	0.708	1.000
InstructMol (Cao et al., 2025)	7B	1D + 2D	0.536	0.967	10.851	0.776	0.878	0.741	1.000
HIGHT (Chen et al., 2025)	7B	1D + 2D	0.037	0.869	23.759	0.590	0.394	0.340	0.993
PRESTO (Cao et al., 2024)	7B	1D + 2D	0.355	0.836	10.647	0.646	0.726	0.624	0.973
UniMoT (Guo et al., 2025)	7B	1D + 2D	0.611	0.980	8.297	0.836	0.911	0.807	1.000
Omni-Mol (Hu et al., 2025)	2B	1D + 2D	0.733	0.980	5.550	0.895	0.947	0.870	1.000
KnowMol (Yang et al., 2025)	7B	1D + 2D	0.752	0.986	5.662	0.889	0.943	0.877	1.000
StructMol (SELFIES)	1B	1D	0.893	**0.989**	1.492	0.964	0.981	0.959	1.000
StructMol (SMILES)	1B	1D	**0.903**	0.976	**1.353**	**0.973**	**0.986**	**0.967**	0.992
Retrosynthesis									
Mol-Instructions (Fang et al., 2023)	8B	1D	0.333	0.842	17.642	0.704	0.815	0.646	1.000
InstructMol (Cao et al., 2025)	7B	1D + 2D	0.407	0.941	13.967	0.753	0.852	0.714	1.000
HIGHT (Chen et al., 2025)	7B	1D + 2D	0.008	0.863	28.912	0.564	0.340	0.309	1.000
PRESTO (Cao et al., 2024)	7B	1D + 2D	0.275	0.902	14.433	0.655	0.737	0.619	0.980
UniMoT (Guo et al., 2025)	7B	1D + 2D	0.478	0.974	11.634	0.810	0.909	0.771	1.000
Omni-Mol (Hu et al., 2025)	2B	1D + 2D	0.570	0.960	8.970	0.864	0.909	0.830	1.000
KnowMol (Yang et al., 2025)	7B	1D + 2D	0.598	**0.975**	8.363	0.856	0.912	0.829	1.000
StructMol (SELFIES)	1B	1D	0.639	0.964	6.546	0.897	0.931	0.869	1.000
StructMol (SMILES)	1B	1D	**0.644**	0.910	**6.510**	**0.909**	**0.932**	**0.881**	0.992
Description-guided Molecule Generation									
LLama (Fang et al., 2023)	7B	1D	0.000	0.003	59.864	0.005	0.000	0.000	0.003
Vicuna (Fang et al., 2023)	7B	1D	0.000	0.006	60.356	0.006	0.001	0.000	1.000
MolT5 (Edwards et al., 2022)	0.2B	1D	0.112	0.546	38.276	0.400	0.538	0.295	0.773
Mol-Instructions (Fang et al., 2023)	8B	1D	0.025	0.521	38.742	0.358	0.520	0.221	1.000
UniMoT (Guo et al., 2025)	7B	1D + 2D	**0.237**	0.698	27.782	0.543	0.651	0.411	1.000
Omni-Mol (Hu et al., 2025)	2B	1D + 2D	0.120	**0.824**	**23.590**	0.562	**0.721**	0.442	1.000
KnowMol (Yang et al., 2025)	7B	1D + 2D	0.083	0.797	30.702	0.570	0.693	0.426	1.000
StructMol (SELFIES)	1B	1D	0.081	0.778	31.848	0.583	0.706	0.426	1.000
StructMol (SMILES)	1B	1D	0.085	0.615	32.214	**0.592**	0.713	**0.449**	0.993

Table 1: Results for molecule generation tasks. **Bold** and underlined mark the best and second-best scores. 1D and 2D refer to 1D molecular representations and 2D molecular graphs.

FTS (RDK) (Schneider et al., 2015), and Morgan FTS (Mor) (Rogers and Hahn, 2010b), to compare structural similarity with reference molecules. For molecule captioning, we report BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Denkowski and Lavie, 2014) scores. For property prediction, we use the scaffold split provided by DeepChem (Ramsundar et al., 2019) and report ROC-AUC scores for classification tasks and Mean Absolute Error (MAE) for regression tasks.

4.2 Molecular Generation

Table 1 presents the experimental results across a range of molecule generation tasks, including forward reaction prediction, retrosynthesis, and description-guided molecule generation. Since our model is designed to simultaneously generate both SMILES and SELFIES for a given input, we report the evaluation metrics for each representation separately.

StructMol consistently outperforms prior baselines on reaction-related tasks, specifically forward reaction prediction and retrosynthesis. Forward reaction prediction entails predicting product molecules from chemical equations, whereas retrosynthesis corresponds to the inverse process. Notably, our model achieves substantial gains over Omni Mol, a recent state-of-the-art approach utilizing the identical backbone. Across these tasks, StructMol improves the Exact Match score by more than 18%p on average and reduces the Levenshtein distance by 3.3 points compared to Omni Mol. These indicate that our model generates molecules that are significantly closer to the ground truth at both the string and structural levels.

These results are particularly compelling given the architectural efficiency of StructMol. While most competing methods rely on larger model sizes and auxiliary 2D structural information, our model relies exclusively on 1D molecular representations. This superiority suggests that explicitly learning substructure-level behaviors via SubMol-Instructions enables a more accurate modeling of chemically meaningful transformations, leading to improved generation performance without complex graph encoders.

In the task of description-guided molecule generation, StructMol also demonstrates competitive performance. Despite its compact model size and the absence of 2D molecular graphs, our approach

Method	# P	HOMO↓	LUMO↓	$\Delta\epsilon$ ↓	Avg↓
LLM-based Generalists					
Vicuna (2022)	7B	0.7367	0.8641	0.5152	0.7510
Mol-Instruction (2023)	7B	0.0210	0.0210	0.0203	0.0210
InstructMol (2025)	7B	0.0048	0.0050	0.0061	0.0050
HIGHT (2025)	7B	0.0056	0.0065	0.0077	0.0066
UniMoT (2025)	7B	0.0042	0.0047	0.0055	0.0049
Omni-Mol (2025)	2B	0.0038	0.0047	0.0049	0.0044
StructMol	1B	**0.0031**	**0.0037**	**0.0035**	**0.0034**

Table 2: Results for molecular property prediction regression tasks (MAE) on QM9 benchmark (Fang et al., 2023). **Bold** indicates the best results.

Method	# P	BBBP	Tox21	MUV	HIV	BACE	Avg.
Specialists							
KV-PLM (2022)	0.1B	70.5	49.2	61.7	71.8	78.5	66.3
MolFM (2023)	0.1B	72.9	77.2	76.0	78.8	83.9	77.8
Uni-Mol (2023)	0.4B	72.9	79.6	**81.7**	**80.8**	**85.7**	80.1
MolBridge (2025)	0.1B	**77.6**	**84.7**	76.8	77.8	84.5	**80.3**
LLM-based Generalists							
Galatica (2022)	130B	66.1	68.9	-	74.5	61.7	-
InstructMol (2025)	7B	70.0	74.7	74.6	68.9	82.3	74.1
HIGHT (2025)	7B	61.8	67.4	51.1	63.3	77.1	64.1
UniMoT (2025)	7B	71.4	76.4	**76.0**	78.5	83.7	77.2
KnowMol (2025)	7B	69.2	68.7	61.6	**81.8**	69.2	70.1
StructMol	1B	**72.8**	**81.9**	69.7	81.0	**88.2**	**78.7**

Table 3: Results for molecular property classification tasks (ROC-AUC) on MoleculeNet benchmark. **Bold** and underlined mark the best and second-best scores.

achieves structural similarity scores that are comparable to, or better than, those of models incorporating richer structural modalities. This further validates that substructure-aware supervision is effective in enabling the generation of consistent and structurally precise molecules.

4.3 Molecule Property Prediction

To examine whether StructMol can capture structure–property relationships, we evaluate its performance on a diverse set of molecular property prediction tasks. Specifically, we conduct regression tasks using the QM9 dataset provided by Mol-Instructions, as well as property classification tasks from the MoleculeNet benchmark. The results are reported in Tables 2 and 3, respectively. Notably, the classification tasks involve unseen data that are not included in SubMol-Instructions, allowing us to assess the generalization of the learned molecular knowledge.

In molecular property regression tasks, we observe trends similar to those found in molecular generation. Despite its relatively small model size, our approach achieves an average improvement of at least 22% over prior baselines. These results suggest that explicitly learning the relations defined in SubMol-Instructions helps the model better understand molecular properties and capture their associations with underlying molecular structures.

In molecular property classification tasks, StructMol outperforms existing molecule LLMs on average, surpassing UniMoT by more than 1.5%p. Moreover, on the HIV and BACE benchmarks, our model achieves higher performance than specialist models that are pre-trained on large-scale molecular datasets, indicating that explicit behavior-level supervision can help narrow the gap between specialist and generalist models.[7]

Method	# P	B-2↑	B-4↑	R-1↑	R-2↑	R-L↑	M↑
Specialists							
MolT5 (2022)	0.7B	0.59	0.50	0.65	0.51	0.59	0.61
MolXPT (2023)	0.3B	0.59	0.50	0.66	0.51	0.59	0.62
Atomas (2025)	0.2B	0.63	0.55	0.69	0.56	0.63	-
MolBridge-Gen (2025)	0.2B	**0.67**	**0.60**	**0.72**	**0.60**	**0.67**	**0.69**
LLM-based Generalists							
Mol-Instruction (2023)	7B	0.24	0.17	0.33	0.20	0.28	0.27
InstructMol (2025)	7B	0.47	0.37	0.56	0.39	0.50	0.50
HIGHT (2025)	7B	0.50	0.40	0.57	0.39	0.50	0.52
Omni-Mol (2025)	2B	0.52	0.44	0.60	0.44	0.54	0.57
StructMol	1B	**0.59**	**0.52**	**0.64**	**0.53**	**0.60**	**0.60**

Table 4: Results of molecule captioning task on the ChEBI-20 test set. **Bold** indicates the best performance.

4.4 Molecule Captioning

Table 4 presents the results of the molecule captioning task on the ChEBI-20 test set. During fine-tuning, we additionally incorporated molecule–text pairs from the pre-training stage, adopting the filtering protocol of Edwards et al. (2021) to restrict captions to a maximum of 20 words. We compare StructMol with both specialist models and LLM-based generalist models using standard captioning metrics.

Among generalist models, StructMol consistently outperforms prior baselines, despite using only 1D molecular representations. In addition, while specialist models still have higher performance, StructMol can successfully narrow the gap between the two approaches. This indicates that explicitly learning the detailed structure-description relationships is more important than merely scaling model size or incorporating additional modalities for molecular understanding.

4.5 Ablation Study

To analyze the impact of the components from StructMol on fine-grained molecular understanding, we conduct an ablation study focusing on dual-view molecule alignment and tuning on

[7] Further details on the differences between specialist and generalist models are provided in Appendix D.

Method	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Forward Reaction Prediction							
Llama-3.2-1B-Instruct	0.723	0.982	6.095	0.879	0.938	0.860	1.000
+ Dual-view Molecule Alignment	0.739	0.983	5.864	0.887	0.939	0.865	1.000
+ SubMol-Instructions	**0.893**	**0.989**	**1.492**	**0.964**	**0.981**	**0.959**	1.000
Retrosynthesis							
Llama-3.2-1B-Instruct	0.526	0.955	10.403	0.831	0.892	0.792	1.000
+ Dual-view Molecule Alignment	0.559	0.959	9.280	0.852	0.909	0.817	1.000
+ SubMol-Instructions	**0.639**	**0.964**	**6.546**	**0.897**	**0.932**	**0.869**	1.000

Table 5: Ablation study of StructMol on Forward Reaction Prediction and Retrosynthesis test set.

SubMol-Instructions. Table 5 reports the results of sequentially adding each component to the Llama-3.2-1B-Instruct backbone on Forward Reaction Prediction and Retrosynthesis. Introducing dual-view molecule alignment yields consistent but modest improvements across both tasks, with an average gain of 2.7%. This suggests that aligning complementary 1D molecular representations helps stabilize molecular understanding and provides a better foundation for downstream learning, as further analyzed in Appendix C.

Most importantly, incorporating our newly proposed dataset leads to substantial performance improvements, with an average gain of 14.9% over the previous setting. This indicates that explicitly learning substructure-level functional behaviors plays a dominant role in enhancing fine-grained molecular understanding, enabling the model to generate molecules that are more accurate at both the string and structural levels.

4.6 Visualization of The Embedding Space

To examine whether StructMol encodes fine-grained structure–property signals beyond surface-level patterns, we sample molecules from the QM9 dataset and visualize their representations using t-SNE (Maaten and Hinton, 2008). We extract molecule representations from the model of §3.2.2 without any fine-tuning on property prediction tasks, and color each molecule according to its property value.

As shown in Figure 3, molecules sharing the same substructures tend to form distinct clusters. Notably, these clusters are not determined solely by structural similarity but also reflect variations in property values. For example, although the blue and red molecules in the figure share an identical substructure, they are assigned to different clusters. This suggests that the model captures property-sensitive variations within a shared scaffold by accounting for contextual substructure-level behaviors learned through SubMol-Instructions. Further discussions are in Appendix F.

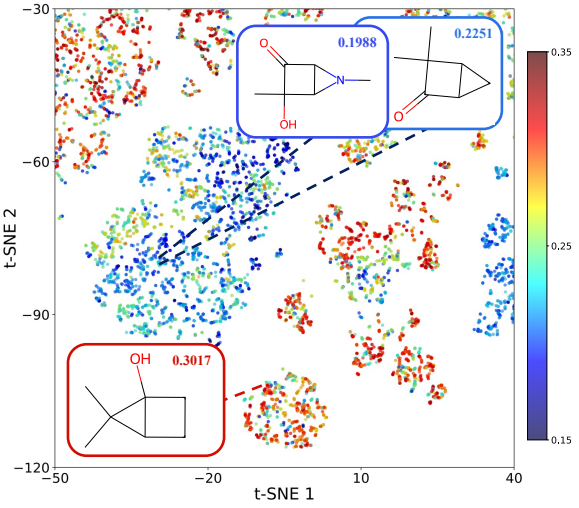


Figure 3: Visualization of StructMol embeddings for molecules sampled from QM9 dataset (HOMO-LUMO gap). Property values are represented using a color gradient. Full image is provided in Figure 5.

5 Conclusion

We have introduced SubMol-Instructions, a substructure-aware instruction-tuning dataset providing explicit supervision on molecular substructure behaviors across diverse chemical tasks. Building on this dataset, we have proposed StructMol, which leverages a dual-view strategy combining SMILES and SELFIES to construct a coherent global molecular representation while capturing substructure-level cues. Across a range of molecular tasks, StructMol achieved consistent improvements and outperformed strong baselines, despite relying solely on 1D representations and a compact model size. Overall, our results suggest that explicitly learning substructure behaviors is crucial for enhancing fine-grained molecular understanding.

Limitations

We demonstrate that explicitly modeling the functional behaviors of substructures across diverse tasks can effectively improve the molecular understanding of LLM. Nevertheless, several limitations remain.

Limited relation coverage. Substructures can exhibit a wide range of functional behaviors beyond the relations defined in this work. As an initial attempt to explicitly model that information, we focus on a limited set of simple and interpretable relations, and show that even these coarse signals can substantially improve molecular understanding. We believe that extending this framework to incorporate richer and more fine-grained relational categories would further enhance a model's chemical reasoning ability.

Limited scalability. Due to computational constraints, our experiments are restricted to models at the scale of approximately 1B parameters. Prior work (Hu et al., 2025) has shown that larger models tend to exhibit stronger molecular understanding and achieve higher performance across tasks. Exploring the scalability of our approach with larger backbone models, therefore, remains an important direction for future work.

References

Mikhail Andronov, Varvara Voinarovska, Natalia Andronova, Michael Wand, Djork-Arné Clevert, and Jürgen Schmidhuber. 2023. Reagent prediction with a molecular transformer improves reaction data quality. *Chemical Science*, 14(12):3235–3246.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

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*, pages 354–379, Abu Dhabi, UAE. Association for Computational Linguistics.

He Cao, Yanjun Shao, Zhiyuan Liu, Zijing Liu, Xiangru Tang, Yuan Yao, and Yu Li. 2024. Presto: Progressive pretraining enhances synthetic chemistry outcomes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10197–10224.

Yongqiang Chen, Quanming Yao, Juzheng Zhang, James Cheng, and Yatao Bian. 2025. Hierarchical graph tokenization for molecule-language alignment. In *Forty-second International Conference on Machine Learning*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. 2019. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377.

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.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(5):1273–1280.

Carl Edwards, Chi Han, Gawon Lee, Thao Nguyen, Sara Szymkuć, Chetan Kumar Prasad, Bowen Jin, Jiawei Han, Ying Diao, Ge Liu, Hao Peng, Bartosz A. Grzybowski, Martin D. Burke, and Heng Ji. 2025. mclm: A modular chemical language model that generates functional and makeable molecules. *Preprint*, arXiv:2505.12565.

Carl Edwards, Tuan 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*, pages 375–413, Abu Dhabi, United Arab Emirates. 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.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*.

Shikun Feng, Lixin Yang, Yanwen Huang, Yuyan Ni, Weiying Ma, and Yanyan Lan. 2023. Unimap: universal smiles-graph representation learning. *arXiv preprint arXiv:2310.14216*.

Shuhan Guo, Yatao Bian, Ruibing Wang, Nan Yin, Zhen Wang, and Quanming Yao. 2025. Unified molecule-text language model with discrete token representation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9205–9213.

Chengxin Hu, Hao Li, Yihe Yuan, Zezheng Song, Chenyang Zhao, and Haixin Wang. 2025. Omni-mol: Multitask molecular model for any-to-any modalities. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. 2025. Mol-llama: Towards general understanding of molecules in large molecular language model. *arXiv preprint arXiv:2502.13449*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2021. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. 2020. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024.

Tu Le, V Chandana Epa, Frank R Burden, and David A Winkler. 2012. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical reviews*, 112(5):2889–2919.

Miguelangel Leon, Yuriy Perezhohin, Fernando Peres, Aleš Popovič, and Mauro Castelli. 2024. Comparing smiles and selfies tokenization for enhanced chemical language modeling. *Scientific Reports*, 14(1):25016.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xuan Lin, Qingrui Liu, Hongxin Xiang, Daojian Zeng, and Xiangxiang Zeng. 2025. Enhancing chemical reaction and retrosynthesis prediction with large language model and dual-task learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 7554–7562. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023. 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.

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*, 62(6):1376–1387.

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

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Frederic P Miller, Agnes F Vandome, and John McBrewster. 2009. Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Hyuntae Park, Yeachan Kim, and SangKeun Lee. 2025. Bridging the gap between molecule and textual descriptions via substructure-aware alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23470–23490.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J Kim. 2024a. Llamo: Large language model-based molecular graph assistant. *Advances in Neural Information Processing Systems*, 37:131972–132000.

Jun-Hyung Park, Hyuntae Park, Yeachan Kim, Woosang Lim, and SangKeun Lee. 2024b. Moleco: Molecular contrastive learning with chemical language models for molecular property prediction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–420.

Qizhi Pei, Rui Yan, Kaiyuan Gao, Jinhua Zhu, and Lijun Wu. 2025. 3d-molt5: Leveraging discrete structural information for molecule-text modeling. In *The Thirteenth International Conference on Learning Representations*.

Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. 2019. *Deep Learning for the Life Sciences*. O'Reilly Media. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

David Rogers and Mathew Hahn. 2010a. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

David Rogers and Mathew Hahn. 2010b. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. Get your atoms in order - an open-source implementation of a novel and robust molecular canonicalization algorithm. *J. Chem. Inf. Model.*, 55(10):2111–2120.

Chufan Shi, Yixuan Su, Cheng Yang, Yujiu Yang, and Deng Cai. 2023. Specialist or generalist? instruction tuning for specific NLP tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15336–15348, Singapore. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Jin-Mao Wei, Xiao-Jie Yuan, Qing-Hua Hu, and Shu-Qin Wang. 2010. A novel measure for evaluating classifiers. *Expert Systems with Applications*, 37(5):3799–3809.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

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.

Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and 1 others. 2023. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1):2585.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and 1 others. 2019. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760.

Zaifei Yang, Hong Chang, Ruibing Hou, Shiguang Shan, and Xilin Chen. 2025. Knowmol: Advancing molecular large language models with multi-level chemical knowledge. *arXiv preprint arXiv:2510.19484*.

Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024a. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391*.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. 2024b. Multimodal molecular pretraining via modality blending. In *The Twelfth International Conference on Learning Representations*.

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*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*.

Appendix

A Implementation Details

Our backbone LLM is Llama-3.2-1B-Instruct (Dubey et al., 2024)[8]. For efficient training of `StructMol`, we employ LoRA (Hu et al., 2022) and explore various configurations to determine optimal hyperparameters. We use LoRA ranks of 64 and 128. The learning rate is set to 1×10^{-4} with a warmup proportion of 0.075. The maximum sequence length is set to 1200 for both training and evaluation. We train the dual-view molecular alignment step (Step 1) for 3 epochs, the `SubMol-Instructions` phase (Step 2) for 6 epochs, and fine-tune on each downstream task for 10 epochs. For downstream tasks, we use a batch size of 64, and for all other phases, a batch size of 128. During training, the model is required to generate both SMILES and SELFIES representations of the molecule simultaneously. All training is conducted using LLaMA-Factory (Zheng et al., 2024).

B Further Details of Substructural Categories

As introduced in Section 3.1, we define explicit relationships between molecular substructures and task-specific entities to operationalize substructural cues across different chemical tasks.

For chemical reaction tasks, including forward reaction prediction and retrosynthesis, task entities correspond to chemical equations and their associated products. Substructures are extracted from both reactants and products, and their roles are assigned based on occurrence patterns across the reaction. Substructures appearing in both sides are labeled as Preserved, those appearing only in the reactants are labeled as Consumed, and those appearing only in the products are labeled as Emergent. In this category, the tail entity is represented by the chemical equation.

For molecule captioning and description-guided molecule generation, we construct relationships based on structural inclusion. Following prior work showing that simple inclusion-based signals suffice for learning fine-grained local alignment (Park et al., 2025), we extract substructures from molecules and chemical keyphrases from textual descriptions. We then establish both intermodal inclusion relationships, such as substructure–description and keyphrase–molecule relations. Chemical keyphrases are extracted using ChemDataExtractor.

For property prediction tasks, we explicitly model structural factors that contribute to specific molecular properties (Xiong et al., 2019). In this setting, the tail entity corresponds to scalar property values, including HOMO, LUMO, and the HOMO-LUMO gap, and substructures that contribute to each property are labeled as Enrolled.

In addition, we establish intra-modal inclusion relationships between each molecule and its extracted substructures, explicitly capturing substructure–molecule associations.

Relation Type	# QA Pairs	# Substructures
Included (intra-modal)	912,011	223,332
Included (inter-modal)	204,570	160,398
Consumed	243,298	17,119
Preserved	234,247	19,794
Emergent	210,472	18,713
Enrolled	6,432	71,170
Total	1,810,030	510,526

Table 6: Distribution of QA pairs and substructures derived from BRICS decomposition in our dataset.

C Impact of Dual-view Molecular Alignment

To evaluate the effectiveness of our dual-view molecular alignment, we measure model performance after the alignment stage, before any instruction tuning or downstream finetuning. We conduct this analysis using the Llama-3.2-1B-Instruct backbone and train it on two tasks where structural understanding is most critical: forward reaction prediction and retrosynthesis. As shown in Table 7, incorporating both SMILES and SELFIES during alignment consistently improves exact-match accuracy, BLEU, Levenshtein distance, and fragment-level similarity metrics (RDK, MACCS, Morgan). These results indicate that dual-view alignment enables the model to internalize more faithful structural cues, resulting in generated molecules that more closely match the ground-truth structures.

D Specialist & Generalist

A central challenge in molecular modeling lies in balancing the broad applicability of **LLM-Based Generalist Models** with the high precision of **Specialist Models**. While specialist architectures excel in specific tasks due to intensive pre-training,

[8] meta-llama/Llama-3.2-1B-Instruct

Method	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Forward Reaction Prediction							
SELFIES-Caption	0.674	0.979	7.771	0.853	0.921	0.827	1.000
SMILES-SELFIES-Caption	0.739	0.983	5.864	0.887	0.939	0.865	1.000
Retrosynthesis							
SELFIES-Caption	0.516	0.956	10.819	0.831	0.895	0.792	1.000
SMILES-SELFIES-Caption	0.559	0.959	9.280	0.852	0.909	0.817	1.000

Table 7: Analysis results of our dual-view approach on Forward Reaction Prediction and Retrosynthesis test set.

generalist models provide a unified framework for diverse applications (Shi et al., 2023). Following established protocols (Cao et al., 2025; Chen et al., 2025; Yang et al., 2025), we benchmark both categories to contextualize StructMol's performance. Our results demonstrate that StructMol successfully defies the conventional trade-off: it consistently surpasses LLM-based generalists and achieves parity with highly optimized specialist baselines.

E Analysis on Filtered Triples

To assess the effectiveness of our graph embedding–based noisy sample filtering method, we analyze the triples removed during preprocessing. We first train the backbone aligned model for 10 epochs using only the filtered triples, and then fine-tune it on forward reaction prediction and retrosynthesis tasks. As shown in Table 8, the model trained solely on the filtered samples collapses and yields substantially degraded performance. In contrast, when we randomly sample the same number of triples from the remaining data and train the model under identical settings, the backbone performance improves. This result suggests that the filtered triples indeed contain harmful noise that hinders molecular understanding.

To further characterize which samples are removed, we conduct a qualitative analysis and present representative examples in Figure 4. We find that a large portion of the filtered triples arise from erroneous relations caused by incorrect substructure decomposition by RDKit, consistent with issues reported in prior work (Park et al., 2025). Notably, inclusion and enrolled relations are most frequently removed. As illustrated in examples (1) and (2), some triples appear to express valid inclusion relations at a superficial level, as the fragment seems to describe part of the molecule. However, closer inspection reveals that these cases correspond to semantic correspondence or functional effect relations rather than true structural inclusion. Such spurious relations are therefore identified as noisy and filtered out by our method. Overall, these observations confirm that our filtering strategy effectively removes erroneous supervision, leading to higher-quality training triples.

F Further Discussions on Visualization

For the visualization, we analyze the molecular embeddings learned by StructMol using the QM9 dataset (HOMO-LUMO gap) curated from Mol-Instructions (Fang et al., 2023). Molecules are divided into five groups based on the distribution of property values, and 10,000 molecules are sampled from each group. We extract molecular representations from StructMol without fine-tuning on property prediction tasks and visualize them using t-SNE. Each molecule is colored according to its property value.

As shown in Figure 5, the embedding space exhibits clustering patterns. Several clusters correspond to recurring structural motifs and are associated with similar ranges of property values. For instance, region (A) mainly consists of molecules with a shared heterocyclic motif and relatively low values around 0.16, while region (B) contains molecules sharing a cyclopropane substructure with higher values above 0.3.

We also observe cases where molecules with similar scaffolds are separated into distinct clusters due to differences in their property values. In region (C), molecules sharing a bridged bicyclic structure are split into multiple clusters, indicating that the representation space is not organized solely by structural similarity. Overall, the visualization suggests that StructMol representations reflect both substructural patterns and property-dependent variations.

Method	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Forward Reaction Prediction							
Llama-3.2-1B-Instruct	0.723	0.982	6.095	0.879	0.938	0.860	1.000
w/ Filtered Samples	0.024	0.776	26.906	0.364	0.570	0.309	1.000
w/ Remain Samples	0.846	2.545	2.545	0.947	0.971	0.936	1.000
Retrosynthesis							
Llama-3.2-1B-Instruct	0.526	0.955	10.403	0.831	0.892	0.792	1.000
w/ Filtered Samples	0.000	0.810	29.579	0.376	0.539	0.322	1.000
w/ Remain Samples	0.576	0.959	7.935	0.878	0.918	0.845	1.000

Table 8: Analysis of our filtering method on Forward Reaction Prediction and Retrosynthesis tasks.

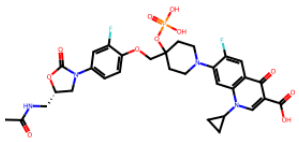
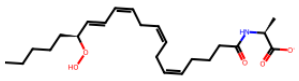
	Head	Relation	Tail
1	Aminoacyl	Included (Inter-modal) (Reason: The tail may affects aminoacyl-tRNA synthetases, instead of being structurally included in the tail.)	 CC(=O)NC[C@H]1CN(c2ccc(OCC3(OP(=O)(O)O)CCN(c4cc5c(cc4F)c(=O)c(C(=O)O)cn5C4CC4)CC3)c(F)c2)C(=O)O1
2	N-acyl-L-alpha-amino acid	Included (Inter-modal) (Reason: The tail corresponds to the head, rather than to a head included in the tail.	 CCCCC[C@@H](/C=C/C=C\C/C=C\C/C=C\CCCC(=O)N[C@@H](C)C(=O)[O-])OO

Figure 4: Examples of filtered chemical knowledge triples during the generation of `SubMol-Instructions`.

Relation Category	Instruction Template
Included (intra-modal)	Decompose the molecule into fragments.\n`{tail}` Assemble the original molecule from these fragments.\n`{head}`
Included (inter-modal)	Generate cross-modal fragments that are included within the following molecule or description.\n`{tail}`
Preserved	Generate fragments that would likely be preserved during the transformation of the following reaction.\n`{tail}`
Consumed	Generate fragments that would likely be consumed in the following reaction.\n`{tail}`
Emergent	Generate fragments that could newly emerge as a result of the following reaction.\n`{tail}`
Enrolled (HOMO)	Generate fragments that are enrolled in determining the following HOMO energy value.\n`{tail}`
Enrolled (LUMO)	Generate fragments that affect the following LUMO energy value.\n`{tail}`
Enrolled (HOMO-LUMO Gap)	Generate fragments that influence the following HOMO-LUMO gap value.\n`{tail}`

Table 9: Examples of instruction templates.

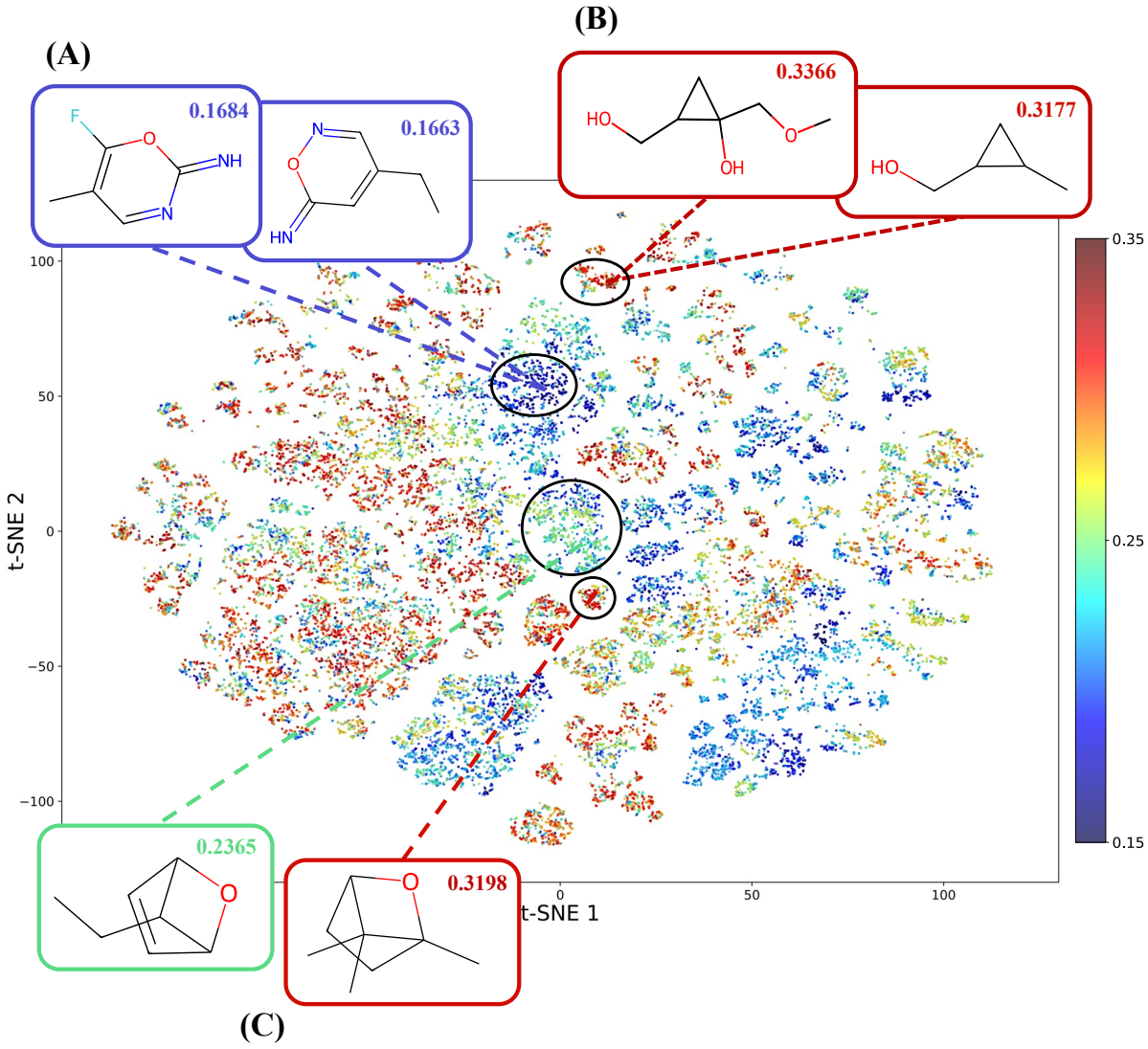


Figure 5: t-SNE visualization of `StructMol` embeddings.