Structural Reasoning Improves Molecular Understanding of LLM

Yunhui Jang
KAIST
yunhuijang@kaist.ac.kr

Jaehyung Kim
Yonsei University
jaehyungk@yonsei.ac.kr

Sungsoo Ahn
KAIST
sungsoo.ahn@kaist.ac.kr

Abstract

Recently, large language models (LLMs) have shown significant progress, approaching human perception levels. In this work, we demonstrate that despite these advances, LLMs still struggle to reason using molecular structural information. This gap is critical because many molecular properties, including functional groups, depend heavily on such structural details. To address this limitation, we propose an approach that sketches molecular structures for reasoning. Specifically, we introduce **M**olecular **S**tructural **R**easoning (**MSR**) framework to enhance the understanding of LLMs by explicitly incorporating the key structural features. We present two frameworks for scenarios where the target molecule is known or unknown. We verify that our MSR improves molecular understanding through extensive experiments.

1 Introduction

Large language models (LLMs; Touvron et al., 2023; OpenAI and et al., 2024; Raffel et al., 2020) have demonstrated remarkable performance across various tasks. To leverage their potential in chemistry, several prior works (Edwards et al., 2022; Christofidellis et al., 2023a; Fang et al., 2024; Pei et al., 2023) have proposed chemical LLMs (i.e., specialized LLMs pre-trained on both natural language and molecular representations) for molecular tasks such as molecule captioning, description-based molecule generation (Edwards et al., 2022), and retrosynthesis (Fang et al., 2024).

However, chemical LLMs still struggle to fully understand the molecular structure (Ganeeva et al., 2024; White et al., 2023). This is critical since structure-based reasoning plays an important role in many molecular tasks. For instance, chemists often consider a molecule toxic if it contains a phenol group, as phenoxyl radicals can form and interact with biological membranes (Hansch et al., 2000). This becomes even more evident in real-world applications of LLMs. As demonstrated in Figure 1a, injecting accurate structural information into the model can slightly improve its ability to generate correct molecules. This highlights the importance of explicitly incorporating structural reasoning into LLMs.

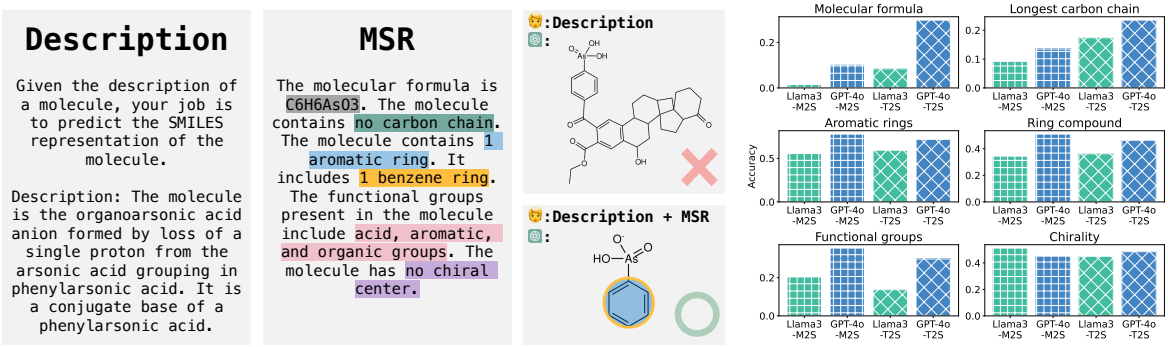
To address this aspect, we consider a framework for LLMs to first reason about the molecular structure for molecular tasks, similar to how LLMs improve arithmetic and commonsense tasks through intermediate reasoning steps (Wei et al., 2022; Kojima et al., 2022). A naïve approach is to prompt LLMs to infer the structural information before generating an answer. However, we find this to be ineffective since even the state-of-the-art LLMs (OpenAI and et al., 2024; Touvron et al., 2023) struggle to accurately capture essential molecular structures, as described in Figure 1b and Section 2.

In this paper, we propose MSR, a simple yet general framework for **M**olecular **S**tructural **R**easoning that progressively sketches the structural features of molecules to improve molecular task performance. To this end, we identify key structural elements crucial for the reasoning of LLMs to solve molecular tasks. Moreover, we propose fine-tuning procedures that employ external tools to identify the molecular structural information.

In particular, our frameworks to fine-tune LLMs for molecular structural reasoning are designed for both molecular and non-molecular inputs. A framework consists of *reasoning module* and an *answering module*. The reasoning module generates structural information to enhance the understanding of the molecule. The answering module generates the final answer based on the original input and the output of the reasoning module. The overall frameworks are illustrated in Figure 4.

To be specific, we consider two types of reasoning framework, inspired by how humans generally form knowledge via analysis and synthesis (Kant, 1899). On the one hand, *analysis* refers

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21016–21036



(a) Incorporating MSR improves GPT-4o in molecule generation. (b) LLMs' capability of structural inference.

Figure 1: **Overview of LLMs with structural information**. (a) Each color in MSR represents a structural component. The top molecule is incorrectly generated using only the description while the bottom is correctly generated by incorporating the description and MSR. (b) Despite the importance of structural information, even recent LLMs struggle to accurately infer key structural details from molecular representations such as SMILES (Molecule-to-structure; M2S) or given descriptions (Text-to-structure; T2S).

to breaking down complex information into fundamental components for better understanding. In molecular tasks, analytic reasoning applies when the molecule is provided as input, allowing the model to decompose its structure for meaningful insights. Specifically, we utilize external tools like RDKit (Landrum et al., 2024) as the reasoning module to precisely extract structural information from the molecule.

On the other hand, *synthesis* constructs a whole from its constituent parts. This aligns with molecular tasks where the molecule must be generated from non-molecular input, e.g., textual description, requiring the model to infer structural information and reconstruct the entire molecule. In detail, for synthetic reasoning, we fine-tune the LLMs as the reasoning module that generates MSR (Ho et al., 2023; Fu et al., 2023; Magister et al., 2023) based on the given input. We additionally incorporate a novel *matching-ratio-based rejection sampling* into the answering module, to ensure that the generated molecule aligns with MSR, using the external tools for validation.

We empirically show that incorporating MSR into chemical LLMs (Edwards et al., 2022; Christofidellis et al., 2023a) and general LLMs (Touvron et al., 2023; OpenAI and et al., 2024) both lead to consistent performance improvements in three molecular tasks: molecule-to-text, retrosynthesis, and text-to-molecule. In particular, chemical LLMs outperform the considered baselines when combined with our MSR framework. In summary, our contributions are as follows:

- We identify and evaluate the limits of LLMs in inferring molecular structural information.
- We propose MSR, a simple yet broadly applicable molecular reasoning framework that progressively sketches molecular structures.
- We introduce an analytic reasoning for MSR when the input molecule is given, leveraging external tools for structural identification.
- We develop a synthetic reasoning for MSR when the molecule is in the desired output, incorporating fine-tuning for the reasoning module and a novel matching ratio-based rejection sampling procedure for the answering module.
- We validate the effectiveness of MSR by demonstrating consistent performance improvements across chemical and general LLMs.

2 Recent large language models do not understand structural information

Here, we demonstrate that even the recent LLMs, i.e., GPT-4o (OpenAI and et al., 2024) and Llama3-8B-Instruct (Touvron et al., 2023), fail to infer important structural information from the given inputs, such as molecular representations (e.g., SMILES (Weininger, 1988)) and the text descriptions (Edwards et al., 2021). Notably, such tasks can be considered straightforward for individuals with a bachelor's degree in chemistry.

Our analysis is inspired by how chemists reason about the structure to analyze a molecule. They progressively identify the molecular structure, starting with primary elements like rings and long carbon chains before identifying finer details such as

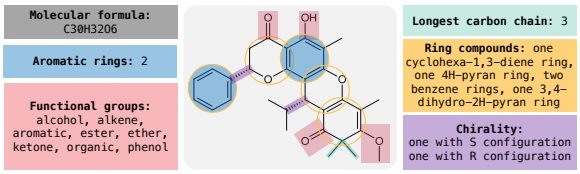


Figure 2: **The six key structural information: molecular formula, longest carbon chain length, aromatic rings, ring compounds, functional groups, and chirality.** The same color indicates the structural information and the corresponding component of the molecule.

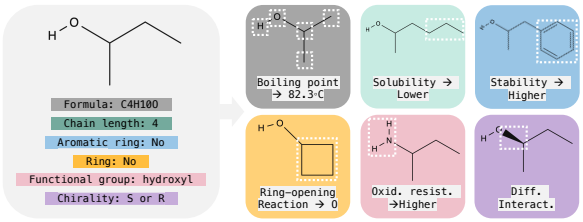


Figure 3: **Illustration of the importance of structural information.** This example shows how replacing each structural information (dashed box) alters the molecule. Colors correspond to the structural elements in Figure 2.

functional groups. Reflecting on this behavior, we define six significant key structural elements for chemical reasoning as illustrated in Figure 2. In detail, these six key structural components include (1) molecular formula, (2) longest carbon chain length, (3) aromatic rings, (4) ring compounds, (5) functional groups, and (6) chirality.

Molecular formula. The molecular formula provides essential information about a molecule's composition, specifying the number and type of atoms present. This information is critical because, for example, it directly determines the molecular weight. To illustrate, although 2-Butanol ($C_4H_{10}O$) and 2-Propanol (C_3H_8O) are composed of the same type of atoms, i.e., carbon, hydrogen, and oxygen, their differing molecular formulas result in distinct molecular weights (74.1g/mol for 2-Butanol and 60.1g/mol for 2-Propanol). These differences lead to the change in boiling points, 99.4°C and 82.3°C, respectively, as shown in the gray part of Figure 3.

Longest carbon chain. The longest carbon chain (excluding atoms in ring systems) forms the molecular backbone where functional groups are attached. The length of this chain significantly influences properties like solubility. For example, extending the carbon chain of 2-Butanol from four to six carbons creates 2-Hexanol, which exhibits reduced solubility. This is illustrated in the green section of Figure 3.

Aromatic rings. Aromatic rings (e.g., benzene and pyridine) play a critical role in determining the stability and electronic properties. For instance, adding a benzene ring to 2-Butanol yields 1-Phenyl-2-Propanol, which has enhanced stability and greater oxidation resistance. This transformation is shown in the blue section of Figure 3.

Ring compounds. Similar to the longest carbon chain, ring structures often serve as the backbone where functional groups are attached. The ring system significantly affects molecular behavior and reactions. For example, although 2-Butanol and Cyclobutanol share the same number of carbons and oxygen, the ring in Cyclobutanol introduces a tendency toward ring-opening reactions, as depicted in the yellow section of Figure 3.

Functional groups. Functional groups, e.g., hydroxyl, amino, ester, etc., play a pivotal role in determining the chemical reactivity. For example, alcohols with a hydroxyl group (-OH) are prone to oxidize more while the molecules with an amino group ($-NH_2$) are generally resistant to oxidation under mild conditions. A single replacement of a hydroxyl (-OH) group in 2-Butanol with an amino ($-NH_2$) group leads to 2-Butanamine, which has increased oxidation resistance, as described in the red part of Figure 3.

Chiral centers. Chirality refers to the stereochemical property of a molecule that makes it non-superimposable on its mirror image, leading to different chemical behaviors. The chirality is determined by the chiral centers and their configurations, i.e., R- and S-configuration [1], which describe the spatial arrangement of the groups around the chiral centers. This leads to different interactions between other molecules with chirality. For instance, (R)-2-Butanol and (S)-2-Butanol may interact differently with other chiral substances. This is described in the purple part of Figure 3.

Despite their significance, we observe that even recent LLMs often fail to accurately infer crucial structural details from the molecule or their text description. Specifically, as shown in Figure 1b, both GPT-4o and LlaMA3-8B-Instruct fail to capture the structural information accurately when the molecule is provided (Molecule-to-structure; M2S) or the description is provided (Text-to-structure; T2S). Notably, we provide detailed experimental settings and prompts in Appendix A.1.

[1]The names of R and S come from the Latin word *Rectus* and *Sinister*, which means right and left, respectively.

Component	Expression	Description
Molecular formula	*The molecular formula is* $X_1N_1 \cdots X_MN_M$.	X_m: m-th atom type N_m: # of m-th atoms
Longest carbon chain length	*The longest carbon chain is N carbons long.*	N: the length of the longest carbon chain
Aromatic rings	*The molecule contains N aromatic rings.*	N: # of aromatic rings
Ring compounds	*It includes N_1 X_1 rings, $\cdots$, $N_M X_M$ ring(s).*	N_m: # of m-th ring X_m: IUPAC name of m-th ring
Functional groups	*The functional groups include $X_1, \cdots$, and X_M group.*	X_m: the name of functional group
Chirality	*The molecule has N chiral centers: N_S with S configuration and N_R with R configuration.*	N_S: # of chiral centers of S config. N_R: # of chiral centers of R config. $N = N_S + N_R$

Table 1: **The description of each component of MSR.**

First, when provided with a molecule (M2S), both GPT-4o and LlaMA3-8B-Instruct struggled to achieve high accuracy. Even in their best-performing case, counting the number of aromatic rings, their accuracies remained low, at approximately 50% and 75%, respectively. Similarly, when given a detailed text description (T2S), both models still failed to achieve high accuracy. This implies that LLMs struggle to fully understand the molecular structures, whether presented as molecular representations or text descriptions. These observations highlight the potential benefits of explicitly incorporating structural reasoning to enhance molecular comprehension.

3 MSR: Molecular Structural Reasoning

Here, we present our framework for enhancing LLMs' understanding of molecules through **M**olecular **S**tructural **R**easoning (MSR). MSR incorporates six key structural elements as reasoning for LLMs, following a two-stage process (Zhang et al., 2024): a reasoning stage and an answering stage. First, a *reasoning module* generates MSR, providing supplementary structural information for a better understanding of the molecule. Next, an *answering module* generates the final output using the input augmented with the generated MSR. The framework is illustrated in Figure 4.

To address various tasks, we categorize the MSR framework based on whether the molecule is provided as input (*analytic reasoning*) or must be inferred as output (*synthetic reasoning*). In summary, for analytic reasoning, one decomposes complex molecules into fundamental structural components to better understand their structure. For synthetic reasoning, one constructs the entire molecule from its constituent structural components.

3.1 Overview of MSR

Here, we introduce MSR, a molecular structural reasoning framework that enhances language models' understanding of molecules. Each component of MSR corresponds to one of the six structural elements introduced in Section 2. The expression and corresponding description of the reasoning for each structural component in MSR are provided in Table 1. Additionally, a concrete example illustrating MSR is shown in Figure 2.

3.2 Analytic reasoning

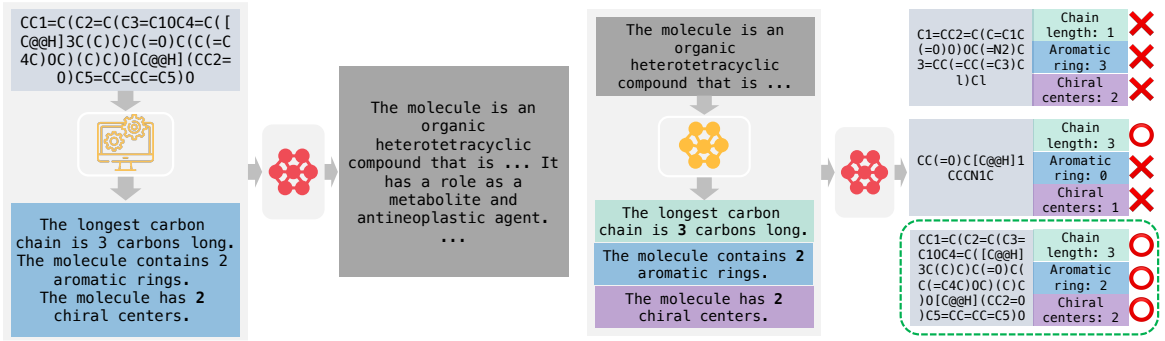
In MSR, analytic reasoning refers to decomposing a given input molecule into smaller structural components for enhanced comprehension. When the input molecule is available, one can utilize a deterministic reasoning module for its decomposition. Our approach integrates MSR by (1) employing external tools like RDKit (Landrum et al., 2024) to extract precise structural information as a reasoning module, and (2) fine-tuning the answering module LLM with the generated rationale as an additional input. The overall workflow is described in Figure 4a.

Reasoning module. In the analytic reasoning scenario, we employ external tools to extract precise structural information from the input molecule. This process eliminates uncertainty, as the structural information is deterministic for a given molecule. Next, this information serves as MSR, which guides the answering module.

Answering module. With the molecule and its corresponding MSR as input, we fine-tune the chemical LLMs to generate the desired output of the molecule. In our experiments, we mainly consider MolT5 (Edwards et al., 2022) and ChemT5 (Christofidellis et al., 2023a), as the answering module.

3.3 Synthetic reasoning

Synthetic reasoning refers to composing structural information to construct an entire molecule. When the input molecule is unavailable, the relevant structural information must first be inferred before generating the final molecule. To address this, we fine-tune a reasoning module to generate MSR, which is then attached to the input and utilized by the answering module to generate the final molecule, as illustrated in Figure 4b.



(a) Analytic reasoning (b) Synthetic reasoning

Figure 4: **Overview of MSR fine-tuning framework.** Analytic reasoning applies when the input molecule is available, while synthetic reasoning applies when it is not. Light gray boxes denote the molecules (SMILES); gray boxes denote related description; colored boxes represent MSR. The yellow and the red modules perform reasoning and answering, respectively. In (a), yellow module indicates an external tool. In (b), colors indicate MSR and the corresponding structural elements; here, the third molecule is chosen after *matching ratio-based rejection sampling* according to its highest matching ratio (3/3).

	BL.2	BL.4	RO.1	RO.2	RO.L	ME.
Baselines (without reasoning)						
Meditron-7B	0.792	0.576	0.797	0.602	0.575	0.757
Mol2Lang-VLM	0.777	0.563	0.786	0.591	0.565	0.741
BioT5+	0.798	0.579	0.812	0.617	0.584	0.777
Chemical LLMs (fine-tuning)						
MolT5-small	0.709	0.512	0.745	0.558	0.544	0.701
+MSR	0.780	0.565	0.807	0.613	0.585	0.757
MolT5-base	0.738	0.535	0.750	0.559	0.539	0.718
+MSR	0.805	0.592	0.864	0.677	0.642	0.822
MolT5-large	0.769	0.556	0.777	0.580	0.557	0.743
+MSR	**0.832**	**0.622**	**0.914**	**0.743**	**0.691**	**0.878**

Table 2: **Molecule-to-text performance for L+M val.** BL., RO., and ME. stand for BLEU, ROUGE, and METEOR, respectively.

Reasoning module. We fine-tune the chemical LLMs to generate MSR similar to prior works that fine-tune LLMs to generate chain-of-thoughts (Ho et al., 2023; Fu et al., 2023; Magister et al., 2023). Unlike analytic reasoning, where external tools can precisely extract structural information, the reasoning module in synthetic reasoning must infer this information from the input.

Notably, we selectively retain only reliable structural components before incorporating them into the answering module. One considers the component to be reliable if it achieves sufficiently high reasoning accuracy across the entire dataset. This selection process also leverages the deterministic nature of structural information, allowing a quantitative evaluation of the reasoning module's capability in generating each type of rationale, as presented in Section 4.3.

Answering module. Similar to the analytic reasoning scenario, we fine-tune chemical LLMs to generate an appropriate molecule from the input and its corresponding MSR. To further enhance structural consistency between the generated molecule and MSR, we propose a *matching ratio-based rejection sampling* method.

Specifically, the model first generates k candidate molecules using beam search. Then, for each candidate, one computes the matching ratio, which quantifies the consistency between the generated molecule's structural components and those in MSR. Finally, the molecule with the highest matching ratio is selected as the final output, ensuring the consistency between the rationale and the generated answer.

Again, this approach leverages the deterministic nature of the molecular structural information, allowing us to easily obtain the information with external tools and compare them between the rationale and the generated molecule. Notably, the search process differs from prior works (Wang et al., 2023; Xi et al., 2023; Sun et al., 2024) that search over rationale-answer pairs since our method focuses on searching the answer that coincides with a given rationale.

4 Experiments

In this section, we present our experiments on two frameworks: analytic reasoning and synthetic reasoning. For the analytic reasoning framework, we consider molecule-to-text and retrosynthesis tasks. For the synthetic reasoning framework, we address

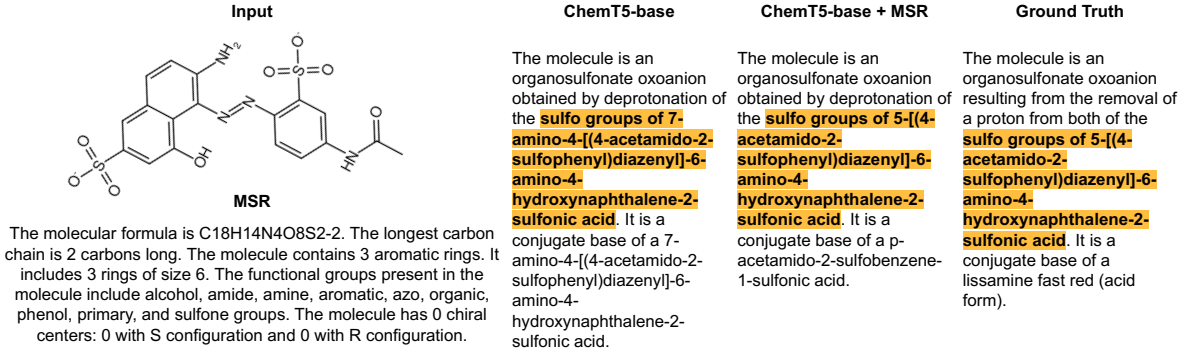


Figure 5: **An example of generated samples for molecule-to-text.** We observe that MSR improves the accuracy of detailed molecular information (highlighted in yellow). We provide more examples in Appendix B.

	BL.2	BL.4	RO.1	RO.2	RO.L	ME.
Baselines (without reasoning)						
T5-base	0.511	0.423	0.607	0.451	0.550	0.539
MolXPT	0.594	0.505	0.660	0.511	0.597	0.626
BioT5	0.635	0.556	0.692	0.559	0.633	0.656
Chemical LLMs (fine-tuning)						
MolT5-base	0.540	0.457	0.634	0.485	0.578	0.569
+MSR	0.592	0.507	0.667	0.523	0.606	0.619
MolT5-large	0.594	0.508	0.654	0.510	0.594	0.614
+MSR	**0.645**	**0.567**	**0.699**	**0.568**	**0.639**	**0.666**
ChemT5-small	0.553	0.462	0.633	0.481	0.574	0.583
+MSR	0.601	0.513	0.664	0.519	0.603	0.624
ChemT5-base	0.580	0.490	0.647	0.498	0.586	0.604
+MSR	0.639	0.560	0.687	0.553	0.626	0.657
General LLMs (without fine-tuning)						
Llama3	0.211	0.117	0.367	0.183	0.308	0.257
+MSR	0.259	0.158	0.401	0.208	0.324	0.341
GPT-4o	0.232	0.128	0.389	0.183	0.307	0.291
+MSR	0.286	0.174	0.405	0.199	0.313	0.341

Table 3: **Molecule-to-text performance for ChEBI-20.**

	BL.2	BL.4	RO.1	RO.2	RO.L	ME.
General LLM (without fine-tuning)						
Mol-Instruct.	0.217	0.143	0.337	0.196	0.291	0.254
+MSR	**0.347**	**0.275**	**0.601**	**0.518**	**0.593**	**0.520**

Table 4: **Molecule-to-text performance for Mol-Instructions.**

the text-to-molecule task. For clarity, in all tables, the teal color indicates improvements over the vanilla model, and the best results are highlighted in **bold**.

4.1 Analytic reasoning: Molecule-to-text

The molecule-to-text task aims to generate a precise and informative textual description that accurately represents the given molecule.

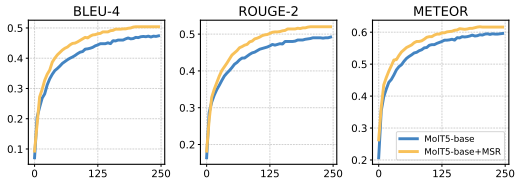


Figure 6: **Fast performance improvement with MSR.**

Dataset. We employ three datasets for the molecule-to-text task: (1) the recent L+M dataset (Edwards et al., 2024), (2) the widely used ChEBI-20 dataset (Edwards et al., 2021), and the (3) Mol-instructions dataset (Fang et al., 2024). Each dataset consists of 182,331, 33,010, and 298,319 pairs of SMILES (or SELFIES) and their text descriptions, respectively. We use the same splits used in prior works.

Baselines. We evaluate the performance of MSR with chemical and general LLMs. On the one hand, we employed two chemical LLMs: MolT5 (Edwards et al., 2022) and Text+Chem T5 (ChemT5; Christofidellis et al., 2023a). On the other hand, we employed three general LLMs: Llama3-8B-Instruct (Touvron et al., 2023), GPT-4o (OpenAI and et al., 2024)[2], and Mol-Instructions (Fang et al., 2024). Additionally, we include T5 (Raffel et al., 2020), MolXPT (Liu et al., 2023), BioT5 (Pei et al., 2023), Meditron-7B (Chen et al., 2023b), Mol2Lang-VLM (Tran et al., 2024), and BioT5+ (Pei et al., 2024) as baselines to compare absolute performance.

Experimental setup and metrics. Chemical LLMs are trained following the process described in Section 3.2. For general LLMs without any domain-specific instruction tuning (Llama3 and GPT-4o), we cannot guarantee that the generated

[2]We used `gpt-4o-2024-05-13`.

Models	BL.	Ex.	Le. ↓	MA.	RDK	Mo.	Val.
General LLM (without fine-tuning)							
Mol-Instruct.	**0.705**	0.009	31.23	0.283	0.487	0.230	**1.000**
+ MSR	0.502	**0.016**	**31.21**	**0.315**	**0.493**	**0.273**	**1.000**

Table 5: **Retrosynthesis performance for Mol-Instructions.** BL., Ex., and Le. indicate BLEU, Exact, and Levenshtein distance. MA., RDK, and Mo. indicate MACCS, RDK, and Morgan fingerprint metrics. Val. indicates the validity.

descriptions align with our training data. Therefore, we apply 10-shot in-context learning by attaching MSR in the same manner as for chemical LLMs. Additionally, for Mol-Instructions, we prompt with instructions enriched with MSR. Note that we additionally consider the molecular weight and IUPAC name components used by M. Bran et al. (2024), as they slightly improved the performance.

We evaluate the performance by comparing the generated description with the ground truth using six metrics: BLEU2, BLEU4 (Papineni et al., 2002), ROUGE1, ROUGE2, ROUGEL (Banerjee and Lavie, 2005), and METEOR (Banerjee and Lavie, 2005). We provide detailed experimental settings and prompts in Appendix A.2.

Results. We report the results in Table 2, Table 3, and Table 4. We observe that adding MSR consistently improves performance across both chemical and general LLMs. Notably, in Table 3, ChemT5-base+MSR achieves performance comparable to BioT5 (without MSR), despite BioT5 being pretrained on a larger dataset. Furthermore, Table 2 shows that integrating MSR with MolT5-base or MolT5-large yields superior performance compared to baseline models. We provide examples of generated samples in Figure 5 and Figure 16. In addition, our method exhibits faster performance improvement, as illustrated in Figure 6.

4.2 Analytic reasoning: Retrosynthesis

The retrosynthesis task aims to generate the corresponding set of reactant molecular representations based on a given product molecular representation.

Dataset and baselines. We employ the dataset and the model used by Mol-instructions (Fang et al., 2024). The dataset consists of 129,684 product and reactant pairs.

Experimental setup and metrics. As the input molecule (i.e., product) is given for the retrosynthesis task, we follow the framework proposed in Section 3.2. The performance is evaluated by comparing the generated molecules with the ground truth with eight metrics: SMILES comparison metrics (BLEU, Exact, and Levenshtein distance (Miller et al., 2009)), fingerprint similarity metrics (MACCS FTS (Durant et al., 2002), RDK FTS (Schneider et al., 2015), and Morgan FTS (Rogers and Hahn, 2010)), a molecular distribution metric (Fréchet ChemNet Distance (FCD) (Preuer et al., 2018)), and the validity of the molecule.

Models	Fo.	Ch.	Ar.	Ri.	Fu.	Ch.	We.	Na.
Chemical LLMs (MSR *fine-tuning) - L+M*								
MolT5-small	0.048	0.235	0.783	0.781	0.849	0.647	0.418	0.248
MolT5-base	0.426	0.527	0.825	0.813	0.889	0.807	0.615	0.309
MolT5-large	0.221	0.317	0.820	0.809	0.872	0.691	0.529	0.576
Chemical LLMs (MSR *fine-tuning) - MolT5*								
MolT5-base	0.458	0.922	0.926	0.930	0.957	0.798	0.606	0.512
ChemT5-small	0.447	0.920	0.930	0.926	0.954	0.788	0.634	0.495
ChemT5-base	0.475	0.925	0.931	0.930	0.960	0.799	0.641	0.525
General LLMs (MSR *few-shot learning) - MolT5*								
Llama3	0.084	0.174	0.593	0.362	0.137	0.450	0.435	0.015
GPT-4o	0.298	0.235	0.718	0.464	0.298	0.485	0.728	0.040

Table 6: **Reasoning accuracy for each structural information.** Fo., Ch., Ar., Ri., Fu., Ch., We., Na., stand for molecular formula, longest carbon chain length, aromatic rings, ring compounds, functional groups, chiarlity, molecular weight, and IUPAC name, respectively.

Results. We report the results in Table 5, showing that incorporating MSR improves performance across all metrics except BLEU. This highlights its effectiveness in enhancing complex tasks. Notably, while we report BLEU for consistency with prior work, it is less critical than other metrics, as it evaluates string-based accuracy rather than molecular structure alignment.

4.3 Synthetic reasoning: Text-to-molecule

The text-to-molecule task is the inverse of molecule-to-text, aiming to generate a molecular representation based on a given textual description.

Dataset. We employ two datasets for the text-to-molecule task: (1) L+M (Edwards et al., 2024) and (2) ChEBI-20 (Edwards et al., 2021). We followed the same settings used in Section 4.1.

Baselines. Two popular chemical LLMs, including MolT5 (Edwards et al., 2022) and ChemT5 (Christofidellis et al., 2023a), serve as our baselines. Notably, we exclude general LLMs

	BL.	Ex.	Le. ↓	MA.	RDK	Mo.	FCD↓	Val.
Baselines (without reasoning)								
Meditron-7B	0.694	0.010	46.49	0.772	0.693	0.501	2.46	0.996
Lang2Mol-Diff	0.543	0.000	55.87	0.606	0.332	0.328	38.09	**1.000**
BioT5+	**0.731**	0.010	41.47	0.781	0.709	0.515	3.29	**1.000**
Chemical LLMs (fine-tuning)								
MolT5-small	0.566	0.000	56.34	0.642	0.581	0.374	NaN	0.805
+MSR	0.730	0.002	41.15	0.798	0.712	0.514	2.82	0.995
MolT5-base	0.684	0.000	44.79	0.760	0.652	0.475	NaN	**1.000**
+MSR	0.706	0.052	40.18	0.825	0.762	0.548	**1.45**	0.997
MolT5-large	0.564	0.000	55.40	0.757	0.650	0.395	17.50	0.994
+MSR	0.710	**0.111**	**39.54**	**0.837**	**0.783**	**0.560**	1.54	0.999

Table 7: **Text-to-molecule performance for L+M val.**

	BL.	Ex.	Le. ↓	MA.	RDK	Mo.	FCD↓	Val.
Baselines (without reasoning)								
T5-base	0.762	0.069	24.95	0.731	0.605	0.545	2.48	0.660
MolXPT	-	0.215	-	0.859	0.757	0.667	0.45	0.983
BioT5	0.867	0.413	15.10	0.886	0.801	0.734	0.43	**1.000**
Chemical LLMs (fine-tuning)								
MolT5-base	0.769	0.081	24.46	0.721	0.588	0.529	2.18	0.772
+MSR	0.863	0.385	13.91	0.918	0.843	0.783	0.29	0.983
MolT5-large	0.854	0.311	16.07	0.834	0.746	0.684	1.20	0.905
+MSR	**0.886**	0.391	12.98	0.906	0.822	0.765	0.35	0.947
ChemT5-small	0.739	0.157	28.54	0.859	0.736	0.660	0.07	0.776
+MSR	0.874	0.381	13.22	0.918	0.845	0.787	0.29	0.976
ChemT5-base	0.750	0.212	27.39	0.874	0.767	0.697	**0.06**	0.792
+MSR	0.878	**0.421**	**12.76**	**0.924**	**0.856**	**0.804**	0.26	0.982

Table 8: **Text-to-molecule performance for ChEBI-20.**

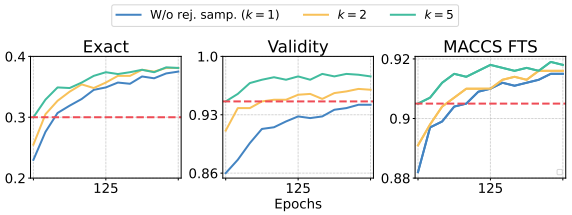


Figure 7: **Impact of k in rejection sampling.** Dotted lines indicate the initial performance of $k = 5$.

from this evaluation due to their insufficient reasoning accuracy as shown in Table 6. In detail, their low accuracy implies that their reasoning cannot guide the answer appropriately, even in a few-shot learning setting. For completeness, we provide the results for general LLMs in Appendix B.3. Additional baselines are consistent with those in Section 4.1 other than Lang2Mol-Diff (Nguyen et al., 2024).

Experimental setup and metrics. We follow the framework proposed in Section 3.3. We provide detailed experimental settings and prompts in Appendix A.3. The performance is evaluated using the same metrics described in Section 4.2.

Reasoning accuracy. We first measure the reasoning accuracy to filter out low-accuracy components that may misguide the answer. The detailed computation process is in Appendix A.3. The reasoning accuracies are provided in Table 6. Our results show that our fine-tuned reasoning modules exhibit superior accuracy compared to larger general LLMs, underscoring their ability to understand molecular structures effectively. However, they still struggle with certain structural elements, such as molecular formula, molecular weight, and IUPAC name, with additional challenges in carbon chain length and chirality in the L+M dataset. Consequently, we exclude these components.

Results. The results are reported in Table 7 and Table 8. Incorporating MSR into the molecular description always improved performance. In particular, integrating MSR into the ChemT5-base achieves state-of-the-art performance compared to the recent baselines, validating its efficacy. Surprisingly, our MSR even improves the performance of smaller models beyond that of the vanilla larger models, e.g., MolT5-base+MSR showed superior performance to MolT5-large. We provide examples of generated samples in Appendix B.1.

4.4 Ablation study

We perform ablation studies on matching ratio-based rejection sampling and each structural component. Here, we utilize ChemT5-small on the ChEBI-20 dataset. Due to limited space, additional ablation study results, including a comparison with ChemCrow (M. Bran et al., 2024), prior work on the reasoning for chemistry tasks, and extra structural component, are provided in Appendix B.4.

Matching ratio-based rejection sampling. We discuss the efficacy of matching ratio-based rejection sampling and the impact of the number of samples k in text-to-molecule. We compare the results of without ($k = 1$) and with the rejection sampling ($k \in \{2, 5\}$). As demonstrated in Figure 7, the rejection sampling improves performance by encouraging the output to follow the MSR. Notably, increasing k beyond 5 does not further improve performance, implying that $k = 5$ is sufficient.

Structural component. To verify the effectiveness of each component, we evaluated the performance of molecule-to-text using each structural information component individually. We provide the results in Figure 8. Incorporating each single component resulted in better performance compared to the baseline model without any reasoning. Notably, combining all the proposed structural elements yielded the best results, validating the effectiveness of our comprehensive approach.

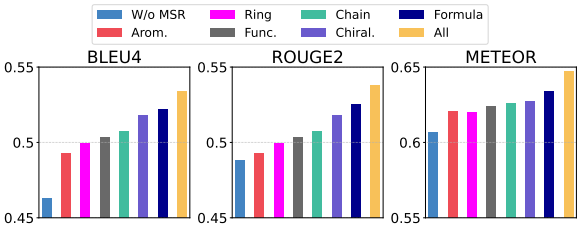


Figure 8: **Impact of each structural component.**

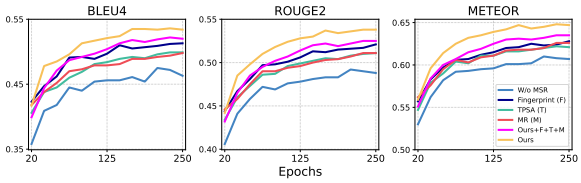


Figure 9: **The impact of additional molecular descriptors.**

Additional molecular descriptors. In addition to the proposed six structural components, we conducted experiments using three more advanced molecular descriptors: the Morgan fingerprint and two electronic properties—topological polar surface area (TPSA) and molar refractivity (MR). Specifically, the Morgan fingerprint encodes local substructures within a specified radius; TPSA represents the sum of the surface areas of all polar atoms and their attached hydrogen atoms; and MR quantifies the total polarizability of a molecule.

To verify the effectiveness of each additional descriptor, we evaluated the performance of molecule captioning using ChemT5-small. We provide the results in Figure 9. We observed that incorporating all three additional descriptors together did not further improve the performance of MSR, although applying each additional descriptor individually improved performance. This validates the importance of structural information and the sufficiency of our proposed structural components.

5 Related work

Large language models for chemistry. General LLMs often struggle to solve basic chemistry problems and molecular tasks (White et al., 2023; Castro Nascimento and Pimentel, 2023; Guo et al., 2023). To address this issue, prior works have introduced chemical LLMs by pre-training models on molecule-related texts (Edwards et al., 2022; Christofidellis et al., 2023b; Liu et al., 2023; Pei et al., 2023), through instruction tuning (Fang et al., 2024; Cao et al., 2023), and using retrieval-based in-context learning (Li et al., 2024). Our work focuses on reasoning processes that are broadly applicable to these chemical and general LLMs.

Reasoning of LLMs. Generating intermediate reasoning before arriving at a final answer (Wei et al., 2022; Kojima et al., 2022) improves the overall quality of generated answers. However, the ability to perform complex reasoning remains limited to huge models (>100B parameters).

To address this challenge, various approaches have been introduced to distill knowledge from larger language models to smaller ones (<10B). Specifically, Ho et al. (2023); Fu et al. (2023); Magister et al. (2023) employed the larger models as teacher models to generate rationales for fine-tuning smaller student models. Nevertheless, even recent LLMs struggle to generate appropriate rationales that demonstrate a correct understanding of molecular structures (as described in Figure 1a and Section 2), restricting the efficacy of LLMs in generating rationales for molecular tasks.

Reasoning for chemistry. Recently, a few works have extended the reasoning of LLMs to address chemistry problems. For instance, Ouyang et al. (2024) proposed employing the program-of-thoughts (PoT; Chen et al., 2023a) to handle chemical question-answering tasks. Additionally, Jin et al. (2024) presented the protein chain-of-thought (ProCoT) to replicate the signaling pathways in the protein-protein interaction (PPI) problem. Despite these advances, none of these works are generally applicable to various molecular tasks. We note that M. Bran et al. (2024) provided a reasoning approach comparable to ours, but their rationales are less focused on molecular structures, e.g., rationales based on tools like *LitSearch/WebSearch*, *PatentCheck*, *ReactionPlanner*, and *SMILES2Price*. Moreover, it shows limited performance improvement in molecule generation and molecule captioning tasks, as observed in Appendix B.4.

6 Conclusion

We introduced MSR, a molecular structural reasoning framework that enhances LLMs' understanding of molecules by explicitly incorporating key structural features. Our investigation revealed recent LLMs' limitations in inferring structural information, emphasizing the need for explicit reasoning. Fine-tuning chemical LLMs with MSR led to consistent improvements across three molecular tasks, highlighting the effectiveness of domain-specific models for molecular reasoning.

Broader impacts

Our work contributes to the development of more interpretable and reliable models for molecular applications. By incorporating explicit molecular reasoning, our framework has the potential to enhance molecular understanding and improve decision-making in areas such as drug discovery, materials science, and chemical synthesis. However, as with any AI-driven molecular generation system, there are potential risks and ethical concerns. For instance, the generation of harmful or toxic compounds poses significant safety challenges. Additionally, over-reliance on AI-generated molecular reasoning without expert validation could lead to unintended consequences in scientific and industrial applications.

Limitations

One limitation of MSR is its reliance on the accuracy of structural information in synthetic reasoning. While external tools like RDKit provide precise structural information for molecule-forward reasoning, errors in molecule-backward reasoning (where structural features must be inferred) could degrade performance. However, appropriate filtering based on reasoning accuracy can prevent this to some extent. Additionally, we assume that the given molecular representations are accurate when we extract the structural information. However, real-world data can be noisy or incomplete. Extending MSR to handle uncertain molecular inputs via self-correction remains an open challenge.

Reproducibility

All experimental code related to this paper is available at https://github.com/yunhuijang/MSR. Detailed insights regarding the experiments, encompassing dataset and model specifics, are available in Section 4. For intricate details like hyperparameters, consult Appendix A.

Acknowledgement

This work was partly supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019- II190075, Artificial Intelligence Graduate School Support Program(KAIST)), National Research Foundation of Korea(NRF) grant funded by the Ministry of Science and ICT(MSIT) (No. RS-2022- NR072184), GRDC(Global Research Development Center) Cooperative Hub Program through the National Research Foundation of Korea(NRF) grant funded by the Ministry of Science and ICT(MSIT) (No. RS-2024-00436165), and the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-02304967, AI Star Fellowship(KAIST)).

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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. *Preprint*, arXiv:2311.16208.

Cayque Monteiro Castro Nascimento and AndréSilva Pimentel. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023a. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023b. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6140–6157. PMLR.

Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. 2002. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280.

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, Qingyun Wang, Lawrence Zhao, and Heng Ji. 2024. L+M-24: Building a dataset for Language+Molecules @ ACL 2024. In *Proceedings of the 1st Workshop on Language + Molecules (L+M 2024)*, pages 1–9, Bangkok, Thailand. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR.

Veronika Ganeeva, Andrey Sakhovskiy, Kuzma Khrabrov, Andrey Savchenko, Artur Kadurin, and Elena Tutubalina. 2024. Lost in translation: Chemical language models and the misunderstanding of molecule structures. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12994–13013.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Corwin Hansch, Susan Mckarns, Carr Smith, and David Doolittle. 2000. Comparative qsar evidence for a free-radical mechanism of phenol-induced toxicity. *Chemico-Biological Interactions*, 127:61–72.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. 2024. ProLLM: Protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. In *First Conference on Language Modeling*.

I. Kant. 1899. *Critique of pure reason*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Rachel Walker, Daniel Probst, Kazuya Ujihara, Axel Pahl, guillaume godin, Juuso Lehtivarjo, tadhurst cdd, François Bérenger, and Jonathan Bisson. 2024. rdkit/rdkit: 2024_09_1 (q3 2024) release beta.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, page 1–13.

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.

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.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

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.

Nguyen Doan Hieu Nguyen, Nhat Truong Pham, Duong Thanh Tran, and Balachandran Manavalan. 2024. Lang2mol-diff: A diffusion-based generative model for language-to-molecule translation leveraging SELFIES representation. In *ACL 2024 Workshop Language + Molecules*.

OpenAI and Josh Achiam et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. 2024. Structured chemistry reasoning with large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Jinhua Zhu, and Rui Yan. 2024. Enhanced biot5+ for molecule-text translation: A three-stage approach with data distillation, diverse training, and voting ensemble. In *ACL 2024 Workshop Language + Molecules*.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1123, Singapore. Association for Computational Linguistics.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fréchet chemnet distance: A metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

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. *Journal of chemical information and modeling*, 55(10):2111–2120.

Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. 2024. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4074–4101, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Duong Thanh Tran, Nhat Truong Pham, Nguyen Doan Hieu Nguyen, and Balachandran Manavalan. 2024. Mol2lang-VLM: Vision- and text-guided generative pre-trained language models for advancing molecule captioning through multimodal fusion. In *ACL 2024 Workshop Language + Molecules*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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.

Andrew D. White, Glen M. Hocky, Heta A. Gandhi, Mehrad Ansari, Sam Cox, Geemi P. Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, and Willmor J. Peña Ccoa. 2023. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2:368–376.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Self-Polish: Enhance reasoning in large language models via problem refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11383–11406, Singapore. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

Appendix

Organization The appendix is organized as follows: We first present the experimental details such as hyperparameters and prompts in Appendix A. Then we provide the additional experimental results including the generated samples and additional ablation studies in Appendix B. Next, we described the usage of AI assistants and scientific artifacts in Appendix C and Appendix D, respectively.

A Experimental details

In this section, we provide the details of the experiments. All experimental code related to this paper is available at https://github.com/yunhuijang/MSR and our experiments are based on a single run. Additionally, we used the packages including rouge-score==0.1.2 and nltk==3.8.1.

A.1 Structure information analysis

Here, we describe the detailed settings for the analysis in Section 2. To evaluate the understanding of two recent LLMs: Llama3-8B-Instruct (Touvron et al., 2023) and GPT-4o (OpenAI and et al., 2024), we prompt the LLMs to infer the structural information from the given molecular SMILES string and text description of the molecule.

Prompts given text description of molecules. First, we asked LLMs to infer the structural information from the text description of the molecule, with the prompt described in Figure 10.

Prompts given SMILES string. Next, we asked LLMs to infer the structural information from the SMILES string, with the prompt described in Figure 11.

Prompts for M2S

Head prompt: You are now working as an excellent expert in chemistry and drug discovery.
Given the SMILES representation of a molecule, your job is to predict the structural information of the molecule.
The structural information of the molecule caption includes the molecular formula, the length of the longest carbon chain, the number of aromatic rings, the IUPAC name of all the rings, all the functional groups, the number of chiral centers with S and R configurations each, the molecular weight, the IUPAC name of the molecule.
The functional group and ring IUPAC names should be on the list. The number of chiral centers should also be format {"S": , "R": }.
Your response should only be in the JSON format following {"molecular formula": , "functional group": , "longest carbon chain length": , "aromatic ring": , "ring IUPAC name":, "chiral": {"S": , "R": }, "weight": , "IUPAC name": }.
THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE. DO NOT CHANGE THE JSON KEY NAMES.

Input prompt: Input: [SMILES]

Figure 10: **Prompts for structure information analysis given SMILES string.**

Prompts for T2S

Head prompt: You are now working as an excellent expert in chemistry and drug discovery.
Given the caption of a molecule, your job is to predict the structural information of the molecule.
The molecule caption is a sentence that describes the molecule, which mainly describes the moleculeś structures, properties, and production.
The structural information of the molecule caption includes the molecular formula, the length of the longest carbon chain, the number of aromatic rings, the IUPAC name of all the rings, all the functional groups, the number of chiral centers with S and R configurations each, the molecular weight, the IUPAC name of the molecule.
The functional group and ring IUPAC names should be on the list. The number of chiral centers should also be format {"S": , "R": }.
Your response should only be in the JSON format following {"molecular formula": , "functional group": , "longest carbon chain length": , "aromatic ring": , "ring IUPAC name":, "chiral": {"S": , "R": }, "weight": , "IUPAC name": }.
THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE. DO NOT CHANGE THE JSON KEY NAMES.

Input prompt: Input: [Description]

Figure 11: **Prompts for structure information analysis given text description.**

A.2 Molecule-to-text

Here, we describe the detailed settings for the experiments of molecule-to-text in Section 4.1. Note that we used four A100-80GB GPUs.

Hyperparameters. The hyperparameters for the specialist models are provided in Table 9. Note that MolT5-large was not trained for the same epochs as the other models due to limited resource.

Prompts. The prompts used for the generalist models are described in Figure 12. We primarily followed the prompt presented by (Li et al., 2024).

Hyperparameter	MolT5-base	MolT5-large	ChemT5-small	ChemT5-base
Batch size	8	4	8	8
Learning rate	$2e^{-4}$	$2e^{-4}$	$6e^{-4}$	$6e^{-4}$
Epochs	250	220	250	250
Warmup ratio	0	0	0.1	0.1
Weight decay	0.01	0.01	0	0
Lr scheduler	linear	linear	linear	linear

Table 9: **Hyperparameters for molecule captioning.**

A.3 Text-to-molecule

Here, we described the detailed settings for the experiments of text-to-molecule in Section 3.3. Note that we also used four A100-80GB GPUs.

Hyperparameters. The hyperparameters for the reasoning and answering module for the specialist models are provided in Table 10 and Table 11, respectively. Note that MolT5-large was not trained for the same number of epochs as the other models due to limited time constraints.

Hyperparameter	MolT5-base	ChemT5-small	ChemT5-base
Batch size	8	8	8
Learning rate	$1e^{-3}$	$6e^{-4}$	$6e^{-4}$
Epochs	250	250	250
Warmup ratio	0.1	0	0
Weight decay	0	0	0
Lr scheduler	cosine	linear	linear

Table 10: **Hyperparameters for the reasoning module of text-based molecule generation.**

Hyperparameter	MolT5-base	MolT5-large	ChemT5-small	ChemT5-base
Batch size	8	4	8	8
Learning rate	$1e^{-3}$	$1e^{-3}$	$6e^{-4}$	$6e^{-4}$
Epochs	250	140	250	250
Warmup ratio	0.1	0.1	0	0
Weight decay	0	0	0	0
Lr scheduler	cosine	cosine	linear	linear

Table 11: **Hyperparameters for the answering module of text-based molecule generation.**

Reasoning accuracy The accuracies for molecular formula, longest carbon chain length, number of aromatic rings, chirality, and IUPAC names are computed by exact match. The accuracies for ring compounds and functional groups are computed by the ratio of intersection between the set of true and generated CoTs. Lastly, the accuracy for molecular weight is considered correct if the generated weight is within 95% to 105% of the true weight.

Prompts. The prompts used for the generalist models are described in Figure 13. We also primarily followed the prompt presented by (Li et al., 2024).

Prompts for molecule2text

Head prompt: You are now working as an excellent expert in chemistry and drug discovery.
Given the caption of a molecule, your job is to predict the SMILES representation of the molecule.
The molecule caption is a sentence that describes the molecule, which mainly describes the molecule's structures, properties, and production.
You can infer the molecule SMILES representation from the caption.
Before you infer the molecule SMILES representation, YOU SHOULD FIRST GENERATE the molecular formula, the length of the longest carbon chain, the number of aromatic rings, the IUPAC name of all the rings, all the functional groups, the number of chiral centers with S and R configurations each, the molecular weight, the IUPAC name of the molecule.

Example 1: Instruction: Given the caption of a molecule, predict the SMILES representation of the molecule.
Input: [Description][MSR]
Your output should be: {"molecule": <SMILES>}
...
Example k: Instruction: Given the caption of a molecule, predict the SMILES representation of the molecule.
Input: [Description][MSR]
Your output should be: {"molecule": <SMILES>}

You should FIRST generate the structural information following the examples above, and then provide the JSON format of the molecule SMILES based on that.
NOTE THAT THE SMILES REPRESENTATION MUST BE IN THE JSON format above {"molecule": }. THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR JSON. DO NOT CHANGE THE JSON KEY NAME.

Input prompt: Input: [Description]

Figure 12: **Prompts for the generalist models in molecule captioning task.**

Prompts for text2molecule

Head prompt: You are now working as an excellent expert in chemistry and drug discovery.
Given the SMILES representation of a molecule and structural description of the molecule, your job is to predict the caption of the molecule.
The molecule caption is a sentence that describes the molecule, which mainly describes the molecule's structures, properties, and production.

Example 1:
Instruction: Given the SMILES representation of a molecule, predict the caption of the molecule.
Input: [SMILES][MSR]
Your output should be: {"caption": <Description>}
...
Example k:
Instruction: Given the SMILES representation of a molecule, predict the caption of the molecule.
Input: [SMILES][MSR]
Your output should be: {"caption": <Description>}

Your response should only be in the JSON format above; THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

Input prompt: Input: [SMILES]<MSR >

Figure 13: **Prompts for generalist models in text-based molecule generation task.**

A.4 Ablation study

Here, we describe the detailed settings for the ablation study.

Prompts for ChemCrow. The prompts used for ChemCrow (M. Bran et al., 2024) are described in Figure 14 and Figure 15. Notably, it was not able to apply few-shot learning for ChemCrow as it was not applicable as the original prompt proposed in ChemCrow does not include any few-shot setting.

`Prompts for molecule2text with ChemCrow`

Head prompt: Given the SMILES representation of a molecule and structural description of the molecule, your job is to predict the caption of the molecule.
"Final Answer" follows the format: Final Answer: {"caption": }

Input prompt: The SMILES representation of the molecule is as follows: : [SMILES]

Figure 14: **Prompts for molecule captioning with ChemCrow.**

`Prompts for text2mol with ChemCrow`

Head prompt: Given the caption of a molecule, your job is to predict the SMILES representation of the molecule.
The molecule caption is a sentence that describes the molecule, which mainly describes the molecule's structures, properties, and production.
You can infer the molecule SMILES representation from the caption.
"Final Answer" follows the format: Final Answer: {"molecule": }

Input prompt: The caption is as follows: [Description]

Figure 15: **Prompts for text-based molecule generation with ChemCrow.**

B Additional experimental results

In this section, we provide additional experimental results including several concrete examples of generated samples.

B.1 Molecule-to-text

Here, we show the samples of molecule captioning, i.e., generated text descriptions of given molecules in Figure 16. Notably, we show the generated samples from base-sized models for fair comparison.

B.2 Retrosynthesis

Here, we show the samples of retrosynthesis, i.e., generated reactants of given product in Figure 17.

B.3 Text-to-molecule

Here, we show the samples of text-based molecule generation, i.e., generated molecules for the given text description in Figure 18. Notably, we show the generated samples from base-sized models for fair comparison.

Additionally, we provide the results of generalist models in Table 12. Note that it is natural to show no consistent enhancement for generalist models as they lack reasoning ability as shown in Table 6.

Models	**BL.**	**Ex.**	**Le.** ↓	**MA.**	**RDK**	**Mo.**	**FCD**↓	**Val.**
Generalists (10-shot learning)								
Llama3	0.251	0.007	117.30	0.586	0.352	0.276	13.11	0.629
+MSR	0.259	0.008	109.77	0.579	0.279	0.344	4.47	0.669
GPT-4o	0.521	0.079	40.87	0.797	0.496	0.583	3.67	0.881
+MSR	0.509	0.088	41.68	0.783	0.498	0.571	1.57	0.846

Table 12: **Text-based Molecule Generation Performance for generalist models.**

B.4 Ablation study

Comparison to ChemCrow. To validate the efficacy of our MSR, we compare our method with ChemCrow (M. Bran et al., 2024), which has employed CoTs for various chemical tasks. The comparative results are provided in Table 13 and Table 14. One can observe that ChemCrow shows limited performance in both molecule captioning and text-based molecule generation tasks. It is notable that the comparison may not be entirely appropriate, as ChemCrow is primarily designed for practical synthesis tasks, as the reviewer mentioned. Nevertheless, we included comparisons with ChemCrow to provide additional insights, as they share a similar motivation: enriching large language models (LLMs) with a chemistry-aware chain-of-thoughts.

Models	BL.-2	BL.-4	RO.-1	RO.-2	RO.-L	MET.
ChemCrow (GPT-4o)	0.162	0.078	0.299	0.097	0.211	0.225
Ours (GPT-4o)	0.249	0.139	0.386	0.179	0.300	0.303
Ours (ChemT5-base)	0.639	0.560	0.687	0.553	0.626	0.657

Table 13: **Comparison to ChemCrow in molecule captioning.**

Models	**BL.**	**Ex.**	**Le.** ↓	**MA.**	**RDK**	**Mo.**	**FCD**↓	**Val.**
ChemCrow (GPT-4o)	0.306	0.194	56.46	0.772	0.632	0.555	2.31	0.851
Ours (GPT-4o)	0.509	0.088	41.68	0.783	0.498	0.571	1.57	0.846
Ours (ChemT5-base)	0.878	0.421	12.76	0.924	0.856	0.804	0.26	0.982

Table 14: **Comparison to ChemCrow in text-based molecule generation.**

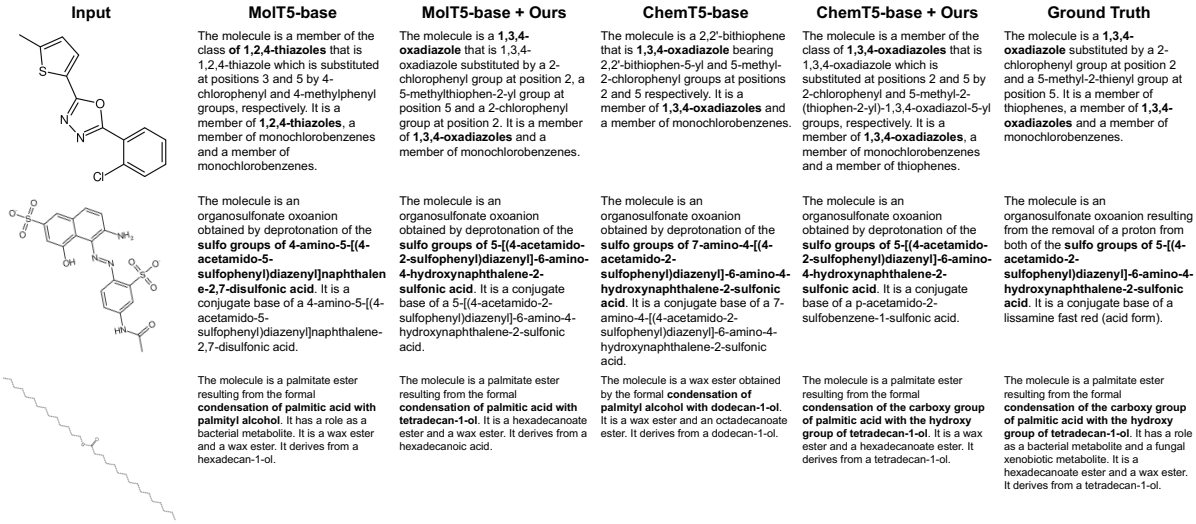


Figure 16: **The generated samples of molecule captioning.**

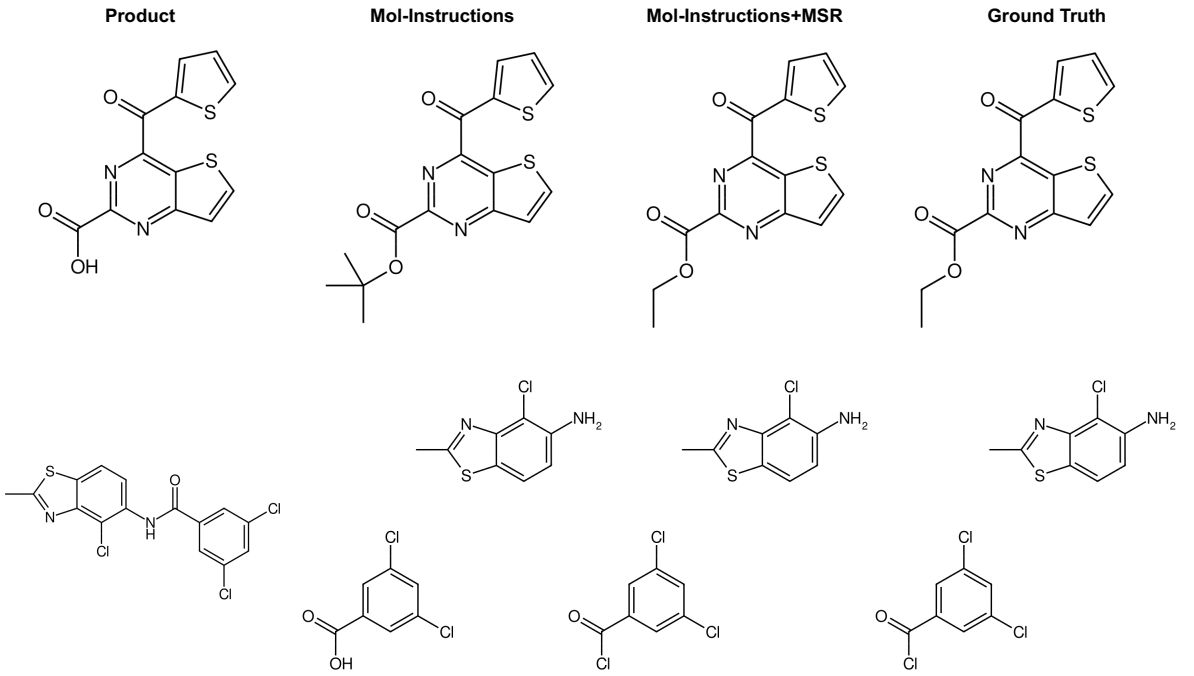


Figure 17: **The generated samples of retrosynthesis.**

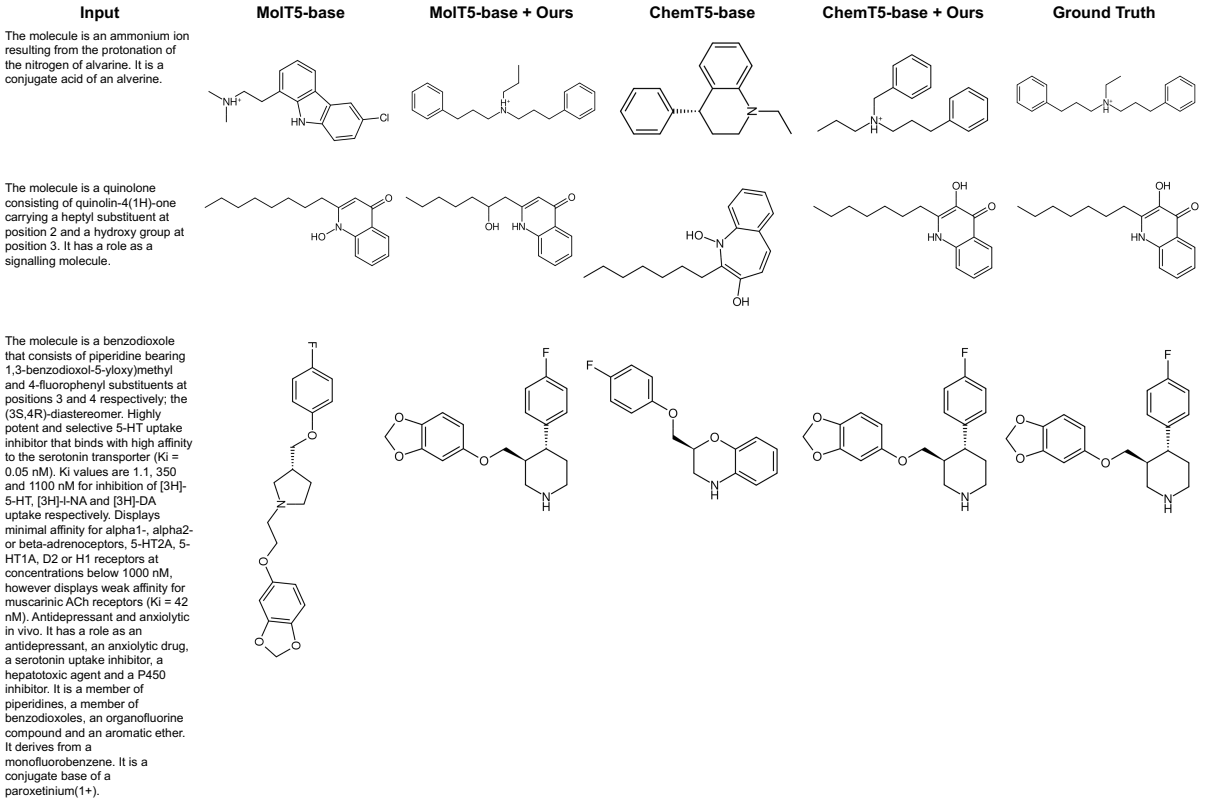


Figure 18: **The generated samples of text-based molecule generation.**

C Usage of AI assistants

In preparing this work, we utilized AI-based writing assistants to refine sentence structure, correct grammatical errors, and enhance readability. These tools were employed only for rephrasing and language improvements, ensuring that the technical content, methodology, and experimental findings remained entirely authored by the researchers. The use of AI assistance was limited to editorial enhancements without influencing the originality or scientific contributions of the paper.

D Scientific Artifacts

The License for artifacts. All datasets and software tools used in this work adhere to their respective licenses. Specifically, we employed publicly available datasets such as ChEBI-20 and L+M under their permitted usage terms. Additionally, external tools like RDKit were used following their open-source license. We release our trained models and code in https://github.com/yunhuijang/MSR under an appropriate open-source license to facilitate reproducibility.

Artifact use consistency with intended use. The datasets and tools utilized in our study were used in accordance with their intended purpose. For example, ChEBI-20 and L+M datasets were originally developed for molecule captioning and generation tasks, aligning with our research objectives. Similarly, RDKit was employed for molecular structure analysis as intended by its developers.

Documentation of artifacts. We provide details in https://github.com/yunhuijang/MSR.