MOLSTRUCT: A BENCHMARK FOR MULTI-SCALE MOLECULAR STRUCTURE UNDERSTANDING, REASONING, AND PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models have shown growing capabilities in molecular understanding and structure editing, yet their ability to reason about how molecular interventions shape dynamic behavior remains largely unexplored. In this work, we introduce MOLSTRUCT, a benchmark for molecular structural reasoning built around an intervention-relaxation pipeline. MOLSTRUCT covers molecular structure at three levels of granularity: atoms and bonds, functional groups, and whole molecules and evaluates four capabilities: structural perception, inverse change reasoning, stability reasoning, and structural prediction, in both text-based and multimodal settings. Experiments across multiple open-source and proprietary models show that current models perform reasonably well on perception, but struggle substantially with inverse reasoning, stability reasoning and predicting post-intervention relaxed outcomes. The results indicate that models are relatively stronger at recognizing molecular structures than at judging their stability or predicting how they relax. We further find that increasing model scale or reasoning depth provides only limited improvement, highlighting molecular stability reasoning as an important unresolved challenge for current foundation models.

1 INTRODUCTION

Understanding molecular behavior requires reasoning about how molecular structures respond to interventions. In molecular modeling, an edited structure is typically followed by structural relaxation, or geometry optimization, in which atomic coordinates adjust toward a lower-energy configuration. This process can substantially change the resulting geometry, so the same intervention may lead to different relaxed structures and stability outcomes depending on the surrounding molecular environment. Evaluating molecular reasoning therefore requires more than testing whether LLMs or MLLMs can recognize static structures or answer isolated chemistry questions. A more complete assessment should examine whether models can reason about how a given intervention changes molecular structure, stability, and the eventual relaxed outcome. As LLMs and MLLMs are increasingly used as scientific assistants, it becomes important to evaluate whether they can perform such structure-based molecular reasoning across both textual and visual representations of molecules.

Existing works have broadened chemical and molecular evaluation from knowledge and property reasoning to multimodal understanding, structure generation, and molecular editing (Alampara et al., 2025; Niyongabo Rubungo et al., 2025; Antunes et al., 2024; Choudhary, 2024). More recent structure-oriented benchmarks have begun to move beyond static molecular understanding: MolLangBench evaluates language-guided molecular recognition, editing, and generation across multiple representations, while AtomWorld evaluates whether LLMs can execute predefined atomistic structure modifications by generating the post-edit structure from an initial structure and an explicit action instruction (Cai et al., 2026; Lv et al., 2025). However, these settings do not capture how molecular structures dynamically change under intervention, or whether models can infer stability and future structural outcomes directly from three-dimensional structure. Many such operations are specified primarily as geometric or symbolic edits, for example moving, inserting, swapping, or rotating atoms according to prescribed spatial rules. The resulting target is typically the immediate edited structure, whereas in molecular systems the intervention may be followed by substantial relaxation and may lead to configurations with very different levels of stability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

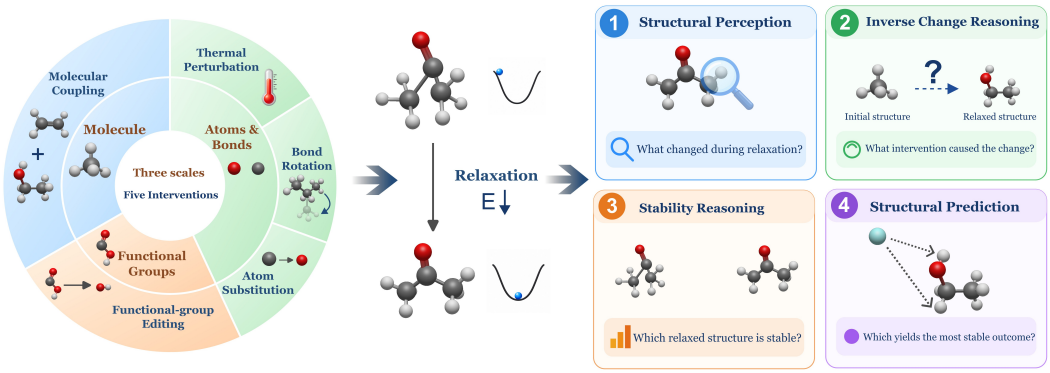


Figure 1: Overview of MOLSTRUCT. The benchmark covers three structural scales: molecular, atom-and-bond, and functional-group, together with five classes of molecular intervention, *i.e.,* molecular coupling, thermal perturbation, atom substitution, bond rotation, and functional-group editing. Each intervention is followed by structural relaxation or simulation, from which MOLSTRUCT constructs four complementary evaluation settings: **structural perception**, **inverse change reasoning**, **stability reasoning**, and **structural prediction**.

This distinction motivates a different question: *Can a foundation model infer molecular stability directly from structure?* Answering this question requires several related capabilities. First, the model must be able to perceive basic three-dimensional changes in molecular geometry; if it cannot identify which atoms, bonds, or local configurations have changed, reliable stability reasoning is unlikely. Second, given multiple relaxed molecular configurations, the model should be able to assess their relative stability from structure alone, without access to energy values. Third, molecular reasoning also has an inverse form: from an observed before-after structural change, the model may need to infer which bond, atom, or functional group was modified. Finally, the strongest form is predictive: before the resulting structure is known, the model must anticipate which intervention, site, or geometric change will lead to a more stable relaxed configuration, or whether a perturbation such as heating will induce bond breaking.

To study these capabilities, we introduce MOLSTRUCT, a benchmark for multi-scale molecular structural reasoning. As illustrated in Fig. 1, MOLSTRUCT spans three structural scales, including molecular, atom-and-bond, and functional-group, and five classes of molecular intervention: molecular coupling, thermal perturbation, atom substitution, bond rotation, and functional-group editing. Building on the resulting intervention, *i.e.,* the relaxation process, the benchmark evaluates four complementary capabilities: **structural perception**, **inverse change reasoning**, **stability reasoning**, and **structural prediction**. It supports both text-based and multimodal evaluation settings, allowing us to compare coordinate-based molecular representations with direct visual access to three-dimensional structure. **Structural perception** evaluates whether models can detect global and local 3D structural changes. **Inverse change reasoning** requires models to recover hidden molecular edits from observed structural differences. **Stability reasoning** presents only relaxed molecular structures and asks models to identify the most stable configuration. **Structural prediction** requires models to anticipate the result of an unseen intervention, such as selecting the molecular assembly site, bond-rotation angle, or functional-group modification that leads to the most stable relaxed structure, or predicting bond breaking under thermal perturbation. Note that a key design principle of MOLSTRUCT is that an intervention is not treated as the endpoint. Instead, benchmark questions are constructed around the structural relaxation or response that follows an intervention, with labels derived from the resulting structures and simulation records. This shifts the evaluation from recognizing or executing a molecular edit to reasoning about its structural consequences.

We systematically evaluate multiple open-weight and closed-source LLMs on MOLSTRUCT. Our experiments show that performance varies substantially across both capabilities and model families. Structural perception is generally more accessible, whereas inverse change reasoning, stability reasoning, and structural prediction remain challenging for many models. In particular, several models exhibit a strong bias in stability reasoning toward selecting the structure closest to the immediate post-intervention state, even when it is the least stable candidate. Image-based inputs generally yield lower performance than CIF-text representations on reasoning and prediction tasks, with sta-

bility reasoning as a notable exception: the visual representation makes this misleading structural similarity harder to identify and can therefore weaken the bias, resulting in higher apparent accuracy. Moreover, larger models and more intensive reasoning provide only limited and inconsistent gains, suggesting that scaling alone may be insufficient to address these challenges. These findings reveal persistent limitations of current LLMs in reasoning about intervention-induced structural changes, molecular stability, and post-intervention outcomes.

Our main contributions are threefold:

- We introduce MOLSTRUCT, a multi-scale benchmark for molecular structural reasoning across molecular, atom-and-bond, and functional-group representations.
- We evaluate molecular reasoning through four complementary capabilities: structural perception, inverse change reasoning, stability reasoning, and structural prediction, with particular emphasis on whether stability can be inferred directly from relaxed molecular structures without access to energy values.
- We provide both text-based and multimodal versions of the benchmark, allowing controlled evaluation of how molecular representations affect structural perception, inverse reasoning, stability reasoning, and prediction.

2 RELATED WORK

LLMs for Chemistry and Molecular Science. Large language models have been increasingly adapted to chemistry and molecular science, extending from text-based chemical knowledge to molecular structure understanding and generation. MoleculeSTM (Liu et al., 2023a) aligns molecular structures with textual descriptions for molecule–text retrieval and text-guided molecular editing, while MolCA (Liu et al., 2023b) integrates molecular graphs with language models for molecular captioning, retrieval, and name prediction. Mol-Instructions (Fang et al., 2024) further provides large-scale instruction data for biomolecular tasks, and LLaMo (Park et al., 2024) introduces a molecular graph, *i.e.,* a language model for molecular description, property prediction, and IUPAC name prediction.

However, existing approaches mainly focus on molecular representation, cross-modal understanding, generation, and property-related tasks, with limited attention to reasoning over three-dimensional structural changes and the stability consequences of molecular interventions.

Benchmarks for Molecular and Structural Reasoning. Recent benchmarks have expanded chemical and molecular evaluation from general knowledge and reasoning to multimodal and structure-aware settings. ChemBench (Mirza et al., 2024), MaCBench (Alampara et al., 2024), ChEBI-20-MM (Liu et al., 2025a), and FGBench (Liu et al., 2025b) cover chemical knowledge, multimodal understanding, molecular representations, and functional-group-level reasoning. More recent work has begun to move beyond static understanding. MolLangBench (Cai et al., 2026) evaluates language-guided molecular recognition, editing, and generation across molecular strings, images, and graphs. AtomWorld (Lv et al., 2025) evaluates the execution of predefined atomistic structure modifications, while EvoMD-LLM (Tang et al., 2026) reformulates reactive molecular dynamics trajectories as symbolic molecular-event sequences for temporal prediction.

However, existing benchmarks provide little evaluation of reasoning directly from three-dimensional molecular structures about intervention-induced change, post-relaxation stability, and future structural outcomes. Our scope comparisons to prior work are shown in Table 1.

3 METHOD

3.1 BENCHMARK CONSTRUCTION

Figure 2 summarizes the construction pipeline of MOLSTRUCT. Starting from stratified QM9 source molecules, we apply controlled molecular interventions to obtain ideal edited structures, which are then evolved through geometry optimization or reactive molecular dynamics. The resulting structures and simulation records are subsequently used for task-specific candidate sampling and ground-truth construction in either CIF-text or image form.

Table 1: Comparison with related chemical and molecular benchmarks. 3D Perc.: image-based 3D structural perception; Traj.: trajectory-derived evaluation; Perturb.: structural editing or controlled perturbations; Evol.: prediction or assessment of structural evolution. Checks indicate explicit coverage in the evaluation tasks.

Benchmark	Capability	3D Perc.	Traj.	Perturb.	Evol.
ChemBench (Mirza et al., 2024)	Chemical Knowledge	×	×	×	×
MaCBench (Alampara et al., 2024)	Multimodal Understanding	✓	×	×	×
ChEBI-20-MM (Liu et al., 2025a)	Molecular Representations	×	×	×	×
FGBench (Liu et al., 2025b)	Functional-Group Reasoning	×	×	×	×
MolLangBench (Cai et al., 2026)	Molecular Recognition, Editing, and Generation	×	×	✓	×
AtomWorld (Lv et al., 2025)	CIF Perception and Structure Editing	×	×	✓	×
EvoMD-LLM (Tang et al., 2026)	Molecular-Event Sequence Prediction	×	✓	×	✓
Ours	**Structural Evolution Reasoning**	✓	✓	✓	✓

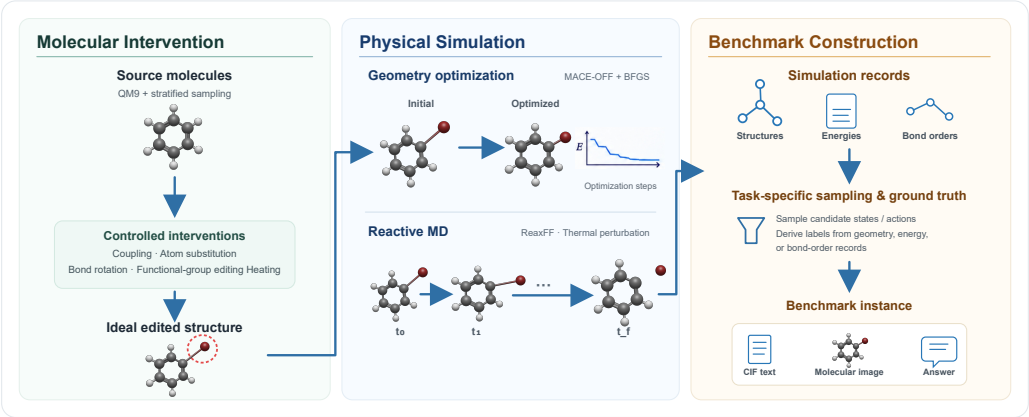


Figure 2: Overview of the MOLSTRUCT benchmark construction pipeline. Source molecules are sampled from QM9 and subjected to controlled molecular interventions to obtain ideal edited structures. Structural responses are generated through geometry optimization or reactive molecular dynamics, producing structures and simulation records used for task-specific sampling and ground-truth construction. The resulting instances are provided as CIF-text or rendered molecular images together with the corresponding evaluation targets.

Molecular Source and Stratified Sampling. We construct MOLSTRUCT primarily from molecules in the QM9 dataset (Ramakrishnan et al., 2014). Rather than sampling molecules uniformly or selecting a fixed prefix of the dataset, we build a stratified source pool to improve structural and chemical diversity. Each molecule is categorized according to five attributes: the number of heavy atoms, heavy-atom element composition, ring count, number of rotatable bonds, and number of heteroatoms. Molecules within each stratum are randomly shuffled using a fixed seed, after which round-robin sampling is performed across strata. This procedure prevents common small molecules and dominant element compositions from overwhelming less frequent structural patterns, while maintaining reproducibility across benchmark construction. Additional source molecules are sampled from the same pool when a particular task requires structural properties that are insufficiently represented in the initial selection.

Molecular Interventions. Starting from a source molecule S^{init}, we apply a controlled intervention a to construct an edited molecular structure,

$$S^{\text{ideal}} = \mathcal{A}\big(S^{\text{init}}, a\big)\,, \tag{1}$$

where $\mathcal{A}$ denotes a deterministic molecular editing operation. We consider five intervention classes: *atom substitution*, *functional-group editing*, *bond rotation*, *molecular coupling*, and *thermal perturbation*. Atom substitution changes the element type of a selected heavy atom while preserving its original index, connectivity, and coordinates. Functional-group editing modifies terminal substituents through operations such as replacement, addition, or deletion. Bond rotation selects a non-ring heavy-atom single bond and rotates one side of the molecule by a prescribed dihedral-angle

Table 2: Task taxonomy and statistics of MOLSTRUCT.

ID	Scientific	CS Capability	Answer Format	Num.	Modality
M1	Molecule	Structural Perception	Single choice	200	Image-only
M2	Molecule	Inverse Change Reasoning	SMILES	400	CIF / Image
M3	Molecule	Inverse Change Reasoning	Atom indices	400	CIF / Image
M4	Molecule	Stability Reasoning	Single choice	400	CIF / Image
M5	Molecule	Structural Prediction	Single choice	400	CIF / Image
Q1	Atom and Bond	Structural Perception	Atom index	200	Image-only
Q2	Atom and Bond	Stability Reasoning	Single choice	400	CIF / Image
Q3	Atom and Bond	Structural Perception	Single choice	200	Image-only
Q4	Atom and Bond	Inverse Change Reasoning	Atom-index pair	400	CIF / Image
Q5	Atom and Bond	Stability Reasoning	Single choice	400	CIF / Image
Q6	Atom and Bond	Structural Prediction	Boolean + bond list	400	CIF / Image
Q7	Atom and Bond	Structural Prediction	Single choice	300	CIF / Image
Q8	Functional Group	Structural Perception	Single choice	200	Image-only
Q9	Functional Group	Inverse Change Reasoning	Structured edits	400	CIF / Image
Q10	Functional Group	Stability Reasoning	Single choice	1000	CIF / Image
Q11	Functional Group	Structural Prediction	Single choice	400	CIF / Image
			Total	**6,100**	

increment. Molecular coupling connects two source molecules by removing one terminal hydrogen from each and introducing a new single bond between the corresponding attachment atoms. Thermal perturbation instead changes the molecular environment by applying a controlled heating protocol.

Structural Relaxation. For structural editing operations, S^{ideal} represents the structure immediately after the intervention has been applied, before the molecular geometry has adapted to the modification. It should therefore be interpreted as an ideal post-edit state rather than a physical reaction intermediate or transition state. For atom substitution, functional-group editing, bond rotation, and molecular coupling, the ideal edited structure is subsequently relaxed using MACE-OFF23-small (Kovács et al., 2025) together with BFGS geometry optimization from ASE (Hjorth Larsen et al., 2017). The relaxation trajectory is denoted as

$$\mathcal{T} = \left\{ S^{(0)}, S^{(1)}, \ldots, S^{(T)} \right\}, \qquad S^{(0)} = S^{\text{ideal}}, \tag{2}$$

where $S^{(T)}$ is the final relaxed structure. Thermal perturbation is handled separately using reactive molecular dynamics simulations in LAMMPS (Thompson et al., 2022) using the ReaxFF `ffield.reax.rdx` parameterization (Strachan et al., 2003), where the molecule is heated according to a predefined temperature protocol and time-resolved bond-order information is recorded. The resulting structures, optimization trajectories, energies, and bond-order records form the simulation outputs used for downstream task construction.

3.2 TASK TAXONOMY

MOLSTRUCT organizes molecular structural reasoning along two complementary axes: *structural scale* and *reasoning capability*. Along the first axis, tasks operate at three levels: **molecular**, **atom-and-bond**, and **functional-group**. Table 2 shows the detailed task taxonomy and statistics of our MOLSTRUCT. The molecular scale concerns global conformation, molecular assembly, and whole-structure outcomes. The atom-and-bond scale focuses on local geometric and connectivity changes, including atomic displacement, bond rotation, and bond breaking. The functional-group scale evaluates chemically meaningful local motifs and their interaction with the surrounding molecular environment. Table 3 further introduces the action space covered by each task family in MOLSTRUCT.

Along the second axis, we group the tasks into four capabilities of increasing difficulty: **structural perception**, **inverse change reasoning**, **stability reasoning**, and **structural prediction**. These capabilities correspond to different ways of querying the same intervention, *i.e.,* the relaxation process, as illustrated in Fig. 1.

Table 3: Action space covered by each task family in MOLSTRUCT.

Question ID(s)	Content	Actions
M1–M5	Molecular coupling	Terminal-H removal and inter-molecular bond formation.
Q1–Q2	Atom substitution	C→N, N→C, N→O, O→C, O→N.
Q3–Q5, Q7	Bond rotation	-120°, -60°, $+60^\circ$, $+120^\circ$, 180°.
Q6	Thermal perturbation	500, 750, 1000, and 1500 K.
Q8–Q11	Functional-group editing	F, Cl, Br, OH, NH_2, CH_3, OCH_3, CN, CHO, COOH, $CONH_2$.

Structural Perception. Structural perception tasks provide all structures required for comparison and ask the model to identify geometric changes directly from observed molecular configurations. At the molecular scale, four observed conformations of the same molecule are compared against a reference structure, and the ground-truth choice is the one with the largest RMS change over all heavy-atom pairwise distances. At the atom-and-bond scale, we include tasks that identify either the atom with the largest displacement after Kabsch alignment or the candidate dihedral with the largest periodic angle change. At the functional-group scale, candidate groups are locally aligned to the surrounding backbone, and the ground truth is the group whose connection direction changes the most. All perception tasks therefore derive their labels directly from geometric differences between provided structures and do not require prediction of an unseen state. The molecular deformation, dihedral-change, and functional-group tasks require selecting a single answer from a set of candidate structures, while atomic-displacement questions require the zero-based index of the target atom.

Inverse Change Reasoning. Inverse change reasoning asks the model to recover hidden molecular interventions from observed structures. At the molecular scale, tasks either infer a missing source molecule from the assembled product or recover the two attachment atoms used to couple two source molecules. At the atom-and-bond scale, the model is given the initial and relaxed structures after a hidden dihedral rotation and must identify the rotated central bond. At the functional-group scale, the model compares the initial and final structures and returns the edited attachment sites together with the minimal removed and added fragments. Ground-truth labels are taken from the stored intervention parameters or, for functional-group edits, from the minimal graph difference relative to the largest unchanged heavy-atom backbone. The required outputs follow the corresponding intervention target: missing molecules are returned as SMILES strings, molecular-coupling tasks return two attachment-atom indices, bond-rotation tasks return an unordered pair of atom indices, and functional-group editing tasks return a structured list of smallest edited sites together with the removed and added fragments. Chemically equivalent valid SMILES are accepted by parsing and canonicalizing the predicted strings with RDKit (rdk) before comparison.

Stability Reasoning. Stability reasoning presents several real relaxed states from the same molecular system and asks the model to identify the most stable configuration. At the molecular, atom-and-bond, and functional-group scales, the candidate structures are sampled from distinct states of a single BFGS relaxation trajectory after molecular coupling, atom substitution or bond rotation, and functional-group editing, respectively. We exclude the earliest portion of the trajectory, require candidates to come from distinct optimization steps, and retain only geometrically and energetically distinguishable states. For candidate structures $\{S_i\}_{i=1}^{K}$ with internally computed potential energies $\{E_i\}_{i=1}^{K}$, the ground-truth labels are

$$y_{\text{stable}} = \arg\min_i E_i. \tag{3}$$

The energy values are used only to derive the labels and are never exposed to the evaluated model, so the task directly evaluates structure-to-stability reasoning. All stability-reasoning tasks require selecting a single answer from a set of candidate structures

Structural Prediction. Structural prediction hides the post-intervention outcomes and asks the model to anticipate them from the initial structure and candidate actions. At the molecular scale, two candidate attachment sites from each of two source molecules form four coupling choices; each product is independently relaxed, and the ground truth is the coupling whose converged endpoint has the lowest potential energy. At the atom-and-bond scale, tasks either compare four candidate dihedral rotations through independent relaxation or predict temperature-induced bond breaking from reactive molecular dynamics. At the functional-group scale, the same substituent is placed

at four non-equivalent sites, each product is independently relaxed, and the ground-truth site is the one yielding the most stable endpoint. For all stability-prediction tasks, the relaxed candidate outcomes and their energies remain hidden from the model. Molecular coupling, bond-rotation, and functional-group-site prediction are formulated as as single-choice questions over a set of candidate structures. Thermal prediction instead returns whether bond breaking occurs together with the set of broken initial bonds.

3.3 TEXT AND MULTIMODAL REPRESENTATIONS.

MOLSTRUCT provides both text-based and multimodal evaluation settings. Internally, molecular structures are stored in SDF format to preserve atom types, coordinates, and bond information. For text-based evaluation, structures are converted to coordinate-only CIF, which contains atomic species and three-dimensional coordinates but does not explicitly provide molecular connectivity, except where required by the task definition. Each atom is labeled by its element symbol and zero-based index, allowing model predictions to refer consistently to specific atoms and bonds.

For multimodal evaluation, molecular structures are rendered in a ball-and-stick representation with atom-specific colors and index labels. Each structure is shown from two complementary three-dimensional viewpoints, and all structures within the same question share a common rendering scale. Structural perception tasks are evaluated in the image-only setting, whereas inverse reasoning, stability reasoning, and prediction tasks are provided in both CIF-text and image-only variants.

3.4 BENCHMARK STATISTICS AND EVALUATION METRICS.

We use task accuracy as the primary evaluation metric. A response is counted as correct after task-specific parsing and normalization, including canonical SMILES comparison for molecular reconstruction and order-invariant matching for unordered bonds or edit sets. For a task containing N examples, accuracy is defined as

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i). \tag{4}$$

We use accuracy as the primary evaluation metric and report results at both the task level and the capability level. For capability-level evaluation, we macro-average the accuracies of all tasks belonging to the same reasoning capability, such that each task contributes equally regardless of its number of samples. For tasks evaluated in both CIF-text and image modalities, results are reported separately for the two representations. Structural perception is evaluated only in the image setting.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate a diverse set of foundation models spanning both open-weight and proprietary model families. The open-weight models include Gemma 4-31B (Team et al., 2026), InternVL3.5-30B-A3B (Wang et al., 2025), DeepSeek-V4.1-Flash (Xu et al., 2026), Qwen3.6-35B-A3B (Qwen Team, 2026), and Qwen3-8B (Yang et al., 2025), while the proprietary models include Gemini 3.7 Flash (Google DeepMind, 2026), Claude Opus 4.7 (Anthropic, 2026), and GPT-5.5 (OpenAI, 2026). The selected models cover a range of model scales, architectural families, and reasoning capabilities, allowing us to compare molecular structural reasoning comprehensively.

For the text-based evaluation, we test all eight models using the CIF-based molecular representation. For the multimodal evaluation, we evaluate all models except Qwen3-8B, which is used only in the text setting. Whenever a model provides an explicit thinking or reasoning mode, we enable it throughout the evaluation to assess its strongest reasoning configuration. All models are evaluated under the same task instructions and output constraints.

Evaluation Protocol. We evaluate all models in a zero-shot setting using standardized task-specific prompts and fixed output formats. Structural perception tasks are evaluated only in the image setting for difficulty reasons, while inverse change reasoning, stability reasoning, and structural

Table 4: Overall performance on MOLSTRUCT. We report task-macro-averaged accuracy (%) for each reasoning capability. T and I denote CIF-text and image modalities, respectively. Structural perception is evaluated only in the image setting.

Model	**Perception**	**Inverse**		**Stability**		**Prediction**		**Avg.**
	I	T	I	T	I	T	I	
Gemma 4-31B	25.5	1.0	1.0	1.8	13.1	20.6	19.8	11.8
InternVL3.5-30B-A3B	19.8	0.4	0.1	5.8	8.0	21.5	16.6	10.3
DeepSeek-V4.1-Flash	23.8	1.0	0.9	4.4	15.4	21.8	17.1	12.1
Qwen3.6	22.9	3.5	1.5	4.1	16.7	18.6	20.0	12.5
Qwen3-8B	–	0.8	–	9.0	–	18.8	–	9.5
Claude Opus 4.7	21.4	0.9	0.5	7.0	24.9	20.4	17.4	13.2
Gemini 3.7 Flash	35.5	65.8	24.8	44.5	25.1	34.9	34.1	37.8
GPT-5.5	35.9	37.1	26.1	27.8	24.4	31.6	30.3	26.7
Expert	-	-	-	-	-	-	-	87.3

prediction are evaluated in both CIF-text and image settings. For tasks evaluated in both modalities, the CIF-text and image versions share the same underlying sample and ground-truth answer.

The evaluated models are given only the information explicitly specified by each task. In particular, potential energies, optimization metadata, simulation trajectories, and ground-truth annotations are never exposed unless they form part of the task input by design. Model outputs are parsed according to the task-specific answer schema, including single-choice indices, atom or bond indices, SMILES strings, structured functional-group edits, and bond-breaking predictions. Invalid or unparsable responses are counted as incorrect and are additionally reflected in the valid response rate.

4.2 OVERALL PERFORMANCE

We first compare the evaluated models from two complementary perspectives. Table 4 reports task-macro-averaged accuracy across the four capabilities in MOLSTRUCT: structural perception, inverse change reasoning, stability reasoning, and structural prediction. For capabilities available in both modalities, CIF-text and image results are reported separately to reveal the effect of molecular representation. The results reveal a substantial performance gap across model families. Gemini 3.7 Flash and GPT-5.5 achieve the strongest overall performance, whereas most open-weight models remain at relatively low accuracy across the benchmark. Across capabilities, structural perception is generally more accessible than stability reasoning and structural prediction, while inverse change reasoning shows large variation across models. Models are relatively more reliable at structural perception, while stability reasoning remains substantially more difficult. Inverse change reasoning is particularly model-dependent: some models perform well in the CIF-text setting, whereas others remain close to chance or fail to recover the edited structure reliably. Structural prediction also remains challenging, especially when the task requires anticipating the relaxed outcome of an intervention.

We further group the tasks by structural scale in Table 5 (a), covering molecules, atoms and bonds, and functional groups. Here, text and image results are combined to provide a compact view of model performance across different molecular scales. Functional-group-level tasks are often easier for the stronger models, while atom-and-bond-level reasoning is still difficult in several settings.

4.3 ANALYSIS AND FINDINGS

Bias toward unrelaxed structures. To further understand the failure modes in stability reasoning, we examine which candidate is selected when models are asked to identify the most stable structure. Interestingly, many models show a strong tendency to select the candidate that remains closest to the directly edited, pre-relaxation structure, even though this candidate corresponds to the least stable state among the provided options. Table 5 (b) shows that this behavior is particularly pronounced for DeepSeek-V4.1-Flash, Gemma 4, Qwen3.6, and Claude Opus 4.7, for which the average selection rate exceeds 65%.

This demonstrates that the models lack the ability to assess structural stability; instead, they simply tend to select an initial structure that resembles the one provided in the prompt. Consequently,

(a) Performance across molecular scales				(b) Preference for unrelaxed structures				
Model	**Molecule**	**Atom & Bond**	**Functional Group**	**M4**	**Q2**	**Q10**	**Q5**	**Avg.**
Gemma 4-31B	9.3	11.0	16.5	72.5	76.0	72.4	78.5	74.9
InternVL3.5-30B-A3B	9.7	7.5	15.9	57.0	74.0	58.6	72.4	65.5
DeepSeek-V4.1-Flash	10.3	10.5	17.0	72.5	83.5	73.2	81.0	77.6
Qwen3.6	10.9	12.5	14.3	67.5	69.5	65.6	69.0	67.9
Qwen3-8B	7.3	11.0	10.0	53.0	52.0	44.4	47.5	49.2
Claude Opus 4.7	10.8	12.5	17.5	62.0	77.5	56.4	72.5	67.1
Gemini 3.7 Flash	35.1	29.4	55.6	41.5	21.5	36.2	39.5	34.7
GPT-5.5	26.2	29.6	42.8	59.0	56.4	48.2	39.5	40.8

Table 5: Analysis of molecular-scale reasoning and stability bias on MOLSTRUCT. **(a)** Task-modality macro-averaged accuracy (%) across three molecular scales. **(b)** Percentage of stability predictions selecting the candidate closest to the directly edited, unrelaxed structure, corresponding to the least stable option among the provided relaxed states. Higher values indicate a stronger tendency to favor the immediate post-intervention geometry.

even though the task involves multiple-choice questions, the accuracy of models other than Gemini fell below that of random guessing. For the same reason, accuracy on the image-based version of the question was higher than on the corresponding CIF version: because the models struggled to evaluate structural similarity based on images, they avoided repeatedly selecting the same incorrect initial answer, resulting in accuracy levels approaching those of random guessing.

Error Decomposition in SMILES Reconstruction. We further analyze the error patterns in the inverse reconstruction task that requires models to recover the missing source molecule as a SMILES string. The failures exhibit a clear hierarchy across model families. For Qwen3.6, 23.5% of predictions are not valid SMILES strings, indicating that a substantial fraction of errors already occurs at the level of output syntax. DeepSeek-V4.1-Flash and Qwen3-8B more often produce syntactically valid SMILES, but the predicted molecules frequently have an incorrect molecular formula. In contrast, stronger models such as Gemini 3.7 Flash usually recover the correct molecular formula, with the remaining errors mainly arising from confusion between constitutional isomers.

Structured-Output Compliance. We also observe substantial differences in models' ability to follow structured-output requirements. On Q9, 82.0% of InternVL3.5 responses and 61.5% of Qwen3-8B responses are unparsable because required fields, particularly the `edits` array, are missing. Qwen3.6 exhibits the same issue in 23.0% of cases, while the other models rarely fail at this parsing stage. Even when explicitly instructed to return only the final answer, several models, *e.g.,* Gemma 4 and Qwen3.6, frequently prepend substantial analysis or explanatory text before the structured prediction. To avoid penalizing otherwise valid predictions for such extra text, we use a permissive parser that extracts the final complete JSON object from the response whenever possible.

5 CONCLUSION

We introduce MOLSTRUCT, a benchmark for evaluating multi-scale molecular structural reasoning through the consequences of controlled molecular interventions. Rather than treating structural editing as the endpoint, MOLSTRUCT evaluates how models reason about the subsequent structural response and relaxation across molecular, atom-and-bond, and functional-group scales. The benchmark covers four complementary capabilities: structural perception, inverse change reasoning, stability reasoning, and structural prediction, under both text-based and multimodal representations. Our experiments on the proposed MOLSTRUCT reveal a clear limitation of current foundation models in multi-scale molecular structure understanding, reasoning and prediction. We hope MOLSTRUCT provides a useful benchmark for developing models that connect molecular structure with physical response, moving foundation models from describing molecular configurations toward reasoning about how those configurations evolve under intervention.

AI USE STATEMENT

We used OpenAI Codex to assist with experimental and evaluation code, including parts of the implementation for benchmark construction and ground-truth computation. The benchmark design and ground-truth generation logic were developed by the authors, and reference answers were computed programmatically from molecular structures, intervention records, and simulation outputs rather than generated through LLM inference. We also used ChatGPT for language polishing and editing of the manuscript. All AI-assisted code and manuscript content were reviewed by the authors, who take responsibility for the final text, claims, code, experimental results, and artifacts presented in this work.

REFERENCES

Rdkit: Open-source cheminformatics. https://www.rdkit.org. Accessed: 2026.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research, 2024. URL https://arxiv.org/abs/2411.16955.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, August 2025. ISSN 2662-8457. doi: 10.1038/s43588-025-00836-3. URL http://dx.doi.org/10.1038/s43588-025-00836-3.

Anthropic. Claude opus 4.7 system card. System Card, 2026.

Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, 2024.

Feiyang Cai, Jiahui Bai, Tao Tang, Guijuan He, Joshua Luo, Tianyu Zhu, Srikanth Pilla, Gang Li, Ling Liu, and Feng Luo. Mollangbench: A comprehensive benchmark for language-prompted molecular structure recognition, editing, and generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Kamal Choudhary. Atomgpt: Atomistic generative pretrained transformer for forward and inverse materials design. *The Journal of Physical Chemistry Letters*, 15(27):6909–6917, 06 2024. ISSN 1948-7185. doi: 10.1021/acs.jpclett.4c01126. URL https://doi.org/10.1021/acs.jpclett.4c01126.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net, 2024. URL https://openreview.net/pdf?id=Tlsdsb6l9n.

Google DeepMind. Gemini 3.7 flash model card. Model Card, August 2026.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Yixuan Pu, Venkat Kapil, William C Witt, Ioan-Bogdan Magdau, Daniel J Cole, et al. Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025.

Pengfei Liu, Jun Tao, and Zhixiang Ren. A quantitative analysis of knowledge-learning preferences in large language models in molecular science. *Nature Machine Intelligence*, pp. 1–13, 2025a.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, Dec 2023a. ISSN 2522-5839. doi: 10.1038/s42256-023-00759-6. URL https://doi.org/10.1038/s42256-023-00759-6.

Xuan Liu, Siru Ouyang, Xianrui Zhong, Jiawei Han, and Huimin Zhao. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *arXiv preprint arXiv:2508.01055*, 2025b.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP*, 2023b. URL https://openreview.net/forum?id=14WRhMNq7H.

Taoyuze Lv, Alexander Chen, Fengyu Xie, Chu Wu, Jeffrey Meng, Dongzhan Zhou, Bram Hoex, Zhicheng Zhong, and Tong Xie. Atomworld: A benchmark for evaluating spatial reasoning in large language models on crystalline materials, 2025. URL https://arxiv.org/abs/2510.04704.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Amir Mohammad Elahi, Mehrdad Asgari, Juliane Eberhardt, Hani M. Elbeheiry, María Victoria Gil, Maximilian Greiner, Caroline T. Holick, Christina Glaubitz, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024.

Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. Llm4mat-bench: benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, 2025.

OpenAI. Gpt-5.5 system card. System Card, April 2026.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J Kim. Llamo: Large language model-based molecular graph assistant. In *NeurIPS*, 2024.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):140022, 2014.

Alejandro Strachan, Adri CT van Duin, Debashis Chakraborty, Siddharth Dasgupta, and William A Goddard III. Shock waves in high-energy materials: The initial chemical events in nitramine rdx. *Physical Review Letters*, 91(9):098301, 2003.

Zhichen Tang, Zhengzheng Dang, Yulin Chen, Jixin Wu, Haiwen Li, and Yanming Wang. EvoMD-LLM: Learning the language of species evolution in reactive molecular dynamics. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 39072–39088, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1947. URL https://aclanthology.org/2026.findings-acl.1947/.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Aidan P Thompson, H Metin Aktulga, Richard Berger, Dan S Bolintineanu, W Michael Brown, Paul S Crozier, Pieter J In't Veld, Axel Kohlmeyer, Stan G Moore, Trung Dac Nguyen, et al. Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer physics communications*, 271:108171, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Appendix

You may include other additional sections here.