# MolOpt-Eval: Can Frontier LLMs Perform Structure-Based Hit-to-Lead Optimization?

**Chengzhu Li** [*1,2] **Haichuan Tan** [*1,3] **Wenyu Zhu** [1] **Bowen Gao** [1,3] **Jiqing Zheng** [4] **Ya-Qin Zhang** [1] **Wei-Ying Ma** [1] **Yanyan Lan** [1,5]

## Abstract

Structure-based molecular optimization—the iterative editing of hit molecules to enhance binding affinity while preserving pocket geometry constraints—is a fundamental task in computer-aided drug discovery. Despite the rapid progress of Large Language Models (LLMs) in scientific reasoning, their efficacy as molecular optimizers remains under-explored. In this work, we first demonstrate through end-to-end experiments that frontier LLMs exhibit subpar optimization performance: affinity gains remain marginal, and high-affinity leads frequently deteriorate upon modification. To diagnose *where* in the reasoning chain these failures originate, we introduce **MolOpt-Eval**, a diagnostic benchmark that decomposes structure-based molecular optimization into three independently evaluable cognitive stages—Structural Perception, Strategy Discovery, and Strategy Execution—supplemented by chain-of-thought quality analysis. Evaluating 14 frontier LLMs across 30 DUD-E protein targets with 10 diagnostic tasks, we reveal that: (i) LLMs achieve high accuracy on 2D molecular structure (micro-F1 = 0.92) and protein fold classification (up to 90%), yet 3D interaction perception drops sharply (F1 $<$ 0.40); (ii) proposed strategies are chemically plausible (>93%) and produce valid molecules (>95% SMILES validity), but fewer than 2.5% achieve their intended interactions when validated through Boltz-2 co-folding; (iii) model capability differences vanish on the hardest tasks, suggesting fundamental paradigm limitations rather than model-specific deficiencies; and (iv) distance sensitivity experiments confirm that LLMs do process spatial information but cannot translate this awareness into accurate interaction predictions. Our findings identify 3D spatial reasoning as the critical bottleneck and provide actionable guidance for developing structure-aware molecular foundation models. The code repository is available at: `https://anonymous.4open.science/r/MolOpt_Eval-1292/`.

## 1. Introduction

Structure-based molecular optimization is a central task in the drug discovery pipeline.(Bleicher et al., 2003; Hughes et al., 2011) During the hit-to-lead stage, medicinal chemists modify an initial hit compound to improve its binding affinity to a target protein under the geometric constraints of its binding pocket. (Cavasotto & W Orry, 2007) Each such modification is the outcome of a structured cognitive process: the chemist first reads the three-dimensional interaction pattern between the current ligand and the pocket, then locates an opportunity within the structural environment where a new interaction can be installed or a weak one strengthened, formulates a chemically feasible modification strategy (*e.g.*, introducing a hydrogen bond donor at a specific atom position to engage a neighboring residue), and finally realizes the strategy as a precise molecular edit. This cognitive chain rests jointly on deep medicinal chemistry knowledge, spatial reasoning, and accurate perception of the structural environment—capabilities that have long remained the domain of human experts. (Bissantz et al., 2010)

Recent LLMs have achieved breakthrough performance in mathematics, code generation, and scientific reasoning, naturally raising the question of whether they can support structure-based molecular optimization. As an initial probe, we conducted an end-to-end hit-to-lead optimization experiment on 30 real protein–ligand complexes from DUD-E

[*]Equal contribution [1]Institute for AI Industry Research, Tsinghua University, Beijing, China [2]Zhili College, Tsinghua University, Beijing, China [3]Department of Computer Science and Technology, Tsinghua University, Beijing, China [4]Department of Chemistry, Tsinghua University,Beijing, China [5]Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, China. Correspondence to: Yanyan Lan <lanyanyan@air.tsinghua.edu.cn>.

*Accepted at the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026)*

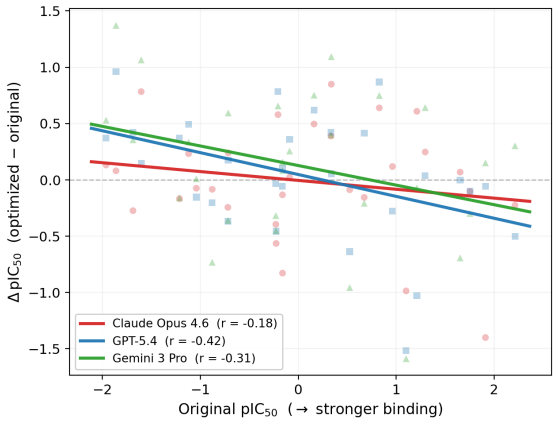


*Figure 1.* End-to-end affinity optimization test. Each point denotes one target–model pair, with the $y$-axis showing the change in Boltz-2 predicted $\mathrm{pIC}_{50}$. The negative regression slopes indicate that improvements concentrate on low-affinity ligands, high-affinity ligands often deteriorate, consistent with regression to the mean rather than directed optimization. Details are provided in Appendix A.

using three frontier LLMs. Given the crystal ligand and structure-derived context of the binding pocket, each model was asked to propose a modified molecule with improved binding affinity; the resulting molecule was then co-folded with the target protein using Boltz-2 to estimate its binding affinity. The results were sobering (Figure 1). Across all models, affinity changes were marginal, and high-affinity ligands often deteriorated after modification. End-to-end optimization failures may arise from distinct sources: inadequate structural perception, chemically or spatially invalid strategy formulation, or failure to execute otherwise reasonable strategies as concrete molecular edits. Disentangling these failure modes is essential, as each reflects a different capability deficit and suggests a different path for improvement.

To this end, we introduce **MolOpt-Eval**, a diagnostic benchmark for evaluating LLM capabilities in structure-based molecular optimization at the level of individual reasoning steps. As illustrated in Figure 2, MolOpt-Eval decomposes the optimization pipeline into three sequential stages: **Structural Perception**, which evaluates understanding of molecular composition, protein structure, and ligand–protein interactions; **Strategy Discovery**, which assesses whether models can propose chemically feasible and spatially grounded modification plans; and **Strategy Execution**, which tests whether models can translate these plans into valid molecular edits with the intended structural effects. In addition, MolOpt-Eval includes an orthogonal **Inference Quality** analysis that examines chain-of-thought (CoT) traces for factual errors, logical inconsistencies, and reasoning–execution misalignment. Together, the stage-wise tasks localize *where* failures occur, while the CoT analysis characterizes *how* the underlying reasoning breaks down. MolOpt-Eval consists of 10 diagnostic tasks across 30 structurally diverse protein targets, and is evaluated on frontier LLMs from the GPT, Claude, Gemini, Kimi, and GLM families under unified experimental conditions.

Applying MolOpt-Eval, we obtain three high-level findings; detailed are provided in Section 3.

1. **LLMs handle individual molecular and protein representations, but fail at ligand–protein interaction reasoning.** Performance remains strong on 2D molecular substructures and protein secondary structure, but drops sharply on 3D ligand–pocket interaction perception.

2. **Chemically plausible edits rarely produce the intended structural effects.** Models can propose reasonable strategies and generate mostly valid molecules, yet the intended ligand–residue interactions almost never materialize after structure-based validation.

3. **LLMs are spatially responsive but not structurally predictive.** Models respond to changes in ligand–pocket distances, but this sensitivity does not translate into reliable prediction or realization of binding interactions.

Together, these findings suggest that the core bottleneck is not general chemical knowledge or SMILES-level editing, but the ability to convert 3D ligand–pocket context into physically meaningful optimization decisions.

Together, these findings identify 3D ligand–pocket reasoning as the central bottleneck in LLM-driven molecular optimization, rather than general chemical knowledge or SMILES-level editing ability. Our contributions are threefold:

- **A diagnostic benchmark for structure-based molecular optimization.** We introduce MolOpt-Eval, which decomposes LLM-driven hit-to-lead optimization into 10 independently evaluable tasks spanning structural perception, strategy discovery, strategy execution, and inference quality.

- **A systematic evaluation of frontier LLMs.** We benchmark models from the GPT, Claude, Gemini, Kimi, and GLM families across 30 structurally diverse DUD-E targets under unified experimental conditions.

- **An empirical characterization of the key bottleneck.** We show that current LLMs can handle many chemistry and molecular-editing operations, but fail to convert 3D ligand–pocket context into physically meaningful optimization decisions.

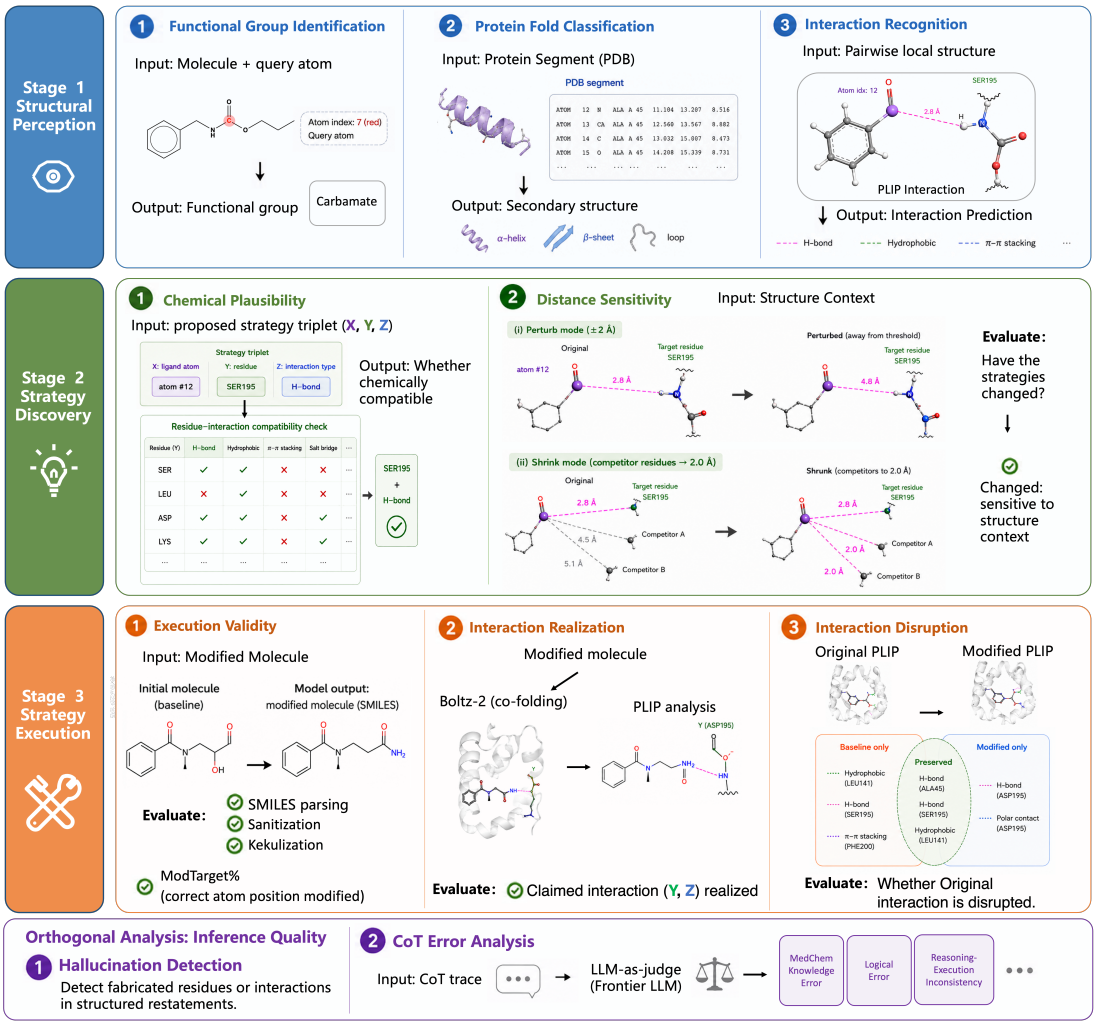


*Figure 2.* Overview of **MolOpt-Eval**. MolOpt-Eval evaluates LLMs for structure-based molecular optimization by decomposing the reasoning pipeline into three sequential stages: structural perception, optimization strategy formulation, and chemical execution. Each stage is evaluated with independent task designs and metrics, enabling fine-grained diagnosis of where and how LLM reasoning breaks down. In addition, inference-quality analysis further examines hallucination and chain-of-thought consistency across the reasoning process.

## 2. MolOpt-Eval Benchmark Design

### 2.1. Overview and Problem Formulation

We focus on *structure-based affinity optimization*: improving the binding affinity of an initial hit molecule to a given protein target under the geometric and physicochemical constraints of its binding pocket. Unlike general molecular property optimization, this task requires not only generating chemically valid molecules, but also reasoning over 3D ligand–pocket interactions that determine binding strength.

We formulate hit-to-lead affinity optimization as a multi-step reasoning process. Given an initial molecule $m_0$, a protein target $P$, and structure-derived ligand–pocket context $C$, the model must perceive the structural environment, formulate an optimization strategy, and execute it as a molecular edit:

$$(m_0, P, C) \xrightarrow{\text{Perception}} S \xrightarrow{\text{Strategy}} A \xrightarrow{\text{Execution}} m^*, \quad (1)$$

where $S$ denotes perceived structural information, $A = \{a_1, \ldots, a_k\}$ denotes candidate modification strategies, and $m^*$ is the optimized molecule.

Accordingly, MolOpt-Eval decomposes the task into three stage-wise components: *Structural Perception*, *Strategy Discovery*, and *Strategy Execution*. In addition, we analyze the reasoning trajectory $T$ through an orthogonal *Inference Quality* module, which diagnoses factual errors, logical inconsistencies, and reasoning–execution misalignment. Each component is detailed in the following subsections.

### 2.2. Data Construction

MolOpt-Eval is constructed from DUD-E, from which we select 30 protein targets spanning diverse therapeutic families and binding-site topologies. For each target, we use the co-crystallized ligand and four additional active compounds selected by molecular-diversity sampling, yielding 150 active protein–ligand complexes as shared inputs across the benchmark. For each complex, we use PLIP to generate programmatic ligand–protein interaction annotations, focusing on four common non-covalent interaction types: hydrogen bonds, hydrophobic contacts, $\pi$-stacking, and salt bridges. These annotations are included in the model input as structure-derived context, providing local ligand–pocket interaction information for each complex.

To evaluate strategy execution independently from strategy generation stage, we additionally construct a standardized set of expert-curated optimization strategies. Each strategy is represented as a triplet $(X, Y, Z)$, specifying the ligand atom position $X$, target residue $Y$, and intended interaction type $Z$. All models receive the same strategy set, ensuring that differences in execution performance reflect molecular editing ability rather than variation in strategy quality. Details of target selection, PLIP annotation criteria, and strategy construction are provided in Appendix B.

### 2.3. Structural Perception

The perception stage corresponds to the mapping $(m_0, P, C) \rightarrow S$: extracting structured information from raw inputs. We decompose $S$ into three components of increasing difficulty—molecular substructure features (functional groups), protein structural elements (secondary structure), and ligand–protein interaction triples of the form $(\text{atom\_idx}, \text{residue}, \text{interaction\_type})$. This yields three evaluation tasks.

**Task 1.1: Functional Group Identification.** Given a molecule represented by its SMILES string, along with atom-level indexing and a queried atom index, the model is required to identify the functional group to which the atom belongs and output it in SMILES format. To ensure objective evaluation, we employ RDKit-based matching to verify whether the predicted functional group corresponds to a valid substructure containing the queried atom. Predictions are considered correct if the generated SMILES can be matched to the ground-truth substructure. This task evaluates the model's ability to parse molecular graphs and reason about local chemical environments.

**Task 1.2: Protein Fold Classification.** Given a segment of a protein structure in PDB format, the model is asked to identify its structural fold type (*e.g.*, alpha-helix, beta-sheet, loop). Ground-truth labels are derived from secondary structure annotations, and model predictions are evaluated using classification accuracy. This task assesses whether the model can interpret protein structural patterns from coordinate-based representations.

**Task 1.3: Interaction Recognition.** Given explicit descriptions of neighboring atoms and residues in a protein–ligand complex, the model is required to determine whether specific intermolecular interactions (*e.g.*, hydrogen bonds, hydrophobic contacts) are present. Ground-truth interaction labels are obtained from PLIP reports. Model predictions are evaluated by comparing predicted interaction types and existence against the reference annotations. This task evaluates the model's ability to reason about spatial proximity and physicochemical compatibility underlying molecular interactions.

### 2.4. Strategy Discovery

The strategy stage corresponds to the mapping $S \rightarrow A = \{a_1, \ldots, a_k\}$. Each strategy $a_i$ is a structured triplet $(X, Y, Z)$, where $X$ specifies the target atom position on the ligand, $Y$ identifies the target protein residue, $Z$ defines the intended interaction type (*e.g.*, hydrogen bond, hydrophobic contact). This triplet representation makes each strategy independently verifiable: $X$ and $Y$ can be checked against the structural context, $Z$ can be validated for chemical compatibility with residue $Y$.

Given a protein–ligand complex and its PLIP interaction report, the model is prompted to propose up to 3 optimization strategies as $(X, Y, Z)$ triplets. Each protein–ligand pair is evaluated across 5 independent runs to capture output variability. The raw strategy outputs then serve as input to two evaluation tasks that assess distinct aspects of strategy quality.

**Task 2.1: Chemical Plausibility.** This task evaluates whether the proposed strategies are chemically valid at the residue level—specifically, whether the target residue $Y$ can plausibly provide the interaction type $Z$. For example, proposing a salt bridge with a residue that carries no formal charge, or a hydrogen bond with a residue lacking donor/acceptor groups, constitutes a plausibility failure. We report the plausibility rate: the fraction of $(Y, Z)$ pairs that are chemically compatible according to residue-level interaction rules.

**Task 2.2: Distance Sensitivity.** This task tests whether the model genuinely processes 3D spatial information or relies on chemical knowledge templates. We introduce two controlled perturbations to the structural inputs. In the *perturb* setting, we shift the distance between the target residue and ligand by $\pm 2$ Å away from the interaction threshold and measure RetainDrop, the decrease in the recommendation

rate of the target residue after perturbation. In the *shrink* setting, we compress competing residue distances to 2.0 Å while keeping the target residue unchanged, and measure whether models shift their recommendations toward the artificially closer competitors (CompPkp). This task captures whether the model exhibits genuine structure-aware reasoning, as opposed to relying on superficial chemical patterns. Detailed perturbation protocols and metrics are provided in Appendix **??**.

### 2.5. Strategy Execution

The execution stage corresponds to the mapping $(m_0, a_i) \rightarrow m_i^*$: for each strategy $a_i = (X, Y, Z)$, the model edits $m_0$ to produce a modified molecule $m_i^*$ intended to install or strengthen interaction $Z$ between atom position $X$ and residue $Y$.

Given the baseline ligand, its PLIP interaction report, and a curated strategy triplet, the model is prompted to output a modified SMILES that implements the strategy. All models receive the same set of strategies to ensure fair comparison. The modified molecules are then evaluated through three tasks that assess progressively deeper aspects of execution quality—from syntactic validity to structural realization.

**Task 3.1: Execution Validity.** We assess whether the generated SMILES corresponds to a chemically valid molecule through a three-stage pipeline: SMILES parsing, sanitization, and Kekulization (RDKit). We additionally report ModTarget%—the fraction of modifications that target the correct atom position specified in the strategy—to measure strategy comprehension independently of chemical validity.

**Task 3.2: Interaction Realization.** This is the most demanding evaluation in the benchmark. Each valid modified molecule is co-folded with its target protein using Boltz-2, and the resulting 3D structure is analyzed with PLIP. We then check whether the LLM-claimed interaction—the specific (residue $Y$, interaction type $Z$) pair—actually appears in the co-folded structure. The realization rate measures the fraction of strategies whose intended interactions materialize in the physics-based prediction.

**Task 3.3: Baseline Disruption.** Molecular modifications should not only create new interactions but also preserve existing favorable ones. We compare the PLIP interaction set of the baseline ligand (co-folded separately) against that of the modified molecule. The retention rate measures the fraction of baseline interactions that are preserved after modification. Low retention indicates that the modification disrupts the existing binding mode, even if it successfully introduces new contacts.

### 2.6. Inference Quality

Beyond stage-wise performance, we analyze the reasoning trajectory $T$ to diagnose *how* model reasoning breaks down. This module complements the preceding tasks, which localize *where* failures occur in the optimization pipeline.

**Task 4.1: Hallucination Detection.** Given a PLIP interaction report and pocket-neighbor descriptions, the model is asked to restate the provided structural information in a structured JSON format. We compare the output against the input context and measure the fraction of fabricated entities, including nonexistent interactions and residues. The resulting hallucination rate quantifies whether models introduce unsupported structural details during information processing.

**Task 4.2: Chain-of-Thought Error Analysis.** We use an LLM-as-judge protocol to evaluate model-generated CoT traces. For each sample, the reasoning chain and final output are presented to a judge model for binary error classification across five categories: medicinal chemistry knowledge, environment information, logic, reasoning–execution consistency, and representation. We report the error rate for each category, with detailed definitions and judging criteria provided in Appendix **??**.

## 3. Experimental Results and Analysis

We evaluate 14 frontier LLMs spanning five families: Claude (Opus 4.6, Sonnet 4.6, plus Thinking variants of each), GPT (GPT-5, GPT-5.4, GPT-5.4 Mini), Gemini (2.5 Flash, 2.5 Pro, 3 Flash, 3 Pro), GLM (GLM-5, GLM-5.1), and Kimi K2.5. All models are accessed through their APIs under identical prompts and default sampling parameters. Strategy execution outputs are structurally validated through Boltz-2 co-folding. CoT quality is assessed using Claude Opus 4.6 as an LLM-as-judge. Full prompt templates, decoding settings, and per-target breakdowns are provided in the Appendix. (Comanici et al., 2025; Singh et al., 2025) We present results organized by pipeline stage. For each stage, we report key metrics in a summary table and highlight the most significant findings.

### 3.1. Stage 1: Structural Perception

**Functional Group Identification.** LLMs demonstrate strong competence in 2D molecular structure recognition. The best model (Claude Opus 4.6) achieves micro-F1 = 0.920, and 10 of 14 models exceed 0.84. All models exhibit higher recall than precision (see Appendix), indicating a systematic tendency toward over-prediction rather than omission—models would rather report a plausible functional group than miss one.

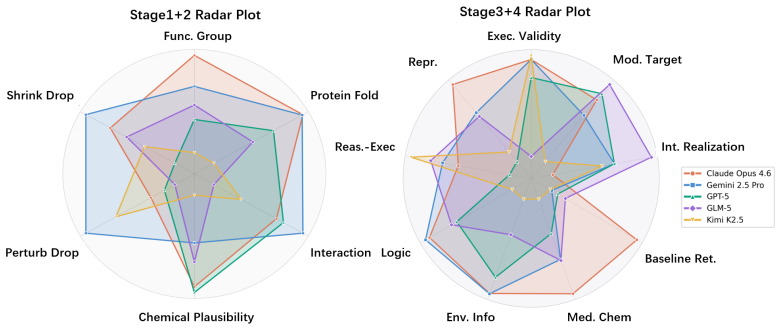


*Figure 3.* Radar plots compare the best-performing model from each vendor across selected metrics in Stage 1&2 (left) and Stage 3&4 (right).

*Table 1.* Results for Stage 1 structural perception and Stage 2 strategy discovery. Func. Group reports micro-F1 for functional group identification, Protein Fold reports accuracy for secondary-structure classification, and Interaction reports macro-F1 for ligand–protein interaction perception. Chem. Plaus. denotes the fraction of chemically compatible residue–interaction pairs. Perturb Drop and Shrink Drop measure changes in recommendation rates under controlled distance perturbations. All values are percentages.

| Model | Stage 1: Structural Perception | | | Stage 2: Strategy Discovery | | |
|---|---|---|---|---|---|---|
| | Func. Group | Protein Fold | Interaction | Chem. Plaus. | Perturb Drop | Shrink Drop |
| Claude Sonnet 4.6 | 89.9 | 87.2 | 36.4 | 98.8 | 43.9 | 46.3 |
| w/ Thinking | 90.0 | 88.3 | 33.3 | 98.9 | 42.4 | 48.1 |
| Claude Opus 4.6 | 92.0 | 90.5 | 27.7 | 99.2 | 38.6 | 57.3 |
| w/ Thinking | 91.8 | 90.2 | 33.6 | 99.3 | 41.6 | 57.5 |
| Gemini 2.5 Flash | 84.1 | 66.5 | 27.7 | 96.6 | 38.9 | 41.5 |
| Gemini 2.5 Pro | 87.4 | 72.1 | 32.4 | 98.9 | 42.7 | 39.7 |
| Gemini 3 Flash | 89.0 | 88.8 | 33.7 | 93.6 | 47.0 | 61.5 |
| Gemini 3 Pro | 90.5 | 90.2 | 39.4 | 96.9 | 49.2 | 64.7 |
| GPT-5.4 Mini | 75.8 | 44.4 | 27.4 | 97.7 | 11.1 | 20.3 |
| GPT-5 | 88.9 | 78.5 | 30.8 | 99.5 | 36.4 | 37.7 |
| GPT-5.4 | 83.5 | 57.0 | 35.5 | 98.9 | 34.8 | 37.2 |
| GLM-5 | 89.6 | 70.1 | - | 97.9 | 34.7 | 52.3 |
| GLM-5.1 | 88.5 | 63.7 | 5.5 | 98.5 | 36.4 | 36.7 |
| Kimi K2.5 | 87.3 | 54.2 | 12.0 | 94.4 | 44.3 | 46.9 |

**Protein Fold Classification.** Claude Opus (90.5%) and Gemini 3 Pro (90.2%) achieve the best overall accuracy, whereas GPT-5.4 Mini lags substantially at 44.4%. Across models, performance is consistently higher on $\alpha$-helix and $\beta$-sheet segments than on coil regions, suggesting that LLMs more readily recognize regular geometric patterns than irregular local structures. Per-class results are provided in the Appendix.

**Interaction Perception.** Interaction perception is the most challenging Stage 1 task. Even the best model, Gemini 3 Pro, reaches only 39.4% macro-F1, with most models clustered between 27–36%. Models exhibit high recall but low precision, indicating systematic over-prediction of ligand–protein interactions. These results identify 3D ligand–pocket interaction reasoning as the main perception bottleneck.

**Key finding: the perception cliff from 2D to 3D.** The three perception tasks reveal a sharp capability transition. Models that achieve micro-F1 $> 0.90$ on functional group identification (a 2D task) and >90% on fold classification (a single-entity 3D task) cannot exceed F1 = 0.40 on interaction perception (a cross-entity 3D task). This pattern is consistent across all model families and scales, identifying 3D ligand–protein interaction understanding as the universal perception bottleneck. The failure is not attributable to hallucination: as shown in §3.4, mainstream models fabricate fewer than 0.4% of structural entities, confirming that the low interaction perception scores reflect genuine reasoning limitations rather than information fabrication.

### 3.2. Stage 2: Strategy Discovery

Given the current PLIP interaction report for each protein–ligand complex, models propose 1–3 optimization strategies,

each a (target_atom, target_residue, interaction_type) triplet. All 14 models achieve >84% format compliance and consistently output the requested three strategies per run (avg 2.99), confirming that strategy generation is a well-followed instruction across model families.

**Chemical Plausibility.** All models achieve >93% chemical plausibility, confirming that LLMs possess sufficient medicinal chemistry knowledge to propose residue–interaction type pairs that are chemically compatible. GPT-5 leads at 99.5%, while Gemini 3 Flash (93.6%) and Kimi K2.5 (94.4%) are at the lower end.

**Distance Sensitivity.** To test whether models genuinely use 3D spatial information, we shift the target residue distance by ±2 Å and measure the change in recommendation rate. Under normal distances, models recommend the target residue 78–90% of the time (BasRate); after perturbation, this drops to 34–53% (PtbRate), yielding an average RetainDrop of ∼0.39. Gemini 3 Pro (0.49) and Gemini 3 Flash (0.47) are the most distance-sensitive, while GPT-5.4 Mini is a clear outlier (Drop = 0.11)—its baseline recommendation rate is already low (64%), and perturbation barely changes its output, suggesting it relies on chemical templates rather than spatial context.

In shrink mode, we compress competitor residue distances to 2.0 Å while preserving the target residue distance, testing whether models are lured by artificially close alternatives. Target recommendation rates drop substantially (RetDrop 0.20–0.65), with Gemini 3 Pro again leading (0.65).

**Key finding: sensing without reasoning.** The distance sensitivity results provide direct evidence that LLMs process 3D spatial information—perturbations shift strategy outputs by ∼40%, ruling out pure template-matching. Yet this spatial awareness does not translate into accurate interaction identification: the best model achieves only F1 = 0.39 on interaction perception (§3.1). The models *sense* 3D context but cannot *reason* from it to predict which interactions will actually form. This dissociation between spatial awareness and spatial reasoning is one of our central findings, suggesting that the bottleneck is not input processing but the inferential step from distances and angles to interaction predictions.

### 3.3. Stage 3: Strategy Execution

**Execution Validity.** *Can LLMs generate valid molecular edits that target the strategy-specified position?* Top-tier models generate highly valid molecular edits: Gemini 3 Pro (99.5%), Claude variants (≥98%), and Kimi K2.5 (98.5%) all produce nearly flawless SMILES. Weaker models lag substantially—GLM-5.1 (75.1%) and GPT-5.4 Mini (81.6%) fail on roughly one in five edits. However, the modification target accuracy (Mod. Target) is strikingly uniform across all models at ∼83%, indicating that strategy comprehension is consistent but imperfect: regardless of model capability, about one in six edits modifies a position other than the one specified by the strategy.

*Table 2.* Results for Stage 3 strategy execution. Exec. Validity reports the SMILES parse, sanitize, and Kekulize pass rate. Mod. Target denotes the fraction of edits that modify the strategy-specified position. Int. Realization measures whether the intended residue–interaction pair appears after Boltz-2 co-folding and PLIP analysis. Baseline Ret. reports the fraction of original PLIP interactions preserved after editing. All values are percentages.

| Model | Exec. Validity | Mod. Target | Int. Realization | Baseline Ret. |
|---|---|---|---|---|
| Claude Sonnet 4.6 | 99.0 | 81.7 | 2.1 | 60.5 |
| w/ Thinking | 99.0 | 81.7 | 2.1 | 60.2 |
| Claude Opus 4.6 | 98.0 | 84.1 | 1.7 | 65.9 |
| w/ Thinking | 99.0 | 81.2 | 2.1 | 60.4 |
| Gemini 2.5 Flash | 86.1 | 83.8 | 2.4 | 60.4 |
| Gemini 2.5 Pro | 98.0 | 83.6 | 2.2 | 60.3 |
| Gemini 3 Flash | 97.5 | 84.2 | 2.2 | 60.3 |
| Gemini 3 Pro | 99.5 | 83.9 | 2.1 | 60.3 |
| GPT-5.4 Mini | 81.6 | 83.3 | 2.0 | 61.6 |
| GPT-5 | 95.5 | 84.3 | 2.2 | 60.7 |
| GPT-5.4 | 93.0 | 82.8 | 2.3 | 60.8 |
| GLM-5 | 84.6 | 84.6 | 2.5 | 61.2 |
| GLM-5.1 | 75.1 | 82.3 | 2.1 | 58.6 |
| Kimi K2.5 | 98.5 | 82.1 | 2.1 | 60.2 |

**Interaction Realization.** *Do the intended interactions actually form in the predicted 3D structure?* This is the most consequential question in the benchmark. After each modified molecule is co-folded with the target protein using Boltz-2 and analyzed by PLIP, we check whether the interaction the LLM intended to create—specified as a (target_residue, interaction_type) pair—actually appears in the predicted binding pose. The result is stark: fewer than 2.5% of intended interactions are realized across all 14 models, with rates ranging from 1.7% to 2.5%. The cross-model variance is less than one percentage point, meaning that no model—from the compact GPT-5.4 Mini to the frontier Claude Opus—demonstrates any meaningful advantage in producing structurally effective molecular edits.

**Baseline Interaction Retention.** *Does the editing preserve the existing favorable interactions of the original ligand?* Molecular editing inevitably alters the interaction profile: on average, approximately 60% of the original PLIP interactions are preserved after editing (macro-retention 58.6–65.9%). This level of retention is expected, as the modifications are designed to introduce new interaction patterns and may naturally shift the binding landscape. For context, Boltz-2 itself introduces stochastic variation: re-folding the same unmodified molecule yields a self-retention of 72.3%, indicating that about 12 percentage points of the observed interaction loss are attributable to the editing rather than structural prediction noise.

**Key finding: valid chemistry, failed structural realization.** The juxtaposition of high execution validity (>95% for top models) and near-zero interaction realization (<2.5%) captures the central challenge. LLMs can produce chemically valid molecular edits that target the correct positions, yet the intended new interactions essentially never materialize in the co-folded 3D structure. This gap between text-level chemical reasoning and 3D structural outcomes is consistent across all model families and scales, confirming that the bottleneck lies not in molecular editing competence but in the fundamental disconnect between language-based optimization reasoning and physics-based structure prediction.

### 3.4. Inference Quality

**Hallucination Detection.** *Do LLMs fabricate structural entities when processing molecular interaction data?* When asked to restate PLIP interaction reports in structured format, mainstream models fabricate fewer than 0.4% of structural entities, with most achieving exactly 0% hallucination rate. The sole outlier is Kimi K2.5, which exhibits an 11.9% hallucination rate—fabricating roughly 1 in 8 reported interactions. This near-zero baseline establishes that the low interaction perception scores in §3.1 reflect genuine reasoning limitations rather than information fabrication: models faithfully report what they see, but struggle to reason correctly about 3D spatial relationships.

*Table 3.* CoT Error Analysis — average error rates (%) across three source tasks (functional group identification, strategy discovery, strategy execution). Five error categories scored by LLM-as-judge (Claude Opus 4.6) as binary (present/absent) per reasoning chain.

| Model | Med. Chem.↓ | Env. Info.↓ | Logic↓ | Reas.–Exec.↓ | Repr.↓ |
|---|---|---|---|---|---|
| Claude Opus 4.6 | 6.9 | 0.8 | 10.4 | 2.3 | 1.4 |
| Claude Opus 4.6 (T) | 9.4 | 2.4 | 8.8 | 5.3 | 6.5 |
| Claude Sonnet 4.6 | 13.5 | 0.7 | 12.3 | 0.9 | 1.5 |
| Claude Sonnet 4.6 (T) | 13.4 | 1.2 | 10.5 | 2.2 | 2.7 |
| Gemini 2.5 Flash | 29.0 | 1.0 | 25.1 | 2.5 | 7.2 |
| Gemini 2.5 Pro | 13.3 | 0.7 | 9.5 | 1.9 | 5.2 |
| Gemini 3 Flash | 9.9 | 3.9 | 9.2 | 3.2 | 14.0 |
| Gemini 3 Pro | 14.0 | 4.0 | 13.3 | 4.7 | 7.4 |
| GPT-5 | 18.1 | 6.5 | 16.6 | 3.6 | 11.8 |
| GPT-5.4 | 29.1 | 22.5 | 31.1 | 22.0 | 12.0 |
| GPT-5.4 Mini | 38.0 | 31.5 | 46.1 | 33.8 | 20.8 |
| GLM-5 | 13.1 | 21.5 | 15.4 | 1.6 | 5.7 |
| GLM-5.1 | 5.2 | 4.8 | 5.4 | 0.1 | 1.3 |
| Kimi K2.5 | 24.5 | 34.0 | 29.2 | 1.1 | 10.5 |

**CoT Error Analysis.** *Do reasoning chain errors scale with task difficulty?* We use Claude Opus 4.6 as an LLM-as-judge to diagnose five categories of reasoning errors across three source tasks with complete 14-model coverage (per-task breakdowns in Appendix). The two dominant error types are medicinal chemistry errors (incorrect chemical knowledge) and logical errors (internally inconsistent reasoning), which together account for the majority of failures. GPT-5.4 Mini exhibits the highest overall error rates (Med. Chem. 38.0%, Logic 46.1%), while GLM-5.1 achieves the lowest (∼5% across all categories). Notably, Kimi K2.5 and GLM-5 show disproportionately high environmental information error rates (34.0% and 21.5%, respectively), indicating that these models frequently reference information not present in the prompt during reasoning. Error rates also correlate with task difficulty: on protein fold classification (the easiest perception task), most models show <5% error rates, while on interaction perception (the hardest), logic errors exceed 40% for most models—confirming that the 3D reasoning failures observed in §3.1 originate from the reasoning process itself, not merely from output-level mistakes.

**Key finding: more reasoning is not always better.** *Does explicit chain-of-thought improve reasoning quality?* In our evaluation, thinking-mode variants do not consistently reduce reasoning errors and, in several cases, exhibit higher error rates than their standard counterparts. On functional group identification, Claude Opus shows an increase in medicinal chemistry errors from 4.0% to 12.8%, and Sonnet from 14.3% to 22.3%. A similar trend is observed for logical errors, with Opus increasing from 4.0% to 6.0% and Sonnet from 4.1% to 14.2%. In the aggregated results, Opus standard achieves a lower medicinal chemistry error rate than Opus Thinking (6.9% vs. 9.4%), with a larger gap on reasoning–execution consistency (2.3% vs. 5.3%). These results suggest that longer reasoning traces may introduce additional opportunities for factual mistakes or logical inconsistencies, especially when the task relies heavily on precise domain knowledge rather than multi-step deliberation. Thus, explicit chain-of-thought should not be assumed to uniformly improve molecular reasoning quality; its benefit appears to depend on the task type and error mode.

## 4. Conclusion

We introduced MolOpt-Eval, a diagnostic benchmark for evaluating LLM capabilities in structure-based affinity optimization at the level of individual reasoning steps. Across 14 frontier LLMs and 30 DUD-E protein targets, we find that current LLMs can recognize many 2D molecular and protein-level structural patterns, but fail systematically when reasoning over 3D ligand–pocket interactions. Although models can propose chemically plausible strategies and generate valid molecular edits, these edits rarely realize their intended interactions after structure-based validation.

These findings suggest that the central challenge for LLM-driven molecular optimization is not general chemical knowledge or SMILES-level editing, but the ability to understand and reason over 3D molecular environments. Future progress will likely require models that integrate spatially grounded representations, protein–ligand geometry, and physics-aware structural feedback more directly into

the optimization process.

## References


Bissantz, C., Kuhn, B., and Stahl, M. A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry*, 53(14):5061–5084, 2010.

Bleicher, K. H., Böhm, H.-J., Müller, K., and Alanine, A. I. Hit and lead generation: beyond high-throughput screening. *Nature reviews Drug discovery*, 2(5):369–378, 2003.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Brown, N., Fiscato, M., Segler, M. H., and Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

Cavasotto, C. N. and W Orry, A. J. Ligand docking and structure-based virtual screening in drug discovery. *Current topics in medicinal chemistry*, 7(10):1006–1014, 2007.

Chen, X., Wu, R., Lan, Y., Ma, T., and Liu, Y. Molevolve: Llm-guided evolutionary search for interpretable molecular optimization. *arXiv preprint arXiv:2603.24382*, 2026.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dey, V., Hu, X., and Ning, X. Large language models for controllable multi-property multi-objective molecule optimization. *molecules*, 1(331,586):433–166, 2025.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

Li, F., Hogg, D. C., and Cohn, A. G. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18500–18507, 2024.

Liu, H., Zeng, Z., Yan, Y., Chen, Y., and Xiao, Y. Drugr: Optimizing molecular drugs through llm-based explicit reasoning. *arXiv preprint arXiv:2602.08213*, 2026.

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Rodionov, F., Eldesokey, A., Birsak, M., Femiani, J., Ghanem, B., and Wonka, P. Floorplanqa: A benchmark for spatial reasoning in llms using structured representations. *arXiv preprint arXiv:2507.07644*, 2025.

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., and Schroeder, M. Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Xu, P., Wang, S., Zhu, Y., Li, J., Qi, G., and Zhang, Y. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

Ye, G., Cai, X., Lai, H., Wang, X., Huang, J., Wang, L., Liu, W., and Zeng, X. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26 (1):bbae693, 2025.

## A. End-to-End Affinity Optimization

**Objective.** We test whether frontier LLMs can perform end-to-end structure-based affinity optimization: given a protein–ligand complex and its current interaction profile, can the model propose a molecular modification that improves predicted binding affinity?

**Setup.** We select 30 protein targets from the DUD-E dataset, spanning kinases, GPCRs, proteases, nuclear receptors, and enzymes. For each target, we use the co-crystallized ligand as the starting molecule. Three frontier LLMs are evaluated: Claude Opus 4.6 (Anthropic), GPT-5.4 (OpenAI), and Gemini 3 Pro (Google).

For each (target, model) pair, the procedure is:

1. **Interaction profiling.** The crystal ligand is analyzed with PLIP to obtain the current non-covalent interaction profile (hydrogen bonds, hydrophobic contacts, $\pi$-stacking, salt bridges).
2. **LLM modification.** The model receives the ligand SMILES, the PLIP interaction report, and pocket residue information, and is prompted to propose a modified SMILES with improved binding affinity.
3. **Baseline co-folding.** The original crystal ligand is co-folded with the target protein using Boltz-2 to obtain a baseline predicted affinity.
4. **Modified co-folding.** The LLM-proposed modified ligand is co-folded identically to obtain a post-optimization predicted affinity.
5. **Comparison.** We compute $\Delta\text{pIC}_{50} = \text{pIC}_{50}^{\text{modified}} - \text{pIC}_{50}^{\text{original}}$, where $\text{pIC}_{50} = -\log_{10}(\text{IC}_{50}/\mu\text{M})$, so positive values indicate improved binding.

**Boltz-2 configuration.** All co-folding runs use Boltz-2 with a single structural model, pre-computed MSA alignments (UniRef30 + ColabFold), and the affinity prediction head enabled. Boltz-2 reports `affinity_pred_value` = $\log_{10}(\text{IC}_{50}/\mu\text{M})$, which we negate to obtain $\text{pIC}_{50}$.

**Results.** Of 90 possible (target, model) pairs, 89 produced valid modified SMILES that could be co-folded (one Gemini modification for BRAF failed RDKit conformer generation).

*Table 4.* Stage 0 summary statistics. $\overline{\Delta}$ = mean $\text{pIC}_{50}$ change; $p_W$ = Wilcoxon signed-rank $p$-value; $r$ = Pearson correlation between original $\text{pIC}_{50}$ and $\Delta\text{pIC}_{50}$; $p_r$ = correlation $p$-value; Impr. = fraction of targets with $\Delta > 0$.

| **Model** | $\overline{\Delta}$ | $p_W$ | $r$ | $p_r$ | **Impr.** |
|---|---|---|---|---|---|
| Claude Opus 4.6 | $-0.008$ | 0.336 | $-0.18$ | 0.336 | 50% |
| GPT-5.4 | $+0.040$ | 0.020 | $-0.42$ | 0.020 | 53% |
| Gemini 3 Pro | $+0.126$ | 0.098 | $-0.31$ | 0.098 | 62% |

No model achieves statistically significant directional improvement. The negative Pearson correlations indicate that affinity changes are anti-correlated with original potency: high-affinity molecules tend to weaken after modification, while low-affinity molecules tend to improve. This pattern is consistent with regression to the mean—the modifications act as random perturbations whose apparent "improvements" on weak binders reflect statistical artifacts rather than directed optimization.

Claude Opus 4.6 exhibits the most symmetric behavior (50% improvement rate across all potency terciles, $r = -0.18$), suggesting its modifications are closest to random perturbations. GPT-5.4 shows the strongest regression-to-mean signature ($r = -0.42$, $p = 0.02$): among the 10 strongest-binding targets, only 30% improve after modification, while among the 10 weakest-binding targets, 70% improve.

**Boltz-2 prediction noise.** The Boltz-2 affinity predictor itself introduces noise: some DUD-E crystal ligands with known nanomolar potency are predicted as weak binders ($\text{pIC}_{50} < 0$, corresponding to $\text{IC}_{50} > 1\ \mu$M). This predictor noise adds variance to the $\Delta\text{pIC}_{50}$ measurements but does not create the systematic regression-to-mean pattern, which requires that the LLM modifications be uncorrelated with the true optimization direction.

# B. Dataset Construction Details

## B.1. Target Selection

MolOpt-Eval draws its protein targets from the DUD-E (Directory of Useful Decoys, Enhanced) database, which provides experimentally confirmed actives and matched decoys for 102 protein targets. We select 30 targets spanning 11 therapeutic protein families to ensure diversity in binding site topology, pocket size, and pharmacological relevance (Table 5).

**Selection criteria.** Targets were selected to satisfy three requirements: (i) coverage of major therapeutic protein families encountered in real drug discovery programs (kinases, GPCRs, proteases, nuclear receptors, enzymes); (ii) availability of a high-resolution co-crystal structure with a bound ligand suitable for PLIP analysis; and (iii) a sufficiently diverse pool of confirmed actives (minimum $\geq$40) to enable meaningful MaxMin diversity sampling.

*Table 5.* The 30 DUD-E protein targets in MolOpt-Eval, organized by protein family. $N_{\text{actives}}$ = total confirmed actives in DUD-E; $N_{\text{posed}}$ = actives with valid 3D poses.

| Family | Target | Full Name | PDB | $N_{\text{actives}}$ | $N_{\text{posed}}$ |
|---|---|---|---|---|---|
| Kinase | abl1 | Abelson kinase 1 | 2HZI | 182 | 181 |
| | cdk2 | Cyclin-dep. kinase 2 | 1H00 | 474 | 474 |
| | egfr | Epidermal GF receptor | 2RGP | 542 | 542 |
| | braf | B-Raf kinase | 3D4Q | 152 | 152 |
| | jak2 | Janus kinase 2 | 3LPB | 107 | 107 |
| Nuclear Receptor | andr | Androgen receptor | 2AM9 | 269 | 269 |
| | esr1 | Estrogen receptor $\alpha$ | 1SJ0 | 383 | 383 |
| | pparg | PPAR-$\gamma$ | 2GTK | 484 | 484 |
| GPCR | cxcr4 | CXCR4 chemokine receptor | 3ODU | 40 | 40 |
| | adrb2 | $\beta_2$-adrenergic receptor | 3NY8 | 231 | 231 |
| | drd3 | Dopamine D3 receptor | 3PBL | 480 | 480 |
| Protease | bace1 | $\beta$-secretase 1 | 3L5D | 283 | 283 |
| | hivpr | HIV-1 protease | 1XL2 | 536 | 536 |
| | fa7 | Coagulation factor VIIa | 1W7X | 114 | 114 |
| | ace | Angiotensin-conv. enzyme | 3BKL | 281 | 281 |
| Phosphatase | ptn1 | Protein tyrosine phosph. 1B | 2QBS | 130 | 130 |
| Oxidoreductase | dyr | Dihydrofolate reductase | 3NXO | 231 | 231 |
| | hdac2 | Histone deacetylase 2 | 3MAX | 185 | 185 |
| | hmdh | HMG-CoA reductase | 3CCW | 87 | 87 |
| Hydrolase | aces | Acetylcholinesterase | 1E66 | 453 | 453 |
| | pde5a | Phosphodiesterase 5A | 1XOZ | 398 | 398 |
| Transferase | comt | Catechol-O-methyltransf. | 3BWM | 41 | 41 |
| | fnta | Farnesyl transferase $\alpha$ | 3E37 | 592 | 592 |
| Ion Channel | gria2 | Glutamate receptor iGluR2 | 3KGC | 158 | 158 |
| Lyase | cah2 | Carbonic anhydrase II | 1BCD | 492 | 492 |
| Chaperone | hs90a | Heat shock protein 90$\alpha$ | 1UYG | 105 | 105 |
| PPI | xiap | X-linked IAP | 2JK1 | 100 | 100 |
| DNA Repair | parp1 | Poly(ADP-ribose) polym. 1 | 3L3M | 508 | 508 |
| Neuraminidase | nram | Influenza neuraminidase | 1B9V | 98 | 98 |
| Isomerase | fpps | Farnesyl pyrophosphate synth. | 1YV5 | 85 | 85 |

## B.2. PLIP Interaction Annotation

Ground-truth interaction annotations are generated programmatically using PLIP (Protein–Ligand Interaction Profiler) with its default geometric and physicochemical criteria. PLIP provides deterministic, reproducible annotations without human subjectivity.

**Core interaction types** For interaction perception evaluation, we retain four interaction types with well-defined geometric criteria:

*Table 6.* PLIP geometric criteria for the four core interaction types used in Task 1.3.

| Type | Distance Criterion | Additional Constraints |
|---|---|---|
| Hydrogen bond | D–A: 2.5–3.5 Å | D–H–A angle $> 120°$ |
| Hydrophobic | C–C: $< 4.0$ Å | Both atoms non-polar; PLIP aromatic/aliphatic filter |
| $\pi$-Stacking | Centroid: 3.3–5.5 Å | Dihedral and offset angle constraints |
| Salt bridge | Charged groups: $< 5.5$ Å | Requires formal charge on both partners |

**Pocket definition.** All residues with any heavy atom within 6.0 Å of any ligand heavy atom are included in the pocket representation provided to LLMs. This standard cutoff balances inclusion of interaction-relevant residues with exclusion of distant structural context that would inflate prompt length without informational benefit.

## C. Stage 1: Detailed Metrics

This section reports the detailed results for the three Stage 1 structural perception tasks.

Table 7 presents the full precision, recall, and F1 metrics for functional group identification. Table 8 reports overall and per-class accuracy for protein fold classification. Table 9 provides the complete micro- and macro-level metrics for ligand–protein interaction perception. Together, these results supplement the summary reported in Table 1 and provide a more fine-grained view of model behavior across perception tasks.

*Table 7.* Functional Group Identification — Full Metrics. All models exhibit higher recall than precision, confirming a systematic tendency toward over-prediction. The gap between micro-F1 and macro-F1 indicates that performance is uneven across functional group categories, with common groups (hydroxyl, amine) identified more reliably than rare ones (sulfonamide, phosphate).

| Model | micro-P | micro-R | micro-F1 | macro-F1 |
|---|---|---|---|---|
| Claude Opus 4.6 | 0.891 | 0.951 | 0.920 | 0.857 |
| Claude Opus 4.6 (T) | 0.889 | 0.949 | 0.918 | 0.857 |
| Claude Sonnet 4.6 | 0.861 | 0.940 | 0.899 | 0.825 |
| Claude Sonnet 4.6 (T) | 0.870 | 0.933 | 0.900 | 0.835 |
| Gemini 2.5 Flash | 0.812 | 0.873 | 0.841 | 0.779 |
| Gemini 2.5 Pro | 0.822 | 0.934 | 0.874 | 0.812 |
| Gemini 3 Flash | 0.845 | 0.941 | 0.890 | 0.824 |
| Gemini 3 Pro | 0.893 | 0.918 | 0.905 | 0.842 |
| GPT-5 | 0.882 | 0.897 | 0.889 | 0.829 |
| GPT-5.4 | 0.810 | 0.862 | 0.835 | 0.781 |
| GPT-5.4 Mini | 0.738 | 0.778 | 0.758 | 0.686 |
| GLM-5 | 0.882 | 0.910 | 0.896 | 0.828 |
| GLM-5.1 | 0.848 | 0.925 | 0.885 | 0.827 |
| Kimi K2.5 | 0.846 | 0.901 | 0.873 | 0.791 |

## D. Stage 2: Detailed Metrics

This section reports the detailed results for Stage 2 strategy discovery, supplementing the summary metrics in Table 1. We first define the distance-sensitivity metrics used in Task 2.2, including the perturb and shrink settings. Table 10 reports the full results for perturb mode, and Table 11 reports the full results for shrink mode. Together, these results characterize whether model-generated strategies respond to controlled changes in ligand–pocket geometry.

**Metric Definitions.** For Task 2.2 (Distance Sensitivity), we define the following metrics:

*Table 8.* Protein Fold Classification — Per-Class Accuracy. Coil (irregular loop) regions are the critical differentiator: models that achieve >95% on helix and sheet can drop below 40% on coil, revealing a bias toward periodic geometric patterns. GPT-5.4 Mini's near-zero coil accuracy (2.5%) suggests it defaults to classifying all fragments as regular secondary structures.

| Model | Overall | Helix | Sheet | Coil |
|---|---|---|---|---|
| Claude Opus 4.6 | 0.905 | 0.896 | 0.919 | 0.907 |
| Claude Opus 4.6 (T) | 0.902 | 0.929 | 0.849 | 0.907 |
| Claude Sonnet 4.6 | 0.872 | 0.948 | 0.756 | 0.856 |
| Claude Sonnet 4.6 (T) | 0.883 | 0.929 | 0.814 | 0.873 |
| Gemini 2.5 Flash | 0.665 | 0.896 | 0.616 | 0.398 |
| Gemini 2.5 Pro | 0.721 | 0.961 | 0.977 | 0.220 |
| Gemini 3 Flash | 0.888 | 0.968 | 0.977 | 0.720 |
| Gemini 3 Pro | 0.902 | 0.903 | 0.942 | 0.873 |
| GPT-5 | 0.785 | 0.987 | 0.988 | 0.373 |
| GPT-5.4 | 0.570 | 0.630 | 0.314 | 0.678 |
| GPT-5.4 Mini | 0.444 | 0.896 | 0.209 | 0.025 |
| GLM-5 | 0.701 | 0.942 | 0.628 | 0.441 |
| GLM-5.1 | 0.637 | 0.797 | 0.726 | 0.426 |
| Kimi K2.5 | 0.542 | 0.682 | 0.209 | 0.602 |

*Table 9.* Interaction Perception — Full Metrics. The universally low precision (macro-P $< 0.29$) confirms systematic over-prediction across all models. GLM-5 achieves exactly 0 on all metrics due to consistent format non-compliance on this task (outputs free text instead of structured JSON). Kimi K2.5's high recall (0.64) but extremely low precision (0.07) indicates it predicts an excessive number of interactions, achieving coverage through volume rather than accuracy.

| Model | mi-P | mi-R | mi-F1 | ma-P | ma-R | ma-F1 |
|---|---|---|---|---|---|---|
| Claude Opus 4.6 | 0.244 | 0.435 | 0.312 | 0.200 | 0.477 | 0.277 |
| Claude Opus 4.6 (T) | 0.266 | 0.538 | 0.356 | 0.247 | 0.559 | 0.336 |
| Claude Sonnet 4.6 | 0.254 | 0.581 | 0.354 | 0.261 | 0.644 | 0.364 |
| Claude Sonnet 4.6 (T) | 0.243 | 0.544 | 0.336 | 0.239 | 0.610 | 0.333 |
| Gemini 2.5 Flash | 0.168 | 0.525 | 0.254 | 0.187 | 0.598 | 0.277 |
| Gemini 2.5 Pro | 0.235 | 0.597 | 0.338 | 0.224 | 0.652 | 0.324 |
| Gemini 3 Flash | 0.225 | 0.634 | 0.333 | 0.228 | 0.704 | 0.337 |
| Gemini 3 Pro | 0.270 | 0.639 | 0.380 | 0.286 | 0.712 | 0.394 |
| GPT-5 | 0.207 | 0.507 | 0.294 | 0.220 | 0.565 | 0.308 |
| GPT-5.4 | 0.246 | 0.594 | 0.348 | 0.251 | 0.652 | 0.355 |
| GPT-5.4 Mini | 0.181 | 0.435 | 0.256 | 0.205 | 0.473 | 0.274 |
| GLM-5 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| GLM-5.1 | 0.031 | 0.096 | 0.047 | 0.037 | 0.122 | 0.055 |
| Kimi K2.5 | 0.061 | 0.584 | 0.110 | 0.068 | 0.640 | 0.120 |

- **Perturb mode.** Let $N$ be the number of independent runs (default 5). For a given target gene and target residue $r^*$:

$$\text{BasRate}(r^*) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[r^* \in \mathcal{S}_i^{\text{bas}}] \tag{2}$$

$$\text{PtbRate}(r^*) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[r^* \in \mathcal{S}_i^{\text{ptb}}] \tag{3}$$

$$\text{RetainDrop} = \text{BasRate} - \text{PtbRate} \tag{4}$$

  where $\mathcal{S}_i^{\text{bas}}$ is the set of residues recommended in run $i$ under normal distances, and $\mathcal{S}_i^{\text{ptb}}$ is the set after shifting the target residue distance by $\pm 2$ Å (direction chosen to move away from the optimal interaction distance). A high RetainDrop indicates that the model's strategy selection is genuinely sensitive to spatial distances.

- **Shrink mode.** Competitor residue distances are compressed to 2.0 Å while the target residue remains at its original

distance:

$$\text{ShkRate}(r^*) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[r^* \in \mathcal{S}_i^{\text{shk}}] \tag{5}$$

$$\text{ShkDrop} = \text{BasRate} - \text{ShkRate} \tag{6}$$

$$\text{CompPkp} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\mathcal{C} \cap \mathcal{S}_i^{\text{shk}} \neq \emptyset] \tag{7}$$

where $\mathcal{C}$ is the set of competitor residues whose distances were shrunk. CompPkp (Competitor Pickup) measures whether the model blindly follows the closest residue.

*Table 10.* Distance Sensitivity (Perturb mode) — Full Metrics. BasRate denotes the baseline recommendation rate for the target residue under normal distances. PtbRate denotes the rate after ±2 Å perturbation. RetainDrop = BasRate − PtbRate, where higher values indicate stronger sensitivity to spatial perturbation.

| Model | BasRate | PtbRate | RetainDrop |
|---|---|---|---|
| Gemini 3 Pro | 0.857 | 0.365 | 0.492 |
| Gemini 3 Flash | 0.876 | 0.406 | 0.470 |
| Kimi K2.5 | 0.788 | 0.345 | 0.443 |
| Claude Sonnet 4.6 | 0.849 | 0.410 | 0.439 |
| Gemini 2.5 Pro | 0.832 | 0.405 | 0.427 |
| Claude Sonnet 4.6 (T) | 0.829 | 0.405 | 0.424 |
| Claude Opus 4.6 (T) | 0.897 | 0.481 | 0.416 |
| Gemini 2.5 Flash | 0.777 | 0.388 | 0.389 |
| Claude Opus 4.6 | 0.868 | 0.482 | 0.386 |
| GLM-5.1 | 0.825 | 0.461 | 0.364 |
| GPT-5 | 0.809 | 0.445 | 0.364 |
| GPT-5.4 | 0.829 | 0.481 | 0.348 |
| GLM-5 | 0.820 | 0.473 | 0.347 |
| GPT-5.4 Mini | 0.644 | 0.533 | 0.111 |

*Table 11.* Distance Sensitivity (Shrink mode) — Full Metrics. Competitor residue distances are compressed to 2.0 Å while the target residue distance is unchanged. BasRate denotes the baseline recommendation rate, ShkRate denotes the rate after shrinking competitor distances, and RetDrop measures the decrease in target-residue recommendation.

| Model | BasRate | ShkRate | RetDrop |
|---|---|---|---|
| Gemini 3 Pro | 0.900 | 0.580 | 0.647 |
| Gemini 3 Flash | 0.904 | 0.678 | 0.615 |
| Claude Opus 4.6 (T) | 0.918 | 0.727 | 0.575 |
| Claude Opus 4.6 | 0.913 | 0.740 | 0.573 |
| GLM-5 | 0.821 | 0.298 | 0.523 |
| Kimi K2.5 | 0.820 | 0.519 | 0.469 |
| Claude Sonnet 4.6 (T) | 0.861 | 0.688 | 0.481 |
| Claude Sonnet 4.6 | 0.881 | 0.683 | 0.463 |
| Gemini 2.5 Flash | 0.803 | 0.570 | 0.415 |
| Gemini 2.5 Pro | 0.856 | 0.769 | 0.397 |
| GPT-5 | 0.832 | 0.679 | 0.377 |
| GPT-5.4 | 0.845 | 0.662 | 0.372 |
| GLM-5.1 | 0.813 | 0.447 | 0.367 |
| GPT-5.4 Mini | 0.698 | 0.615 | 0.203 |

## E. Stage 3: Detailed Metrics

This section reports the detailed results for Stage 3 strategy execution, supplementing the summary metrics in Table 2. We first define the interaction realization and baseline retention metrics used for structure-based validation. Table 12 reports execution validity and modification-target accuracy, Table 13 reports interaction realization after Boltz-2 co-folding and

PLIP analysis, and Table 14 reports baseline interaction retention. Together, these results provide a fine-grained view of whether LLM-generated molecular edits are chemically valid, target the intended position, and produce the desired structural effects.

**Metric Definitions.** For Task 3.2 (Interaction Realization), we define:

$$\text{Realization Rate} = \frac{|\text{Reflected}|}{|\text{wBoltz}|} \tag{8}$$

where |wBoltz| is the number of valid modified molecules that were successfully co-folded by Boltz-2 (i.e., produced a structure output), and |Reflected| is the number of those whose intended interaction—specified as a (target_residue, interaction_type) pair—appears in the PLIP analysis of the co-folded complex. A relaxed variant, ResHitRate, counts cases where the target residue participates in *any* interaction type (not necessarily the one specified).

For Task 3.3 (Baseline Retention):

$$\text{Retention} = \frac{|\mathcal{I}_{\text{base}} \cap \mathcal{I}_{\text{mod}}|}{|\mathcal{I}_{\text{base}}|} \tag{9}$$

where $\mathcal{I}_{\text{base}}$ and $\mathcal{I}_{\text{mod}}$ are the sets of PLIP-detected interactions for the baseline and modified co-folded structures, respectively. MacroRetention averages per-target retention across all evaluated molecules; MicroRetention pools all interactions globally before computing the ratio.

*Table 12.* Execution Validity — Full Metrics. ValidRate = three-stage validation pass rate (parse + sanitize + Kekulize). Changed% = fraction that differ from the input SMILES. ModTarget% = fraction of valid edits that modify the strategy-specified atom position. The uniformity of ModTarget% (∼83% across all models) indicates that strategy comprehension is a model-independent ceiling—about 1 in 6 edits consistently targets the wrong position regardless of model capability.

| Model | Total | Valid | ValidRate | Changed% | ModTarget% |
|---|---|---|---|---|---|
| Gemini 3 Pro | 201 | 200 | 99.5% | 99.5% | 83.9% |
| Claude Opus 4.6 (T) | 201 | 199 | 99.0% | 99.0% | 81.2% |
| Claude Sonnet 4.6 | 201 | 199 | 99.0% | 99.0% | 81.7% |
| Claude Sonnet 4.6 (T) | 201 | 199 | 99.0% | 99.0% | 81.7% |
| Kimi K2.5 | 201 | 198 | 98.5% | 98.5% | 82.1% |
| Claude Opus 4.6 | 201 | 197 | 98.0% | 98.0% | 84.1% |
| Gemini 2.5 Pro | 201 | 197 | 98.0% | 98.0% | 83.6% |
| Gemini 3 Flash | 201 | 196 | 97.5% | 97.5% | 84.2% |
| GPT-5 | 201 | 192 | 95.5% | 95.5% | 84.3% |
| GPT-5.4 | 201 | 187 | 93.0% | 93.0% | 82.8% |
| Gemini 2.5 Flash | 201 | 173 | 86.1% | 86.1% | 83.8% |
| GLM-5 | 201 | 170 | 84.6% | 84.6% | 84.6% |
| GPT-5.4 Mini | 201 | 164 | 81.6% | 77.6% | 83.3% |
| GLM-5.1 | 201 | 151 | 75.1% | 75.1% | 82.3% |

## F. Related Work

**LLMs for molecular optimization.** Large language models have been increasingly applied to molecular tasks. ChemCrow (Bran et al., 2023) augments an LLM with external chemistry tools for synthesis planning and property queries. For molecular optimization specifically, DrugAssist (Ye et al., 2025) fine-tunes an LLM on paired (molecule, improved molecule) examples to propose edits guided by natural-language instructions; DrugR (Liu et al., 2026) introduces explicit chain-of-thought reasoning into the optimization loop; MolEvolve (Chen et al., 2026) couples LLM-guided mutation operators with evolutionary search for interpretable, multi-round optimization; and recent work explores controllable multi-property multi-objective optimization with LLMs (Dey et al., 2025). While these methods report promising end-to-end metrics (*e.g.*, docking score improvement, QED gain), they operate predominantly on 1D/2D molecular representations (SMILES strings, property descriptors) and evaluate aggregate outcomes rather than diagnosing intermediate reasoning steps. In contrast, MolOpt-Eval explicitly tests whether LLMs can *perceive* 3D interactions, *propose* spatially grounded strategies, and *execute* edits that realize intended structural contacts—providing a stage-wise account of where current approaches succeed and fail.

*Table 13.* Interaction Realization — Full Metrics. Reflected = number of strategies whose intended (residue, type) pair appears in the PLIP analysis after Boltz-2 co-folding. ResHit = number where the target residue participates in *any* interaction (relaxed criterion). The uniformly low realization rates across models confirm that interaction realization remains a shared bottleneck rather than a model-specific failure.

| Model | Reflected | ResHit | Real. Rate | ResHitRate |
|---|---|---|---|---|
| GLM-5 | 4 | 11 | 2.50% | 6.88% |
| Gemini 2.5 Flash | 4 | 11 | 2.42% | 6.67% |
| GPT-5.4 | 4 | 12 | 2.27% | 6.82% |
| GPT-5 | 4 | 12 | 2.20% | 6.59% |
| Gemini 3 Flash | 4 | 13 | 2.19% | 7.10% |
| Gemini 2.5 Pro | 4 | 13 | 2.15% | 6.99% |
| Kimi K2.5 | 4 | 13 | 2.14% | 6.95% |
| Gemini 3 Pro | 4 | 13 | 2.13% | 6.91% |
| Claude Sonnet 4.6 | 4 | 13 | 2.13% | 6.91% |
| Claude Sonnet 4.6 (T) | 4 | 13 | 2.13% | 6.91% |
| Claude Opus 4.6 (T) | 4 | 13 | 2.12% | 6.88% |
| GLM-5.1 | 3 | 8 | 2.11% | 5.63% |
| GPT-5.4 Mini | 3 | 10 | 1.95% | 6.49% |
| Claude Opus 4.6 | 3 | 7 | 1.66% | 3.87% |

*Table 14.* Baseline Interaction Retention — Full Metrics. MacroRetention denotes the per-target average retention rate, and MicroRetention pools all interactions globally before computing retention. The 72.3% self-retention of Boltz-2 establishes an approximate ceiling; the $\sim$60% observed retention indicates that part of the interaction loss is attributable to LLM editing rather than structural prediction noise alone.

| Model | MacroRet. | MicroRet. |
|---|---|---|
| Claude Opus 4.6 | 65.9% | 66.0% |
| GPT-5.4 Mini | 61.6% | 63.4% |
| GLM-5 | 61.2% | 62.1% |
| GPT-5.4 | 60.8% | 62.2% |
| GPT-5 | 60.7% | 61.3% |
| Claude Sonnet 4.6 | 60.5% | 62.0% |
| Gemini 2.5 Flash | 60.4% | 61.1% |
| Claude Opus 4.6 (T) | 60.4% | 61.7% |
| Gemini 2.5 Pro | 60.3% | 61.8% |
| Gemini 3 Flash | 60.3% | 61.4% |
| Gemini 3 Pro | 60.3% | 61.7% |
| Claude Sonnet 4.6 (T) | 60.2% | 61.5% |
| Kimi K2.5 | 60.2% | 61.7% |
| GLM-5.1 | 58.6% | 60.7% |

**Benchmarks for molecular ML.** MoleculeNet (Wu et al., 2018) and the Therapeutics Data Commons (Huang et al., 2021) have standardized evaluation for molecular property prediction and ADMET tasks, while GuacaMol (Brown et al., 2019) benchmarks de novo molecular generation along axes such as distribution learning and goal-directed optimization. DUD-E (Mysinger et al., 2012) provides curated active/decoy sets across diverse protein families for virtual screening evaluation, and interaction profilers such as PLIP (Salentin et al., 2015) enable automated annotation of non-covalent contacts from crystal or predicted structures. Recent co-folding models like Boltz-2 (Passaro et al., 2025) further allow *in silico* validation of predicted protein–ligand complexes at near-experimental accuracy. None of these resources, however, evaluate the *cognitive sub-steps* of structure-based optimization—perception, strategy, and execution—as separable capabilities. MolOpt-Eval fills this gap by building on DUD-E targets and combining Boltz-2 co-folding with PLIP interaction analysis as its validation backbone.

**Spatial reasoning in LLMs.** A parallel line of work has probed LLMs' ability to reason about spatial configurations in general domains. StepGame (Li et al., 2024) evaluates multi-hop spatial reasoning over grid-based object arrangements; FloorPlanQA (Rodionov et al., 2025) tests understanding of architectural layouts from structured representations; and SpatialBench (Xu et al., 2025) benchmarks multimodal LLMs on diverse spatial cognition tasks including distance estimation and orientation judgment. A consistent finding across these benchmarks is that LLMs exhibit significant deficits in geometric

and spatial reasoning, even when provided with explicit coordinate information. Our results extend this finding into the molecular domain and sharpen it: LLMs demonstrably *sense* spatial inputs (distance perturbations shift their outputs), yet they cannot translate this sensitivity into accurate 3D interaction predictions—a failure mode with direct consequences for drug design.

## G. LLM Usage Statement

We used large language models to assist with writing, editing, and formatting parts of the manuscript. LLMs were also used to support portions of the code implementation, including code drafting, refactoring, and debugging. All LLM-assisted text was reviewed and revised by the authors, and all LLM-assisted code was manually inspected, tested, and verified before use in the experiments. The scientific claims, experimental design, data analysis, and final conclusions were determined by the authors.

## H. Limitations

Several limitations should be acknowledged. First, Boltz-2 is an imperfect structure predictor—the 72.3% self-retention in re-fold control indicates that structural prediction noise contributes to observed interaction loss. Second, our strategy sets, while covering all 30 targets, represent a finite sampling of the optimization space. Third, the LLM-as-judge framework for CoT evaluation introduces potential bias. Finally, our evaluation is limited to frontier commercial models; open-source and domain-fine-tuned models may exhibit different profiles.