Benchmarking Boltz-2 Docking for Lipid–Protein Complexes

Jackson Stempel[*,†] and Micholas Smith[‡,†]

†Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996, United States

‡Biosciences Division and Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, United States

E-mail: jstempel@vols.utk.edu

Abstract

Lipid–protein interactions are central to membrane biology, yet computational prediction of lipid binding poses remains challenging due to the inherent flexibility of lipid tails and the importance of specific headgroup contacts. Here we present a systematic benchmark comparing Boltz-2, an open-source biomolecular structure prediction model, against AutoDock Vina, a widely used physics-based docking program, on a curated set of 100 lipid–protein complexes; Vina and GNINA are evaluated in a binding-site-known re-docking setup, while Boltz-2 is evaluated as sequence-and-ligand complex prediction. We evaluate predictions using complementary geometric (ligand and headgroup RMSD) and interaction-based (contact overlap) metrics. Boltz-2 substantially outperforms Vina, achieving a median ligand RMSD of 1.26 Å compared to 4.37 Å for Vina's top-ranked pose ($p = 2.2 \times 10^{-12}$ for Boltz-2 vs Vina top-1, Wilcoxon signed-rank test), with 73% (95% CI 64–81) of Boltz-2 predictions within 2 Å of the experimental structure versus 12% (95% CI 7–20) for Vina. Boltz-2 also better recapitulates native protein–headgroup contacts, with a mean environment Jaccard similarity of 0.70 compared to 0.39 for Vina. GNINA improves Vina's top-1 ranking and headgroup placement but remains inferior to Boltz-2, indicating that scoring improvements alone do not close the gap to end-to-end structure prediction. To probe whether Boltz-2's accuracy reflects learned physical principles or memorization of training examples, we performed an adversarial mutagenesis experiment: mutating binding-site residues to glycine (removing side-chain chemistry) or phenylalanine (adding steric bulk) and re-running predictions. Upon mutation, 75–80% of initially accurate predictions degraded substantially (headgroup RMSD shifting from <3 Å to >3 Å), indicating that Boltz-2's lipid placement depends on local binding-site chemistry rather than positional memorization.

Introduction

Lipids are essential biomolecules that serve as structural components of cellular membranes, signaling molecules, and energy storage compounds.[1] Understanding how lipids interact with proteins is crucial for elucidating membrane protein function, lipid transport mechanisms, and the molecular basis of diseases involving lipid dysregulation.[2] Despite advances in structural biology, experimentally determining lipid binding poses remains challenging, motivating the development of computational docking methods.

Molecular docking has become a standard tool for predicting protein–ligand binding modes.[3] Physics-based methods such as AutoDock Vina use empirical scoring functions to evaluate the energetic favorability of candidate poses.[4] CNN-based rescoring methods such as GNINA have been developed to improve pose ranking without changing the underlying sampling procedure, making them a natural comparison point for lipid docking where ranking is often the bottleneck. More recently, deep learning approaches have demonstrated remarkable accuracy in biomolecular structure prediction. AlphaFold2 revolutionized protein structure prediction,[5] and subsequent models including AlphaFold3 extended this capability to protein–ligand complexes.[6] Boltz-2 is an open-source diffusion-based model for biomolecular complex prediction that jointly models protein and ligand conformations.[7,8]

Lipids present unique challenges for docking methods compared to typical drug-like small molecules. First, lipids are highly flexible, with hydrocarbon tails that can adopt many conformations with similar energies. Second, the biologically relevant features of lipid binding are often concentrated in the polar headgroup (phosphate, glycerol, or amine moieties), which forms specific hydrogen bonds and electrostatic interactions with the protein, while the aliphatic tail provides nonspecific hydrophobic contacts. Third, lipid binding sites are frequently located at membrane–protein interfaces or within transmembrane domains, where the local environment differs from typical soluble protein binding pockets.

Previous docking benchmarks have focused predominantly on drug-like molecules,[9,10] leaving lipid–protein docking relatively underexplored. The few studies examining lipid docking have noted the difficulty of accurately predicting tail conformations and the importance of headgroup placement for biological relevance.[11]

In this work, we present a benchmark comparing Boltz-2 and AutoDock Vina on 100 curated lipid–protein complexes. For the docking baselines (Vina and GNINA), each ligand is docked back into its cognate experimental receptor in a re-docking, binding-site-known setting. We evaluate predictions using both geometric metrics (ligand and headgroup RMSD) and interaction-based metrics (overlap of protein–headgroup contacts). While our results reveal substantially better performance from the AI-based method, we discuss important caveats regarding the distinction between re-docking and prospective prediction. Our concrete contributions are fourfold: (i) a curated lipid–protein docking benchmark with standardized evaluation across docking and structure-prediction methods, (ii) a metric suite that separates tail-dominated geometric error from headgroup placement and interaction fidelity, (iii) an analysis that disentangles sampling quality from ranking quality for classical docking baselines, and (iv) an adversarial mutagenesis experiment that probes whether AI-based predictions reflect learned physical principles or memorization of training examples.

Methods

Benchmark Dataset Curation

We assembled a benchmark set of lipid–protein complexes from the Protein Data Bank (PDB).[12] For this study, we define lipids as amphipathic molecules containing both a polar headgroup and one or more hydrocarbon chains, including fatty acids, phospholipids, glycolipids, and sterols.

Starting from PDB structures containing bound lipid ligands, we applied the following curation criteria to ensure unambiguous evaluation:

1. **Single ligand requirement**: Structures with multiple lipid ligands or ambiguous ligand identity were excluded to enable unambiguous atom-wise comparison between predicted and experimental poses.

2. **Single protein chain**: We restricted the benchmark to single-chain protein complexes to simplify alignment.

3. **Chemistry validation**: Ligands were validated using RDKit[13] to ensure correct valence states and recognized elements; structures with chemistry errors were excluded.

4. **Atom mapping coverage (evaluation)**: RMSD calculations required RDKit MCS atom correspondences. Predicted poses were considered valid if $\geq$90% of predicted heavy atoms matched the reference and $\geq$60% of reference heavy atoms were covered; partial matches (reference coverage <90%) were flagged explicitly.

In total, we considered 159 candidate complexes and excluded 58 entries during curation due to RDKit failures, insufficient atom-mapping coverage, or file parsing issues.

After curation, we excluded one outlier target (PDB 3L1N) due to severe protein RMSD in Boltz-2 predictions, leaving 100 complexes for all analyses reported here. The dataset spans diverse lipid types including fatty acids (e.g., palmitate, oleate), phospholipids (e.g., phosphatidylcholine derivatives), glycolipids (e.g., GM3), and sterols (e.g., cholesterol), with ligand sizes ranging from 11 to 102 heavy atoms (mean: 25.2 atoms).

Docking Methods

Boltz-2

Boltz-2 is an open-source deep learning model for biomolecular structure prediction that uses a diffusion-based generative architecture.[7,8] Unlike traditional docking, Boltz-2 predicts protein structure from sequence while jointly modeling the bound ligand conformation. For each benchmark target, Boltz-2 was provided the protein sequence and the ligand identity. MSAs were generated automatically via the mmseqs2 server (accessed January 2026).[14] We used `boltz predict` with MSA-server mode enabled and sampling settings of 200 diffusion steps, 3 recycling steps, and 1 diffusion sample (output format: mmCIF; other parameters at defaults). As a robustness check, we re-ran Boltz-2 predictions on the University of Tennessee ISAAC cluster with substantially increased sampling and recycling settings and observed no material change in aggregate benchmark metrics; we therefore retained the original settings for all reported results.

AutoDock Vina

AutoDock Vina is a physics-based docking program that uses an empirical scoring function combining steric, hydrophobic, hydrogen bonding, and torsional terms.[4] Unlike Boltz-2, Vina docks ligands into a rigid protein receptor.

For each target, we extracted the ligand from the experimental structure and docked it back into the corresponding experimental protein receptor after removing the ligand and crystallographic waters (ions retained); receptor and ligand preparation assumed pH 7.4 and used the PDBQT format required by Vina. The search box was centered on the experimental ligand centroid, and each dimension was set to the ligand's Cartesian extent plus a 6 Å margin (approximately 3 Å padding per side), ensuring the binding site was fully covered while keeping the search volume target-specific. Vina was run with exhaustiveness 8 and configured to output up to 20 poses ranked by predicted binding affinity (estimated free energy of binding in kcal/mol).

To investigate sensitivity to compute input, we re-ran the full Vina docking set on ISAAC at much higher exhaustiveness (256). Aggregate pose accuracy was broadly similar to the baseline: median top-1 ligand RMSD changed from 4.37 Å to 4.29 Å, and median top-20 best RMSD changed from 1.77 Å to 1.78 Å.

We note that this setup is intentionally favorable to docking baselines: the search box is derived from the experimental ligand position, providing a binding-site-known re-docking scenario that gives Vina and GNINA a best-case starting point.

GNINA (Vina with CNN rescoring)

AutoDock Vina can sample accurate lipid poses but often fails to rank them correctly (Results), motivating evaluation of methods that specifically target *ranking quality*. GNINA is a docking program derived from Vina/smina that augments classical scoring with a convolutional neural network (CNN) that can rescore or rerank sampled poses.[15] This makes GNINA a useful intermediate baseline: it retains Vina-like sampling but introduces a learned scoring component intended to improve pose ranking, without changing the overall docking workflow.

In our ISAAC workflow, GNINA is run on the same prepared inputs and search boxes as Vina and configured to output 20 ranked poses. We additionally ran GNINA with CNN rescoring disabled (`cnn_scoring=none`) to isolate the effect of the CNN on ranking. When interpreting results from any multi-pose docking method (Vina or GNINA), it is important to distinguish: (i) **top-1** performance (the program's highest-ranked pose; the practically relevant case) from (ii) **top-K best** performance (the best RMSD among the top K poses; an oracle upper bound that measures whether accurate poses exist somewhere in the candidate set, independent of ranking). We therefore analyze both, and additionally quantify the *ranking gap* (top-1 RMSD minus top-20 best RMSD) as a direct measure of ranking quality. GNINA ligands were supplied as SDFs converted from the prepared PDB ligands to preserve explicit bonding.

Software and Versions

AutoDock Vina runs used AutoDock Vina 23d1252-mod. GNINA runs used GNINA 1.3.1 (Apptainer image, `gnina/gnina:latest`), with CNN rescoring enabled (`cnn_scoring=rescore`) unless otherwise specified.[15] Boltz runs used boltz 2.2.0.

Benchmark evaluation used Python 3.12.11 with RDKit 2025.03.3, BioPython 1.84, Gemmi 0.7.3, NumPy 1.26.4, SciPy 1.13.1, PyYAML 6.0.2, pandas 2.3.1, PandaMap 4.1.0, matplotlib 3.10.7, SciencePlots 2.2.0, and cmocean 4.0.3.

ChimeraX was used for manual verification of calculated values.[16]

Structure Alignment

To enable fair comparison, predicted complexes were aligned to experimental structures using the protein backbone. Chain correspondence was determined by global sequence alignment using BioPython's PairwiseAligner,[17] and corresponding Cα atoms were fit using an iterative outlier-pruned least-squares superposition (2.0 Å cutoff; Kabsch algorithm). The resulting transformation was applied to both protein and ligand coordinates in the predicted complex (Boltz-2). Vina poses were evaluated directly in the experimental receptor coordinate frame and therefore did not require additional protein alignment.

Geometric Metrics

Ligand RMSD

Ligand root-mean-square deviation (RMSD) was calculated over paired heavy atoms after placing predictions in the experimental reference frame (protein-backbone alignment for

Boltz-2; direct evaluation in the experimental receptor frame for Vina). Atom correspondence between predicted and experimental ligands was established using RDKit's maximum common substructure (MCS) mapping, and RMSD was computed over the paired atoms without an additional ligand-only fitting step. As a robustness check for symmetric substructures, we performed a symmetry-aware mapping check and found no cases where the optimal correspondence changed (0/100 targets). Predicted poses were considered valid if the MCS covered $\geq$90% of predicted heavy atoms and $\geq$60% of reference heavy atoms; cases with lower reference coverage were flagged as partial matches. These thresholds were chosen to balance stringency against dataset retention, ensuring reliable mappings without discarding a large fraction of targets.

Headgroup RMSD

Given the biological importance of headgroup positioning, we computed a separate RMSD restricted to headgroup atoms. Headgroup atoms were identified automatically using a functional group-based heuristic:

1. If phosphorus atoms are present, the headgroup comprises all atoms within two bonds of any phosphorus.

2. Otherwise, if nitrogen atoms with degree $\geq$3 exist, atoms within two bonds of such nitrogens define the headgroup.

3. Otherwise, if any carbon has $\geq$2 oxygen neighbors, that carbon and its oxygen neighbors constitute the headgroup.

4. As a fallback, all heteroatoms (O, N, P, S) are considered headgroup atoms.

This rule set is intended to capture the polar functional groups that define lipid identity (e.g., phosphates, charged amines, and carboxylate/ester regions) across chemically diverse ligands, while providing a fully automated and reproducible headgroup definition. Headgroup RMSD was computed over the mapped headgroup atom pairs. Across the 100 targets, the headgroup selection contained a median of 3 atoms (IQR 2; range 1–19). A small number of sterol-like ligands have only a single polar atom, for which headgroup RMSD primarily reflects hydroxyl placement. The headgroup definition rule was triggered by phosphorus atoms in 20 ligands, a cationic nitrogen criterion in 1, a carbonyl/oxygen-neighborhood criterion in 73, and by the heteroatom fallback in 7.

Interaction Metrics

Beyond geometric accuracy, we evaluated how well predictions recapitulate native protein–lipid interactions.

Headgroup Environment Overlap

We identified all protein residues with any heavy atom within 5.0 Å of any headgroup atom in both experimental and predicted structures. This coarse, distance-based metric complements RMSD by evaluating whether the predicted headgroup occupies the correct residue environment even when specific interaction geometries are sparse or ambiguous. The overlap between these residue sets was quantified using the Jaccard similarity coefficient:

$$J = \frac{|R_{\text{ref}} \cap R_{\text{pred}}|}{|R_{\text{ref}} \cup R_{\text{pred}}|} \tag{1}$$

where R_{ref} and R_{pred} are the sets of contacting residues in the reference and predicted structures, respectively.

Typed Interaction Overlap

We analyzed specific interaction types using PandaMap,[18] which detects hydrogen bonds, ionic interactions, salt bridges, and attractive charge interactions. To focus on biologically informative features for lipids, we report Jaccard similarity over typed headgroup–protein interaction pairs.

Multi-Pose Analysis

For Vina, we report both the top-ranked pose (top-1) and "top-K best" performance, where top-K best denotes oracle selection of the lowest ligand RMSD among the first K ranked poses ($K \in \{1, 5, 20\}$). This analysis distinguishes intrinsic sampling capability from ranking accuracy.

Statistical Analysis

All comparisons between methods were performed on the same 100 targets. Our primary hypothesis test compares Boltz-2 versus Vina top-1 ligand RMSD; statistical significance was assessed using the Wilcoxon signed-rank test (paired, non-parametric) with a significance threshold of $\alpha = 0.05$. Other method comparisons are reported descriptively.

Adversarial Binding-Site Mutagenesis

A key concern with AI-based structure prediction is whether models learn transferable physical principles or simply memorize ligand positions from training structures.[19] To probe this for lipid–protein complexes, we designed an adversarial mutagenesis experiment inspired by Masters et al.[19]

For each of the 100 benchmark targets, we identified binding-site residues as those with side-chain heavy atoms within 5.0 Å of any headgroup heavy atom in the experimental structure. We then generated two mutant sequences per target:

1. **Glycine sweep**: all binding-site residues mutated to glycine, removing side-chain chemistry while preserving backbone geometry.

2. **Phenylalanine sweep**: all binding-site residues mutated to phenylalanine, introducing bulky hydrophobic groups that sterically occlude the binding site.

Mutant sequences were run through Boltz-2 with identical ligand definitions and inference settings as the baseline predictions. Predictions were evaluated using the same benchmark pipeline, with a protein-fold filter (protein RMSD ≤ 2.0 Å) to ensure observed ligand displacements reflect local binding-site effects rather than global refolding.

If Boltz-2 places lipids by memorizing expected positions regardless of local chemistry, predictions should remain accurate despite binding-site disruption. Conversely, if predictions depend on learned physical interactions, disrupting the binding site should cause substantial headgroup displacement.

Results

Overall Docking Accuracy

Boltz-2 substantially outperformed AutoDock Vina across all geometric metrics (Table 1, Figure 1). For ligand RMSD, Boltz-2 achieved a mean of 2.20 Å (median 1.26 Å) compared to 5.11 Å (median 4.37 Å) for Vina's top-ranked pose. GNINA with CNN rescoring improved top-1 ligand RMSD (mean 3.46 Å, median 2.31 Å) relative to Vina, while GNINA without CNN rescoring was intermediate (mean 4.57 Å). This difference was highly significant for Boltz-2 versus Vina top-1 ($p = 2.2 \times 10^{-12}$, Wilcoxon signed-rank test). The distributions reveal that Boltz-2 predictions cluster near the experimental structure, with 73% within 2 Å, while only 12% of Vina top-1 poses achieved sub-2 Å accuracy; GNINA-CNN improves this to 44%. Using Wilson 95% confidence intervals, the corresponding success rates are 73% [64–81] for Boltz-2 and 12% [7–20] for Vina top-1.

Notably, Vina top-20 best achieves substantially lower ligand RMSD than top-1 and improves contact overlap, but still trails Boltz-2 across metrics.

Headgroup RMSD followed similar trends, with Boltz-2 achieving a mean of 2.38 Å (median 1.76 Å) versus 5.52 Å (median 4.42 Å) for Vina. This indicates that Boltz-2 not only places the overall ligand more accurately but also better captures the positioning of the

Table 1: Summary of docking performance on 100 lipid–protein complexes. Values shown are mean (median) [IQR]. "Vina top-K best" denotes oracle selection of the lowest ligand RMSD among the first K ranked poses (with all other metrics reported for that same pose). Typed interaction overlap is computed over targets with non-empty reference headgroup interactions (83/100).

	Ligand RMSD (Å)	Headgroup RMSD (Å)	Env. Jaccard	Typed Jaccard
Boltz-2	2.20 (1.26) [1.22]	2.38 (1.76) [1.50]	0.70 (0.76) [0.28]	0.69 (0.80) [0.52]
Vina top-1	5.11 (4.37) [4.18]	5.52 (4.42) [5.54]	0.39 (0.40) [0.60]	0.34 (0.25) [0.67]
GNINA CNN top-1	3.46 (2.31) [2.75]	3.16 (2.15) [2.14]	0.62 (0.70) [0.39]	0.57 (0.60) [0.80]
GNINA no-CNN top-1	4.57 (3.47) [3.97]	5.05 (3.73) [5.46]	0.44 (0.48) [0.62]	0.36 (0.25) [0.67]
Vina top-20 best	2.55 (1.77) [1.13]	2.79 (2.02) [1.20]	0.63 (0.67) [0.34]	0.57 (0.50) [0.67]

functionally critical headgroup region.

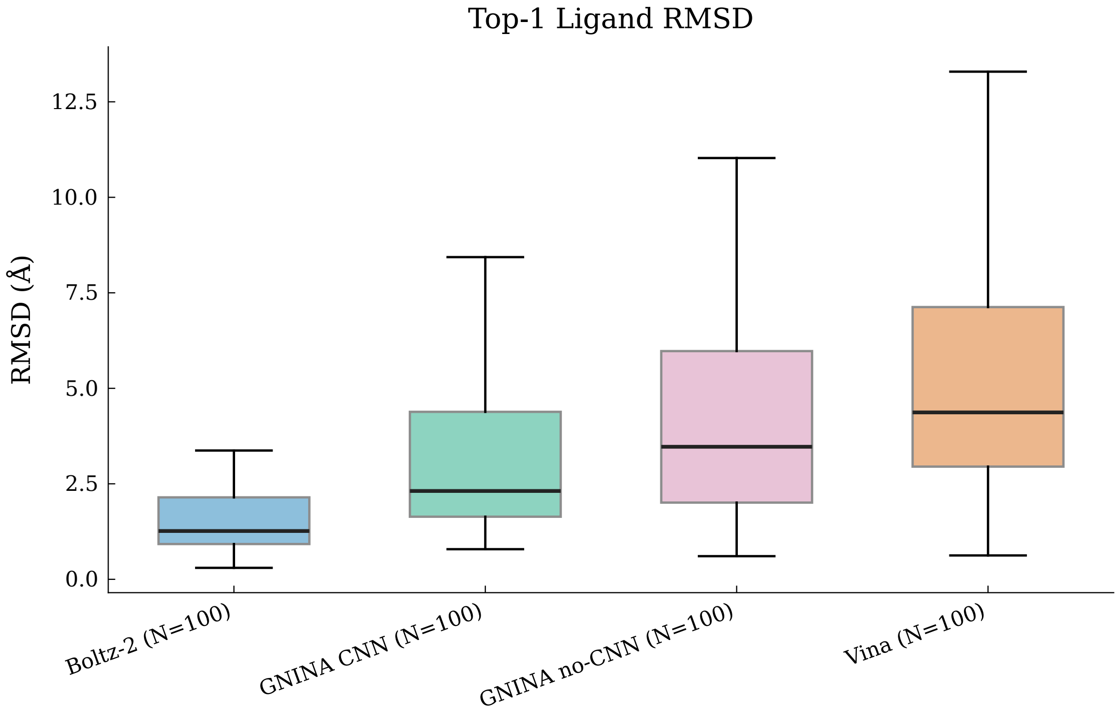


Figure 1: Top-1 ligand RMSD distributions across methods. GNINA-CNN improves Vina accuracy but does not reach Boltz-2 performance.

Per-Target Comparison

Pairwise comparison across all 100 targets reveals that Boltz-2 outperforms Vina on the large majority of complexes (Figure 2). Boltz-2 achieved lower ligand RMSD than Vina top-1 in 83 of 100 cases (83%). Even when comparing against Vina's oracle-selected best pose among 20 candidates, Boltz-2 remained superior in 65 cases (65%). GNINA-CNN outperformed Vina top-1 in 71 of 100 cases, reflecting a consistent (but smaller) improvement in ranking quality.

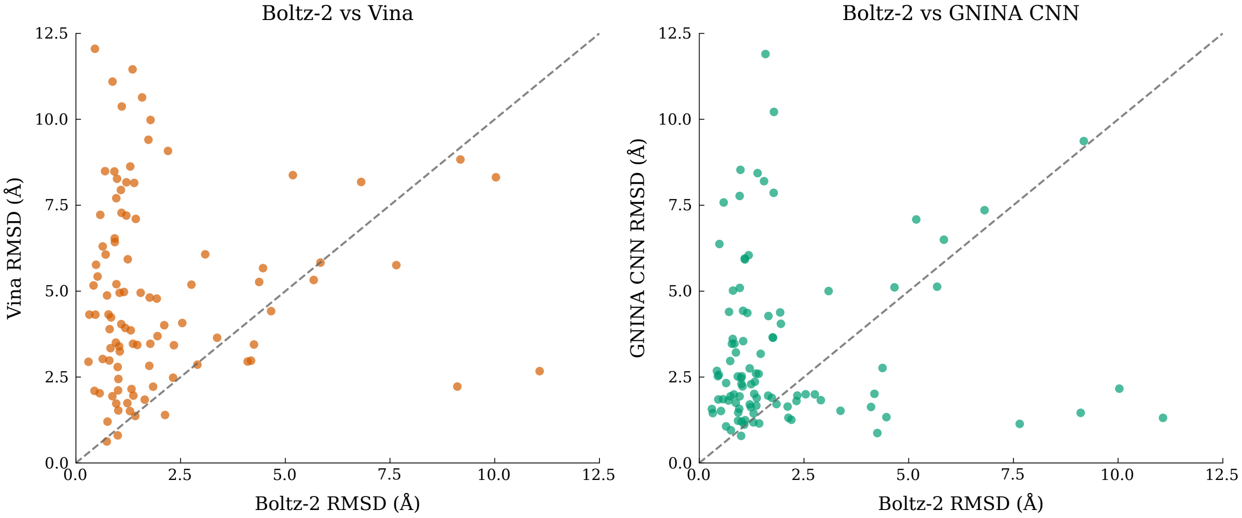


Figure 2: Per-target ligand RMSD comparisons. Left: Boltz-2 vs Vina. Right: Boltz-2 vs GNINA-CNN. Points above the diagonal favor Boltz-2.

Sampling vs Ranking: Vina and GNINA

Analysis of Vina's multi-pose output reveals that accurate poses often exist among lower-ranked candidates. Vina's best-of-20 ligand RMSD (mean 2.55 Å, median 1.77 Å) is far better than its top-1 RMSD (mean 5.11 Å), indicating that sampling is often adequate but ranking is weak.

GNINA shifts the best-of-20 distribution toward lower RMSD compared to Vina, indicating modestly improved sampling (or local refinement) in addition to better ranking. GNINA-CNN achieves a best-of-20 mean ligand RMSD of 1.95 Å (median 1.51 Å) and reduces the ranking gap (top-1 minus best-of-20) from 2.56 Å for Vina to 1.51 Å (Figure 3). GNINA without CNN rescoring shows a gap similar to Vina (2.53 Å), implying that the CNN primarily improves ranking rather than sampling.

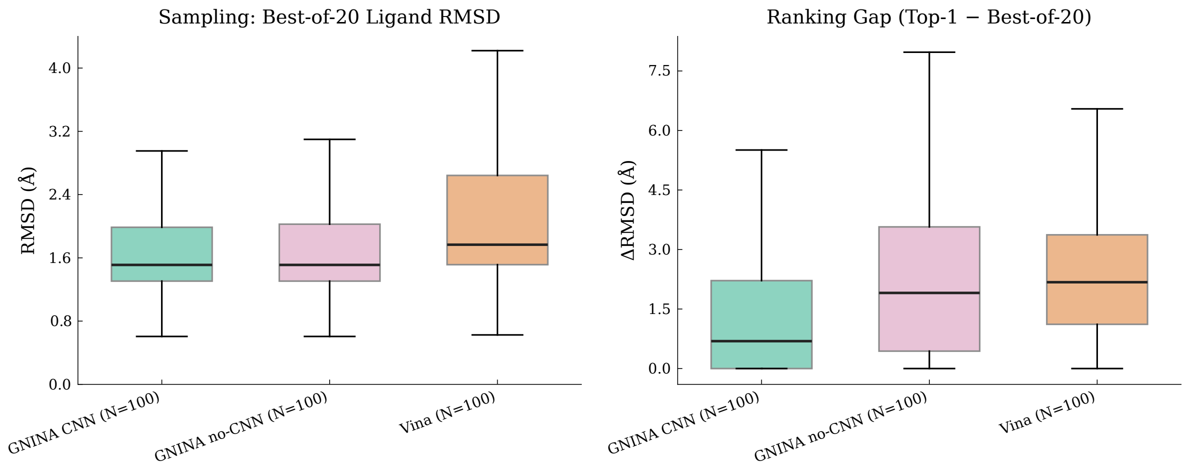


Figure 3: Sampling vs ranking behavior for Vina and GNINA. Left: best-of-20 ligand RMSD distributions (sampling quality; Boltz-2 omitted because it produces a single pose). Right: ranking gap defined as top-1 minus best-of-20 ligand RMSD (positive values indicate ranking error). GNINA-CNN narrows the ranking gap relative to Vina.

Interaction Fidelity

Geometric accuracy translated into improved interaction recovery (Figure 4). Boltz-2 predictions achieved a mean headgroup environment Jaccard of 0.70 (median 0.76), indicating substantial overlap with experimental residue contacts, whereas Vina top-1 reached 0.39 (median 0.40). GNINA-CNN improved headgroup environment overlap to 0.62 (median 0.70), paralleling its improved headgroup RMSD.

Typed interaction analysis showed similar trends. Boltz-2 achieved a mean typed interaction Jaccard of 0.69 (median 0.80), while Vina top-1 reached 0.34 (median 0.25); GNINA-CNN improved to 0.57 (median 0.60). These results indicate that improved pose ranking yields more biologically relevant headgroup contacts.

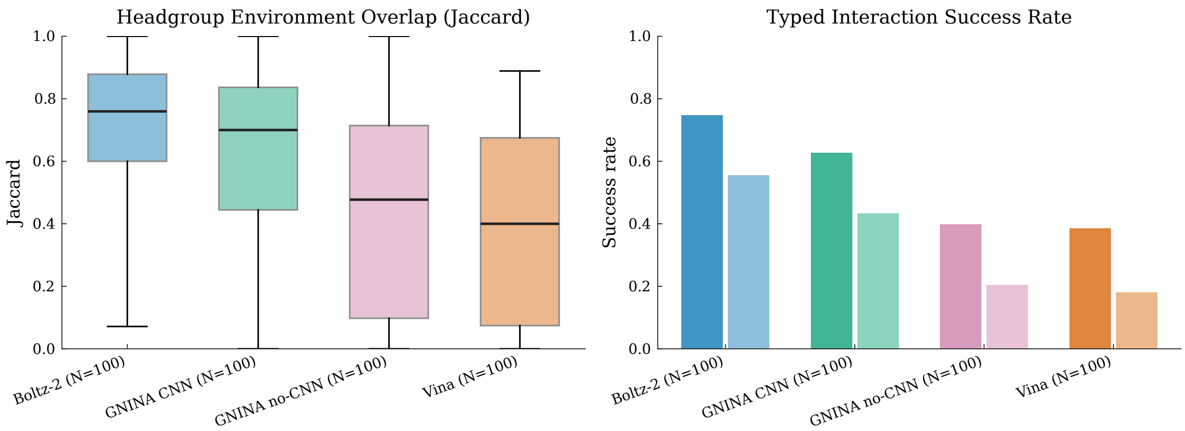


Figure 4: Headgroup interaction fidelity for top-1 poses. Left: headgroup environment overlap (Jaccard) shown as boxplots (median and interquartile range; whiskers indicate 5th–95th percentiles). Right: typed interaction success rate at Jaccard thresholds of 0.50 and 0.75. GNINA-CNN improves Vina's contact recovery but does not reach Boltz-2 performance.

Performance Stratified by Ligand Flexibility

To test whether ligand flexibility impacts docking outcomes, we stratified targets by the number of rotatable bonds (Vina `TORSDOF`, i.e., torsional degrees of freedom) into low (0–20), medium (21–40), and high (≥41) flexibility bins (Table 2). The medium and high bins are small ($n = 13$ and $n = 6$) and should be interpreted as exploratory. While Boltz-2 performance remained stable across flexibility bins, Vina accuracy degraded markedly with increasing torsion count even under oracle top-20 selection.

Table 2: Ligand RMSD stratified by ligand flexibility (Vina `TORSDOF`). Values are median ligand RMSD and success rate (fraction with RMSD ≤ 2 Å) with Wilson 95% confidence intervals; medium/high bins have small sample sizes.

Bin (TORSDOF)	n	Boltz-2	Vina top-1	Vina top-20 best
Low (0–20)	81	1.29; 0.74 [0.64, 0.82]	4.04; 0.14 [0.08, 0.23]	1.71; 0.72 [0.61, 0.80]
Medium (21–40)	13	1.18; 0.69 [0.42, 0.87]	6.07; 0.08 [0.01, 0.33]	3.84; 0.38 [0.18, 0.64]
High (≥41)	6	1.37; 0.67 [0.30, 0.90]	7.49; 0.00 [0.00, 0.39]	2.32; 0.33 [0.10, 0.70]

Failure case analysis of the 10 worst Boltz-2 predictions by ligand RMSD shows that errors are typically accompanied by headgroup displacement and loss of contact overlap: the top-10 median ligand RMSD is 8.38 Å (mean 8.41 Å), with median headgroup RMSD 9.49 Å and median headgroup environment Jaccard 0.10 (mean 0.26). Only 3/10 cases retain headgroup RMSD $\leq 3\,\text{Å}$, and 4/10 have headgroup environment Jaccard ≥ 0.5, indicating that most extreme failures reflect incorrect headgroup placement rather than tail-only deviations. Further inspection of these 10 cases suggests three patterns: (i) most failures are headgroup-dominated (6/10 have headgroup RMSD $> 5\,\text{Å}$ and headgroup environment Jaccard < 0.2), (ii) tail-dominated outliers are uncommon (3/10 retain good headgroup placement yet high ligand RMSD), and (iii) 8/10 are cases where Vina top-1 achieves a lower ligand RMSD, implying these are not uniformly hard targets but specific Boltz-2 misplacements. Two of the 10 are partial atom-mapping matches, and the set spans small and large ligands, indicating failures are not confined to very large lipids.

Adversarial Mutagenesis: Probing Memorization vs. Physics

To test whether Boltz-2's lipid placement reflects learned physical interactions or memorization of training examples, we evaluated predictions on binding-site mutants (Figure 5). Using a stricter fold-preservation cutoff (protein RMSD $\leq 2.0\,\text{Å}$), 89% (Gly) and 91% (Phe) of targets pass the fold filter, confirming that observed ligand displacements largely reflect local binding-site perturbations rather than global refolding.

Wild-type Boltz-2 predictions achieved headgroup RMSD $< 3\,\text{Å}$ in 88% of targets (median 1.76 Å). Upon glycine sweep, this dropped to 20% (median 5.14 Å); upon phenylalanine sweep, to 18% (median 6.86 Å). The phenylalanine sweep produced significantly larger displacements than the glycine sweep ($p = 0.042$, paired t-test), consistent with steric occlusion being more disruptive than simple side-chain removal.

Among the wild-type-correct targets that also pass the fold filter, only 23% (18/79) remained accurate after glycine mutation and 20% (16/81) after phenylalanine mutation. This retention rate—substantially lower than expected under pure memorization—indicates that Boltz-2's lipid headgroup placement depends on the chemical identity of binding-site residues.

A minority of targets (18–22%) retained accurate headgroup placement despite severe binding-site disruption, suggesting either backbone-dominated interactions, intrinsically stable binding modes, or residual memorization effects for a subset of complexes. Consistent with a backbone-dominated explanation, resistant Gly cases (protein RMSD ≤ 2.0 Å) did not show fewer mutations ($p = 0.56$) or more total headgroup contacts in the experimental structures at a 5.0 Å distance cutoff ($p = 0.83$), but did exhibit a higher fraction of backbone-mediated headgroup contacts (median 0.29 vs 0.00; $p = 0.013$).

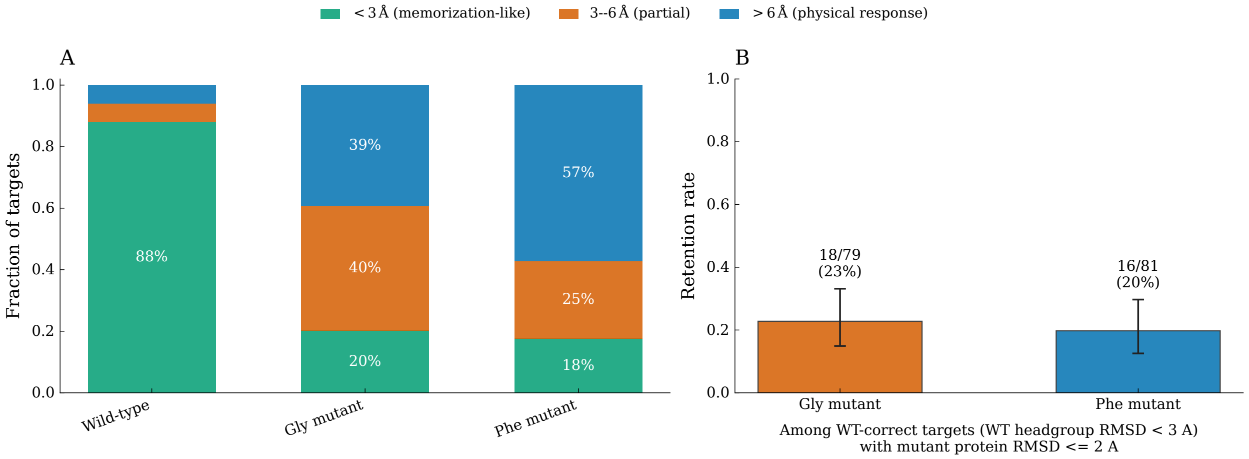


Figure 5: Adversarial binding-site mutagenesis results. (A) Outcome categories based on headgroup RMSD: strong memorization-like (<3 Å), partial response (3 Å to 6 Å), or physical response (>6 Å). Upon mutation, the fraction with accurate headgroup placement drops from 88% to ∼20% among fold-preserving mutants (protein RMSD ≤ 2.0 Å). (B) Retention analysis: among wild-type-correct targets that also satisfy the fold filter (79 Gly; 81 Phe), only 23% (Gly) and 20% (Phe) remain below 3 Å after mutation. Error bars show Wilson 95% confidence intervals.

Discussion

This benchmark shows a clear separation between AI-based complex prediction and classical docking baselines evaluated in a lipid re-docking setting. Boltz-2 consistently produces lower ligand and headgroup RMSD and recovers native headgroup contacts more faithfully, indicating that its advantage is not limited to a single metric but reflects a more coherent pose that better matches the functional interaction geometry.

The sampling-versus-ranking analysis clarifies why docking underperforms in this setting. Vina can often generate a near-native pose within its top-20 candidates, but its scoring function struggles to identify it reliably. GNINA improves ranking and also shifts the best-of-20 distribution toward lower RMSD, suggesting that its scoring and local optimization pipeline yields a better candidate set, yet the gap to Boltz-2 remains substantial. Together, these results indicate that better scoring is helpful but not sufficient on its own; capturing the coupled protein–ligand geometry (including induced fit and headgroup-specific interactions) is a key determinant of success for lipid docking.

The adversarial mutagenesis experiment provides evidence that Boltz-2's lipid placement is chemistry-dependent rather than purely memorization-driven. When binding-site residues are mutated to glycine or phenylalanine, 75–80% of previously accurate predictions degrade substantially, with headgroup RMSD shifting from sub-3 Å to >5 Å on average. The phenylalanine sweep produces larger displacements than the glycine sweep, consistent with steric blockage being more disruptive than simple side-chain removal. This sensitivity to local chemistry suggests that Boltz-2 has learned aspects of protein–lipid interaction physics, not merely where lipids "typically sit" in familiar protein folds. However, the ~20% of targets that retain accurate predictions despite binding-site disruption warrant further investigation; our resistant-case analysis indicates that these are enriched for backbone-mediated headgroup contacts, supporting a backbone-dominated interaction geometry as one plausible explanation.

The headgroup interaction metrics further emphasize the biological relevance of these differences. Methods that place the headgroup accurately are also the ones that recapitulate the protein's headgroup contact environment, supporting the idea that headgroup placement is the most informative proxy for functional correctness in lipid docking. This motivates evaluating lipid docking with interaction-based metrics alongside RMSD, rather than relying solely on tail-heavy geometric measures.

Finally, the results highlight a practical implication: in workflows where only a single ranked pose is used (as is common in screening), the ranking quality is as important as sampling. GNINA narrows that gap but does not eliminate it, whereas Boltz-2 performs well without reliance on multi-pose reranking. For lipid systems where headgroup chemistry and local protein interactions dominate specificity, models that jointly generate protein and ligand conformations appear to offer a substantial advantage in this re-docking benchmark.

Limitations

Several important limitations should be noted.

1. **Re-docking versus prospective prediction.** The docking baselines (Vina/GNINA) are evaluated in a re-docking benchmark where ligands are docked back into their cognate experimental receptors. The structures used are likely present in Boltz-2's training data. While the adversarial mutagenesis experiment provides evidence against pure memorization by showing that predictions depend on binding-site chemistry, it does not fully substitute for prospective evaluation on structures solved after Boltz-2's training cutoff, which would provide stronger evidence of generalization to novel targets.

2. **Comparison constraints.** Boltz-2 can model protein flexibility implicitly, while Vina and GNINA use a rigid receptor. Additionally, Vina/GNINA require a user-defined search box, whereas Boltz-2 predicts the binding site de novo. A more controlled comparison might use flexible receptor docking or provide Vina/GNINA with ensemble receptor structures.

3. **Scope of targets.** Our benchmark is restricted to single-chain proteins with single lipid ligands. Many biologically important lipid–protein interactions involve oligomeric proteins or multiple cooperative lipid binding events, which we did not evaluate.

4. **Scope of methods.** We evaluated Boltz-2, Vina, and GNINA. Other AI-based tools (e.g., AlphaFold3) and additional docking approaches were not included, in part due to limited public access and ligand-handling constraints at the time of analysis.

5. **Ligand representation.** A small subset of GNINA outputs were truncated relative to the experimental ligand (likely introduced during PDBQT→SDF conversion), so RMSD values for those targets reflect the predicted ligand portion rather than the full-length lipid.

6. **Tail accuracy.** While we focused on headgroup accuracy, lipid tail conformations may also be biologically relevant in some contexts (e.g., membrane protein cavities), and our metrics may underweight tail prediction quality.

Conclusion

We present a benchmark comparing AI-based (Boltz-2) and physics-based docking (AutoDock Vina), and we further evaluate GNINA CNN rescoring as an intermediate approach, on 100 curated lipid–protein complexes. Boltz-2 substantially outperforms Vina in both geometric accuracy and interaction fidelity on this benchmark, while GNINA improves Vina's top-1 ranking and headgroup placement but remains below Boltz-2. Adversarial mutagenesis experiments demonstrate that Boltz-2's lipid placement depends on binding-site chemistry: mutating binding-site residues to glycine or phenylalanine causes 75–80% of accurate predictions to degrade, providing evidence against pure memorization of training poses. A minority of targets (∼20%) retain accurate predictions despite binding-site disruption, warranting further investigation into whether these represent backbone-dominated interactions or residual memorization. Future work should evaluate performance on prospectively solved structures to more directly assess generalization. The benchmark dataset, analysis code, and adversarial mutagenesis pipeline are publicly available to facilitate further method development.

Data and Code Availability

The benchmark dataset, analysis code, and prediction inputs are available at `https://github.com/jacksonstempel/lipid_docking_benchmark`. The curated benchmark entry list is provided in `structures/benchmark_entries.csv`, and excluded entries with brief reasons are listed in `excluded_entries.txt`.

Associated Content

Supporting Information Available

Supporting Information: None.

Author Information

Corresponding Author

* E-mail: jstempel@vols.utk.edu

Author Contributions

J.S. implemented the benchmark pipeline and performed the analyses. J.S. and M.S. designed the study, interpreted results, and wrote the manuscript.

Notes

The authors declare no competing interests.

Acknowledgments

Computations were performed in part using the University of Tennessee Advanced Computing Facility (ISAAC). The benchmark and plotting code were developed with assistance from OpenAI's Codex CLI coding agent (GPT-5.2).

References

(1) van Meer, G.; Voelker, D. R.; Feigenson, G. W. Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology* **2008**, *9*, 112–124.

(2) Escribá, P. V.; González-Ros, J. M.; Goñi, F. M.; Kinnunen, P. K.; Vigh, L.; Sánchez-Magraner, L.; Fernández, A. M.; Busquets, X.; Horváth, I.; Barceló-Coblijn, G. Membranes: a meeting point for lipids, proteins and therapies. *Journal of Cellular and Molecular Medicine* **2008**, *12*, 829–875.

(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery* **2004**, *3*, 935–949.

(4) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*, 455–461.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; others Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(6) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; others Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.

(7) Passaro, S.; Corso, G.; Wohlwend, J.; Reveiz, M.; Thaler, S.; Somnath, V. R.; Getz, N.; Portnoi, T.; Roy, J.; Stark, H.; Kwabi-Addo, D.; Beaini, D.; Jaakkola, T.; Barzilay, R. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv* **2025**,

(8) Wohlwend, J.; Corso, G.; Passaro, S.; Reveiz, M.; Leidal, K.; Swanson, W.; Weinstein, H.; Murat, W.; Barzilay, R.; Jaakkola, T. Boltz-1: Democratizing Biomolecular Interaction Modeling. *bioRxiv* **2024**, 2024.11.19.624167.

(9) Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics* **2016**, *18*, 12964–12975.

(10) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative assessment of scoring functions: the CASF-2016 update. *Journal of Chemical Information and Modeling* **2019**, *59*, 895–913.

(11) Sridhar, A.; Ross, G. A.; Sherborne, B. Computational approaches for lipid docking. *Methods in Molecular Biology* **2017**, *1529*, 317–330.

(12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(13) RDKit RDKit: Open-source cheminformatics. `https://www.rdkit.org`, 2024; Accessed: 2024.

(14) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: making protein folding accessible to all. *Nature methods* **2022**,

(15) McNutt, A. T.; Li, Y.; Meli, R.; Aggarwal, R.; Koes, D. R. GNINA 1.3: the next

increment in molecular docking with deep learning. *Journal of Cheminformatics* **2025**, *17*, 28.

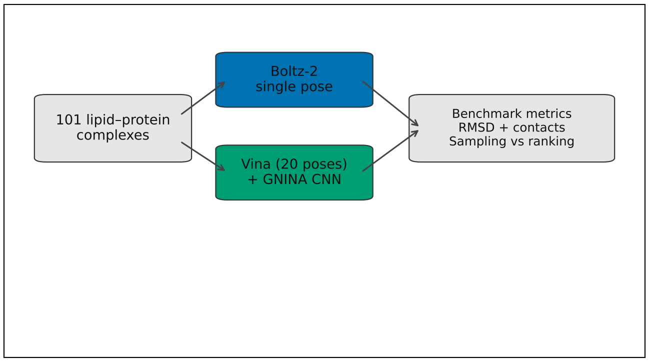
(16) Goddard, T. D.; Huang, C. C.; Meng, E. C.; Pettersen, E. F.; Couch, G. S.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science* **2018**, *27*, 14–25.

(17) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; others Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423.

(18) PandaMap PandaMap: Protein–ligand interaction analysis. `https://github.com/chopralab/pandamap`, 2024; Accessed: 2024.

(19) Masters, M. R.; Mahmoud, A. H.; Wei, Y.; Lill, M. A. Evaluating and mitigating limitations of large language models in clinical decision making. *Nature Communications* **2025**, *16*, 504, Adversarial mutation benchmark for AlphaFold3 ligand binding.

TOC Graphic


101 lipid–protein complexes
Boltz-2 single pose
Vina (20 poses) + GNINA CNN
Benchmark metrics RMSD + contacts Sampling vs ranking