MolLangBench: A Comprehensive Benchmark for Language-Prompted Molecular Structure Recognition, Editing, and Generation

Poster Presenter: **Ling Liu**

Feiyang Cai [1] Jiahui Bai[2] Tao Tang[3] Guijuan He[1] Joshua Luo[1] Tianyu Zhu[1] Srikanth Pilla[4] Gang Li[1] Ling Liu[5] Feng Luo[1]

[1]Clemson University [2]Independent Researcher [3]HitGen [4]University of Delaware [5]Georgia Institute of Technology

Motivation

Key point: *Precise recognition and manipulation of molecular structure are fundamental to chemists to reason about downstream chemical tasks, and should be for AI systems as well.*

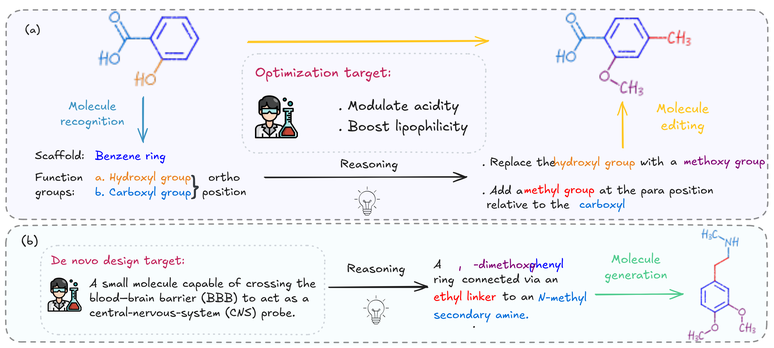


Figure 1. Chemists' workflows for diverse chemical tasks motivate MolLangBench.

Why structure matters for molecule–language models:

- **Property prediction:** Molecular function is determined by structure; reliable prediction requires accurate structural recognition.
- **Molecule optimization / *de novo* discovery:** Beyond recognition, tasks require precise structural manipulation to satisfy design objectives (Figure 1).
- **Reaction reasoning:** Understanding reactants and mechanisms depends on accurate structure recognition, while predicting products requires precise structural manipulation.

Implication for molecule–language modeling:

- Molecular representations should be first aligned with **structure-grounded language descriptions**, while higher-level reasoning is delegated to LLM backbones.
- Without these foundational interface capabilities, reliable chemical reasoning is unlikely.
- Similar to vision–language models, align language with observable visual content first (e.g., "an armchair that looks like an apple") before enabling higher-level reasoning.

Benchmark Overview

MolLangBench evaluates three core **molecule–language interface** tasks: *recognition*, *editing*, and *generation*.

- **Molecular structure recognition**
  - **Task description:** extract structural information from molecular inputs, including topology, connectivity, functional groups, substructures, and stereochemistry (18 subtasks).
  - **Curation:** constructed using automated cheminformatics tools (Figure 2).
  - **Scale:** 200 samples per subtask (with additional data available for training).
- **Language-prompted molecule editing**
  - **Task description:** modify molecular structures following natural-language instructions with precise, deterministic outcomes.
  - **Curation:** multi-stage expert annotation and validation to ensure correctness and unambiguity (Figure 2).
  - **Scale:** core set: 200 samples; extended set: 200 samples.
- **Molecule generation from structural descriptions**
  - **Task description:** generate valid molecular structures from detailed language descriptions.
  - **Curation:** multi-stage expert annotation and validation to ensure correctness and unambiguity (Figure 2).
  - **Scale:** core set: 200 samples; extended set: 200 samples.

Annotation Pipeline

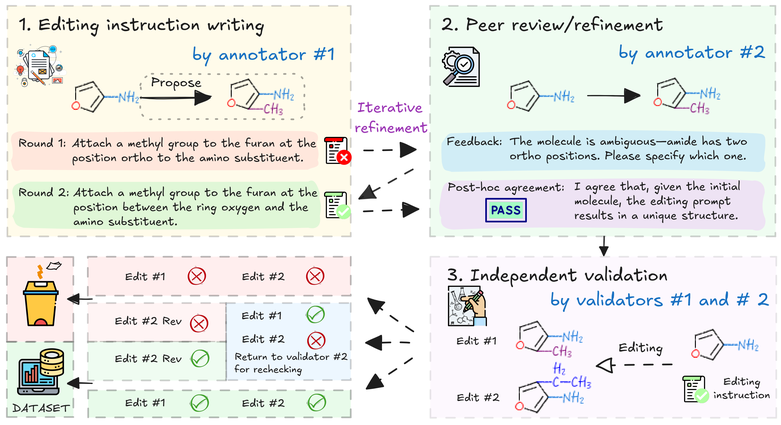
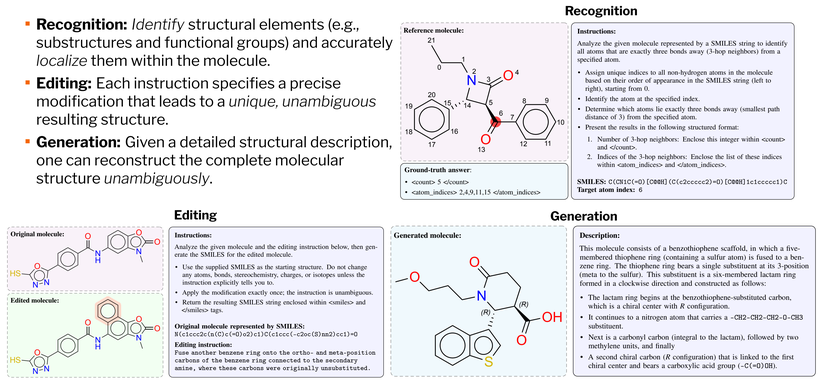


Figure 2. Rigorous annotation and validation pipeline for molecular structure editing and generation tasks.

- **Step 1: Expert annotation**
  - Chemistry-background annotators write editing instructions or structural descriptions from molecular structures.
  - Annotations follow strict guidelines to ensure **clarity**, **completeness**, and **one-to-one mapping**.
- **Step 2: Peer review and refinement**
  - A second annotator reviews and iteratively refines each sample for correctness and adherence to standards.
  - Only agreed-upon annotations proceed to validation.
- **Step 3: Independent validation**
  - Two independent validators reconstruct molecules from annotations using chemical drawing tools.
  - Samples are accepted only if both validators produce an **exact structural match** and confirm **unambiguity**.
- **Annotation statistics** (over 500 hours on expert annotation and validation)
  - **Editing:** ∼30 minutes per sample.
  - **Generation:** ∼60 minutes per sample.

Task Examples

- **Recognition:** *Identify* structural elements (e.g., substructures and functional groups) and accurately *localize* them within the molecule.
- **Editing:** Each instruction specifies a precise modification that leads to a *unique, unambiguous* resulting structure.
- **Generation:** Given a detailed structural description, one can reconstruct the complete molecular structure *unambiguously*.



Baseline Results

The benchmark supports multiple modalities based on molecular representations:

- **Graph:** inherent molecular structure graphs; preliminary evaluation of graph–language models.
- **Linear strings:** SMILES / SELFIES; 16 commercial and 3 chemistry-specific LLMs evaluated.
- **Images:** molecular structure images; evaluated using vision–language models (VLMs).

Table 1. Performance of representative models on molecular structure recognition (average over 18 subtasks), editing, and generation tasks. Recognition results are reported as accuracy / localization accuracy; editing and generation as validity / accuracy. For image-based evaluation, o4-mini is used for recognition and GPT Image 1 for editing and generation. Complete results are provided in Table S4.

Task	DeepSeek-R1	Qwen3 -Max	Claude- Opus-4.1	o3	o3 (SELFIES)	Gemini-2.5 -Pro	GPT-5	o4-mini / GPT Image 1
Recognition	0.721/0.566	0.486/0.186	0.814/0.692	0.877/0.792	0.528/-	0.852/0.753	**0.923/0.862**	0.772/0.700
Editing	0.720/0.485	0.690/0.360	0.950/0.705	0.945/0.785	0.960/0.195	0.930/0.745	**0.945/0.855**	0.510/0.080
Generation	0.400/0.045	0.465/0.000	0.920/0.330	0.670/0.290	0.185/0.000	**0.865/0.430**	0.690/0.430	0.130/0.000

Key Findings

- **Graph modality fails:** Existing graph–language models are task-specific and lack general alignment, leading to near-complete failure on our benchmark (Appendix A.11).
- **Vision underperforms:** Despite intuitive structural representations, VLMs perform worse than LLMs; performance drops sharply for editing and is near zero for generation.
- **LLMs still far from perfect:** LLMs using SMILES representation perform best across all modalities, but even GPT-5 achieves only 86.2% (recognition), 85.5% (editing), and 43.0% (generation) accuracy on tasks that are simple for humans.
- **SELFIES vs. SMILES:** SELFIES shows substantially lower performance than SMILES, with near-zero accuracy in generation tasks (some analysis in Appendix A.16).
- **Chemistry-specific LLMs fail to generalize:** Despite domain specialization, these models perform no better than random guessing on our benchmark.

Relationship Between Structure Understanding and Property Prediction

- Evaluate two MoleculeNet tasks (BBBP, BACE) with two prompting strategies:
  - Direct prediction from SMILES
  - Structure-first: describe structure → predict property
- **Result:** Structure-first prompting improves accuracy by $> 5\%$ across models and tasks.
- **Insight:** Explicit structural reasoning provides useful intermediate signals for property prediction.

Key Takeaways

- **Models still struggle on prerequisite tasks:** Even state-of-the-art AI systems show significant limitations in fundamental molecule–language interface tasks.
- **Current direction may be on the wrong track:** Without reliable **structure recognition and manipulation**, the reasoning capabilities of LLMs cannot be effectively leveraged.
- **Structure first, reasoning next:** Progress requires prioritizing **structure–language alignment** before higher-level chemical reasoning.
- **Structure description data is the bottleneck:** High-quality annotation is costly, time-intensive, and essential for aligning molecular structures with language; to address this, we develop **MolLangData**, a large-scale dataset constructed via a rule-regularized method (`https://huggingface.co/datasets/ChemFM/MolLangData`).

Acknowledgments

- This work was supported by the AIM for Composites EFRC (U.S. DOE, Office of Science, Basic Energy Sciences).
- Ling Liu acknowledges the support from NSF CISE grants, an IBM Faculty Award, and CISCO Edge AI programs.
- We thank Yi Hu for data annotation support and Dr. Yongkai Wu for generously providing access to Azure AI Foundry for our benchmark evaluation.

{feiyang, luofeng}@clemson.edu, ling.liu@cc.gatech.edu