Improving Chemical Understanding of LLMs via SMILES Parsing

Yunhui Jang
KAIST
yunhuijang@kaist.ac.kr

Jaehyung Kim
Yonsei University
jaehyungk@yonsei.ac.kr

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

Abstract

Large language models (LLMs) are increasingly recognized as powerful tools for scientific discovery, particularly in molecular science. A fundamental requirement for these models is the ability to accurately understand molecular structures, commonly encoded in the SMILES representation. However, current LLMs struggle to interpret SMILES, even failing to carry out basic tasks such as counting molecular rings. To address this limitation, we introduce CLEANMOL, a novel framework that formulates SMILES parsing into a suite of clean and deterministic tasks explicitly designed to promote graph-level molecular comprehension. These tasks span from subgraph matching to global graph matching, providing structured supervision aligned with molecular structural properties. We construct a molecular pretraining dataset with adaptive difficulty scoring and pre-train open-source LLMs on these tasks. Our results show that CLEANMOL not only enhances structural comprehension but also achieves the best or competes with the baseline on the Mol-Instructions benchmark.

1 Introduction

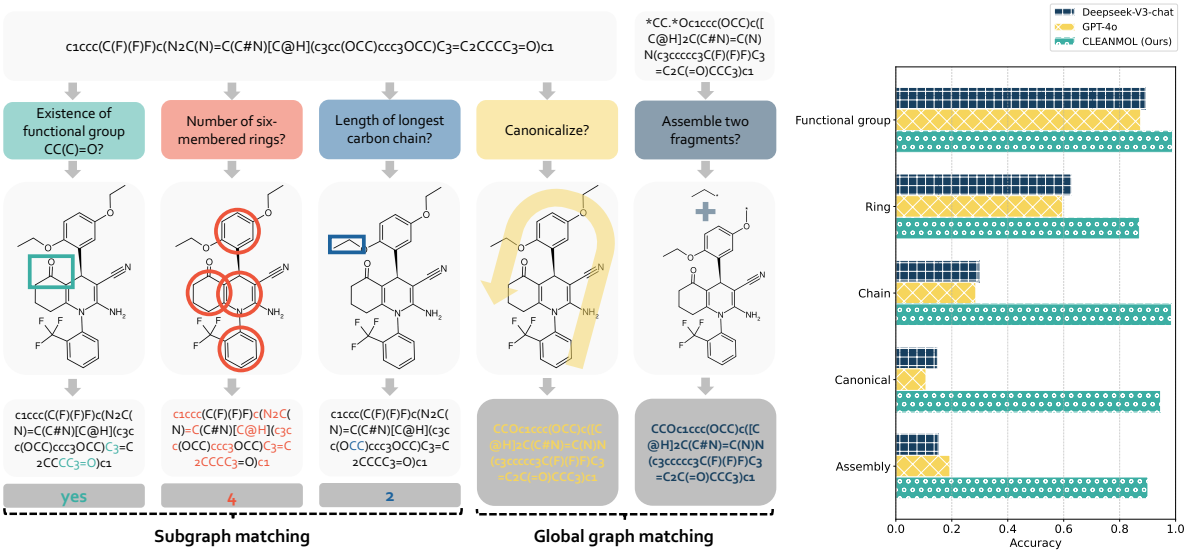
Molecular string representations such as SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020) have become a standard format for applying large language models (LLMs) to chemistry. These one-dimensional strings flatten molecular graphs by traversing atoms and bonds and are syntactically compatible with LLMs (Xia et al., 2025; Taylor et al., 2022; Edwards et al., 2022; Christofidellis et al., 2023a; Pei et al., 2023; Fang et al., 2024). As a result, most molecular LLMs adopt training paradigms from the natural language processing domain, treating molecular strings as sequences of tokens analogous to sentences in natural language.

However, molecular strings follow complex syntactic rules for encoding molecular structures, which LLMs often struggle to interpret. For instance, SMILES grammar includes specific conventions to denote rings and branches—often involving non-contiguous tokens to represent connected substructures. Additionally, SMILES representations must satisfy structural constraints such as proper valency and ring closure. As a result, current LLMs often misinterpret SMILES, which implies a failure to capture the underlying molecule represented by the SMILES string. This is reflected in their inability to perform even basic tasks, such as counting the number of rings or producing consistent outputs for different SMILES strings of the same molecule (Jang et al., 2024; White et al., 2023; Ganeeva et al., 2024). Our experiments revisit such limitations, as shown in Figure 1 and Section 2.2.

One might expect such an understanding would "naturally emerge" from training LLMs on large corpora of SMILES strings for downstream tasks such as molecular generation and retrosynthetic analysis. However, high-quality data is limited and difficult to obtain. Unlike text or image data, which can be gathered at scale via web scrapping, chemical data often require expensive wet lab experiments or simulations for annotation. Although open-source datasets such as USPTO series (Wei et al., 2010; Lu and Zhang, 2022) and MoleculeNet (Wu et al., 2018) exist, their scale remains modest compared to datasets in other domains (Deng et al., 2009; Raffel et al., 2020a; Lozhkov et al., 2024). Consequently, most chemical LLMs often rely on ambiguous and indirect pretraining objectives with non-deterministic and unclear tasks (e.g., masking each token in SMILES and reconstruct them or translation between a molecular string and its description) (Pei et al., 2023; Edwards et al., 2022), or focus on instruction tuning with limited-scale datasets (Fang et al., 2024; Yu et al., 2024).

In response, we propose *SMILES parsing*—a suite of clean, deterministic, and scalable tasks that require models to extract structural information

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15695–15710



(a) **Illustration of SMILES parsing tasks.** (b) **Failure of LLMs on SMILES parsing.**

Figure 1: **Overview of SMILES parsing**. (a) Each column visualizes one of the five SMILES parsing tasks: functional group matching, ring counting, carbon chain length measurement, SMILES canonicalization, and fragment assembly. The highlighted tokens in the SMILES correspond to the substructures involved in each task. (b) Recent LLMs fail for SMILES parsing while the model trained with our CLEANMOL shows improvement.

from molecular strings, as illustrated in Figure 1. We argue that a natural and necessary candidate task for training LLMs to understand the SMILES representation is the extraction of deterministic graph-level information from molecular structures. To address this, we define five SMILES parsing tasks including subgraph matching (e.g., functional group, ring size, and chain length) and global graph matching (e.g., SMILES canonicalization and fragment assembly). Each task provides unambiguous supervision with deterministic answers. Based on these tasks, we construct the CLEANMOL dataset, consisting of 250K molecules annotated via lightweight molecular graph analysis tools such as RDKit (Landrum et al., 2024). Notably, our approach is scalable since the annotations for these tasks do not require any experiment or human annotation, in principle, SMILES parsing can be applied to all the existing molecules in the real world.

To evaluate and demonstrate the benefit of our new CLEANMOL dataset, we also introduce a two-stage training framework: first, the model is pre-trained on the proposed SMILES parsing tasks and then fine-tuned on downstream chemical applications. To enhance data efficiency in the first stage, we propose a task-adaptive data pruning that selects structurally informative molecules and a curriculum learning framework that organizes them from easy to hard order.

We empirically validate our approach by training recent LLM backbones (Grattafiori et al., 2024; Yang et al., 2024) and evaluating them on three downstream tasks from the Mol-Instructions benchmark (Fang et al., 2024), including retrosynthesis, reagent prediction, and forward reaction prediction. Surprisingly, our clean and structure-aware CLEANMOL framework enables the models to achieve state-of-the-art or competitive results on the downstream tasks. This demonstrates that incorporating deterministic structural supervision via SMILES parsing can significantly enhance molecular generation capabilities, even without direct exposure to generation-specific training data.

We summarize our contributions as follows:

- We revisit the limitations of LLMs in interpreting molecular strings, highlighting the structural bottleneck.
- We propose five deterministic and scalable SMILES parsing tasks and introduce the CLEANMOL dataset to bridge the gap between string-level and graph-level molecular understanding of LLMs.
- We design a two-stage training framework incorporating a task-adaptive data pruning and curriculum learning strategy.
- We validate the impact of CLEANMOL by demonstrating a consistent performance improvement across multiple downstream tasks.

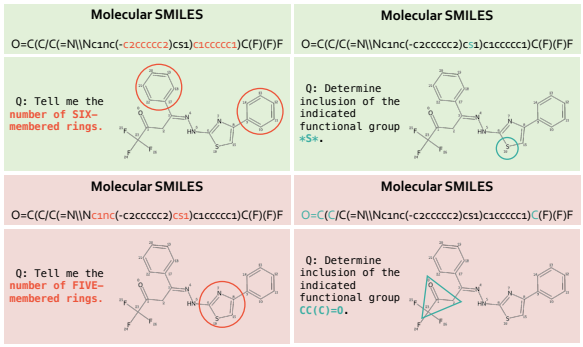


Figure 2: **Complex cases in SMILES parsing.** The top green panels represent relatively simple cases, while the bottom red panels illustrate more complex examples with non-continuous substructures in SMILES. Orange and teal highlights correspond to tasks involving ring counting and functional group matching, respectively.

2 SMILES parsing task

In this section, we introduce five SMILES parsing tasks designed to enhance the mapping between molecular SMILES strings and their corresponding graph structures. We then highlight two key bottlenecks in applying LLMs to molecular tasks: (1) the inability of models to extract structural information from SMILES strings and (2) the lack of high-quality, scalable molecular datasets. To address the first bottleneck, we show that even advanced LLMs such as GPT-4o (OpenAI and et al., 2024) and DeepSeek-V3 (Liu et al., 2024) fail to perform well on simple SMILES parsing tasks, revealing the need for explicit structure-aware supervision. To address the second bottleneck, we explain the limitation of open-source molecular datasets, motivating the need for scalable molecular datasets that can be generated without costly experiments.

2.1 SMILES parsing task description

We define SMILES parsing as a suite of deterministic, scalable, and structure-focused tasks designed to map molecular strings to their corresponding molecular graphs. The tasks fall into two categories—*subgraph matching* and *global graph matching*—as illustrated in Figure 1a. Importantly, all annotations can be generated automatically using open-source chemical tools such as RDKit (Landrum et al., 2024) without any experiment, making the tasks highly scalable. We provide more details in Appendix A.

- **Subgraph matching.** This category includes *functional group matching*, *ring counting*, and *carbon chain length measurement*. Functional group matching determines the presence of a specified functional group. Ring counting identifies the number of rings with specific sizes (e.g., five- or six-membered), and chain length measurement evaluates the length of the longest carbon chain excluding rings. These tasks focus on local subgraphs such as structural motifs, branching, and ring patterns.

- **Global graph matching.** This category consists of *SMILES canonicalization* and *fragment assembly*. Canonicalization involves converting arbitrarily ordered SMILES into a canonical form, which encourages structural invariance to syntactic permutation. Fragment assembly requires the model to combine two SMILES fragments into a single valid molecule, testing its ability to reorganize the global structure from disjoint components.

2.2 Failure of existing LLMs

Although SMILES parsing appears simple from a structural point of view, it poses significant challenges for existing LLMs. Complex cases involving nested rings or hierarchical branching often disrupt token-level patterns, making it difficult for models to resolve SMILES parsing accurately. In detail, as shown in Figure 2, many structural features are represented non-contiguously in SMILES, further complicating the parsing process. Our motivation closely aligns with that of Jang et al. (2024).[1]

We observe that even state-of-the-art general-purpose LLMs, including GPT-4o (OpenAI and et al., 2024) and DeepSeek-V3-Chat (Liu et al., 2024), struggle with SMILES parsing, achieving no more than 60% accuracy across five tasks except for the binary classification (functional group matching), as described in Figure 1b and detailed in Section 4.1. This failure is notable given the strong performance of these models in other domains such as mathematics and code. The inability of these models to handle even basic molecular parsing tasks underscores a critical gap in their structural understanding. It motivates the need for explicit pretraining strategies tailored to molecules.

2.3 Costly high-quality data acquirement

A second challenge lies in acquiring sufficient high-quality training data for molecules. In contrast to textual and visual domains, which benefit from

[1]Unlike Jang et al. (2024), which fine-tunes models directly on structural information and downstream tasks, we pretrain LLMs on SMILES parsing objectives and subsequently fine-tune them for downstream tasks.

SMILES: c1ccc(C(F)(F)F)c(N2C(N)=C(C#N)[C@H](c3cc(OCC)ccc3OCC)C3=C2CCCC3=O)c1

Functional group
Question: Given the SMILES, determine inclusion of the functional group COC.
Answer: Yes
Ring
Question: Calculate the count of SIX-membered rings in the given SMILES string.
Answer: 4
Canonicalization
Question: Give me a canonicalized SMILES that represents the same given molecule.
Answer: CCOc1ccc(OCC)c([C@H]2C(C#N)=C(N)N(c3ccccc3C(F)F)F)C3=C2C(=O)CCC3)c1

Figure 3: **Examples of CLEANMOL dataset.**

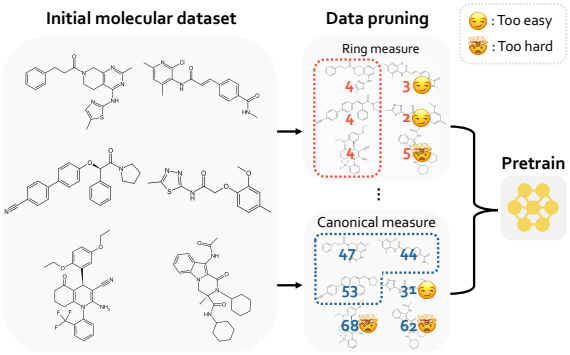


Figure 4: **Overview of molecular data pruning and ranking.** Each number represents the task-specific difficulty score assigned to a molecule, as defined in Table 1. For each parsing task, molecules are ranked based on these scores and we select the mid-difficulty samples.

large-scale web scraping (Deng et al., 2009; Raffel et al., 2020a; Lozhkov et al., 2024), chemical datasets often rely on costly and labor-intensive wet lab experiments or computational simulations. While resources such as the USPTO series (Wei et al., 2010; Lu and Zhang, 2022) and MoleculeNet (Wu et al., 2018) exist, expanding them is expensive and labor-intensive. This highlights the need for scalable alternatives—datasets that can be automatically generated with minimal cost while preserving domain relevance.

3 Training framework of CLEANMOL

In this section, we present our framework to improve the molecular understanding of LLMs using a new dataset, coined CLEANMOL.[2] Our scheme consists of (1) data preparation and (2) a two-stage training procedure. In the data preparation step, we prepare the CLEANMOL dataset with deterministic and scalable SMILES parsing tasks. Next, in the training step, we pre-train LLMs with the CLEANMOL dataset, followed by fine-tuning downstream applications. To improve the pre-training, we also introduce a task-adaptive data pruning and curriculum learning strategy based on task-specific difficulty measures.

[2]Our framework and dataset are both termed CLEANMOL.

Functional group matching	Ring counting	Chain length measurement	SMILES canonicalization	Fragment assembly
# of functional groups	# of rings	# of branches	SMILES length	

Table 1: **Definition of each task-specific difficulty.**

3.1 CLEANMOL data preparation

First, we introduce the CLEANMOL dataset based on the SMILES parsing tasks proposed in Section 2.1. There exist two key advantages of our proposed tasks: determinism and scalability.

In detail, on the one hand, in terms of determinism, our tasks are designed to have a unique and clearly defined answer (i.e., number or canonicalized SMILES) unlike previous pre-training objectives such as masking and translation as detailed in Section 6. This ensures unambiguous supervision during training and facilitates reliable learning.

On the other hand, regarding scalability, as the proposed tasks apply to any valid molecules without any experimental data, they can be expanded to a vast set of molecules. In detail, all annotations can be automatically generated using open-source cheminformatics tools such as RDKit (Landrum et al., 2024), making the dataset extensible to virtually unlimited molecular corpora. We provide the simplified example instructions of SMILES parsing tasks in Figure 3 and more examples including detailed instruction formats in Appendix A.

3.2 Training with CLEANMOL

Once the CLEANMOL dataset is prepared, we adopt a task-specific **data pruning** and **curriculum learning** inspired by recent work on high-quality LLM data curation (Gunasekar et al., 2023;

Task type	Model	Subgraph			Global graph	
		FG	Ring	Chain	Canonical	Assembly
5-shot	Deepseek-V3-chat	0.8912	0.6266	0.2976	0.1484	0.1512
	GPT-4o	0.8750	0.5955	0.2857	0.1078	0.1932
	Llama3.1-8B	0.6725	0.2103	0.2747	0.0027	0.0190
	Qwen2.5-7B	0.7704	0.4148	0.1139	0.0022	0.0131
	Galactica-6.7B	0.5000	0.0732	0.1511	0.0000	0.0046
SFT	Llama3.1-8B (Single)	0.9414	0.8612	0.9859	0.9356	0.8858
	Llama3.1-8B (Multi)	0.9891	0.8707	0.9851	**0.9463**	**0.9010**
	Qwen2.5-7B (Single)	0.9891	0.8674	**0.9907**	0.7593	0.3371
	Qwen2.5-7B (Multi)	**0.9901**	**0.8750**	0.9902	0.9262	0.8835

Table 2: **SMILES parsing performance.** FG stands for the functional group. Background indicates the improvement of multi-task learning compared to the single-task learning and the best results are highlighted in **bold**.

Marion et al., 2023; Ankner et al., 2024) to further enhance pre-training with CLEANMOL. As illustrated in Figure 4, our approach involves: (1) subsampling sufficiently informative molecules, and (2) constructing a curriculum by ranking these examples from simple to complex using task-specific difficulty measures.

The difficulty measures are defined for each parsing task as summarized in Table 1. For instance, in the chain length measurement task, molecules with extensive branches often lead to SMILES where relevant subgraph atoms appear far apart in the string, increasing parsing difficulty. By excluding extremely easy or hard molecules (i.e., subsample molecules with mid-level difficulties) and organizing the training data from simple to complex, our approach aligns with curriculum learning principles (Bengio et al., 2009) and leads to improved performance, as validated in Section 4.2.

Next, we adopt a two-stage training pipeline to effectively integrate SMILES parsing into LLM. In the first stage, we perform pre-training on the pruned CLEANMOL dataset using supervised fine-tuning. This allows the model to acquire core structural understanding and compositional knowledge of molecular graphs. In the second stage, we further fine-tune this trained model on downstream molecular tasks. By initializing with a model that has already learned to parse molecular structures, downstream adaptation becomes more accurate.

4 Experiments: SMILES parsing tasks

In this section, we evaluate the effectiveness of our proposed SMILES parsing task as a pre-training signal for LLMs. The parsing task is formally defined in Section 2.1. We demonstrate that recent LLMs, while not inherently proficient in SMILES parsing, can acquire this capability through targeted training. We provide experimental settings including prompts and resources in Appendix B.

4.1 LLMs can learn SMILES parsing

As described in Section 2.2, SMILES parsing poses a significant challenge for general-purpose LLMs, despite its foundational importance for molecular understanding. Our experiments reveal that LLMs lack the inductive bias to naturally understand the molecular structure encoded in SMILES strings. However, we show that through supervised fine-tuning (SFT), LLMs can learn to accurately parse and interpret SMILES representations.

Dataset. We construct a CLEANMOL benchmark consisting of 50K molecules per SMILES parsing task, totaling 250K examples across five tasks. The molecules are subsampled from the ZINC250k (Irwin et al., 2012) training dataset using our proposed molecular data pruning strategy described in Section 3.2, which excludes extremely easy or hard molecules to enhance the molecular pre-training. Additionally, for the test dataset, we randomly selected 10K molecules from the ZINC250K test split and fixed this subset across all experiments.

Baselines. We evaluate the parsing capabilities of four general-purpose LLMs—Deepseek-V3-Chat (Liu et al., 2024), GPT-4o (OpenAI and et al., 2024), LLaMA3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct (Yang et al., 2024)—and one chemistry-specific LLM, Galactica-6.7B (Taylor et al., 2022). To assess the basic molecular understanding of general-purpose LLMs, we apply 5-shot prompting to Deepseek and GPT-4o, which are not publicly trainable and thus cannot be fine-tuned. Similarly, we apply 5-shot prompting to Galactica, a chemistry-specific LLM pre-trained on molecular corpora, to evaluate its zero-shot capabilities without further supervision. In contrast, for LLaMA and Qwen, which are open-weight general-purpose LLMs, we perform supervised fine-tuning using our SMILES parsing dataset to examine whether explicit structure-aware

Pruning type	Subgraph			Global graph		
	FG	Ring	Chain	Canonical	Assembly	Average
Random	0.9921	0.9212	0.9886	0.7845	0.7352	0.8843
Length	0.9910	0.8531	0.9785	0.8519	0.8044	0.8958
Molecular pruning (top)	0.9902	0.8123	0.9716	0.9446	0.7487	0.8934
Molecular pruning (bottom)	0.9729	0.6995	0.9597	0.5514	0.5186	0.7404
Molecular pruning (middle, **ours**)	0.9901	0.8750	0.9902	0.9262	0.8835	**0.9330**

Table 3: **Effect of molecular data pruning on Qwen2.5-7B-Instruct.** "Random" and "Length" refer to baselines using random sampling and SMILES length as proxies for difficulty. "Top," "middle," and "bottom" denote subsamples consisting of the most difficult, moderately difficult, and easiest molecules, respectively, based on task-specific difficulty heuristics.

training can bridge the gap in molecular comprehension. Notably, we explore two experimental settings: *single-task*, where a separate model is trained for each parsing task, and *multi-task*, where a single model is jointly trained on all five tasks.

Metrics. We evaluate performance using accuracy, as SMILES parsing tasks are deterministic and each input has a well-defined answer.

Results. The results are presented in Table 2. We observe that recent general-purpose LLMs (GPT-4o and Deepseek) and even a chemical LLM (Galactica) perform poorly on SMILES parsing, revealing their limited molecular comprehension. This validates that *the primary bottleneck in applying LLMs to molecular domains lies not in the absence of chemical knowledge, but in the lack of basic molecular structural understanding—specifically, the ability to parse and interpret SMILES strings*. In contrast, fine-tuned LLaMA and Qwen models show substantial improvements, demonstrating that SMILES parsing can be effectively learned through training. Moreover, all tasks—except for chain length measurement—achieved higher accuracy in the multi-task setting, suggesting that transferable structural understanding across tasks contributes to improved performance.

4.2 Effect of molecular data pruning

We further investigate the impact of our molecular data pruning strategy on parsing performance. As detailed in Section 3.2, this technique aims to curate a training set that maximizes informativeness. The results, shown in Table 3, demonstrate that our pruning method improves performance, suggesting that data quality plays a critical role in teaching LLMs the implicit grammar of SMILES.

4.3 Ablation study

Here, we conduct an ablation study to validate the impact of the increase in dataset size in our proposed CLEANMOL dataset. In detail, we evaluate

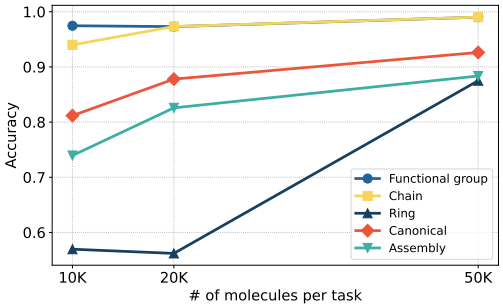


Figure 5: **Data scale analysis for SMILES parsing.**

the accuracy of the SMILES parsing task for 10K, 20K, and 50K data settings per task in the same setting in Section 4.1. We provide the results in Figure 5. Here, we observed that increasing the dataset size consistently improves SMILES parsing performance, with particularly dramatic gains in the ring counting and fragment assembly tasks. This validates the expandability of our framework.

5 Experiments: Downstream tasks

In this section, we evaluate the effect of pre-training LLMs on CLEANMOL dataset across three molecular generation downstream applications. We provide the experimental settings in Appendix B and additional experimental results in Appendix C.

Our results demonstrate that incorporating CLEANMOL as a pre-training strategy consistently improves performance across diverse downstream molecular tasks. These findings provide strong empirical support for our central hypothesis: clean and structurally faithful SMILES parsing serves as an effective and transferable learning signal for LLMs. Notably, CLEANMOL achieves state-of-the-art or competitive performance despite being pre-trained without any task-specific data, underscoring the strength and generality of our approach.

5.1 Molecular generation

The molecular generation task aims to generate molecules given prompts, including retrosynthesis, reagent prediction, and forward reaction prediction.

Models	Exact.	BLEU	Levenshtein ↓	MACCS FTS	RDK FST	Morgan FTS	Validity
Task 1: Retrosynthesis							
Text+Chem T5	0.141	0.765	24.04	0.685	0.765	0.585	0.698
Mol-Instructions (Lla.2)	0.009	0.705	31.23	0.283	0.487	0.230	-
Mol-Instructions (Lla.3)	0.333	0.842	17.64	0.704	0.815	0.646	-
Mol-Instructions (Lla.3.1)*	0.255	0.890	17.76	0.813	0.690	0.644	-
InstructMol-GS	0.407	0.941	13.97	0.753	0.852	0.714	-
Llama3.1-8B	0.456	0.944	10.22	0.895	0.837	0.801	0.979
+ Mol-Instructions (SFT)*	0.541	0.955	8.25	0.915	0.878	0.843	-
+ CLEANMOL	**0.581**	**0.959**	**7.86**	**0.923**	**0.890**	**0.856**	**0.998**
Qwen2.5-7B	0.460	0.946	10.11	0.897	0.849	0.809	0.910
+ CLEANMOL	0.554	0.958	8.26	0.915	0.880	0.844	0.995
Task 2: Reagent prediction							
Text+Chem T5	0.000	0.255	49.32	0.039	0.186	0.052	0.313
Mol-Instructions (Lla.2)	0.044	0.224	23.17	0.237	0.364	0.213	-
Mol-Instructions (Lla.3)	0.101	0.648	18.33	0.412	0.521	0.375	-
Mol-Instructions (Lla.3.1)*	0.085	0.676	22.40	0.505	0.398	0.356	-
InsturctMol	0.129	0.610	19.66	0.444	**0.539**	0.400	-
Llama3.1-8B	0.124	0.625	17.31	0.538	0.433	0.398	**0.999**
+ Mol-Instructions (SFT)*	0.142	0.678	17.14	0.562	0.467	0.430	-
+ CLEANMOL	**0.147**	**0.687**	16.89	**0.564**	0.472	**0.434**	**0.999**
Qwen2.5-7B	0.120	0.649	17.76	0.533	0.431	0.395	-
+ CLEANMOL	0.128	0.685	**16.58**	0.557	0.455	0.415	0.975
Task 3: Forward reaction prediction							
Text+Chem T5	0.236	0.782	13.63	0.523	0.630	0.505	0.967
Mol-Instructions (Lla.2)	0.045	0.654	27.26	0.313	0.509	0.262	-
Mol-Instructions (Lla.3)	0.503	0.883	13.41	0.756	0.863	0.708	-
Mol-Instructions (Lla.3.1)*	0.402	0.907	13.11	0.848	0.718	0.679	-
InstructMol-GS	0.536	0.967	10.85	0.776	0.878	0.741	-
Llama3.1-8B	0.794	0.981	2.47	0.965	0.938	0.926	0.988
+ Mol-Instructions (SFT)*	0.888	**0.990**	**1.33**	**0.983**	**0.967**	**0.961**	-
+ CLEANMOL	**0.890**	**0.990**	1.37	0.980	0.966	0.959	**0.996**
Qwen2.5-7B	0.833	0.986	2.08	0.972	0.947	0.943	0.987
+ CLEANMOL	0.874	0.989	1.56	0.980	0.963	0.956	0.959

Table 4: **Molecular generation performance.** Background indicates the improvement compared to vanilla model. Asterisks (*) denote reproduced results and - in validity represents the SELFIES-based methods which guarantees the perfect validity. For each metric, the best and second-best result is highlighted with **bold** and underline.

Dataset. We use the Mol-Instructions dataset (Fang et al., 2024), which covers three molecule generation tasks. Specifically, retrosynthesis predicts the possible precursors that lead to a given target molecule. Next, the reagent prediction task requires the generation of suitable catalysts, solvents, or ancillary reagents for a given chemical reaction. Lastly, forward reaction prediction involves the generation of a plausible product from given reactants and reagents. We follow the data splits provided in Mol-Instructions.

Baselines. We evaluate CLEANMOL by integrating it with two base models: LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024), to test whether CLEANMOL consistently improves performance. Notably, the vanilla base models are fine-tuned on each downstream task without pre-training. For an absolute performance comparison, we include three baselines: Text+Chem T5 (Christofidellis et al., 2023a), Mol-Instructions (Fang et al., 2024) and InstructMol (Cao et al., 2023). Additionally, we include a variant of Mol-Instructions denoted as Mol-Instructions (SFT), which is first instruction-tuned on the same dataset size as our CLEANMOL dataset (250K) and then further fine-tuned on each downstream task. This ensures a fair comparison for both the model and the training data size.

Metrics. We assess the performance by comparing the generated molecules with the ground truth based on eight metrics. These include SMILES string-based metrics (Exact match, BLEU (Papineni et al., 2002), and Levenshtein distance (Miller et al., 2009)), molecular fingerprint similarities (MACCS (Durant et al., 2002), RDK (Schneider et al., 2015), and Morgan (Rogers and Hahn, 2010)), distributional similarity via Fréchet ChemNet Distance (FCD) (Preuer et al., 2018), and the validity of generated molecules.

Results. The results are summarized in Table 4. Incorporating CLEANMOL consistently improves performance across all backbones, demonstrating the effectiveness of SMILES parsing tasks in en-

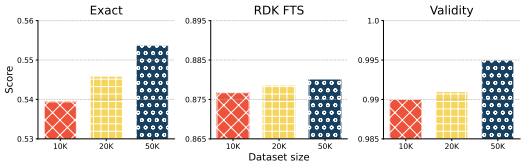


Figure 6: **Data scale analysis for retrosynthesis.**

hancing molecular language modeling. These improvements suggest that pre-training on clean and deterministic CLEANMOL dataset facilitates the model's structural understanding required for generation tasks. Notably, integrating CLEANMOL into LLaMA3.1-8B-Instruct achieves state-of-the-art—or at least comparable—performance to Mol-Instructions (SFT), despite using no molecular generation data during pre-training.

5.2 Ablation study

Here, we evaluate the effect of CLEANMOL dataset size on retrosynthesis performance using 10K, 20K, and 50K molecules per parsing task following the setup in Section 5.1. As described in Figure 6, the performance grows with data scale, demonstrating CLEANMOL 's scalability. As SMILES parsing requires no costly experiment, this framework easily extends to large molecular corpora.

6 Related work

LLMs for chemistry. General-purpose LLMs often struggle with fundamental chemistry tasks, particularly those requiring molecular structure understanding (White et al., 2023; Castro Nascimento and Pimentel, 2023; Guo et al., 2023). To address this gap, several studies have proposed chemically specialized LLMs. Some approaches pre-train LLMs on molecular and biomedical corpora to inject domain-specific knowledge (Edwards et al., 2022; Christofidellis et al., 2023b; Liu et al., 2023a; Pei et al., 2023). Others explore instruction tuning on curated molecular tasks (Fang et al., 2024; Cao et al., 2023), or leverage retrieval-augmented prompting to improve few-shot performance (Li et al., 2024). While these methods aim to inject domain knowledge, they often neglect the need for grounding models in basic molecular understanding. In contrast, we emphasize clean and deterministic structural supervision through well-defined SMILES parsing tasks, which can complement existing methods and integrate with instruction tuning or domain adaptation.

Pre-training of LLMs for chemistry. Effective pre-training tasks should be well-structured and sufficiently simple to support generalizable learning. In chemistry, many works adopt NLP-inspired objectives such as masked language modeling (MLM) (Devlin et al., 2019) and sequence-to-sequence translation (Raffel et al., 2020b), applied to SMILES (Weininger, 1988) or SELFIES (Krenn et al., 2020). Edwards et al. (2022) used separate MLM pretraining on molecular and textual data, while later studies (Pei et al., 2023; Christofidellis et al., 2023b) combined MLM with molecule–text translation. Liu et al. (2023a) embedded SMILES in natural language prompts, and other works incorporated 2D or 3D geometry (Li et al., 2023; Ji et al., 2024; Zhou et al., 2023).

Despite these advancements, most strategies introduce unambiguous supervision signals due to the non-determinism of molecular representations. For example, in masked SMILES prediction, multiple chemically valid tokens can fill the same masked position, leading to a noisy training signal. This undermines training effectiveness and limits the model's ability to learn robust understanding. To address this issue, we provide clean and deterministic SMILES parsing tasks as pre-training tasks.

Data pruning in LLMs. Data pruning refers to selecting an informative subset of training data, which is crucial for reliable LLM training (Gunasekar et al., 2023). Most data pruning methods rely on rule-based filters (Wenzek et al., 2020; Raffel et al., 2020a), perplexity scores (Marion et al., 2023; Ankner et al., 2024), or LLM embeddings (Tirumala et al., 2023). However, these metrics are ill-defined for molecular strings, where perplexity and embeddings do not reflect the structural information of the corresponding molecules. To address this, we introduce task-specific difficulty measures and data pruning strategies for molecules.

7 Conclusion

In this paper, we revisit the key limitation in applying LLMs to chemistry: the inability to interpret the structures encoded in SMILES. To address this, we propose CLEANMOL, a framework that introduces deterministic and scalable SMILES parsing tasks to provide unambiguous structural supervision. Our experiments show that CLEANMOL significantly enhances molecular structural understanding and improves performance across multiple downstream tasks. These results highlight the value of incorporating clean and structure-aware objectives into LLMs to support more robust applications.

Broader Impact

Our work contributes to the development of structurally grounded models for molecular applications. By introducing a structured, clean, and scalable set of SMILES parsing tasks, we aim to equip LLMs with a stronger inductive bias toward molecular structure understanding. This can enhance downstream applications such as drug discovery, materials design, and reaction prediction by improving the fidelity and reliability of molecular reasoning. However, as with any generative AI system in chemistry, potential misuse remains a concern. The capacity to generate toxic, harmful, or restricted compounds necessitates careful integration of safety measures and expert oversight.

Limitations

Limited structural information. Our SMILES parsing tasks focus on graph-level molecular structures and do not incorporate 3D conformational information, which is essential for many biological and physicochemical applications. Additionally, while our tasks are deterministic and scalable, they do not capture more nuanced chemical features such as stereochemistry, electronic effects, or reactivity patterns, which often require context beyond 2D topological graphs.

Language-specific scope. Our experiments are conducted exclusively in English and do not explore the applicability of the method across other languages, including morphologically rich or typologically diverse ones. Given that behaviors can vary across languages due to linguistic structure and training data distributions, the generalizability of our approach to multilingual settings remains an open question.

Model and dataset scale. Due to computational constraints, our experiments are limited to language models with up to 7.5B–8B parameters. It remains to be seen whether our framework scales effectively to larger models (e.g., 70B or beyond). Moreover, our pretraining is performed on a relatively modest dataset of 250K molecules, and while we observe consistent improvements, further studies on larger-scale datasets are necessary to assess the robustness and scalability of the approach.

Acknowledgments

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), the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-02304967, AI Star Fellowship(KAIST)), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-02653113, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale).

References

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

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.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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.

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*.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. Zinc: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–1768.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. 2024. Chain-of-thoughts for molecular understanding. *arXiv preprint arXiv:2410.05610*.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. 2024. Exploring molecular pretraining model at scale. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

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, and 11 others. 2024. rdkit/rdkit: 2024_09_1 (q3 2024) release beta.

Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. 2023. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1):7568.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

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

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, and 1 others. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

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.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

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.

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

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, 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. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. 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.

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*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995.

Tloen. 2023. Alpaca-lora. `https://github.com/tloen/alpaca-lora`.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

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.

David Weininger, Arthur Weininger, and Joseph L Weininger. 1989. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences*, 29(2):97–101.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, Zekun Guo, Yeqi Bai, Pan Deng, Yaosen Min, Ziheng Lu, Hongxia Hao, Han Yang, Jielan Li, Chang Liu, and 27 others. 2025. Nature language model: Deciphering the language of nature for scientific discovery. *Preprint*, arXiv:2502.07527.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*.

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

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

A Detailed description of SMILES parsing tasks

A.1 Subgraph matching

This category includes *functional group matching*, *ring counting*, and *carbon chain length measurement*. These tasks are designed to focus on local substructures within the molecular graph, such as common functional motifs, ring systems, and chain connectivity. Each task formulation is deterministic and lends itself to clear evaluation.

Functional group matching. Functional group matching evaluates whether a specified functional group is present in a given molecule. To ensure determinism, we cast this task as a binary classification problem: the model must predict "yes" or "no" based on the presence of the target group. An example of the instruction format is shown in Figure 7.

```
Functional group matching
```

Answer only in 'Yes' or 'No' without any other information.

Question: Does the molecule represented by the SMILES string contain the specified functional group? Respond with 'Yes' or 'No'.
SMILES: [SMILES]
FUNCTIONAL GROUP: [Functional group SMILES]

ANSWER: [Yes/No]

Figure 7: **An instruction format of functional group matching.**

Ring counting. Ring counting asks the model to determine the number of rings of a specific size (e.g., five- or six-membered) in the molecule. This task tests the model's ability to track topological cycles through non-contiguous token spans in SMILES. The instruction format is illustrated in Figure 8.

```
Ring counting
```

Answer only with the corresponding integer number without any other information.

Question: Assess the SMILES below and report how many rings consist of [RING SIZE] atoms. Give me the integer only.
SMILES: [SMILES]
SIZE OF RINGS: [RING SIZE]

ANSWER: [NUMBER OF RINGS]

Figure 8: **An instruction format of ring counting.**

Chain length measurement. This task requires the model to identify the length of the longest acyclic carbon chain in the molecule, excluding atoms that are part of rings. It challenges the model to distinguish between linear and branched motifs and to reason about connectivity beyond localized tokens. Such chains often span long syntactic distances in SMILES, making the task non-trivial. The instruction format is shown in Figure 9.

```
Chain length measurement
```

Answer only with the corresponding integer number without any other information.

Question: Report the size of the largest carbon-only chain not contained within a ring in the molecule represented by this SMILES. Answer with an integer only.
SMILES: [SMILES]

ANSWER: [LENGTH OF CHAIN]

Figure 9: **An instruction format of chain length measurement.**

A.2 Global graph matching

This category includes tasks that operate on a global level: *SMILES canonicalization* and *fragment assembly*. Unlike subgraph matching, these tasks require full-graph interpretation, where success depends on integrating information across the entire molecular structure.

This category consists of *SMILES canonicalization* and *fragment assembly*.

SMILES canonicalization. Canonicalization involves transforming a randomly ordered SMILES string into its canonical form following the canonicalization rules (Weininger et al., 1989). In detail, these rules typically involve assigning a unique ranking to atoms based on graph invariants (e.g., atomic number, connectivity, bond types), selecting the lexicographically smallest traversal path, and applying consistent numbering for ring closures. This task encourages the model to learn structural invariance under permutation and reinforces a graph-level understanding of molecular identity. The task format is provided in Figure 10.

`SMILES canonicalization`

Answer only with the corresponding SMILES string without any other information.

Question: Give me a canonicalized SMILES string that represents the same molecule as the given one.

SMILES: [SMILES]
ANSWER: [CANONICAL SMILES]

Figure 10: **An instruction format of SMILES canonicalization.**

Fragment assembly. Fragment assembly evaluates whether the model can reconstruct a full molecule from two disconnected SMILES fragments. This task tests global molecular coherence and the model's ability to resolve attachment points into a chemically valid structure. The instruction format of the instruction is shown in Figure 11.

B Experimental details

In this section, we provide the details of the experiments. All experimental code related to this paper is available at `https://anonymous.4open.science/r/CLEANMOL` and our experiments are based on a single run. We use NVIDIA A100-80GB GPUs. We also apply low rank adaptation (Hu et al., 2022) and report results from a single run. Our implementations are based on the `transformers` library (Wolf et al., 2020), the `trl` library (von Werra et al., 2020), the `accelerate` library (Gugger et al., 2022), and `unsloth` library

`Fragment assembly`

Answer only with the corresponding SMILES string without any other information.

Question: Connect the following two SMILES fragments into a unified structure at their reactive sites.
SMILES: [FRAGMENT 1, FRAGMENT 2]

ANSWER: [SMILES]

Figure 11: **An instruction format of SMILES assembly.**

(Daniel Han and team, 2023). Additionally, we used the packages including rouge-score==0.1.2 and nltk==3.8.1.

B.1 SMILES parsing

Here, we describe the detailed settings for the SMILES parsing experiments in Section 4, including the pre-trainig step with SMILES parsing tasks.

Hyperparameters. The hyperparameters for all the models are provided in Table 5. We share the same hyperparameter for all the SMILES parsing tasks and base models. Notably, the model trained with SMILES parsing tasks is used as the pre-trained model for downstream tasks in Section 5.

	Hyperparameter
Batch size	16
Learning rate	$5e^{-4}$
Epochs	1
Warmup ratio	0.01
Weight decay	0.1
Lr scheduler	cosine
Gradient accumulation steps	1
Repetition penalty	1
Temperature	0.2
Lora r	64
Lora alpha	16
Lora dropout	0.05

Table 5: **Hyperparameters for SMILES parsing.**

B.2 Downstream tasks

Here, we describe the detailed settings for the downstream task experiments in Section 5.

Hyperparameters. The hyperparameters for all the models are provided in Table 5. We share the

same hyperparameter for all downstream tasks and base models. Notably, for the reproduced Mol-instructions (Fang et al., 2024) models, we follow the hyperparameters given in the original paper.

	Hyperparameter
Batch size	16
Learning rate	$5e^{-4}$
Epochs	1
Warmup ratio	0.01
Weight decay	0.1
Lr scheduler	cosine
Gradient accumulation steps	1
Repetition penalty	1
Temperature	0.2
Lora r	64
Lora alpha	16
Lora dropout	0.05

Table 6: **Hyperparameters for downstream tasks.**

C Additional experimental results

In this section, we provide additional experimental results including additional downstream tasks and ablation study.

C.1 Molecular property prediction

The molecular property classification task aims to predict binary labels for intrinsic physical or chemical properties, such as blood-brain barrier permeability or toxicity.

Dataset. We use the MoleculeNet (Wu et al., 2018) dataset, focusing on three binary classification tasks: BACE, HIV, and Clintox. The BACE task predicts whether a molecule can inhibit human β-secretase 1 (BACE-1). The HIV task involves predicting the ability of compounds to inhibit HIV replication. The Clintox task assesses whether a compound is likely to fail clinical trials due to toxicity. We follow the splits provided in MoleculeNet.

Baselines. We evaluate CLEANMOL by integrating it with two base models: LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024). For an absolute performance comparison, we include additional baselines: MolCA (Liu et al., 2023b), LlasMol (Yu et al., 2024) and InstructMol (Cao et al., 2023).

Metrics. We evaluate the performance using accuracy, which denotes the overall proportion of correct predictions.

Model	BACE	HIV	Clintox
MolCA (1D+2D)	0.798	–	0.895
$\text{LlasMol}_{\text{Mistral}}$	–	0.967	0.931
InstructMol-GS	0.821	0.689	–
LLaMA3.1-8B	0.507	0.971	0.946
+ CLEANMOL	0.639	0.971	0.946
Qwen2.5-7B	0.533	0.969	0.946
+ CLEANMOL	0.638	0.971	0.946

Table 7: **Molecular property classification performance on the MoleculeNet dataset.**

Results. We report the results in Table 7. We observe that models pre-trained with CLEANMOL achieve consistent gains, confirming that the structural alignment learned during SMILES parsing transfers effectively to property classification tasks.

C.2 Molecular property regression

The molecular property regression task focuses on predicting continuous-valued molecular properties.

Dataset. We again use the Mol-Instructions (Fang et al., 2024) dataset. We target quantum mechanics properties: HOMO energy, LUMO energy, and the energy gap (HOMO–LUMO difference). We also follow the same split.

Baselines We evaluate CLEANMOL by integrating it with two base models: LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen-2.5-7B-Instruct (Yang et al., 2024). For an absolute performance comparison, we include additional baselines: Alpaca (Tloen, 2023), Baize (Xu et al., 2023), Vicuna (Chiang et al., 2023), Galactica (Taylor et al., 2022), and Mol-Instructions (Fang et al., 2024). Here, the Mol-Instructions (SFT) follows the same training strategy described in Section 5.1.

Metrics. We use mean absolute error (MAE) to evaluate prediction accuracy.

Results. We report the results in Table 8. The results indicate that models pre-trained on SMILES parsing consistently outperform baselines, demonstrating that structural information learned via parsing enhances quantitative property prediction.

C.3 Ablation study

SMILES parsing component. Here, we conduct an ablation study on the contribution of each SMILES parsing task. We report the results for the retrosynthesis task using LLaMA3.1-8B. We provide the results in Appendix C.3. The results

Model	MAE
Alpaca	322.109
Baize	261.343
Vicuna	860.051
Galactica	0.568
Mol-Instruct. (Lla.2)	0.013
Mol-Instruct. (Lla.3)	15.059
Mol-Instruct. (Lla.3.1)*	0.011
Mol-Instruct. (SFT)*	**0.005**
LLaMA3.1-8B	**0.005**
+ CLEANMOL	**0.005**
Qwen2.5-7B	15.923
+ CLEANMOL	**0.005**

Table 8: **Molecular property regression performance on the Molinstructions dataset.**

	Exact.	BLEU	Lev. ↓	MACCS	RDK	Morgan	Valid.
FG	0.541	0.955	8.66	0.914	0.874	0.840	0.997
Ring	0.537	0.955	8.46	0.914	0.877	0.841	0.993
Chain	0.562	0.956	8.21	0.917	0.880	0.850	0.991
Canonical	0.550	0.955	8.52	0.915	0.877	0.844	0.995
Assembly	0.542	0.955	8.60	0.913	0.880	0.842	0.996
All (CLEANMOL)	0.581	0.959	7.86	0.923	0.890	0.856	0.998

Table 9: **Ablation study of CLEANMOL component.**

show that the carbon chain length measurement task (Chain) contributes the largest performance improvements. Notably, the full combination of all five parsing tasks in CLEANMOL achieves the best overall performance, validating the effectiveness of our comprehensive, multi-task pretraining strategy.

	Exact.	BLEU	Lev. ↓	MACCS	RDK	Morgan	Valid.
Top	0.550	0.957	8.46	0.915	0.877	0.843	0.998
Bottom	0.555	0.956	8.42	0.916	0.875	0.845	0.996
Middle (Ours)	0.581	0.959	7.86	0.923	0.890	0.856	0.998

Table 10: **Ablation study of data pruning.**

Molecular data pruning. Next, we conduct an ablation study to observe the impact of molecular data pruning on downstream tasks. To address this, we evaluated the performance of the retrosynthesis task with Llama3.1-8B with top, middle (ours), and bottom data pruning strategies. We provide the results in Appendix C.3 The results show that our middle data pruning strategy shows the best downstream task performance, validating the effectiveness of our strategy.

D Usage of AI assistants

In preparing this work, we used AI-based writing assistants to improve sentence structure, correct grammatical errors, and enhance overall readability. These tools were employed solely for language refinement and did not contribute to the development of technical content, research methodology, or experimental analysis. All scientific ideas, results, and conclusions presented in the paper were conceived and authored entirely by the researchers. The use of AI assistance was restricted to editorial purposes and did not affect the originality or intellectual contributions of the work.

E Scientific Artifacts

The License for artifacts. All datasets and software tools used in this study comply with their respective licenses. Specifically, we utilized publicly available datasets such as ZINC250K (Irwin et al., 2012) and Mol-Instructions (Fang et al., 2024) in accordance with their usage terms. External tools such as RDKit were employed under their permissive open-source license. To support transparency and reproducibility, we release our trained models and source code at `https://anonymous.4open.science/r/CLEANMOL` under an appropriate open-source license.

Artifact use consistency with intended use. All datasets and tools were used in a manner consistent with their intended use. For instance, the Mol-Instructions dataset (Fang et al., 2024)—originally designed for molecule generation and property prediction—was employed for aligned downstream tasks in our study. Likewise, RDKit was used exclusively for molecular structure analysis and data preprocessing, as intended by its developers.