LAO-EVAL: A DIAGNOSTIC BENCHMARK FOR LLM CAPABILITIES IN STRUCTURE-BASED LIGAND AFFINITY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based small-molecule ligand affinity optimization—the iterative editing of hit compounds to improve protein-binding affinity under binding-pocket constraints—is a fundamental task in computer-aided drug discovery. Despite rapid progress in scientific reasoning, the capabilities of large language models (LLMs) across the component steps of this task remain poorly understood. We first show through end-to-end experiments that frontier LLMs achieve only marginal affinity improvements and frequently degrade higher-affinity starting ligands. To identify where these failures arise, we introduce **LAO-Eval**, a diagnostic benchmark for **Ligand Affinity Optimization** that decomposes the reasoning process into three independently evaluable stages: Structural Perception, Strategy Discovery, and Strategy Execution, supplemented by chain-of-thought quality analysis. We evaluate 14 frontier LLMs on 10 diagnostic tasks spanning 30 DUD-E protein targets. Our results show that: (i) LLMs perform well on 2D molecular structure recognition (micro-F1 = 0.92) and protein secondary-structure classification (accuracy up to 90%), but their performance drops sharply on 3D ligand–protein interaction perception (F1 < 0.40); (ii) proposed strategies are usually chemically plausible (>93%), and most models generate valid molecules in more than 95% of completed responses, yet at most 25% of intended interactions are realized after Boltz-2 co-folding and structural analysis; (iii) performance differences across model families narrow substantially on interaction realization, indicating a shared limitation of current LLM paradigms; and (iv) controlled distance perturbations show that LLMs respond to spatial information but cannot reliably translate it into interaction predictions or structurally effective edits. These findings identify 3D ligand–pocket reasoning as the principal bottleneck in LLM-based ligand affinity optimization and motivate the development of molecular foundation models with spatially grounded representations and physics-aware structural feedback. Code, data, and reproduction instructions are included in the supplementary reproducibility archive.

1 INTRODUCTION

Structure-based optimization of small-molecule ligand affinity is a central task in the drug discovery pipeline Bleicher et al. (2003); Hughes et al. (2011). During hit-to-lead, medicinal chemists modify a hit compound to improve its binding affinity while respecting the pocket's geometric and physicochemical constraints Cavasotto & W Orry (2007): interpreting the current ligand–pocket interaction pattern, identifying an opportunity to introduce or strengthen an interaction, formulating a chemically feasible strategy (*e.g.*, adding a hydrogen-bond donor at a specific atom to engage a nearby residue), and implementing it as a precise edit. This chain jointly relies on medicinal chemistry knowledge, spatial reasoning, and accurate perception of the local structural environment—capabilities traditionally reserved for human experts Bissantz et al. (2010).

Recent LLMs show strong performance in mathematics, code generation, and scientific reasoning, raising the question of whether they can perform structure-based ligand affinity optimization. As an initial probe, we ran an end-to-end hit-to-lead experiment with three frontier LLMs on 30 protein–ligand complexes from DUD-E Mysinger et al. (2012): given a co-crystallized ligand and

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.

its binding-pocket context, each model proposed a modification intended to improve affinity, which we then co-folded with Boltz-2 Passaro et al. (2025) and compared to the original. The results were sobering (Figure 1): affinity improvements were marginal across all three models, and modifications to higher-affinity ligands frequently reduced it. Such end-to-end failures can arise from inaccurate perception, chemically infeasible or spatially ungrounded strategies, or failure to execute an otherwise reasonable strategy—distinguishing among these is essential, since each calls for a different fix.

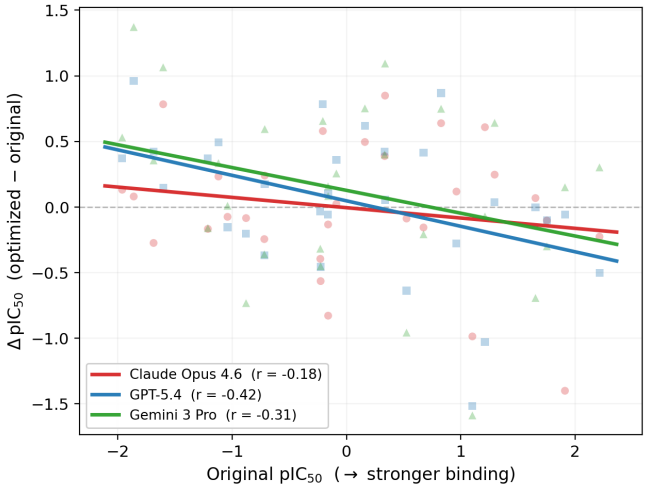


Figure 1: End-to-end affinity optimization test. Each point denotes one target–model pair, with the y-axis showing the change in Boltz-2 predicted pIC_{50}. The negative regression slopes indicate that improvements concentrate on low-affinity ligands; high-affinity ligands often deteriorate, consistent with regression to the mean rather than directed optimization (Appendix D discusses the x/y-axis coupling and Boltz-2 noise underlying this pattern).

To this end, we introduce **LAO-Eval**, a diagnostic benchmark for structure-based **L**igand **A**ffinity **O**ptimization that assesses LLM capabilities at the level of individual reasoning steps. As illustrated in Figure 2, LAO-Eval decomposes the pipeline into three sequential stages—**Structural Perception** (molecular substructures, protein secondary structure, and ligand–protein interactions), **Strategy Discovery** (identifying chemically feasible, spatially grounded modification opportunities), and **Strategy Execution** (translating strategies into valid edits with the intended structural effect)—plus an orthogonal **Inference Quality** analysis of CoT traces for factual, logical, and reasoning–execution errors. The stage-wise tasks localize *where* failures occur; the CoT analysis characterizes *how* reasoning breaks down. LAO-Eval comprises 10 diagnostic tasks across 30 protein targets and evaluates 14 frontier LLMs from the GPT, Claude, Gemini, Kimi, and GLM families under unified conditions.

Applying LAO-Eval, we obtain three high-level findings (Section 4): (1) models understand isolated molecular and protein representations well but struggle to reason jointly about ligand–pocket interactions; (2) chemically plausible, syntactically valid edits rarely realize their intended structural effect—at most one in four does, after structure-based validation; and (3) models respond to spatial information but cannot reliably reason from it into accurate interaction prediction or realization.

Together, these findings identify 3D ligand–pocket reasoning, rather than general chemical knowledge or SMILES-level editing, as the central bottleneck in LLM-driven structure-based ligand affinity optimization. Our contributions are threefold: (i) **LAO-Eval**, a diagnostic benchmark decomposing the optimization pipeline into 10 independently evaluable tasks spanning perception, strategy discovery, execution, and inference quality, localizing failures to specific components; (ii) a **systematic evaluation** of 14 frontier LLMs from five families across 30 structurally diverse DUD-E targets under unified conditions; and (iii) an **empirical characterization** of 3D ligand–pocket reasoning as current LLMs' primary limitation in this task, motivating more spatially grounded molecular models.

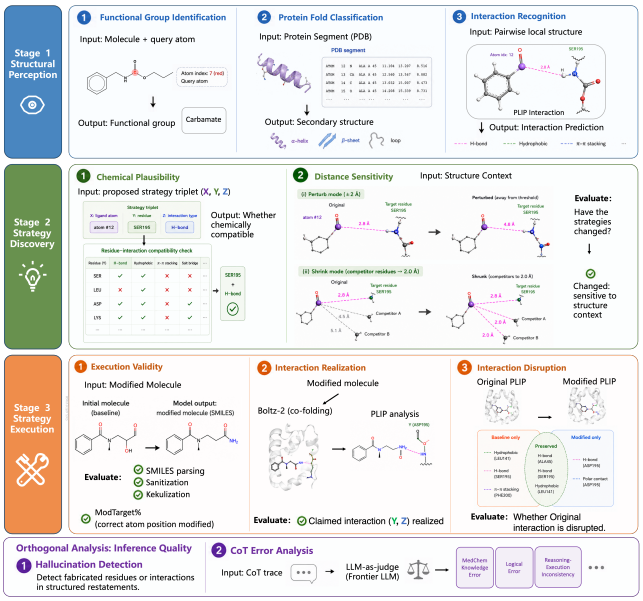


Figure 2: Overview of **LAO-Eval**: three sequential stages (structural perception, strategy formulation, chemical execution), each with independent task designs and metrics, plus an orthogonal inference-quality analysis of hallucination and CoT consistency. Full-size version in Appendix A.

2 Related Work

LLMs for molecular optimization. LLMs have been increasingly applied to molecular tasks: ChemCrow M. Bran et al. (2024) augments an LLM with chemistry tools for synthesis planning; DrugAssist Ye et al. (2025) fine-tunes on paired (molecule, improved molecule) examples; DrugR Liu et al. (2026) adds explicit chain-of-thought to the optimization loop; MolEvolve Chen et al. (2026) couples LLM-guided mutation with evolutionary search; recent work Dey et al. (2025) explores controllable multi-property optimization; and CIDD Gao et al. (2025) couples LLM chain-of-thought planning to a dedicated 3D interaction-aware molecular generator rather than direct SMILES editing. These methods report promising end-to-end metrics but operate on 1D/2D representations or fine-tuned/tool-augmented pipelines rather than a general chat-completion interface, and evaluate aggregate outcomes rather than intermediate reasoning steps; none exposes an interface compatible with LAO-Eval's task-level protocol, so we do not evaluate them directly and scope our headline claims to general-purpose LLMs prompted directly and zero-shot (Appendix J).

Benchmarks for molecular ML. MoleculeNet Wu et al. (2018) and the Therapeutics Data Commons Huang et al. (2021) standardize property/ADMET evaluation, and GuacaMol Brown et al. (2019) benchmarks de novo generation. DUD-E Mysinger et al. (2012) provides curated virtual-screening sets, PLIP Salentin et al. (2015) automates contact annotation, and Boltz-2 Passaro et al. (2025) enables *in silico* affinity prediction—but none evaluates the *cognitive sub-steps* of structure-based optimization as separable capabilities.

Spatial reasoning in LLMs. StepGame Li et al. (2024), FloorPlanQA Rodionov et al. (2026), and SpatialBench Xu et al. (2025) reveal geometric/spatial reasoning deficits even with explicit coordinates; our results extend this into the molecular domain, where LLMs *sense* spatial inputs but cannot translate this into accurate 3D interaction predictions.

3 LAO-EVAL BENCHMARK DESIGN

3.1 OVERVIEW AND PROBLEM FORMULATION

We focus on *structure-based ligand 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, formulated 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:

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

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

Accordingly, LAO-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.

3.2 DATA CONSTRUCTION

LAO-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 plus four additional active compounds, each pre-docked into the same binding site as the crystal ligand using Glide; from the pool of DUD-E actives with a successfully generated Glide pose, we select four via MaxMin diversity sampling (seed 42) over molecular fingerprints, yielding 150 active protein–ligand complexes as shared inputs. For each complex, PLIP Salentin et al. (2015) generates programmatic interaction annotations across four non-covalent types (hydrogen bonds, hydrophobic contacts, π-stacking, salt bridges), included in the model input as structure-derived context.

To evaluate strategy execution independently from strategy generation, we additionally construct a standardized set of expert-curated optimization strategies, each a triplet (X, Y, Z) specifying the ligand atom position X, target residue Y, and intended interaction type Z, shared across all models so execution differences reflect editing ability rather than strategy quality. Details of target selection, PLIP criteria, and strategy construction are in the appendix.

3.3 STRUCTURAL PERCEPTION

This stage maps $(m_0, P, C) \to S$, decomposing S into three components of increasing difficulty—molecular substructure features (functional groups), protein structural elements (secondary structure), and ligand–protein interaction triples $(\text{atom_idx}, \text{residue}, \text{interaction_type})$—yielding three tasks.

Task 1.1: Functional Group Identification. Given a molecule's SMILES string with atom-level indexing and a queried atom index, the model identifies the functional group at that atom, output in SMILES format; RDKit-based matching verifies whether the prediction is a valid substructure containing the queried atom.

Task 1.2: Protein Fold Classification. Given a segment of a protein structure in PDB format, the model identifies its structural fold type (*e.g.*, alpha-helix, beta-sheet, loop). Ground-truth labels are derived from secondary structure annotations; predictions are scored by classification accuracy.

Task 1.3: Interaction Recognition. Given explicit descriptions of neighboring atoms and residues in a protein–ligand complex, the model determines whether specific intermolecular interactions (*e.g.*, hydrogen bonds, hydrophobic contacts) are present, against ground-truth labels from PLIP reports.

3.4 STRATEGY DISCOVERY

This stage maps $S \to A = \{a_1, \ldots, a_k\}$: each strategy a_i is a triplet (X, Y, Z) (X = target ligand atom, Y = target residue, Z = intended interaction type), independently verifiable since X/Y can be

checked against context and Z validated for compatibility with Y. Given a complex and its PLIP report, the model proposes up to 3 such triplets, over 5 independent runs per pair to capture variability; the raw outputs feed two evaluation tasks.

Task 2.1: Chemical Plausibility. Whether the proposed strategies are chemically valid at the residue level—whether target residue Y can plausibly provide interaction type Z. We report the plausibility rate: the fraction of (Y, Z) pairs chemically compatible under residue-level interaction rules.

Task 2.2: Distance Sensitivity. Tests whether the model genuinely processes 3D spatial information or relies on chemical templates, via two controlled perturbations. In *perturb*, we shift the target-residue distance by ± 2 Å away from the interaction threshold and measure RetainDrop, the drop in target-residue recommendation rate. In *shrink*, we compress competing residues to 2.0 Å (target unchanged) and measure whether recommendations shift toward the artificially closer competitors. Protocols and metrics are detailed in the appendix.

3.5 Strategy Execution

This stage maps $(m_0, a_i) \rightarrow m_i^*$: for each strategy $a_i = (X, Y, Z)$, the model edits m_0 to install or strengthen interaction Z between atom X and residue Y, given the baseline ligand, its PLIP report, and the curated triplet—the same set for every model, for a fair comparison. Modified molecules are evaluated through three tasks.

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.

Task 3.2: Interaction Realization. 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: Interaction Disruption. We compare the PLIP interaction set of the baseline ligand against that of the modified molecule. The retention rate measures the fraction of baseline interactions preserved after modification. Low retention indicates that the modification disrupts the existing binding mode.

3.6 Inference Quality

Beyond stage-wise performance, we analyze the reasoning trajectory T to diagnose *how* reasoning breaks down.

Task 4.1: Hallucination Detection. The model restates a given PLIP interaction report and pocket-neighbor description in structured JSON; we compare against the input and measure the fraction of fabricated entities.

Task 4.2: Chain-of-Thought Error Analysis. An LLM-as-judge protocol evaluates model-generated CoT traces: for each sample, the reasoning chain and final output are independently shown to two judge models for binary error classification across five categories (medicinal chemistry, environment information, logic, reasoning–execution consistency, representation). Inter-judge binary agreement serves as a reliability check; definitions and criteria are in the appendix.

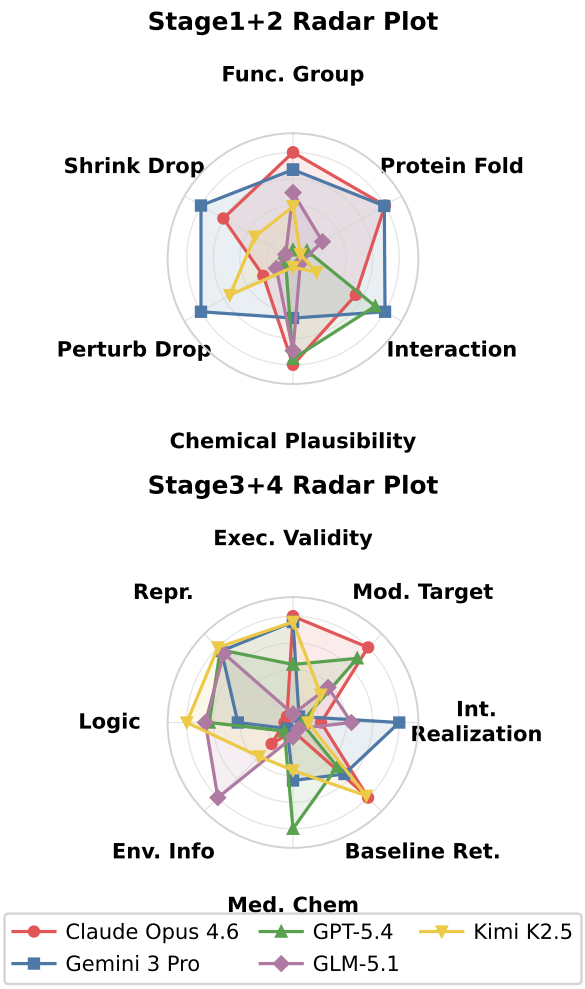


Figure 3: Radar plots compare the best-performing model from each vendor (Claude Opus 4.6, Gemini 3 Pro, GPT-5.4, GLM-5.1, Kimi K2.5) on selected metrics in Stage 1&2 (top) and Stage 3 & CoT (bottom). Axes are min-max normalized per plot across these five models (radius 0.08–1.00); the four CoT error-rate axes are inverted (100 − rate) beforehand so larger radius always means better, as on every other axis. No values are missing.

4 Experimental Results and Analysis

Figure 3 previews this cross-model comparison. We evaluate 14 frontier LLMs spanning five families—Claude, GPT Singh et al. (2025), Gemini Comanici et al. (2025), GLM, and Kimi, plus several Thinking-mode variants (full roster in the appendix)—through *identical, model-agnostic prompt templates*, with strategy execution validated via Boltz-2 co-folding. CoT quality is assessed by two independent LLM judges, Claude Opus 4.8 and GPT-5.6-sol (83.6% mean inter-judge agreement; §4.4); these two also appear as extra, non-headline rows in Tables 1–2 evaluated as ordinary subjects. Prompt templates are in the supplementary archive.

4.1 Stage 1: Structural Perception

Functional Group Identification. LLMs demonstrate strong competence in 2D molecular structure recognition: the best model (Claude Opus 4.6) reaches micro-F1 = 0.920, and 11 of 14 exceed 0.85, all with higher recall than precision (Appendix), indicating over-prediction rather than omission.

Table 1: Results for Stage 1 (structural perception) and Stage 2 (strategy discovery); all values are percentages. Func. Group = micro-F1 for functional group ID; Protein Fold = secondary-structure classification accuracy; Interaction = macro-F1 for ligand–protein interaction perception; Chem. Plaus. = fraction of chemically compatible residue–interaction pairs; Perturb/Shrink Drop = change in recommendation rate under distance perturbation. †GLM-5 outputs free-text instead of the required format on Task 1.3, yielding no parseable predictions. ‡Claude Opus 4.8 and GPT-5.6-sol also serve as CoT judges (§4.4); reported here as ordinary subjects for completeness but not among the "14 frontier LLMs" of the paper's headline claims (Appendix C).

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
Claude Opus 4.8‡	92.0	80.2	23.2	99.3	23.6	16.9
GPT-5.6-sol‡	89.1	77.7	21.0	99.1	43.6	25.9

Protein Fold Classification. Claude Opus 4.6 (90.5%) and Gemini 3 Pro (90.2%) lead, while GPT-5.4 Mini lags at 44.4%; performance is consistently higher on α-helix/β-sheet than on coil regions, suggesting easier recognition of regular geometric patterns.

Interaction Perception. Even the best model, Gemini 3 Pro, reaches only 39.4% macro-F1 (most models: 27–36%), again with high recall but low precision, indicating systematic over-prediction.

Key finding: a cross-entity ligand–pocket reasoning bottleneck, not a general 3D deficit. The cliff is not simply 2D vs. 3D: Task 1.2 already supplies coordinate-based 3D input and reaches up to 90% accuracy, so models handle 3D input *within a single entity* (the protein alone). The gap appears specifically when a task requires reasoning *jointly* over the ligand and pocket together: models exceeding 0.90 micro-F1 on functional groups and 90% on secondary structure cannot top F1 = 0.40 on interaction perception, which relates ligand atoms to protein residues. This holds across all model families and scales, and is not attributable to hallucination: mainstream models fabricate fewer than 0.4% of structural entities (§4.4), so low interaction-perception scores reflect genuine reasoning limitations.

4.2 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).

Chemical Plausibility. All models achieve >93% chemical plausibility, indicating that LLMs reliably satisfy this generic residue–interaction type compatibility check. GPT-5 leads at 99.5%, while Gemini 3 Flash (93.6%) and Kimi K2.5 (94.4%) are at the lower end. This check is coarse—generic (Y, Z) compatibility, not context-sensitive reasoning quality—and should not be read as endorsing

medicinal-chemistry reasoning in general: the finer-grained CoT medicinal-chemistry error rates in §4.4 are substantially higher (31.4–73.4%), so the two should be read as distinct capabilities.

Distance Sensitivity. Under normal distances, most models recommend the target residue 78–90% of the time (BasRate; GPT-5.4 Mini is a low outlier at 64.4%); after perturbation this drops to 34–53% (PtbRate), an average RetainDrop of ∼39%, led by Gemini 3 Pro (49.2%) and Gemini 3 Flash (47.0%), while GPT-5.4 Mini is again an outlier (11.1%). In shrink mode, target recommendation rates drop substantially further (RetDrop 20.3–64.7%), again led by Gemini 3 Pro.

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 (§4.1). The models *sense* 3D context but cannot *reason* from it to predict which interactions will actually form.

4.3 Stage 3: Strategy Execution

Table 2: Results for Stage 3 execution and two representative CoT error categories (full 5-category breakdown in appendix); all values are percentages. Exec. Validity = SMILES parse/sanitize/Kekulize pass rate over completed responses (0–22.9% truncated per model, excluded; counts in Appendix Table 14); Mod. Target = fraction of edits at the strategy-specified position; Int. Realization = whether the intended residue–interaction pair appears post-Boltz-2/PLIP; Baseline Ret. = fraction of original PLIP interactions preserved; Med. Chem./Logic = CoT judge error rates (Claude Opus 4.8); Coverage = fraction with an extractable reasoning trace. †Coverage <50%: GLM-5/GLM-5.1 rates rest on a small, non-representative sample. ‡Claude Opus 4.8 and GPT-5.6-sol also serve as CoT judges above; reported as ordinary subjects for completeness, not among the "14 frontier LLMs" of the headline claims (Appendix C). §Opus 4.8's Med. Chem./Logic values are *self*-judged, unlike every other row; Appendix C shows self-judged rates run roughly half of an independent judge's on the same traces, so read as a lower bound.

Model	Strategy Execution				CoT Error Analysis		
	Exec. Val.	Mod. Target	Int. Real.	Baseline Ret.	Med. Chem.	Logic	Coverage
Claude Sonnet 4.6	98.8	83.3	22.3	48.5	73.4	52.2	94.7
w/ Thinking	96.8	84.1	19.4	47.7	56.6	32.1	66.9
Claude Opus 4.6	98.9	85.1	21.5	46.1	64.2	57.5	97.4
w/ Thinking	99.4	84.2	21.4	45.3	55.2	45.6	92.8
Gemini 2.5 Flash	91.0	80.7	24.1	46.2	73.4	55.4	96.9
Gemini 2.5 Pro	100.0	82.9	18.5	45.0	52.2	29.9	99.5
Gemini 3 Flash	97.9	83.6	19.1	45.4	42.1	24.6	71.6
Gemini 3 Pro	98.5	83.2	25.1	44.8	47.6	33.1	98.9
GPT-5.4 Mini	96.0	82.8	19.1	45.8	32.6	17.0	72.9
GPT-5	92.5	83.8	20.0	44.3	45.1	23.8	100.0
GPT-5.4	95.5	84.8	20.6	44.4	31.4	18.2	99.9
GLM-5†	87.6	82.7	24.1	45.1	57.7	5.6	45.3
GLM-5.1†	92.0	84.0	22.9	42.3	62.2	16.1	9.1
Kimi K2.5	98.5	83.8	20.9	46.0	51.0	6.4	88.6
Claude Opus 4.8‡	100.0	81.8	23.2	47.4	26.9§	22.0§	99.9
GPT-5.6-sol‡	99.0	83.8	18.3	46.3	16.3	8.2	99.9

Execution Validity. Once truncated responses are excluded from the denominator (see caption; per-model rates in the appendix), nearly all models produce highly valid edits: Gemini 2.5 Pro reaches 100%, 10 of 14 exceed 95%, and GLM-5 is the floor at 87.6%. Mod. Target is strikingly uniform across models at ∼83%, indicating consistent but imperfect strategy comprehension.

Interaction Realization. Between 18.5% and 25.1% of intended interactions are realized across all 14 models (Gemini 3 Pro highest at 25.1%, Gemini 2.5 Pro lowest at 18.5%). This 6.6-point spread is modest relative to other Stage 3 metrics and tracks neither model family nor scale, so no model—from compact GPT-5.4 Mini to frontier Claude Opus—shows a clear edge in producing structurally effective edits.

Baseline Interaction Retention. On average, $\sim$45% of the original PLIP interactions survive editing (macro-retention 42.3–48.5%). Boltz-2 itself is noisy even absent editing: re-folding an unmodified molecule agrees with itself only 72.3% of the time (self-retention), and a single fold agrees with the *experimental crystal structure* only 45.7% of the time (crystal fidelity, over 122 baseline complexes)—nearly indistinguishable from the post-edit retention rate (discussed below).

Key finding: valid chemistry, unreliable structural realization. High execution validity (>95% for top models) alongside low interaction realization (18.5–25.1%) captures a persistent challenge: LLMs reliably produce chemically valid, correctly targeted edits, yet in three-quarters or more of cases the intended interaction still fails to materialize in the co-folded structure.

Note on structural validation reliability. Co-folding models are known to introduce noise Škrinjar et al. (2025); Wan et al. (2026). We mitigate this using crystal structures directly wherever available and an identical Boltz-2 pipeline for all 14 models (chosen over AlphaFold 3 Abramson et al. (2024) for open weights and API access), so bias is shared and comparisons remain valid; the benchmark is oracle-agnostic and stronger predictors can be substituted freely. A quantified noise ceiling is discussed in the appendix (Limitations).

4.4 Inference Quality

Hallucination Detection. Asked to restate PLIP interaction reports in structured format, mainstream models fabricate fewer than 0.4% of structural entities (most: exactly 0%; sole outlier Kimi K2.5 at 11.9%), confirming that the low interaction-perception scores in §4.1 reflect genuine reasoning limitations rather than fabrication.

CoT Error Analysis. Full per-model results and coverage are in the appendix. Two independent LLM judges (Claude Opus 4.8, GPT-5.6-sol) agree on error presence/absence 83.6% of the time on average; rates are computed only over samples with an extractable trace, so GLM-5.1 (9.1% coverage) and GLM-5 (45.3%) rest on small, potentially non-representative samples. Among models with adequate coverage (>60%), medicinal-chemistry and logical errors dominate, led by Claude Sonnet 4.6/Gemini 2.5 Flash (73.4% medicinal-chemistry) and Claude Opus 4.6 (57.5% logical), while GPT-5.4/GPT-5.4 Mini are lowest (31.4%/17.0%). Contrary to the intuition that longer chains accumulate more errors, thinking-mode variants consistently produce *fewer*—e.g., Claude Opus's medicinal-chemistry error rate falls from 57.0% to 40.0%, Sonnet's from 67.8% to 43.2%, with logical errors following suit—suggesting explicit reasoning improves rather than dilutes output quality.

5 Conclusion

We introduced LAO-Eval, a diagnostic benchmark for evaluating LLM capabilities in structure-based ligand affinity optimization. Across 14 general-purpose frontier LLMs, prompted directly and zero-shot, and 30 DUD-E targets, models perform well on 2D molecular understanding, chemical strategy formulation, and molecular editing, but struggle with cross-entity ligand–pocket reasoning, so chemically valid edits rarely realize their intended interactions. This bottleneck motivates molecular models that more directly integrate protein–ligand geometry and physics-aware structural feedback; we scope this conclusion to the directly-prompted, general-purpose LLMs we evaluate, leaving fine-tuned or LLM-plus-3D-generator pipelines (Section 2) as an open question.

Recommended metrics for model comparison. We recommend **Protein Fold Accuracy**, **Perturb RetainDrop**, **Shrink Drop**, and **Medicinal-Chemistry Error Rate** as primary comparison metrics, since they show substantial cross-model variation; Chemical Plausibility and Execution Validity are better read as absolute capability indicators, as most models already perform similarly on them.

AI USE STATEMENT

We used generative AI tools (LLMs) as the *subjects of study*: all reported results reflect their outputs under the protocols in Sections 3–4 and the appendix. Separately, we used generative AI tools to assist with manuscript writing/editing and portions of the code implementation; all such text and code were reviewed, tested, and verified by the authors, who take responsibility for the final content.

ETHICS STATEMENT

This work involves no human subjects or personally identifiable data. Protein–ligand structures are drawn from the public resource DUD-E (Mysinger et al., 2012); interactions are annotated with the open-source tool PLIP (Salentin et al., 2015). We see no direct dual-use concern: our central finding is that current frontier LLMs are *unreliable* at realizing intended ligand–pocket interactions, which argues against unsupervised reliance on them for structure-based design.

REPRODUCIBILITY STATEMENT

Dataset construction is detailed in Appendix E; per-task data/compute scale in Appendix E.4; metric definitions in Sections 3.2–3.5 and Appendices G–H; prompt excerpts in Appendix I; and measurement-noise sources (Boltz-2 repeatability, judge self-preference) in Appendix J. Full prompts, outputs, and code are in the supplementary archive.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

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

Konrad H Bleicher, Hans-Joachim Böhm, Klaus Müller, and Alexander I Alanine. Hit and lead generation: beyond high-throughput screening. *Nature reviews Drug discovery*, 2(5):369–378, 2003.

Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3): 1096–1108, 2019.

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

Xiangsen Chen, Ruilong Wu, Yanyan Lan, Ting Ma, and Yang Liu. Molevolve: Llm-guided evolutionary search for interpretable molecular optimization. *arXiv preprint arXiv:2603.24382*, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Vishal Dey, Xiao Hu, and Xia Ning. Large language models for controllable multi-property multi-objective molecule optimization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20996–21023, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1145. URL https://aclanthology.org/2025.findings-emnlp.1145/.

Bowen Gao, Yin Huang, Yuyan Liu, Wenlong Xie, Bowei He, Haichuan Tan, Wei-Ying Ma, Yan-Qing Zhang, and Yanyan Lan. Cidd: Collaborative intelligence for structure-based drug design empowered by llms. *Advances in Neural Information Processing Systems*, 38:15293–15341, 2025.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H. Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/4c56ff4ce4aaf9573aa5dff913df997a-Abstract-round1.html.

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

Fangjun Li, David C Hogg, and Anthony G Cohn. 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.

Haoran Liu, Zheni Zeng, Yukun Yan, Yuxuan Chen, and Yunduo Xiao. Drugr: Optimizing molecular drugs through llm-based explicit reasoning. *arXiv preprint arXiv:2602.08213*, 2026.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

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

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Fedor Rodionov, Abdelrahman Eldesokey, Michael Birsak, John Femiani, Bernard Ghanem, and Peter Wonka. FloorplanQA: A benchmark for spatial reasoning in LLMs using structured representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bcN12UkftY.

Sebastian Salentin, Sven Schreiber, V Joachim Haupt, Melissa F Adasme, and Michael Schroeder. Plip: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Peter Škrinjar, Jérôme Eberhardt, Janani Durairaj, and Torsten Schwede. Have protein-ligand co-folding methods moved beyond memorisation? *BioRxiv*, pp. 2025–02, 2025.

Shunzhou Wan, Xibei Zhang, Xiao Xue, and Peter V Coveney. On the reliability of ai methods in drug discovery: Evaluation of boltz-2 for structure and binding affinity prediction. *arXiv preprint arXiv:2603.05532*, 2026.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, Gege Qi, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. Drugassist: A large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, 2025.

A Full-Size Benchmark Overview Figure

Figure 2 is reproduced here at full size for legibility.

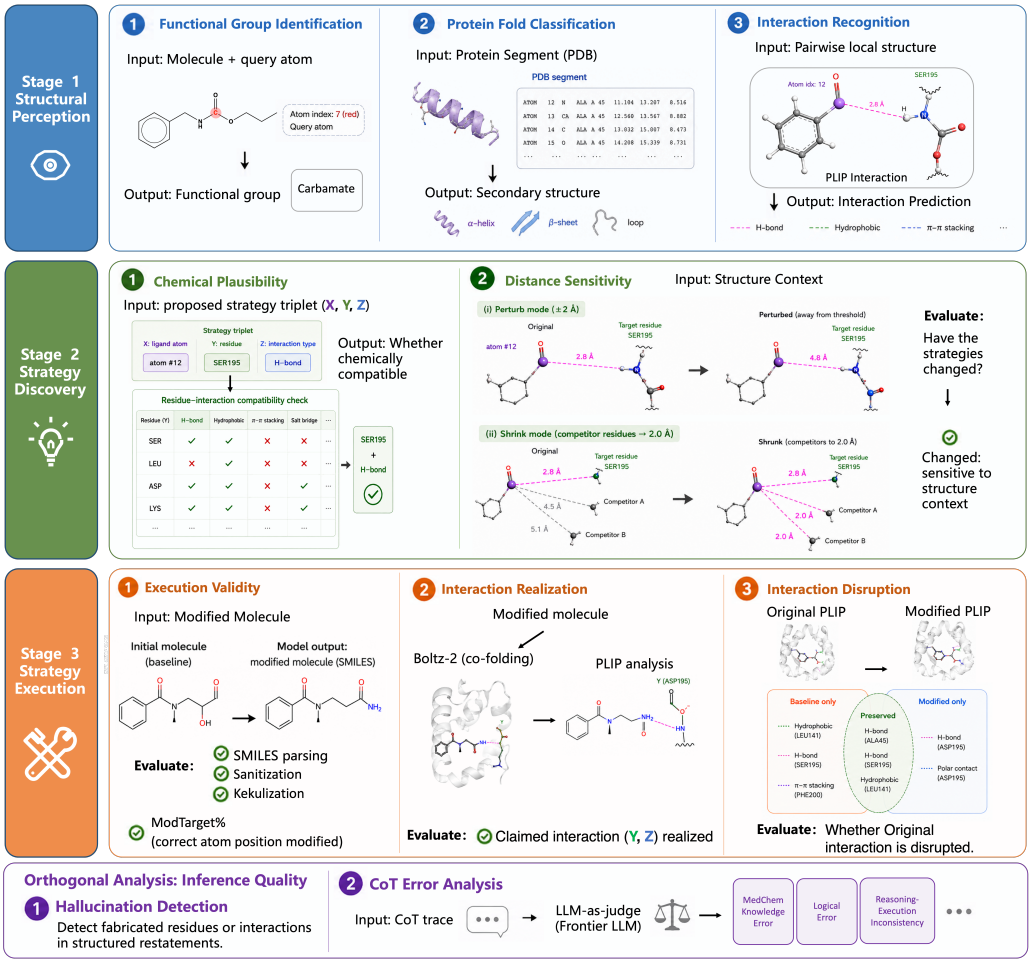


Figure 4: Full-size version of Figure 2: overview of **LAO-Eval**. LAO-Eval evaluates LLMs for structure-based ligand affinity optimization by decomposing the reasoning pipeline into three sequential stages: structural perception, optimization strategy formulation, and chemical execution. Each stage 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.

Note on two additional models. The Stage 1–3 tables here and in the main paper (Tables 1–2) each include two supplementary rows, marked ‡, for Claude Opus 4.8 and GPT-5.6-sol. These two models serve as the CoT judge models used throughout the paper's inference-quality analysis (Section C); we additionally evaluate them under the exact same protocol, sample counts, and target coverage used for the 14 models reported elsewhere, so that their performance can be directly compared on the same footing. They are not included among the "14 frontier LLMs" referenced in the paper's headline claims, since one of them is also a judge and their inclusion would blur the judge/subject separation those claims rely on; the appendix tables below additionally provide per-class and precision/recall detail beyond what the main-paper tables show for all 16 rows. In every numeric table in Sections F–H, **bold** marks the highest and lowest value in each column, computed across all 16 rows shown (the 14 main models plus the two ‡ rows), since those tables use fully automated metrics (exact-match, PLIP, RDKit) with no judge model involved and the two additional models can be compared on an equal footing. The one exception is Table 4 in Section C, where Claude Opus 4.8 is itself one of the two rows being scored; there, bold is restricted to the 14 main models only (see that table's caption).

B TASK SPECIFICATION MATRIX

Table 3 summarizes, for every task in LAO-Eval, the representation supplied to the model, the quantity withheld and scored, the source of the reference annotation, and the sampling unit underlying each reported rate. Tasks 2.1 and 3.1 are post-hoc scorings of upstream model outputs rather than separate elicitations and so require no additional LLM calls (see Appendix E.4).

Table 3: Task specification matrix. "Input origin" indicates whether the scored quantity in a task's input is oracle-provided (fixed by the benchmark) or model-generated (produced by the subject model in this task or an upstream one). Sampling units (denominators) match Appendix E.4.

Task	Input representation	Scored quantity (withheld)	Reference source	Input origin
1.1	SMILES + queried atom index	functional group at queried atom	RDKit substructure match	oracle molecule
1.2	PDB-format coordinate segment	local secondary-structure label	structure-derived secondary-structure annotation	oracle structure
1.3	pre-digested text description of neighboring atoms/residues	presence of specific interaction type(s)	PLIP report on crystal structure	oracle structure
2.1	Strategy Discovery output, i.e. model-proposed (X, Y, Z)	chemical compatibility of (Y, Z)	residue–interaction compatibility rules	model-generated (upstream)
2.2	PLIP report + perturbed/shrunk residue distances	change in target-residue recommendation rate	programmatic distance recomputation	oracle geometry; model-generated recommendation
3.1	baseline SMILES + curated strategy (X, Y, Z)	validity/target-atom accuracy of edited SMILES	RDKit parse/sanitize/Kekulize + MCS atom tracking	oracle strategy; model-generated edit
3.2	same	whether curated (Y, Z) appears post-edit	Boltz-2 co-folding + PLIP	oracle strategy + oracle structure predictor
3.3	same	retention of baseline PLIP interactions	Boltz-2 co-folding + PLIP	oracle strategy + oracle structure predictor
4.1	PLIP report + pocket-neighbor description	fabricated entities in restatement	string/entity match vs. input	oracle input; model-generated restatement
4.2	reasoning traces from Tasks 1.1, 2.x, 3.x	presence of 5 error categories	LLM-as-judge (2 independent judges)	model-generated

C INFERENCE QUALITY: ORTHOGONAL ANALYSIS

Hallucination Detection. 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. This near-zero baseline establishes that the low interaction perception scores in Stage 1 (Structural Perception) of the main paper reflect genuine reasoning limitations rather than information fabrication.

CoT Error Analysis. Medicinal-chemistry and logical errors are the two most frequently detected error categories. Among models with adequate coverage, Claude Sonnet 4.6 and Gemini 2.5 Flash exhibit the highest medicinal-chemistry error rate (73.4% each) and Claude Opus 4.6 the highest logical-error rate (57.5%), while GPT-5.4 and GPT-5.4 Mini are lowest on these two categories, respectively (31.4% and 17.0%). The GLM results should be interpreted cautiously because reasoning-trace coverage is only 45.3% for GLM-5 and 9.1% for GLM-5.1; their category-level error rates are therefore estimated from small, potentially non-representative subsets of responses and should

Table 4: CoT Error Analysis — error rates across three source tasks (functional group identification, strategy discovery, strategy execution), scored by two independent LLM judges (Claude Opus 4.8 and GPT-5.6-sol) as binary (present/absent) per reasoning chain; values shown are from Claude Opus 4.8 (see inter-judge agreement below). Rates are computed only over samples with an extractable reasoning trace (Coverage). All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi). †Coverage <50%: these models frequently emit no reasoning trace at all before their final answer, so their rates rest on a small, potentially non-representative sample and should not be compared directly against high-coverage models. ‡Claude Opus 4.8 and GPT-5.6-sol are also evaluated here as ordinary subject models (Section F, note on two additional models). §For the Claude Opus 4.8 row specifically, the "judge" column is the subject model itself: these values are a *self*-evaluation, not an independent third-party judgment as for every other row (including the GPT-5.6-sol row, still judged by Claude Opus 4.8). See the self-evaluation-optimism discussion below the Inter-Judge Agreement paragraph. **Bold** marks the highest and lowest value in each column among the 14 main-paper models (consistent with the CoT Error Analysis paragraph below); the two supplementary rows are shown for reference but excluded from this marking, since Claude Opus 4.8 is a judge here and its own row is a self-evaluation.

Model	Med. Chem.	Env. Info.	Logic	Reas.–Exec.	Repr.	Coverage
Claude Sonnet 4.6	**73.4**	24.0	52.2	2.3	7.1	94.7
Claude Sonnet 4.6 (T)	56.6	11.9	32.1	**1.2**	7.0	66.9
Claude Opus 4.6	64.2	15.6	**57.5**	5.6	26.2	97.4
Claude Opus 4.6 (T)	55.2	13.3	45.6	4.5	26.4	92.8
Gemini 2.5 Flash	**73.4**	**27.9**	55.4	12.3	14.8	96.9
Gemini 2.5 Pro	52.2	10.2	29.9	3.4	28.6	99.5
Gemini 3 Flash	42.1	18.1	24.6	2.4	19.7	71.6
Gemini 3 Pro	47.6	19.2	33.1	3.6	4.1	98.9
GPT-5.4 Mini	32.6	16.7	17.0	**23.2**	5.1	72.9
GPT-5	45.1	18.1	23.8	3.6	**31.4**	**100.0**
GPT-5.4	**31.4**	18.6	18.2	17.5	3.7	99.9
GLM-5†	57.7	16.8	**5.6**	4.5	6.4	45.3
GLM-5.1†	62.2	**3.1**	16.1	**1.2**	4.9	**9.1**
Kimi K2.5	51.0	12.7	6.4	3.6	**2.9**	88.6
Claude Opus 4.8‡§	26.9	7.3	22.0	6.5	2.5	99.9
GPT-5.6-sol‡	16.3	1.6	8.2	7.7	4.5	99.9

not be treated as evidence of superior reasoning quality. In particular, among the extractable GLM-5.1 traces, the medicinal-chemistry error rate is 62.2%, among the highest of all 14 models. For reference, the two supplementary CoT-judge models (Table 4, bottom two rows) show a different pattern: GPT-5.6-sol's Claude-Opus-4.8-judged error rates are numerically the lowest in the table on medicinal chemistry (16.3%) and environment information (1.6%), while Claude Opus 4.8's own row is judged by itself and is therefore not comparable to any other row in this table (see the self-evaluation discussion below).

Inter-Judge Agreement. To validate the LLM-as-judge protocol, every sample is scored independently by both judges (Claude Opus 4.8 and GPT-5.6-sol), and we report their binary agreement rate — whether the two judges agree on error presence/absence for a given category and sample. Averaged across all 14 models, 3 source tasks, and 5 error categories, the mean binary agreement rate is 83.6% (per-model-task range: 71.6–100%), supporting the reliability of the reported error rates. For the two supplementary models, the same per-model-task agreement rate could be computed for 5 of the 6 (model, task) pairs — the GPT-5.6-sol-judged Claude-Opus-4.8/strategy-execution pair failed at the API level and was excluded, consistent with how missing judge calls are handled everywhere else in this pipeline. Over those 5 pairs the mean is 78.6% (range 67.8–85.5%), slightly below but broadly consistent with the 14-model range; as with the main 14 models, Strategy Discovery is the lowest-agreement task for both supplementary models (67.8% for GPT-5.6-sol, 71.2% for Claude Opus 4.8).

Does a judge go easy on itself? Because the Claude Opus 4.8 row of Table 4 is judged by Claude Opus 4.8 itself, we can directly compare its self-judged error rates against the same reasoning traces scored by the independent judge, GPT-5.6-sol (functional group identification and strategy discovery only, since the self/GPT-5.6-sol pair failed for strategy execution as noted above). The two views disagree sharply and in one direction: self-judged rates are 26.9% vs. 59.8% (GPT-5.6-sol-judged) for medicinal chemistry, 7.3% vs. 30.5% for environment information, and 22.0% vs. 46.5% for logical errors — in every case, Claude Opus 4.8 finds roughly half as many errors in its own reasoning as an independent judge does. We can run the same comparison for GPT-5.6-sol, whose row is genuinely independently judged (by Claude Opus 4.8) in Table 4: its self-judged rates are 27.3% (medicinal chemistry), 36.5% (environment information), and 12.0% (logical), compared to 16.3%, 1.6%, and 8.2% when Claude Opus 4.8 judges it — here self-judgment finds *more* errors than the external judge, the opposite direction from Claude Opus 4.8's self-judgment. This asymmetry indicates that self-preference in LLM-as-judge is not a generic, model-agnostic effect of judging one's own output: at least one of our two judge models (Claude Opus 4.8) shows a large, consistent leniency bias toward itself, while the other (GPT-5.6-sol) does not exhibit the same bias, and is if anything harsher on itself, on the categories that matter most here. Given this, the Claude Opus 4.8 row in Table 4 should be read as a lower bound on that model's true error rate rather than a like-for-like entry alongside the other 15 rows, all of which are judged by a party other than the subject model.

Key finding: thinking-mode inference is associated with fewer judged errors, but the comparison should be read cautiously. Across the two Claude model pairs, thinking-mode variants show lower error rates in most categories: on functional group identification, Claude Opus's medicinal-chemistry error rate falls from 57.0% to 40.0%, and Sonnet's from 67.8% to 43.2%, with logical errors following the same direction for both judges. The pattern is not universal, however — Claude Opus's representation-error rate rises slightly (26.2%→26.4%; Table 4) — and we observed substantial run-to-run variance when repeating this CoT evaluation pipeline on identical (model, task, judge) configurations. Individual category-level comparisons, especially for lower-magnitude categories like representation, should therefore be interpreted as noisy rather than as precise, reproducible effect sizes; we report the pattern as suggestive that explicit reasoning is associated with fewer judged errors overall, not as a uniform, causally established effect.

D End-to-End Affinity Optimization

Objective. We test whether frontier LLMs can perform end-to-end structure-based ligand 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, GPT-5.4, and Gemini 3 Pro.

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.
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.
5. **Comparison.** We compute $\Delta\text{pIC}_{50} = \text{pIC}_{50}^{\text{modified}} - \text{pIC}_{50}^{\text{original}}$, where positive values indicate improved binding.

Results. Of 90 possible (target, model) pairs, 89 produced valid modified SMILES that could be co-folded.

Table 5: Stage 0 summary statistics. $\overline{\Delta}$ = mean pIC_{50} change; r = Pearson correlation between original pIC_{50} and ΔpIC_{50}; Impr. = percentage of targets with $\Delta > 0$.

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

No model shows a consistent, directionally positive mean improvement in pIC_{50}: Claude Opus 4.6's mean change is slightly negative (-0.008), while Gemini 3 Pro ($+0.126$) and GPT-5.4 ($+0.040$) show small positive means, and Impr. (the fraction of targets with $\Delta > 0$) stays close to chance (50–62%) for all three models. 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. We note, however, that ΔpIC_{50} and original pIC_{50} share the baseline prediction as a common term, so some part of the observed negative association is a mathematical consequence of that coupling together with Boltz-2 prediction noise on repeated folds, rather than purely a behavioral effect of LLM editing. Isolating the two would require no-edit reprediction controls (repeatedly re-folding the *unmodified* ligand and regressing the resulting noise against baseline affinity) and is left to future work; the pattern reported here should therefore be read as suggestive of regression-to-the-mean-like behavior rather than as a fully deconfounded causal estimate.

E Dataset Construction Details

E.1 Target Selection

LAO-Eval draws its protein targets from the DUD-E database. We select 30 targets spanning 15 protein families and functional classes to ensure diversity in binding-site topology, pocket size, and pharmacological relevance (Table 6).

Selection criteria. Targets were selected to satisfy three requirements: (i) coverage of major therapeutic protein families; (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.

Pose generation for additional actives. Beyond the co-crystallized ligand, each target's remaining DUD-E confirmed actives are docked into the same binding site (identical grid center, taken from the target's co-crystal structure) using Glide. Actives for which Glide did not return a valid pose are excluded from the sampling pool; N_{posed} in Table 6 reports the resulting per-target pool size. Four actives are then drawn from this pool via MaxMin diversity sampling (RDKit `MaxMinPicker` over Morgan/ECFP4-like fingerprints under Tanimoto distance; seed 42, fixed across all targets), together with the co-crystallized ligand giving the 5 hits (150 across 30 targets) used throughout the benchmark.

E.2 PLIP Interaction Annotation

Ground-truth interaction annotations are generated programmatically using PLIP with its default geometric and physicochemical criteria.

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.

E.3 Expert-Curated Strategy Construction

Phase 1: LLM-assisted candidate generation. For each of the 30 protein targets, candidate optimization strategies were first generated using Claude Opus 4.6 under a structured prompt requiring

Table 6: The 30 DUD-E protein targets in LAO-Eval, organized by protein family. N_{actives} = total confirmed actives in DUD-E; N_{posed} = actives with a successfully generated Glide docking pose (the pool from which 4 are MaxMin-sampled per target; see Pose generation for additional actives above).

Family	Target	Full Name	PDB	N_{actives}	N_{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 α	1SJ0	383	383
	pparg	PPAR-γ	2GTK	484	484
GPCR	cxcr4	CXCR4 chemokine receptor	3ODU	40	40
	adrb2	β_2-adrenergic receptor	3NY8	231	231
	drd3	Dopamine D3 receptor	3PBL	480	480
Protease	bace1	β-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 α	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 90α	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

Table 7: 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 $> 120^\circ$
Hydrophobic	C–C: < 4.0 Å	Both atoms non-polar
π-Stacking	Centroid: 3.3–5.5 Å	Dihedral and offset constraints
Salt bridge	Charged groups: < 5.5 Å	Requires formal charge on both partners

each output to conform to the (X, Y, Z) triplet format. Each candidate was then screened using residue-level chemical plausibility rules applied in Task 2.1.

Phase 2: Expert spatial review. The remaining candidates were reviewed by two domain experts with backgrounds in structural biology and medicinal chemistry. Each candidate was evaluated against a 3D molecular environment visualization rendered with a 6 Å pocket cutoff. Experts applied three rejection criteria:

- **Severe steric clash**: the proposed modification at atom position X would introduce prohibitive steric overlap with surrounding pocket residues.

- **Extreme ligand–residue geometry**: the spatial arrangement between atom position X and target residue Y is geometrically incompatible with the intended interaction.
- **Interaction-angle infeasibility**: even if distance is permissive, the local geometry does not satisfy the angular requirements of interaction type Z.

Strategies rejected on any criterion by either expert were discarded; the final set consists only of candidates that both experts independently accepted. We did not separately quantify inter-annotator agreement between the two experts; this is noted as a limitation in the supplementary Limitations section. The final curated set comprises 201 strategies across 130 of the 150 (target, hit) pairs (the remaining 20 pairs yielded no candidate that survived both review phases); this is the source of the **Total** = 201 column in Table 14 and of the 201 Strategy Execution calls per model in Table 8 below.

E.4 Per-Task Data and Compute Scale

Table 8 reports, for each task, the fixed input scale shared by all models, the number of LLM API calls one model requires to complete it, the total over the 14 main-paper models, and any Boltz-2/PLIP structural-analysis calls involved. Strategy Discovery and Strategy Execution are listed as shared upstream stages because their outputs are reused, at no additional LLM cost, by the tasks below them (Tasks 2.1–2.2 and 3.1–3.3, respectively).

Table 8: Per-task data and compute scale. **LLM Calls/Model** is the number of API calls one model requires to complete the task; **Total (14 Models)** sums this over the 14 main-paper models (excluding the two supplementary CoT-judge models of Section C, and excluding Stage 0, which used only 3 models). Counts are read directly from per-run `meta.json`/`summary.json` metadata, not estimated.

Task	Fixed Input Scale	Calls/Model	Total (14)	Boltz-2 / PLIP
Stage 0 pilot (§D)	30 targets × 3 models	1/pair	90 (3 models)	30 baseline folds (shared) + 89 modified folds
1.1 Func. Group ID	150 molecules (30 targets × 5 hits)	150	2,100	—
1.2 Protein Fold	358 protein fragments	358	5,012	—
1.3 Interaction Recog.	30 crystal complexes	30	420	30/model
Strategy Discovery (feeds 2.1, 2.2)	150 (target, hit) pairs × 5 runs	750	10,500	—
2.1 Chem. Plausibility	reuses Strategy Discovery output	0	0	—
2.2 Dist. Sens., perturb	≤150 hits × 5 runs*	355–620	5,835	—
2.2 Dist. Sens., shrink	150 hits × 5 runs	750	10,500	—
Strategy Execution (feeds 3.1–3.3)	201 expert-curated strategies	201	2,814	—
3.1 Execution Validity	reuses Strategy Execution output	0	0	—
3.2/3.3 Realization & Disruption	123 baseline complexes (shared, one-time)	0	—	123 baseline (shared) + 142–189/model modified folds (2,469 total)
4.1 Hallucination	30 targets × 3 runs	90	1,260	30/model
4.2 CoT Error Analysis	≤1,101 traces/model (150+750+201), ×2 judges	≤2,202	≤30,800	—

*Perturb-mode coverage depends on whether a model's own most-frequent baseline recommendation for a given hit is present in that hit's perturbed prompt context, so the number of hits reachable for perturbation is model-dependent: it ranges from 71/150 (Claude Opus 4.6) to 124/150 (GLM-5.1) across the 14 models. Shrink mode has no such restriction and always covers all 150 hits.

Aggregate compute. Excluding CoT judging and the two post-hoc-only tasks (2.1, 3.1), a single model requires between 2,684 and 2,949 LLM calls to complete Stages 1–3, depending on its own perturb-mode coverage; summed over the 14 main-paper models this totals 38,441 LLM calls. Task 4.2 additionally requires up to 2,202 judge calls per model (2 judges $\times$ up to 1,101 reasoning traces drawn from Tasks 1.1, Strategy Discovery, and Strategy Execution), i.e. up to $\sim$30,800 across the 14 models; actual judged counts are lower wherever Coverage $<$100% (Table 4), since only responses with an extractable reasoning trace are scored. Tasks 3.2–3.3 additionally require 123 one-time baseline Boltz-2 co-foldings, shared across all models since the unmodified complex does not depend on which model proposed the edit, plus 142–189 modified-molecule co-foldings per model (2,469 total across the 14 models), each followed by a PLIP re-analysis of the resulting structure.

F Stage 1: Detailed Metrics

Table 9: Functional Group Identification — Full Metrics. All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); micro-F1 reproduces the Func. Group column of the main paper's Stage 1/2 table. ‡Claude Opus 4.8 and GPT-5.6-sol are also used as CoT judge models elsewhere in this paper (Section C); they are evaluated here as two additional subject models under the identical protocol used for the other 14 models.

Model	micro-P	micro-R	micro-F1	macro-F1
Claude Sonnet 4.6	86.1	94.0	89.9	82.5
Claude Sonnet 4.6 (T)	87.0	93.3	90.0	83.5
Claude Opus 4.6	89.1	**95.1**	**92.0**	**85.7**
Claude Opus 4.6 (T)	88.9	94.9	91.8	**85.7**
Gemini 2.5 Flash	81.2	87.3	84.1	77.9
Gemini 2.5 Pro	82.2	93.4	87.4	81.2
Gemini 3 Flash	84.5	94.1	89.0	82.4
Gemini 3 Pro	89.3	91.8	90.5	84.2
GPT-5.4 Mini	**73.8**	**77.8**	**75.8**	**68.6**
GPT-5	88.2	89.7	88.9	82.9
GPT-5.4	81.0	86.2	83.5	78.1
GLM-5	88.2	91.0	89.6	82.8
GLM-5.1	84.8	92.5	88.5	82.7
Kimi K2.5	84.6	90.1	87.3	79.1
Claude Opus 4.8‡	**91.7**	92.2	**92.0**	85.1
GPT-5.6-sol‡	90.1	88.2	89.1	82.8

G Stage 2: Detailed Metrics

Metric Definitions. For Task 2.2 (Distance Sensitivity):

- **Perturb mode.**

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

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

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

A high RetainDrop indicates that the model's strategy selection is genuinely sensitive to spatial distances.

Table 10: Protein Fold Classification — Per-Class Accuracy. All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); Overall reproduces the Protein Fold column of the main paper's Stage 1/2 table. ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	Overall	Helix	Sheet	Coil
Claude Sonnet 4.6	87.2	94.8	75.6	85.6
Claude Sonnet 4.6 (T)	88.3	92.9	81.4	87.3
Claude Opus 4.6	**90.5**	89.6	91.9	**90.7**
Claude Opus 4.6 (T)	90.2	92.9	84.9	**90.7**
Gemini 2.5 Flash	66.5	89.6	61.6	39.8
Gemini 2.5 Pro	72.1	96.1	97.7	22.0
Gemini 3 Flash	88.8	96.8	97.7	72.0
Gemini 3 Pro	90.2	90.3	94.2	87.3
GPT-5.4 Mini	**44.4**	89.6	**20.9**	**2.5**
GPT-5	78.5	**98.7**	**98.8**	37.3
GPT-5.4	57.0	**63.0**	31.4	67.8
GLM-5	70.1	94.2	62.8	44.1
GLM-5.1	63.7	79.7	72.6	42.6
Kimi K2.5	54.2	68.2	**20.9**	60.2
Claude Opus 4.8‡	80.2	82.5	88.4	71.2
GPT-5.6-sol‡	77.7	**98.7**	84.9	44.9

Table 11: Interaction Perception — Full Metrics. mi = micro-averaged, ma = macro-averaged (cf. full "micro"/"macro" spelling in Table 9). All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); ma-F1 reproduces the Interaction column of the main paper's Stage 1/2 table (GLM-5 is 0.0 here because it produced no parseable predictions, shown as "–" in the main paper). ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	mi-P	mi-R	mi-F1	ma-P	ma-R	ma-F1
Claude Sonnet 4.6	25.4	58.1	35.4	26.1	64.4	36.4
Claude Sonnet 4.6 (T)	24.3	54.4	33.6	23.9	61.0	33.3
Claude Opus 4.6	24.4	43.5	31.2	20.0	47.7	27.7
Claude Opus 4.6 (T)	26.6	53.8	35.6	24.7	55.9	33.6
Gemini 2.5 Flash	16.8	52.5	25.4	18.7	59.8	27.7
Gemini 2.5 Pro	23.5	59.7	33.8	22.4	65.2	32.4
Gemini 3 Flash	22.5	63.4	33.3	22.8	70.4	33.7
Gemini 3 Pro	**27.0**	**63.9**	**38.0**	**28.6**	**71.2**	**39.4**
GPT-5.4 Mini	18.1	43.5	25.6	20.5	47.3	27.4
GPT-5	20.7	50.7	29.4	22.0	56.5	30.8
GPT-5.4	24.6	59.4	34.8	25.1	65.2	35.5
GLM-5	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
GLM-5.1	3.1	9.6	4.7	3.7	12.2	5.5
Kimi K2.5	6.1	58.4	11.0	6.8	64.0	12.0
Claude Opus 4.8‡	16.6	32.6	22.0	17.6	36.9	23.2
GPT-5.6-sol‡	14.9	30.8	20.1	16.7	32.6	21.0

- **Shrink mode.**

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

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

Table 12: Distance Sensitivity (Perturb mode) — Full Metrics. All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); RetainDrop reproduces the Perturb Drop column of the main paper's Stage 1/2 table. ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	BasRate	PtbRate	RetainDrop
Claude Sonnet 4.6	84.9	41.0	43.9
Claude Sonnet 4.6 (T)	82.9	40.5	42.4
Claude Opus 4.6	86.8	48.2	38.6
Claude Opus 4.6 (T)	89.7	48.1	41.6
Gemini 2.5 Flash	77.7	38.8	38.9
Gemini 2.5 Pro	83.2	40.5	42.7
Gemini 3 Flash	87.6	40.6	47.0
Gemini 3 Pro	85.7	36.5	**49.2**
GPT-5.4 Mini	**64.4**	53.3	**11.1**
GPT-5	80.9	44.5	36.4
GPT-5.4	82.9	48.1	34.8
GLM-5	82.0	47.3	34.7
GLM-5.1	82.5	46.1	36.4
Kimi K2.5	78.8	**34.5**	44.3
Claude Opus 4.8‡	**95.6**	**72.0**	23.6
GPT-5.6-sol‡	94.3	50.7	43.6

Table 13: Distance Sensitivity (Shrink mode) — Full Metrics. All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); RetDrop reproduces the Shrink Drop column of the main paper's Stage 1/2 table. ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	BasRate	ShkRate	RetDrop
Claude Sonnet 4.6	88.1	68.3	46.3
Claude Sonnet 4.6 (T)	86.1	68.8	48.1
Claude Opus 4.6	91.3	74.0	57.3
Claude Opus 4.6 (T)	91.8	72.7	57.5
Gemini 2.5 Flash	80.3	57.0	41.5
Gemini 2.5 Pro	85.6	76.9	39.7
Gemini 3 Flash	90.4	67.8	61.5
Gemini 3 Pro	90.0	58.0	**64.7**
GPT-5.4 Mini	**69.8**	61.5	20.3
GPT-5	83.2	67.9	37.7
GPT-5.4	84.5	66.2	37.2
GLM-5	82.1	**29.8**	52.3
GLM-5.1	81.3	44.7	36.7
Kimi K2.5	82.0	51.9	46.9
Claude Opus 4.8‡	**95.6**	**78.7**	**16.9**
GPT-5.6-sol‡	94.3	68.4	25.9

H STAGE 3: DETAILED METRICS

Metric Definitions. Let $\mathcal{W}_{\mathrm{Boltz}}$ (wBoltz) denote the set of valid modified molecules for which Boltz-2 successfully produces a co-folded structure (i.e., valid SMILES *and* a completed co-folding run; molecules that fail to co-fold are excluded from this set and from all rates below). Let $\mathcal{R}_{\mathrm{int}} \subseteq \mathcal{W}_{\mathrm{Boltz}}$ (Reflected) denote the subset for which PLIP detects the exact intended (residue, interaction-type) pair on the co-folded structure, and let $\mathcal{R}_{\mathrm{res}} \subseteq \mathcal{W}_{\mathrm{Boltz}}$ (ResHit) denote the subset for which PLIP detects *any* interaction with the intended target residue, irrespective of interaction type (so $\mathcal{R}_{\mathrm{int}} \subseteq \mathcal{R}_{\mathrm{res}}$ always). We define:

$$\text{Realization Rate (Real. Rate)} = \frac{|\mathcal{R}_{\mathrm{int}}|}{|\mathcal{W}_{\mathrm{Boltz}}|}, \tag{7}$$

$$\text{ResHitRate} = \frac{|\mathcal{R}_{\mathrm{res}}|}{|\mathcal{W}_{\mathrm{Boltz}}|}, \tag{8}$$

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

For the Execution Validity table, **Changed** is the percentage of all responses (not just valid ones) that produce a valid molecule whose canonical SMILES differs from the starting ligand; a response can only count as Changed if it is also Valid. **ModTarget** is restricted to responses that are both valid and changed, and for which the MCS-based atom-tracking procedure yields a determinate match/no-match verdict for the strategy-specified atom position; responses where atom tracking is indeterminate (e.g., MCS computation fails) are excluded from both the numerator and denominator of ModTarget.

Table 14: Execution Validity — Full Metrics. Truncated counts responses cut off before completing their reasoning trace (excluded from ValidRate's denominator; see Stage 3: Strategy Execution in the main paper). Total/Valid/Truncated are raw sample counts; ValidRate, Changed, and ModTarget are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); ValidRate and ModTarget reproduce the Exec. Val. and Mod. Target columns of the main paper's Stage 3 table. ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	Total	Valid	Truncated	ValidRate	Changed	ModTarget
Claude Sonnet 4.6	201	160	39	98.8	79.6	83.3
Claude Sonnet 4.6 (T)	201	150	46	96.8	74.6	84.1
Claude Opus 4.6	201	176	23	98.9	87.6	**85.1**
Claude Opus 4.6 (T)	201	173	27	99.4	85.6	84.2
Gemini 2.5 Flash	201	183	0	91.0	91.0	**80.7**
Gemini 2.5 Pro	201	201	0	**100.0**	**100.0**	82.9
Gemini 3 Flash	201	191	6	97.9	95.0	83.6
Gemini 3 Pro	201	198	0	98.5	98.5	83.2
GPT-5.4 Mini	201	193	0	96.0	95.5	82.8
GPT-5	201	186	0	92.5	92.5	83.8
GPT-5.4	201	192	0	95.5	95.5	84.8
GLM-5	201	169	8	**87.6**	84.1	82.7
GLM-5.1	201	149	39	92.0	**74.1**	84.0
Kimi K2.5	201	193	5	98.5	96.0	83.8
Claude Opus 4.8‡	201	200	1	**100.0**	99.5	81.8
GPT-5.6-sol‡	201	199	0	99.0	99.0	83.8

I PROMPT TEMPLATES

Fairness guarantee. All 14 models are evaluated under *identical* prompt templates with no model-specific adaptation, system-prompt tuning, or few-shot examples. Complete prompt templates are included in the supplementary reproducibility archive un-

Table 15: Interaction Realization — Full Metrics. Reflected and ResHit are raw sample counts; Real. Rate and ResHitRate are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); Real. Rate reproduces the Int. Real. column of the main paper's Stage 3 table. ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	Reflected	ResHit	Real. Rate	ResHitRate
Claude Sonnet 4.6	33	60	22.3	40.5
Claude Sonnet 4.6 (T)	27	48	19.4	**34.5**
Claude Opus 4.6	35	62	21.5	38.0
Claude Opus 4.6 (T)	34	62	21.4	39.0
Gemini 2.5 Flash	42	66	24.1	37.9
Gemini 2.5 Pro	35	72	18.5	38.1
Gemini 3 Flash	35	68	19.1	37.2
Gemini 3 Pro	47	87	**25.1**	**46.5**
GPT-5.4 Mini	35	68	19.1	37.2
GPT-5	35	68	20.0	38.9
GPT-5.4	37	69	20.6	38.3
GLM-5	38	61	24.1	38.6
GLM-5.1	32	60	22.9	42.9
Kimi K2.5	38	72	20.9	39.6
Claude Opus 4.8‡	43	82	23.2	44.3
GPT-5.6-sol‡	33	69	**18.3**	38.3

Table 16: Baseline Interaction Retention — Full Metrics. All values are percentages. Models are grouped and ordered as in the main paper (Claude, Gemini, GPT, GLM, Kimi); MacroRet. reproduces the Baseline Ret. column of the main paper's Stage 3 table. ‡Also used as a CoT judge model (Section C); see Table 9 for details.

Model	MacroRet.	MicroRet.
Claude Sonnet 4.6	**48.5**	**49.1**
Claude Sonnet 4.6 (T)	47.7	48.6
Claude Opus 4.6	46.1	46.9
Claude Opus 4.6 (T)	45.3	46.2
Gemini 2.5 Flash	46.2	46.7
Gemini 2.5 Pro	45.0	46.1
Gemini 3 Flash	45.4	45.7
Gemini 3 Pro	44.8	45.3
GPT-5.4 Mini	45.8	46.3
GPT-5	44.3	45.6
GPT-5.4	44.4	45.1
GLM-5	45.1	45.8
GLM-5.1	**42.3**	**42.3**
Kimi K2.5	46.0	46.4
Claude Opus 4.8‡	47.4	48.2
GPT-5.6-sol‡	46.3	47.8

der `stage1_perception/`, `stage2_discovery/prompt_template.json`, and `stage3_execution/prompt_template.json`.

Stage 1: Structural Perception (Task 1.1 representative). *System prompt:*

```
You are a medicinal chemistry expert. Analyze ligand SMILES
and identify the functional group(s) present at each atom
position.
```

Stage 2: Strategy Discovery. *Output constraint (key excerpt):*

```
Generate EXACTLY 3 opportunities. Each opportunity MUST
follow this exact template:
"identify <ligand_atom> as an opportunity to <action>
<interaction_type> (with <Residue>)"
<action> MUST be exactly one of: introduce OR strengthen.
```

Stage 3: Strategy Execution. *Key constraint (key excerpt):*

```
Apart from changes strictly required to implement the
strategies, all other parts of the molecule must remain
unchanged.
```

J LIMITATIONS

Several limitations should be acknowledged. First, Boltz-2 is an imperfect structure predictor: refolding an unmodified baseline twice agrees with itself only 72.3% of the time (self-retention, an empirical estimate of fold-to-fold repeatability), and a single baseline fold agrees with the experimental crystal structure only 45.7% of the time (crystal fidelity, mean over 122 of the 150 target–hit baseline complexes; 27 were excluded because Boltz-2 co-folding did not complete successfully for that pair, and 1 further complex was excluded because PLIP detected no reference interactions in the crystal structure itself) — nearly indistinguishable from the $\sim$45% post-edit Baseline Interaction Retention rate reported in the main paper. This means that Baseline Interaction Retention cannot cleanly separate editing-induced disruption from variability introduced by Boltz-2. Accordingly, the observed interaction loss (1 − retention) should be interpreted as an upper bound on editing-specific disruption rather than as a precise estimate; a confound-robust retention metric is left to future benchmark iterations. Second, our strategy sets, while covering all 30 targets, represent a finite sampling of the optimization space, and while a subset of candidate strategies was reviewed by two independent experts for spatial feasibility, we did not separately quantify inter-annotator agreement between them. Third, the LLM-as-judge framework for CoT evaluation introduces potential bias; in addition, we observed non-trivial run-to-run variance when repeating CoT evaluation on identical (model, task, judge) configurations, so individual category-level error rates — particularly for lower-magnitude categories — should be read as indicative rather than precisely reproducible. Finally, our evaluation is limited to frontier commercial models; open-source and domain-fine-tuned models may exhibit different profiles.

We note four additional scope limitations raised during peer review. First, Interaction Realization (Task 3.2) checks only whether the intended (residue, interaction-type) pair (Y, Z) appears in the co-folded structure; it does not additionally verify that the realized interaction originates from the specific ligand atom position X named in the strategy, so an edit that installs interaction (Y, Z) through an unintended atom, or that leaves a pre-existing (Y, Z) contact undisturbed, is indistinguishable from a precise, intended realization under this metric. Second, Strategy Execution (Stage 3) evaluates models on expert-curated strategies rather than on each model's own Stage 2 strategy proposals; this isolates execution ability by design (Section 3.2), but it also means our results do not directly characterize the fully autonomous discover-and-execute pipeline, and we cannot rule out that strategy-formulation quality is an additional, independent source of end-to-end failure beyond the execution bottleneck we identify. Third, the 83.6% mean inter-judge agreement reported for the CoT error-analysis protocol (Section C) is a binary agreement rate between two LLM judges, not agreement against a human-annotated gold-standard subset; we did not collect human labels for a held-out sample of reasoning chains, so category-level error rates should be read as internally consistent across judges

rather than externally validated against expert judgment. Fourth, our headline claims are scoped to general-purpose frontier LLMs prompted directly and zero-shot, as described in Section 4; we do not evaluate specialized, fine-tuned molecular-optimization systems (e.g., DrugAssist (Ye et al., 2025), DrugR (Liu et al., 2026), MolEvolve (Chen et al., 2026)) or LLM-and-3D-generator hybrid pipelines (e.g., CIDD (Gao et al., 2025)) under LAO-Eval, because these systems are not exposed through a general chat-completion interface compatible with our task-level, prompt-only protocol and would each require bespoke integration; we discuss this scope boundary further in Section 2.