PROBABILISTIC EVALUATION OF INDUCED-FIT EFFECTS IN FRAGMENT-BASED DRUG OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fragment-based lead optimization often involves relatively small chemical modifications, yet these changes can induce substantial rearrangements in ligand and receptor conformations and alter the preferred bound state among competing conformational minima. Accurately capturing and evaluating these ligand-dependent structural responses therefore remains a central challenge in structure-guided optimization. Here, we present a generation-and-evaluation framework for the probabilistic evaluation of a given bound structure. It first generates an ensemble of conformations around the input using flow matching and then evaluates their conditional log-probability scores. The generator learns to softly restrain scaffold positions near the initial ligand-fragment positions, focusing on sampling around the known binding context while allowing ligand and pocket conformations to adapt. Benchmarking on lead optimization tasks demonstrates that our generator provides more robust sampling than existing methods, producing valid and diverse conformations while accurately reproducing key interactions and induced-fit effects observed in experimental structures. The evaluator shows a meaningful correlation with the activity trends of congeneric ligand series without explicit affinity supervision of the generative model. Overall, these results suggest that the proposed strategy can serve as a competitive evaluation module in structure-based drug optimization workflows, replacing or complementing widely used docking and binding-affinity estimation methods.

1 INTRODUCTION

Structure-based drug design (SBDD) aims to identify and optimize small molecules by using three-dimensional information about protein–ligand binding sites (Anderson, 2003). SBDD typically works in multiple steps. Focusing on the early stage, first a "hit" molecule is identified that weakly binds to the receptor. In the subsequent lead optimization step, the hit molecule with known (or predicted) pose to the receptor is further optimized for its binding affinity and molecular properties. Therefore, the lead optimization step becomes a comparative problem (Griffen et al., 2011; Young and Leeson, 2018). The key structural question is about estimating how a proposed substitution shifts the population of many conformational states in the known binding context, including the most probable conformational state, the corresponding receptor response, and protein–ligand contacts across the ensemble.

Many structure-based lead-optimization workflows evaluate the impact of fragment substitution on a single static complex structure with fixed receptor side chains. These approaches have two critical limitations. The first limitation comes from fixed atomic structures. In principle, chemical modifications are often small, but receptor structural responses can be larger, encouraging new favorable interactions while avoiding clashes or voids at the binding pocket (Frauenfelder et al., 1991; Henzler-Wildman and Kern, 2007; Stank et al., 2016). The second limitation comes from using a single structure. Ligand binding is naturally an ensemble effect, in which multiple ligand/pocket conformational states can compete, and their relative preference can change upon ligand modification. Plausible ensemble structures around the known binding context, which we refer to as the *local ensemble*, collectively contribute to low binding free energy. A contact repeatedly appearing across the local ensemble should support such a local interaction, and vice versa. Therefore, proper sampling of a local ensemble and its probabilistic evaluation is required for the accurate estimation of the effect of a fragment substitution.

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

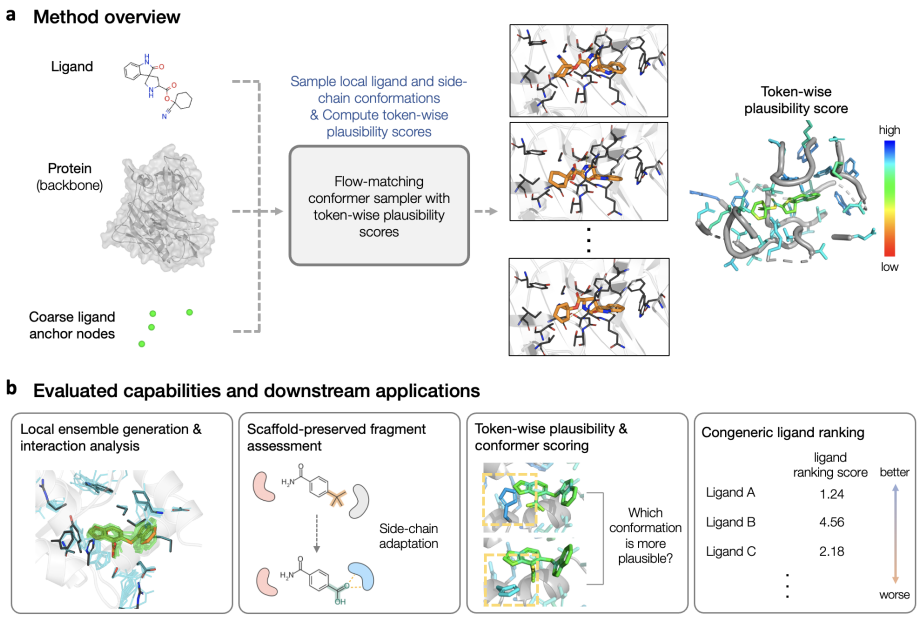

Figure 1: Conceptual overview of local structural evaluation for fragment-based ligand optimization. (a) Sampling-and-scoring workflow. Starting from a ligand, protein backbone, and coarse ligand anchor nodes, the flow-matching conformer sampler generates local ligand and side-chain conformational states and assigns token-wise log-probability-based scores to generated structures. (b) Evaluated capabilities and downstream applications tested in this work.

To address these limitations, we present a flow-matching approach which sets the pose evaluation problem in lead optimization as a probabilistic local ensemble-level sampling and evaluation problem. The framework is designed around three capabilities. First, **local ensemble generation**: starting from the initial complex, it samples a local ensemble around the assumed binding mode, exposing alternative contacts while retaining the input context. Second, **a learned soft scaffold-preserving restraint**: supplying the input complex and coarse ligand anchor nodes as conditioning, the model learns to follow them at sampling time as an implicit soft positional restraint. This places the method between hard-constrained and unconstrained generator, keeping the shared scaffold in the binding region while allowing mild pocket reorganization. Third, **probabilistic plausibility scoring**: the flow-matching formulation (Lipman et al., 2023) with continuous-normalizing-flow log-probability computation (Chen et al., 2018), assigns each generated structure a log-probability, which is decomposed into ligand atoms, fragments, and pocket residues. Together, these components make our method an interpretable evaluator for SBDD, covering broad tasks such as assessing if a proposed substitution preserves the binding context, identifying plausible interaction states, selecting representative conformers, and providing ensemble-based ranking within congeneric ligand series.

2 Related Works

2.1 Generative Models for SBDD and Local Ensemble Analysis

SBDD has evolved from traditional molecular docking and scoring functions (Trott and Olson, 2010; Jain, 2006) to deep learning-based approaches that generate ligand structures, predict protein–ligand poses, or model joint protein–ligand complexes (Corso et al., 2023; Qiao et al., 2024; Anishchenko et al., 2025), and more recently to equivariant diffusion, flow matching, and co-folding architectures (Schneuing et al., 2024; 2025; Abramson et al., 2024). However, these outputs are usually reduced to one or a few representative structures, and modeled variability less represents local induced fit effect. This is limiting for lead optimization. Recent works for modeling conformational distributions (Jing et al., 2022; Arts et al., 2023) are also less suitable for assessing plausibility of local

ensembles. PLACER is especially relevant because it models ligand and side-chain placement in a local protein context like ours (Anishchenko et al., 2025), but it was not tested for scaffold-preserving fragment substitution. For the same fragment substitution, PLACER aims at reconstructing plausible local complexes, whereas our method aims at estimating conditional density of ensemble structures with an ensemble and token-level probability estimation.

2.2 Log-Probability in Continuous Normalizing Flows and Flow Matching

Continuous normalizing flows provide an exact change-of-variables identity for the log-probability along the sampling ODE (Chen et al., 2018), and flow matching offers a scalable, simulation-free way to train the underlying vector field (Lipman et al., 2023). In practice, this log-probability is evaluated with numerical integration and trace estimation. Beyond producing samples, the resulting log-probability estimate has been used as a model-derived measure of how well a configuration conforms to the learned distribution. In molecular modeling, normalizing-flow likelihoods have been used to reweight sampled configurations toward a target Boltzmann distribution (Noé et al., 2019), and score- and flow-based generative models have been interpreted as learned plausibility or energy signals rather than only as samplers (Song et al., 2021; Arts et al., 2023). More directly, recent all-atom generative models use generative likelihood estimates as ranking criteria for design candidates (Kong et al., 2026). We sought to build on this view but differ in granularity and scope. We decompose the flow-matching log-probability to the token level (Section 3.2) and use it for local, within-target analysis of protein–ligand interactions, including conformer selection, residue-and fragment-level attribution, and exploratory ligand prioritization within congeneric ligand series.

3 Methods

Given a fixed protein context and coarse ligand anchor nodes, our framework samples ligand and nearby side-chain conformations in a cropped binding site and estimates their conditional log-probability under the learned local distribution. The sampling ODE and the continuous-normalizing-flow change of variables are integrated together, with the divergence estimated using a Hutchinson trace estimator (Hutchinson, 1989). We interpret the resulting value as a model-derived *local structural plausibility score*. The full derivation and estimator are given in Section A.

3.1 Probabilistic Protein–ligand Ensemble Generation

Let $x = (x_P, x_L)$ contain the local protein and ligand coordinates and $\mathcal{C}$ the fixed context. The protein backbone is fixed while ligand atoms and nearby side chains are updated. We use the linear optimal-transport path and an x_1-parameterized vector field (Lipman et al., 2023),

$$x_t = (1-t)x_0 + tx_1, \qquad t \in [0,1], \tag{1}$$

$$v_\theta(t, x_t, \mathcal{C}) = \frac{\widehat{x}_1 - x_t}{1-t}, \qquad \widehat{x}_1 = f_\theta(t, x_t, \mathcal{C}). \tag{2}$$

The field is integrated with a small number of Euler steps. Ligand fragments define coarse anchor nodes, and updated coordinates are initialized from $x_0 \sim \mathcal{N}(\mu(\mathcal{C}), \sigma_0^2 I)$, using the assigned anchor for ligand atoms and the residue C_α for side-chain atoms. This acts as a learned soft restraint: the shared chemotype remains near its input binding mode while substituted fragments and pocket side chains can adapt. Perturbing the anchors can be used to control the locality–diversity trade-off.

The denoising network is a GVP-based equivariant graph neural network (Jing et al., 2021) over ligand atoms, protein atoms, and coarse anchors, with a cross-product coordinate update for oriented local geometry. It is trained to predict $\widehat{x}_1$ with coordinate, distance, and clash losses; variable-size and off-pocket crops diversify local environments. Architecture, conditioning, losses, integration, and augmentation are detailed in Sections B.1–B.3, with the backbone ablation in Section E.1.

3.2 Token-level decomposition of conditional log-probability

To separate ligand and pocket contributions, we decompose the conditional log-probability over N atomic nodes. Let q_n be node n's contribution to the factorized anchor-centered prior and let

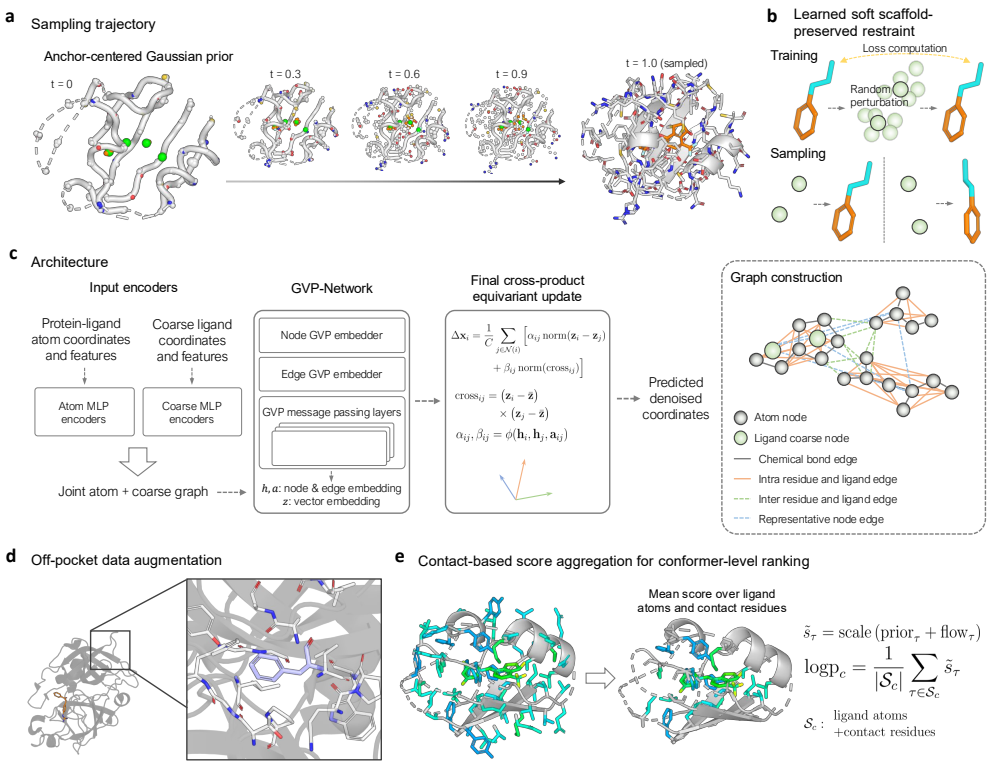

Figure 2: Model design for local protein–ligand conformer sampling. (a–b) Fixed protein context and coarse ligand anchors condition local sampling from an anchor-centered prior. (c) A joint atom–coarse GVP-GNN predicts the denoised endpoint. (d) Off-pocket crops augment local environments. (e) Node-wise log-probability contributions define `logp_contact`.

$f_n(1) = -\int_0^1 \nabla_{x_n} \cdot v_{\theta,n}(t, x_t, \mathcal{C})\, dt$. The raw node score is

$$s_n = q_n + f_n(1), \qquad \sum_{n=1}^{N} s_n \approx \log p_\theta(x_1 \mid \mathcal{C}). \tag{3}$$

The raw scores s_n form an additive node-wise decomposition. For downstream comparison, we clip and linearly scale them to stabilized values $\widetilde{s}_n$, then average these values over the side-chain atoms of each protein residue or over all ligand atoms. The node-wise trace estimator, prior decomposition, and scaling are detailed in Section A.2.

Contact-focused conformer score. For each target, a fixed residue set $\mathcal{R}$ is the union of residues within 4.5 Å of the ligand in any sampled conformer. Conformer k is ranked by

$$\log p_{\text{contact}}(k) = \frac{1}{2}\left[\left\langle \widetilde{s}^{\text{prot}}(r,k)\right\rangle_{r\in\mathcal{R}} + \widetilde{s}^{\text{lig}}(k)\right], \tag{4}$$

where the protein term averages stabilized side-chain scores within each contact residue and then across $\mathcal{R}$, and the ligand term averages over ligand atoms. The same $\mathcal{R}$ is used for every conformer of a target. The exact set definition is given in Section A.2.

3.3 Congeneric Ligand Ranking from Local Ensemble Features

For exploratory prioritization within congeneric series, we represent each ligand by local ensemble features derived from differences in stabilized token scores between its bound ensemble and matched reference ensembles. The unbound reference translates the ligand and its anchors out of the pocket, preserving the anchor-centered prior while removing contacts. The dummy reference preserves the

sampled ligand-shaped occupancy but removes chemical identity. We aggregate bound–unbound ligand differences, bound–dummy protein contact-shell differences, and nonpolar buried-surface-area features across conformers using bound-state plausibility weights. Feature normalization parameters and the coefficients of the linear ranking score are fitted only on the training target series, and the resulting score is evaluated within held-out congeneric series. Exact reference construction, score differences, aggregation, and features are given in Sections D.1 and D.2.

4 EXPERIMENTS

4.1 GENERATED CONFORMATIONS RECOVER KNOWN INTERACTIONS AND ARE PHYSICALLY VALID

We first evaluate local protein–ligand ensemble generation on the PLINDER test set (Durairaj et al., 2024). This benchmark asks, given the experimental complex as input, whether a method can generate physically feasible local ensembles that retain the protein–ligand interaction pattern. Accordingly, all generated samples are evaluated without score-based selection. The main metrics are PoseBusters ligand validity (PB-valid) (Buttenschoen et al., 2024) and protein–ligand clash rate for physical plausibility, lDDT-PLI and interaction-fingerprint similarity for local interaction recovery, side-chain χ accuracy for side-chain recovery, and ensemble diversity for non-degenerate sampling. To fairly assess several baselines strongly benefiting from crystallographic input coordinates, we initialize their initial poses from a locally perturbed complex (details in Supplementary Section C.3).

Table 1: Structural validity and ensemble quality for local protein–ligand generation.

Category	Method	Ligand update	Side-chain update	PB-valid (%)* ↑	Clash ↓	lDDT-PLI ↑	IFP sim. ↑	Diversity ↑	χ_1/χ_2 acc. ↑
Ours	Full model	yes	yes	98.0 ± 11.8	0.00 ± 0.03	0.82 ± 0.08	0.65 ± 0.12	1.27 ± 0.28	0.85 ± 0.08 / 0.74 ± 0.11
	w/o relaxation	yes	yes	94.0 ± 21.5	0.06 ± 0.15	0.81 ± 0.09	0.64 ± 0.11	1.23 ± 0.27	0.85 ± 0.08 / 0.75 ± 0.11
	w/o side-chain update	yes	no	92.8 ± 24.0	0.04 ± 0.15	0.86 ± 0.09	0.65 ± 0.11	0.59 ± 0.35	–
	w/o ligand conformer update	no	yes	93.7 ± 24.2	0.05 ± 0.14	0.95 ± 0.02	1.00 ± 0.00	1.03 ± 0.16	0.86 ± 0.08 / 0.75 ± 0.10
	2 Å coarse-anchor perturbation	yes	yes	95.1 ± 19.8	0.00 ± 0.03	0.79 ± 0.08	0.61 ± 0.11	1.42 ± 0.31	0.85 ± 0.08 / 0.74 ± 0.10
Classical	Random perturb. + minim.	yes	yes	90.1 ± 15.8	0.02 ± 0.02	0.29 ± 0.15	0.42 ± 0.12	2.20 ± 0.91	0.35 ± 0.36 / 0.50 ± 0.32
	Rosetta fixbb repacking	no	yes	96.3 ± 19.0	0.02 ± 0.12	0.95 ± 0.03	1.00 ± 0.00	0.16 ± 0.18	0.90 ± 0.09 / 0.81 ± 0.13
	Vina local search	yes	no	96.9 ± 17.0	0.02 ± 0.09	0.52 ± 0.13	0.40 ± 0.14	2.52 ± 0.80	–
	Vina flexible docking (random init)	yes	limited	96.8 ± 17.2	0.00 ± 0.00	0.41 ± 0.10	0.31 ± 0.11	2.88 ± 0.85	0.42 ± 0.19 / 0.37 ± 0.21
ML baseline	DiffBindFR	yes	yes	93.4 ± 23.5	0.01 ± 0.01	0.35 ± 0.08	0.34 ± 0.10	4.68 ± 1.30	0.80 ± 0.21 / 0.74 ± 0.19
	DiffDock-Pocket	yes	yes	94.5 ± 18.5	0.01 ± 0.02	0.61 ± 0.18	0.48 ± 0.17	1.18 ± 0.79	0.53 ± 0.14 / 0.53 ± 0.15
	PLACER	yes	yes	53.1 ± 32.7	0.03 ± 0.04	0.69 ± 0.15	0.51 ± 0.17	1.86 ± 0.88	0.92 ± 0.07 / 0.85 ± 0.10
	ApoDock	partial	yes	93.9 ± 23.5	0.00 ± 0.01	0.40 ± 0.09	0.37 ± 0.10	2.77 ± 0.86	0.83 ± 0.10 / 0.71 ± 0.12

*Mean percentage of generated poses per target that pass all PoseBusters `mol`-mode checks.

Table 1 shows that the main separation across methods appears in the interaction-sensitive metrics. Our full model reaches high lDDT-PLI (0.82) and IFP similarity (0.65) while still sampling a diverse joint ligand–side-chain ensemble. These values are slightly worse than the maximal possible values from starting with perfect ligand conformations ("Rosetta fixbb repacking" and "Ours w/o ligand conformer update"), but can be thought as a more realistic proxy for a thermal ensemble because not all contacts are always fully retained (e.g., exposed contacts are often broken) in room temperature. Other docking and ML methods (Vina (Trott and Olson, 2010), DiffBindFR (Zhu et al., 2024), DiffDock-Pocket (Plainer et al., 2023), and ApoDock (Luo et al., 2025)) produce PB-valid poses but considerably lower lDDT-PLI and IFP similarity (both metrics typically < 0.6). On the other hand, PLACER generally gives good interaction-based recovery metrics but shows poor PB-validity regardless of initialization strategy (Supplementary Figure 9 and Supplementary Tables 12, 13).

We then investigated the impact of two options in our model: (i) relaxation after generation and (ii) sampling diversity control. In our model, relaxation leaves the interaction metrics essentially unchanged (0.81 to 0.82 lDDT-PLI and 0.64 to 0.65 IFP similarity) and only removes short-range steric artifacts. The "random-perturbation-plus-minimization baseline" fails to recover the binding geometry (0.29 lDDT-PLI and 0.42 IFP similarity), demonstrating that relaxation alone is insufficient to explain the interaction recovery. Second, our model can control diversity through coarse-anchor perturbation. A 2 Å RMS perturbation increases ensemble diversity from 1.27 to 1.42, with a modest reduction in lDDT-PLI from 0.82 to 0.79 and IFP similarity from 0.65 to 0.61, while retaining 95.1%

Table 2: Congeneric cross-docking on the full Q-BioLiP benchmark and an induced-fit-required subset. The subset requires both a source-to-target edited-site side-chain RMSD of at least 1 Å and χ_1 accuracy below 0.8. Metrics are averaged over all generated samples for each directed task and then with equal weight per congeneric group. Values are congeneric-group-balanced mean ± standard deviation. Boldface marks the best value within each evaluation set.

Method	Scaffold restraint	Pocket side chains	Scaffold RMSD ↓	Scaffold IFP ↑	Frag. RMSD ↓	Frag. IFP ↑	Frag. lDDT-PLI ↑
Full Q-BioLiP benchmark							
AutoDock Vina†	none	fixed	5.14 ± 1.29	0.62 ± 0.10	6.90 ± 2.11	0.25 ± 0.13	0.40 ± 0.11
rDock†	hard	fixed	1.92 ± 0.94	0.81 ± 0.10	3.43 ± 1.72	0.46 ± 0.20	0.63 ± 0.12
SigmaDock‡	none	fixed	3.29 ± 2.26	0.74 ± 0.13	4.69 ± 2.71	0.42 ± 0.23	0.56 ± 0.16
GALigandDock†	none	flexible	6.54 ± 1.81	0.49 ± 0.11	7.46 ± 2.93	0.23 ± 0.13	0.37 ± 0.12
PLACER‡	none	flexible	3.33 ± 2.43	0.73 ± 0.13	4.67 ± 2.42	0.41 ± 0.21	0.57 ± 0.15
Ours‡	soft (learned)	fixed	**1.19 ± 0.57**	**0.87 ± 0.08**	**2.58 ± 1.23**	**0.55 ± 0.22**	**0.72 ± 0.11**
		flexible	1.22 ± 0.54	0.84 ± 0.08	2.62 ± 1.25	0.52 ± 0.20	0.70 ± 0.10
Induced-fit-required subset (13 tasks; 11 groups)							
AutoDock Vina†	none	fixed	5.87 ± 1.22	0.55 ± 0.11	7.82 ± 2.18	0.18 ± 0.09	0.32 ± 0.07
rDock†	hard	fixed	2.52 ± 0.87	0.74 ± 0.14	4.76 ± 1.98	0.31 ± 0.15	0.48 ± 0.13
SigmaDock‡	none	fixed	3.96 ± 2.07	0.70 ± 0.14	5.39 ± 2.79	0.35 ± 0.22	0.47 ± 0.15
GALigandDock†	none	flexible	6.60 ± 1.90	0.46 ± 0.09	7.15 ± 2.52	0.19 ± 0.07	0.34 ± 0.08
PLACER‡	none	flexible	3.41 ± 1.78	0.66 ± 0.14	5.78 ± 2.88	0.29 ± 0.14	0.44 ± 0.15
Ours‡	soft (learned)	fixed	1.25 ± 0.60	**0.84 ± 0.11**	2.65 ± 1.56	0.58 ± 0.23	0.68 ± 0.10
		flexible	**1.24 ± 0.54**	0.81 ± 0.12	**2.48 ± 1.58**	**0.59 ± 0.19**	**0.69 ± 0.08**

†Classical method. ‡Machine-learning-based method. The induced-fit-required evaluation uses the common fragment-RMSD-evaluable tasks. Scaffold restraint denotes how the shared scaffold is retained: *hard* = explicit tethered/positional scaffold constraint; *soft (learned)* = learned scaffold-retentive restraint from conditioning on coarse ligand anchor nodes, with no hard coordinate constraint; *none* = no scaffold-specific restraint.

PB-valid poses (Table 1). A broader sweep over anchor perturbation and ligand-prior variance shows the same diversity–interaction recovery trade-off (Supplementary Table 11).

4.2 Scaffold-preserved generation better represents induced-fit effects

We next evaluate ligand pose accuracy following fragment substitution while preserving the scaffold identity. The benchmark asks whether models can recover the fragment-substituted complex structure starting from the pre-substitution structure, in particular (i) the shared scaffold pose accuracy, (ii) substituted fragment placement, and (iii) pocket side-chain response to the substitution. The congeneric ligand cross-docking benchmark set (Supplementary Section C.2) contains two subsets: one from Q-BioLiP (Wei et al., 2024) with experimental structures, and the other from MotifLeadDB (Kim et al., 2026) with synthetic modeled structures.

Table 2 shows that our model gives the best values across all scaffold- and fragment-specific metrics. The comparison to other tools shows the strength of our soft scaffold-restraining. rDock (Ruiz-Carmona et al., 2014) benefits from an explicit scaffold constraint but is less accurate for fragment placement and interaction recovery. Conversely, PLACER can adapt pocket side chains, but without an explicit scaffold-preserved restraint it loses more of the productive core geometry. Focusing on a subset in which considerable side-chain adaptation occurs (rows under "induced-fit-required subset" in the table), flexible sampling improves four of five metrics. Quantitative analyses on ligand-induced side-chain adaptations and more detailed analyses on the induced-fit-required subset are reported in Supplementary Figure 8 and Supplementary Table 10, respectively.

We then asked the benefit of ensemble-based evaluation in this cross-docking setting. Interactions observed consistently across the generated ensemble were more likely to be present in the target reference complex, and their recovery improved as more conformers were included (Figure 3a).

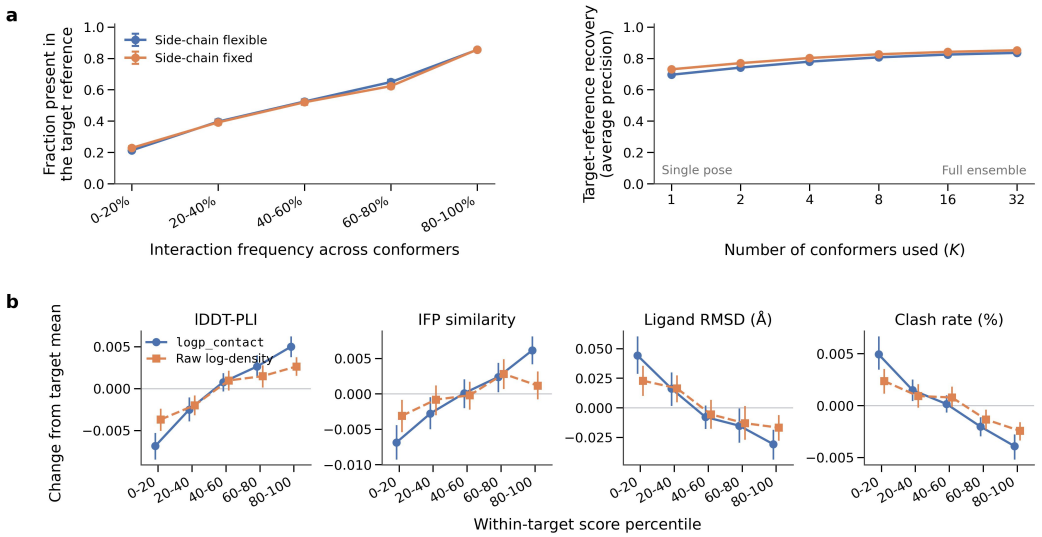

Figure 3: Ensemble generation and score-based conformer evaluation. **a**, Target-reference recovery improves as the consistency of interactions (left) or number of conformers used for analyses increases. **b**, Within-complex conformer-quality changes across score-percentile bins on the PLINDER test set; the y-axis reports the mean deviation of each metric from that complex's ensemble-mean metric. Blue denotes `logp_contact`, and orange denotes the raw joint log-density. Error bars show 95% bootstrap confidence intervals.

These trends were consistent for paired side-chain-flexible and side-chain-fixed ensembles. Thus, evaluating multiple local states provides a more reliable view of protein–ligand interactions than relying on a single pose.

4.3 Token-wise log-probability selects interaction-consistent conformers

We next evaluate score-based selection using the same congeneric benchmark. Each method uses its own scoring function to rank candidate poses. For our model, we use contact-focused score (denoted `logp_contact`; see Section 3.2 for details), which favors interaction-consistent conformers from the generated ensemble. In Supplementary Figure 7a, samples selected by the `logp_contact` show low scaffold- and fragment-RMSD and high fragment lDDT-PLI, although improvement by selection is modest due to the overall good quality of samples.

To see how the score works for the conformer selection in a target, we further examine `logp_contact` on 32-conformer ensembles generated for ligand targets from the PLINDER test set. As shown in Figure 3b, `logp_contact` consistently correlates with improved lDDT-PLI and IFP similarity, as well as reduced ligand RMSD and clash rate. The raw joint log-density generally follows the same direction, while the contact-focused score more clearly separates the highest-scoring conformers for interaction recovery and clash reduction. Qualitative test set examples are consistent with these trends (Supplementary Figure 12). Together, these analyses support the use of token-wise log-probability as a structural selection and interpretation tool while preserving the all-sample ensemble analysis in Table 2.

4.4 Local ensemble-based score is correlated with observed activity values across congeneric ligands

Finally, for downstream analysis, we test one use of our ensemble-based method to prioritize ligands within congeneric series. Using the scoring protocol introduced in Section 3.3, we check its correlation with experimental binding measurements from the FEP relative-ranking benchmark (Ross et al., 2023) (detailed in Supplementary Section D). Note that our generative model was not trained for

this purpose; only the final linear ranking score was fitted across ligands using aggregated local ensemble-derived features (Supplementary Figure 5a).

Our score is compared with the AutoDock Vina docking score and the Boltz-2 predicted-affinity score (Passaro et al., 2025) (Supplementary Figure 5d). Across the 39 common test target series, our method improves over Vina in mean pairwise ranking accuracy, but Boltz-2 remains stronger on both metrics. Per-target correlations for the test series are shown in Supplementary Figure 14. In the PTP1B example (Supplementary Figure 5b,c), the proposed score is positively associated with experimental binding free energy within the target series (Spearman $\rho = 0.72$, pairwise accuracy $= 0.77$), and the attribution maps show which ligand atoms and pocket residues contribute favorably. The BRD4(1) ASH106 series provides a failure case (Spearman $\rho = -0.52$, pairwise accuracy $= 0.25$, $n = 8$), illustrating that the activity ranking is poor when relevant contributions are absent from the model. A feature-subset ablation is reported in Supplementary Table 7.

It is promising that the weakly-supervised flow-matching likelihood correlates with the experimental activity of congeneric ligands. We emphasize that the score should not be interpreted as an absolute binding free energy, because of its dependency on the representation, sampling protocol, finite ensemble size, and the model's learned structural preferences. We expect future study (e.g., supervised learning) may elucidate its better usage for scalable binding affinity estimation.

5 Discussion

This work should be interpreted primarily as a study of probabilistic local protein–ligand structure prediction for fragment-based lead optimization. The restrained sampling regime is intentionally designed to address the central question: how a fragment modification changes the ensemble population around bound conformation. By focusing on this regime, the model can provide scalable analyses of generated or experimentally derived candidates prior to more expensive evaluation.

The current formulation makes several simplifying assumptions. The model is conditioned on an initial complex and therefore inherits some dependence on the quality of the starting pose. It focuses on local ligand and side-chain degrees of freedom while keeping the protein backbone fixed, so slow backbone motions, large induced-fit rearrangements, explicit solvent effects, and long-timescale kinetics are only indirectly captured. The generated samples are independently sampled conformers rather than a physical trajectory, and should not be interpreted as molecular dynamics frames.

These limitations also define useful extensions. A natural next step is to test the evaluator role described above in prospective fragment-optimization workflows, where candidate analogs are proposed iteratively and structural ensemble signals are used to filter, rescore, or refine them. Additional work could incorporate backbone flexibility, solvent-aware features, better uncertainty estimates for finite ensembles, and calibration against physics-based free-energy calculations. In this sense, the present work provides a probabilistic conformer-sampling and local-evaluation foundation for future small-molecule design and optimization tools.

AI use statement

Generative AI tools were used to assist with implementation, data processing, and language editing for manuscript preparation. All AI-assisted code, data-processing steps, and manuscript content were reviewed and verified by the authors, who take full responsibility for the final content of this work.

Reproducibility statement

The appendix documents the model architecture, training objectives, sampling and relaxation protocol, benchmark construction, baseline configurations, evaluation criteria, and ligand-ranking procedure. It also reports the fixed seed protocol, data splits, ablations, and additional per-target results needed to assess the robustness of the findings. The implementation code and experiment configurations are provided in the anonymous supplementary material.

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*, 630:493–500, 2024.

Amy C Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787–797, 2003.

Ivan Anishchenko, Yakov Kipnis, Indrek Kalvet, Guangfeng Zhou, Rohith Krishna, Samuel J Pellock, Anna Lauko, Gyu Rie Lee, Linna An, Justas Dauparas, Frank DiMaio, and David Baker. Modeling protein–small molecule conformational ensembles with PLACER. *Proceedings of the National Academy of Sciences*, 122(45):e2427161122, 2025. doi: 10.1073/pnas.2427161122.

Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zuhlsdorff, Cecilia Frank, and Frank Noé. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. In *International Conference on Learning Representations (ICLR)*, 2023.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15:3130–3139, 2024. doi: 10.1039/D3SC04185A.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3itjR9QxFw.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023.

Janani Durairaj, Julia Gabel, Bernardo P de Almeida, Jose Jimenez Luna, David Marchand, Jiri Sponer, Christoph Gorgulla, Miquel Duran-Frigola, Marwin Segler, Lewis H Mervin, Matthew Ragoza, Nikolaus Stiefl, Daniel Reker, Teodoro Laino, Philippe Schwaller, Michael Moret, and Michael M Bronstein. Plinder: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, 2024. doi: 10.1101/2024.07.17.603955.

Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, Rafal P Wiewiora, Bernard R Brooks, and Vijay S Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7):e1005659, 2017. doi: 10.1371/journal.pcbi.1005659.

Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

Ed Griffen, Andrew G Leach, Graeme R Robb, and Daniel J Warner. Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry*, 54(22):7739–7750, 2011. doi: 10.1021/jm200452d.

Katherine Henzler-Wildman and Dorothee Kern. Protein dynamics: from single molecules to ensembles. *Nature*, 450(7172):964–972, 2007.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806.

Ajay N Jain. Scoring functions for protein-ligand docking. *Current Protein and Peptide Science*, 7 (5):407–420, 2006.

Bowen Jing, Stephan Eismann, Patricia Pranam, Raphael Suriana, and Ron O Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations (ICLR)*, 2021.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24240–24253, 2022.

Nawoon Kim, Byunghyun Bae, Nuri Jung, Chaok Seok, and Hahnbeom Park. MotifLeadDB: A hierarchical structural data set for congeneric ligand binding activity change. *Journal of Chemical Information and Modeling*, 66(11):6376–6386, 2026. doi: 10.1021/acs.jcim.6c00128.

Xiangzhe Kong, Junwei Chen, Ziting Zhang, Gaodeng Li, Qingyuan Zhu, Lei Wei, Mingyu Li, Yan Shi, Weiyang Dai, Zishen Zhang, Wenjuan Tan, Rui Jiao, Xiaolun Wang, Jiqing Zheng, Ziyang Yu, Qilong Wu, Zhiye Guo, Li Zhang, Wentao Li, Qiaojing Huang, Tian Zhu, Xiaowo Wang, Wenbing Huang, Yuli She, Jian Zhang, Yang Liu, Kai Liu, and Jianzhu Ma. Programming biomolecular interactions with all-atom generative model. *bioRxiv*, 2026. doi: 10.64898/2026.03.12.711044.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations (ICLR)*, 2023.

Ding Luo, Xiaoyang Qu, Dixin Lu, Yiqiu Wang, Lina Dong, and Binju Wang. Ligand-conditioned side chain packing for flexible molecular docking. *Journal of Chemical Theory and Computation*, 2025. doi: 10.1021/acs.jctc.4c01636.

Andrew T. McNutt, Yanjing Li, Rocco Meli, Rishal Aggarwal, and David Ryan Koes. GNINA 1.3: The next increment in molecular docking with deep learning. *Journal of Cheminformatics*, 17:28, 2025. doi: 10.1186/s13321-025-00973-x.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, and Stephan Thaler. Boltz-2: Towards accurate and efficient binding affinity prediction, 2025. Preprint.

Michael Plainer, Marcella Toth, Simon Dobers, Hannes Stärk, Gabriele Corso, Céline Marquet, and Regina Barzilay. DiffDock-Pocket: Diffusion for pocket-level docking with sidechain flexibility. In *Machine Learning in Structural Biology, NeurIPS Workshop*, 2023.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller, and Anima Anandkumar. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1–13, 2024.

Sereina Riniker and Gregory A. Landrum. Better informed distance geometry: Using what we know to improve conformation generation. *Journal of Chemical Information and Modeling*, 55(12): 2562–2574, 2015. doi: 10.1021/acs.jcim.5b00654.

Gregory A. Ross, Chao Lu, Guido Scarabelli, Steven K. Albanese, Evelyne Houang, Robert Abel, Edward D. Harder, and Ingle Wang. The maximal and current accuracy of rigorous protein–ligand binding free energy calculations, 2023. Preprint, version 2.

Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A. Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E. Hubbard, and S. David Morley. rDock: A fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Computational Biology*, 10(4):e1003571, 2014. doi: 10.1371/journal.pcbi.1003571.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332. PMLR, 2021.

Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Rökkum Jamasb, Ilia Igashov, Weitao Du, Carla P. Gomes, Tom L. Blundell, Pietro Lió, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024. doi: 10.1038/s43588-024-00737-x.

Arne Schneuing, Ilia Igashov, Adrian W. Dobbelstein, Thomas Castiglione, Michael M. Bronstein, and Bruno Correia. Multi-domain distribution learning for de novo drug design. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=g3VCIM94ke.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Antonia Stank, Daria B. Kokh, Jonathan C. Fuller, and Rebecca C. Wade. Protein binding pocket dynamics. *Accounts of Chemical Research*, 49(5):809–815, 2016. doi: 10.1021/acs.accounts.5b00516.

Oleg Trott and Arthur J Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, Basile I M Wicky, Nikita Hanikel, Samuel J Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vazquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Hong Wei, Wenkai Wang, Zhenling Peng, and Jianyi Yang. Q-BioLiP: A comprehensive resource for quaternary structure-based protein–ligand interactions. *Genomics, Proteomics & Bioinformatics*, 22(1):qzae001, 2024. doi: 10.1093/gpbjnl/qzae001.

Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Fast protein backbone generation with SE(3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

Robert J. Young and Paul D. Leeson. Mapping the efficiency and physicochemical trajectories of successful optimizations. *Journal of Medicinal Chemistry*, 61(15):6421–6467, 2018. doi: 10.1021/acs.jmedchem.8b00180.

Jintao Zhu, Zhonghui Gu, Jianfeng Pei, and Luhua Lai. DiffBindFR: an SE(3) equivariant network for flexible protein–ligand docking. *Chemical Science*, 15(21):7926–7942, 2024. doi: 10.1039/D3SC06803J.

Appendix

Contents

A Log-probability estimation and token-level decomposition . . . 13
A.1 Joint forward sampling and log-probability estimation . . . 13
A.2 Token-level decomposition and contact-focused scoring . . . 14

B Model and Training Details . . . 15
B.1 Model and training setup . . . 15
B.2 Training loss definitions . . . 16
B.3 Sampling, relaxation, and evaluation protocol . . . 17

C Benchmark Datasets and Baselines . . . 18
C.1 Training and PLINDER conformer benchmark data . . . 18
C.2 Scaffold-preserved congeneric ligand cross-docking benchmark construction . . . 18
C.3 Benchmark baselines . . . 19

D Ligand Ranking Score and Protocol . . . 20
D.1 Algorithm for exploratory ligand prioritization . . . 21
D.2 Feature aggregation and fitted score . . . 22

E Ablation Studies . . . 23
E.1 Backbone architecture ablation . . . 23
E.2 Impact of time distribution . . . 25

F Additional Experimental Results . . . 25
F.1 MotifLeadDB scaffold-preserved benchmark . . . 25
F.2 Effect of side-chain flexibility in congeneric cross-docking . . . 27
F.3 Sampling diversity control . . . 28
F.4 PLACER robustness and PB-validity analysis . . . 29
F.5 Interpretation of log-probability . . . 30
F.6 Within-group affinity–structure relationships . . . 33
F.7 Ligand prioritization across the FEP benchmark . . . 35

A LOG-PROBABILITY ESTIMATION AND TOKEN-LEVEL DECOMPOSITION

This section derives the log-probability estimator used in our sampling procedure and its decomposition into token-level attributions. Unlike settings that compute $p_1(x_1)$ by integrating an ODE in reverse from x_1 to x_0, we compute $\log p_1(x_1)$ during forward sampling by augmenting the sampling ODE with an additional scalar state. All quantities below are conditional on a fixed local context $\mathcal{C}$, which enters only through the anchor-centered prior mean and the vector field $v_\theta(t, x; \mathcal{C})$. Once $\mathcal{C}$ is fixed, this is an ordinary time-dependent field in the updated coordinates x, so the divergence is taken with respect to x alone and the standard change-of-variables identity applies to $p_t(x \mid \mathcal{C})$. We omit $\mathcal{C}$ from the notation below.

A.1 JOINT FORWARD SAMPLING AND LOG-PROBABILITY ESTIMATION

Probability-flow identity. Let $p_t(x)$ be a time-dependent density over $x \in \mathbb{R}^d$, and let $v_t(x)$ be a time-dependent vector field that transports probability mass according to the continuity equation:

$$\partial_t p_t(x) + \nabla \cdot \big(p_t(x)\, v_t(x)\big) = 0. \tag{5}$$

We define the flow map ϕ_t induced by v_t as:

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)), \qquad \phi_0(x) = x. \tag{6}$$

Expanding the divergence term in Eq. (5) yields:

$$\partial_t p_t + v_t \cdot \nabla p_t + p_t\, (\nabla \cdot v_t) = 0. \tag{7}$$

Dividing by $p_t(x)$ gives the evolution of the log-probability:

$$\partial_t \log p_t(x) + v_t(x) \cdot \nabla \log p_t(x) + \nabla \cdot v_t(x) = 0. \tag{8}$$

By the chain rule, the total time derivative of the log-probability along the flow path $x_t = \phi_t(x)$ is:

$$\frac{d}{dt} \log p_t(\phi_t(x)) = \partial_t \log p_t(\phi_t(x)) + \Big(\nabla \log p_t(\phi_t(x))\Big)^\top \frac{d}{dt}\phi_t(x). \tag{9}$$

Substituting $\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$ from Eq. (6) into Eq. (9), we obtain:

$$\frac{d}{dt} \log p_t(\phi_t(x)) = \partial_t \log p_t(\phi_t(x)) + v_t(\phi_t(x)) \cdot \nabla \log p_t(\phi_t(x)). \tag{10}$$

Combining Eq. (10) with the PDE in Eq. (8) yields the instantaneous change of log-probability:

$$\frac{d}{dt} \log p_t(\phi_t(x)) = -\nabla \cdot v_t(\phi_t(x)). \tag{11}$$

Integrating Eq. (11) from $t = 0$ to $t = 1$ gives:

$$\log p_1(\phi_1(x)) - \log p_0(\phi_0(x)) = -\int_0^1 \nabla \cdot v_t(\phi_t(x))\, dt. \tag{12}$$

Since $\phi_0(x) = x$, letting $x_0 \sim p_0$ (an anchor-centered Gaussian in our model) and $x_1 = \phi_1(x_0)$ yields:

$$\log p_1(x_1) = \log p_0(x_0) - \int_0^1 \nabla \cdot v_t(\phi_t(x_0))\, dt. \tag{13}$$

Augmented ODE for joint sampling and log-probability estimation. To compute the integral in Eq. (13) *simultaneously with forward sampling*, we introduce an auxiliary scalar state $f(t)$ defined by:

$$\frac{d}{dt} f(t) = -\nabla \cdot v_t(\phi_t(x_0)), \qquad f(0) = 0. \tag{14}$$

Then $f(1) = -\int_0^1 \nabla \cdot v_t(\phi_t(x_0))\, dt$, allowing us to compute:

$$\log p_1(x_1) = \log p_0(x_0) + f(1). \tag{15}$$

Equivalently, we solve the following *augmented ODE*:

$$\frac{d}{dt}\begin{bmatrix} \phi_t(x_0) \\ f(t) \end{bmatrix} = \begin{bmatrix} v_t(\phi_t(x_0)) \\ -\nabla \cdot v_t(\phi_t(x_0)) \end{bmatrix}, \qquad \begin{bmatrix} \phi_0(x_0) \\ f(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}. \tag{16}$$

This yields both the generated sample $x_1 = \phi_1(x_0)$ and a numerical log-probability estimate via Eq. (15) in a single forward integration pass.

Stochastic divergence estimation. In high dimensions, computing the full divergence $\nabla \cdot v_t(x) = \mathrm{tr}(\nabla_x v_t(x))$ requires $\mathcal{O}(d^2)$ operations, which is prohibitively expensive. We therefore use the Hutchinson trace estimator:

$$\nabla \cdot v_t(x) = \mathrm{tr}\big(\nabla_x v_t(x)\big) = \mathbb{E}_z\big[z^\top \nabla_x v_t(x)\, z\big], \qquad \mathbb{E}[z] = 0, \quad \mathbb{E}[zz^\top] = I, \tag{17}$$

which can be evaluated efficiently using vector–Jacobian products (vJP) with $\mathcal{O}(d)$ cost. We use Rademacher probes in our experiments. In practice, we draw a single or a small number of probes per integration step to obtain an unbiased estimate of the divergence term in Eq. (16).

A.2 Token-level decomposition and contact-focused scoring

This subsection details the node-wise divergence decomposition summarized in the main text. For N atomic nodes, the divergence decomposes as

$$\nabla \cdot v_\theta(t, x) = \sum_{n=1}^{N} \nabla_{x_n} \cdot v_{\theta,n}(t, x). \tag{18}$$

Combining this identity with Eq. (13) and moving the finite sum outside the integral gives

$$\log p_\theta(x_1 \mid \mathcal{C}) = \log p_0(x_0 \mid \mathcal{C}) - \sum_{n=1}^{N} \int_0^1 \nabla_{x_n} \cdot v_{\theta,n}(t, x_t, \mathcal{C})\, dt. \tag{19}$$

Node-wise Hutchinson estimator. We draw a single global Hutchinson probe $z = (z_1, \ldots, z_N)$, where each $z_n \in \mathbb{R}^3$, and partition the resulting global trace estimate into node-wise contributions. Writing $J_v = \nabla_x v_\theta(t, x)$, we define

$$\widehat{d}_n(t, x) = z_n^\top [J_v^\top z]_n. \tag{20}$$

These contributions satisfy

$$\sum_{n=1}^{N} \widehat{d}_n(t, x) \;=\; z^\top J_v z, \qquad \mathbb{E}_z[\widehat{d}_n(t, x)] \;=\; \mathrm{tr}(\nabla_{x_n} v_{\theta,n}(t, x))\,. \tag{21}$$

Thus their sum is exactly the global Hutchinson estimate for a given probe, while each node term is an unbiased stochastic attribution of its diagonal Jacobian block. The velocity $v_{\theta,n}(t, x)$ is evaluated as a function of the full state x, so contextual dependencies are retained in the forward computation. The divergence contribution, however, uses only the diagonal Jacobian block $\partial v_{\theta,n}/\partial x_n$, as required by the definition of divergence. We integrate node-wise accumulators alongside sampling,

$$\frac{d}{dt} f_n(t) = -\widehat{d}_n(t, x_t), \qquad f_n(0) = 0, \tag{22}$$

so that $\sum_{n=1}^{N} f_n(1)$ recovers the global divergence term and $\log p_1(x_1) \approx \log p_0(x_0) + \sum_{n=1}^{N} f_n(1)$.

Per-node prior contribution. Because the anchor-centered Gaussian prior factorizes over atoms and coordinates, the prior log-probability itself decomposes into per-node contributions,

$$q_n = \sum_{d=1}^{3} \log \mathcal{N}\big(x_{0,nd}; \mu_{nd}(\mathcal{C}), \sigma_0^2\big), \qquad \sum_{n=1}^{N} q_n = \log p_0(x_0), \tag{23}$$

where $x_{0,nd}$ and $\mu_{nd}(\mathcal{C})$ are the d-th coordinate of node n and its prior mean. Combining this per-node prior with Eq. (22) gives the node-wise score $s_n = q_n + f_n(1)$ of Eq. (3), which avoids broadcasting the global scalar $\log p_0(x_0)$ and preserves the additive identity $\sum_{n=1}^{N} s_n \approx \log p_1(x_1)$.

Stabilized scores for ranking. The additive identity applies to the raw scores s_n. For ranking, we suppress extreme numerical outliers with clipping and linear scaling,

$$\begin{aligned} \bar{s}_n &= \mathrm{clip}(s_n, a, b), \\ \widetilde{s}_n &= 100\, \frac{\bar{s}_n - a}{b - a}, \qquad a = -15, \quad b = 1. \end{aligned} \tag{24}$$

The interval $[a, b]$ is an empirical stability range, not a theoretical bound on a continuous log-probability. Accordingly, the clipped scores $\widetilde{s}_n$ are used as stabilized relative scores and are not claimed to retain the additive joint-log-probability identity.

Table 3: Input node and edge features of the local protein–ligand graph. Hydrogens are removed before featurization. The 3D atomic coordinates of every node additionally supply the geometric (vector) channels of the GVP backbone, in addition to the scalar/one-hot features listed here.

Graph entity	Feature	Encoding
Ligand atom (node)	Element type	one-hot, {C, N, O, S, B, Br, Cl, P, I, F, H} + unknown (12)
	Degree	integer (number of bonded neighbors)
	Total H count	integer
	Aromaticity	binary
	Formal charge	integer
	Conjugation	binary (atom in a conjugated bond)
Ligand bond (edge)	Bond type	one-hot, {single, double, triple, aromatic} (4)
	Conjugation	binary
Coarse ligand anchor node	Heavy-atom count	integer
	H-bond acceptors	integer
	H-bond donors	integer
	Aromatic-ring atoms	integer
	Aliphatic-ring atoms	integer
	Formal charge	integer
Protein atom (node)	Element type	one-hot, same 11-element vocabulary + unknown (12)
	Residue type	one-hot, 20 standard amino acids + unknown (21)
	Buriedness	scalar (count of Cα within 6 Å, normalized)
Protein bond (edge)	Bond type	one-hot, {single, double, triple, aromatic} (4)

Contact-focused conformer score. For each target, we define one fixed contact-residue set as the union of residues that contact the ligand in any of the K sampled conformers,

$$\mathcal{R} = \bigcup_{k=1}^{K} \left\{ r : \min_{i \in r,\, j \in L_k} \left\| x^{i,k} - x^{j,k} \right\|_2 \leq 4.5\,\text{Å} \right\}, \tag{25}$$

where L_k is the set of ligand atoms in conformer k. The same $\mathcal{R}$ is used for every conformer of that target. For conformer k, $\widetilde{s}^{\text{prot}}(r,k)$ is the mean stabilized score over the side-chain atoms of residue r, and $\widetilde{s}^{\text{lig}}(k)$ is the mean stabilized score over ligand atoms. Their equally weighted mean gives the contact-focused score in Eq. (4).

B Model and Training Details

B.1 Model and training setup

The conformer sampler is implemented as a flow-matching model over a local protein–ligand graph. Given an input protein structure and a ligand structure, the sampler builds features for atom-level ligand nodes, coarse ligand anchor nodes, and atom-level protein nodes and edges, as summarized in Table 3. The protein is cropped to the binding site using all residues with at least one atom within a cutoff distance d_{bs} of the ligand coordinates. Unless otherwise specified, $d_{\text{bs}} = 8.0$ Å. Our method also supports an explicit residue crop, which overrides distance-based cropping.

The sampled state concatenates ligand atoms, protein atoms in the local pocket, and optional hetero molecules. Ligand atoms are always updated. Protein backbone atoms are kept fixed, while protein side-chain atoms are updated. Optional hetero molecules are included as conditioning context but are not updated. The input coordinates are centered by subtracting the center of mass of the non-updated context atoms, and the generated coordinates are transformed back to the original coordinate frame before writing outputs.

The current configuration uses a cross-product GVP-GNN backbone for the conformer sampler. In the default experimental configuration, the backbone has 12 message-passing layers, scalar/vector hidden dimensions of 196 and 32 for nodes, scalar/vector hidden dimensions of 196 and 1 for edges, a 6 Å

interaction edge cutoff, radial basis distance features with 16 bins, attention enabled, self-conditioning enabled, and optional input-coordinate features enabled. The flow-matching model uses a prior noise scale of 0.2 and a logit-normal training time distribution.

Crop sampling and off-pocket augmentation. During training the local crop radius is drawn uniformly per example: binding-site crops use $[4.5, 8.5]$ Å and off-pocket crops use $[6, 10]$ Å, exposing the network to multiple pocket sizes and contact densities. With probability 0.05 (i.e. a binding-site crop ratio of 0.95) a training example is replaced by an *off-pocket* crop centered on a randomly chosen non-binding-site residue rather than on the ligand. Off-pocket crops contain no anchored ligand and omit the ligand-specific anchor terms, while neighboring co-bound molecules and ions are retained as context, so the same side-chain denoising objective drives the model to learn generic local protein packing and side-chain environments away from annotated ligand contacts.

Self-conditioning. Following self-conditioning strategies used in diffusion and protein-structure generation models (Chen et al., 2023; Watson et al., 2023), the model first predicts a denoised endpoint $\widehat{x}_1$ from the current noisy state x_t with probability 0.5 during training. This first prediction is computed under `no_grad` and is then provided as an additional conditioning signal for the second prediction used in the loss. The self-conditioned coordinates are unnormalized, non-updated atoms are reset to their ground-truth coordinates, and the resulting structure is passed to the network as auxiliary structural context.

During sampling, the previous denoised endpoint prediction is carried forward between Euler steps. At each step, the network receives the current state, the coarse ligand anchor nodes, the input-coordinate mask, and the previous self-conditioned prediction if self-conditioning is active. If self-conditioning is disabled, this cached prediction is cleared after each step. This design stabilizes local sampling by giving the model a memory of its own previous endpoint estimate, while avoiding gradient flow through the auxiliary prediction during training.

Ligand conformer conditioning. The sampler uses coarse ligand anchor nodes derived from the input ligand geometry. Fragment atom indices map each anchor to its corresponding ligand atoms. During sampling, ligand atom noise is initialized around the corresponding anchor coordinates, and side-chain atom noise is initialized around residue reference positions using the C_α coordinate as the local anchor. This provides a soft restraint that retains the established binding context while allowing ligand and side-chain conformations to adapt under the learned flow.

A separately generated reference conformer can also be provided as an optional atom-level coordinate condition. These coordinates are passed to the GVP-GNN backbone as coordinate features, with a binary mask indicating which coordinates are provided. During training, the reference conformer is randomly dropped with probability 0.5, so the model learns to operate with or without explicit coordinate conditioning. At inference time, the reference conformer is optional and provides an additional local conformer-conditioning signal when available.

During training, we intentionally perturb the coarse ligand anchor nodes. Without this perturbation in early experiments, generated ligand atoms followed the anchor positions too closely. Adding moderate anchor noise encourages the model to use the local protein environment as well as the anchor condition, improving robustness while retaining the intended local binding region.

At sampling time, diversity can be increased by perturbing the coarse anchors to a prescribed RMS magnitude, by multiplying the baseline ligand-prior per-coordinate standard deviation $\sigma_0 = 0.2$ Å, or by combining both controls. Supplementary Table 11 reports a controlled sweep of these settings.

B.2 Training loss definitions

The endpoint and full training objectives are

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{Z(t)} \sum_{i=1}^{N_{\mathrm{u}}} w_i \|\widehat{x}_{1,i} - x_{1,i}\|_2^2, \qquad \mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{FM}} + \mathcal{L}_{\mathrm{geom}}. \tag{26}$$

The flow-matching loss in Eq. (26) is computed only on coordinates that are sampled by the model. Let N_{u} denote the number of updated atoms—ligand atoms and protein side-chain atoms—in the

local graph. With $\bar{t} = \min(t, t_{\max})$, the implementation uses $Z(t) = 3N_{\mathrm{u}}(1-\bar{t})^2$, which matches the x_1-parameterized vector field in Eq. (2). Because $v_\theta - \dot{x}_t = (\widehat{x}_1 - x_1)/(1-t)$ under the OT path in Eq. (1), dividing the endpoint MSE by $(1-\bar{t})^2$ recovers the standard velocity-matching loss, up to the constant $3N_{\mathrm{u}}$ and the per-atom weights w_i. Following FrameFlow (Yim et al., 2023), the cap $\bar{t} = \min(t, t_{\max})$ prevents this factor from diverging as $t \to 1$. We set $w_i = 8$ for ligand atoms and $w_i = 1$ for protein side-chain atoms.

The auxiliary term $\mathcal{L}_{\mathrm{geom}}$ is a weighted sum of distance-preservation terms. For a selected set of node pairs $\mathcal{P}$, define

$$d_{ij} = \|x_{1,i} - x_{1,j}\|_2, \qquad \widehat{d}_{ij} = \|\widehat{x}_{1,i} - \widehat{x}_{1,j}\|_2.$$

The basic pairwise-distance penalty is

$$\mathcal{L}_{\mathrm{dist}}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left(d_{ij} - \widehat{d}_{ij}\right)^2, \tag{27}$$

with normalization performed per complex in the implementation. Pair sets are restricted to nodes from the same complex and to local pairs within the corresponding cutoff. The loss uses several choices of $\mathcal{P}$: all local atom pairs, coarse fragment/residue centroid pairs, inter-residue pairs, and ligand-internal pairs.

Two additional variants emphasize chemically relevant short-range geometry. The Fourier distance loss applies the same comparison after embedding distances with a fixed Fourier basis $\Phi(\cdot)$:

$$\mathcal{L}_{\mathrm{Fourier}}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left\| \Phi(d_{ij}) - \Phi(\widehat{d}_{ij}) \right\|_2^2.$$

The clash penalty uses the same distance error as Eq. (27), but increases the pair weight when the predicted distance $\widehat{d}_{ij}$ is below a short-distance threshold. Overall,

$$\begin{aligned}\mathcal{L}_{\mathrm{geom}} = {} & \lambda_{\mathrm{atom}}\mathcal{L}_{\mathrm{atom}} + \lambda_{\mathrm{coarse}}\mathcal{L}_{\mathrm{coarse}} + \lambda_{\mathrm{inter}}\mathcal{L}_{\mathrm{inter}} + \lambda_{\mathrm{lig}}\mathcal{L}_{\mathrm{lig}} \\ & + \lambda_{\mathrm{Fourier}}\mathcal{L}_{\mathrm{Fourier}} + \lambda_{\mathrm{clash}}\mathcal{L}_{\mathrm{clash}}.\end{aligned}$$

These terms do not change the target endpoint. They regularize the predicted endpoint so that local geometry, ligand internal structure, and pocket contacts remain physically plausible during denoising.

B.3 Sampling, relaxation, and evaluation protocol

For each input complex, the sampler generates K independent local conformers. When a baseline can return multiple poses, we generate and evaluate the same number of poses per input complex. The sampled ensemble is evaluated as generated. Model scores are reserved for the separate top-ranked analyses and are not used to filter the conformers reported in the main benchmark tables. All stochastic sampling runs use fixed random seeds.

At inference time, we integrate the probability flow ODE defined by Eq. (2) over the updated ligand and side-chain atoms. Chemical features, the protein backbone, and any fixed scaffold or residue coordinates are inherited from the input complex. Unless otherwise stated, sampling uses 25 Euler steps.

OpenMM (Eastman et al., 2017) minimization is used only as a post-sampling relaxation step. Bond orders and stereochemistry are restored from the input ligand before minimization, and the sampled ligand together with the local pocket is relaxed for the benchmark setting of 200 minimization steps. This relaxation is intended to remove small geometric strain introduced by sampling. It is not used as an independent search procedure, and the same relaxation convention is applied consistently within each reported comparison.

The protocol records both the generated structures and the quantities needed for downstream interpretation, including per-conformer model scores and optional token-wise log-probability attributions. These annotations support the ranking and visualization analyses, while the main benchmark tables evaluate the full sampled sets.

C Benchmark Datasets and Baselines

C.1 Training and PLINDER conformer benchmark data

The model is trained on curated PLINDER protein–ligand complexes. We retain complexes with resolved ligand coordinates, usable local protein environments, and crystallographic resolution at or near 3.5 Å or better. We use the train, validation, and test splits provided by PLINDER. The local conformer-quality benchmark in Table 1 uses the PLINDER testset. This testset contains 841 systems, which expand to 909 ligand targets after treating each valid ligand chain as a separate local sampling target. Each benchmark item contains an input protein structure, a target-ligand structure, and optional co-bound molecules and ions near the binding site. They are loaded only when they fall within the binding-site cutoff and are used as static conditioning context.

C.2 Scaffold-preserved congeneric ligand cross-docking benchmark construction

The scaffold-preserved benchmark is designed for local lead optimization rather than global redocking. We therefore curate congeneric ligand groups in which a common scaffold is expected to retain a similar binding pose, and evaluate directed source-to-target cross-docking tasks within each group. For a source ligand A and a target ligand B, the method receives the source holo complex and a target-ligand input conformer whose pair common scaffold is placed on the source scaffold coordinates. The reference target holo complex is used only for evaluation. We separately evaluate the pair common scaffold, the added fragment, and the local pocket response because these components correspond to the design questions in scaffold decoration and fragment-growing workflows.

Input-conformer construction and anchor definition. For each directed source-to-target task, the target-ligand input conformer is constructed without using the target's experimental bound pose. The shared scaffold is identified by maximum-common-substructure matching between the source and target ligands, and the scaffold atoms are set to the source holo coordinates. The remaining substituted atoms are produced by an RDKit ETKDG embedding (Riniker and Landrum, 2015) of the full target topology, then aligned through the common scaffold. The constructed congeneric dataset excludes pairs whose common scaffold is too far from the source scaffold after this procedure, using a scaffold RMSD cutoff of 1.5 Å. The coarse ligand anchor nodes used by the model are the fragment centers of mass of this input conformer, and the featurizer applies the same fragment-center rule to scaffold and substituted fragments alike. For this benchmark, perturbation of coarse ligand anchor nodes is disabled, so the coarse ligand anchor nodes for the substituted fragment are exactly the generated input positions rather than perturbed or copied target coordinates. Consequently, the only experimentally determined coordinates entering the model are the source ligand scaffold pose and the source holo receptor, while the experimentally determined target-ligand pose is used only for evaluation.

The primary scaffold-preserved set is built from experimentally determined Q-BioLiP complexes. We keep relevant structures with resolution at most 3.0 Å and, for this experimentally determined benchmark, require release dates after May 25, 2024 where release metadata are available. We require at least two distinct ligands per group, require the aligned binding sites to share sufficient pocket residues, and apply the scaffold-preservation criterion above as an applicability filter. This selects ligand pairs appropriate for local lead optimization, rather than claiming that all ligand series preserve their scaffold pose.

Because the experimentally determined Q-BioLiP subset is intentionally conservative, we also construct a larger supplementary set from MotifLeadDB High-Core, the high-quality subset of MotifLeadDB. MotifLeadDB provides higher coverage of congeneric ligands but relies on modeled structures, so we use it as a high-confidence supplementary stress test and interpret it separately from the experimentally determined Q-BioLiP benchmark. The same directed congeneric ligand cross-docking benchmark construction, scaffold-drift filter, and evaluation metrics are applied to both datasets.

Dataset-level statistics are summarized in Figure 4. These diagnostics show the ligand and scaffold size distributions, source–target chemical similarity within groups, and pocket side-chain rearrangements near the substitution site.

Table 4: Scaffold-preserved benchmark datasets. Q-BioLiP is the primary experimentally determined split. MotifLeadDB is used as a larger high-confidence supplementary split. Counts are reported after final task-input construction.

Dataset	Groups	Group members	Mean members per group	Directed cross-docking tasks
Q-BioLiP (experimental)	48	109	2.22	135
MotifLeadDB (high confidence)	814	2113	2.59	3377

Table 5: Benchmark protocol summary. This table documents which degrees of freedom each method updates, whether scaffold or fragment locality is used, whether the same candidate budget K is used, and whether relaxation is applied.

Method	Ligand update	Side-chain update	Local/scaffold restraint	Candidates	Relaxation
Random perturb. + minim.	yes	yes	local	K	yes
Side-chain repacking	no	yes	pocket	K	optional
Vina local search	yes	no	docking box	K	no
Vina flexible docking (random side-chain init)	yes	limited	docking box	K	no
DiffDock-Pocket	yes	yes	pocket crop	K	no
PLACER	yes	yes	input complex	K	no
Ours	yes	yes	coarse ligand anchors	K	optional

C.3 Benchmark baselines

Baseline methods are configured to generate the same number of poses or conformers as our model. On the induced-fit-required Q-BioLiP subset in Table 10, AutoDock Vina was configured to generate 32 poses per task but returned 9–32; all returned poses were retained for evaluation. Rigid baselines keep the receptor fixed, while flexible or repacking baselines are categorized according to whether they update ligand coordinates, side-chain coordinates, or both. For fairness, Table 1 uses all generated poses/conformers from each method rather than selecting the best pose by score.

Per-method score-based selection criteria. In the score-based selection analyses (Figure 3b and 7), each method ranks poses using its own reference-free score, with no access to the experimental reference structure. AutoDock Vina uses its docking energy, rDock uses its total score, SigmaDock uses a reference-free affinity score (GNINA (McNutt et al., 2025)), GALigandDock uses its internal score, PLACER uses its self-predicted pose RMSD, and our model uses `logp_contact` (defined in Section 3.2).

For methods that can take strong advantage of crystallographic input coordinates, we perturb the local starting complex rather than giving the exact native ligand and side-chain arrangement whenever this is compatible with the method. In these perturbed-input settings, pocket side chains within 8 Å of the ligand are initialized with random side-chain χ angles. For DiffDock-Pocket and the PLACER perturb control, ligand rotatable bonds are also initialized with Gaussian torsion noise (standard deviation 20°, clipped to 45°). The random-init Vina flexible receptor row uses side-chain randomization only (no ligand torsion perturbation), matching how that baseline consumes its ligand input. PLACER is reported in the main table under its standard input setting, and an otherwise matched perturbed-input PLACER control is reported separately in Supplementary Table 12. Rosetta fixbb is treated as a side-chain recovery control. The native ligand pose and backbone are fixed, while nearby side chains are removed and repacked.

The "Random perturb. + minim." baseline has a different purpose from the fairness perturbation controls above. It is an intentionally stronger perturb-and-relax control (ligand torsion noise with larger scale and backbone-initialized side-chain rebuild) used to test whether local relaxation alone is sufficient to recover near-native interactions.

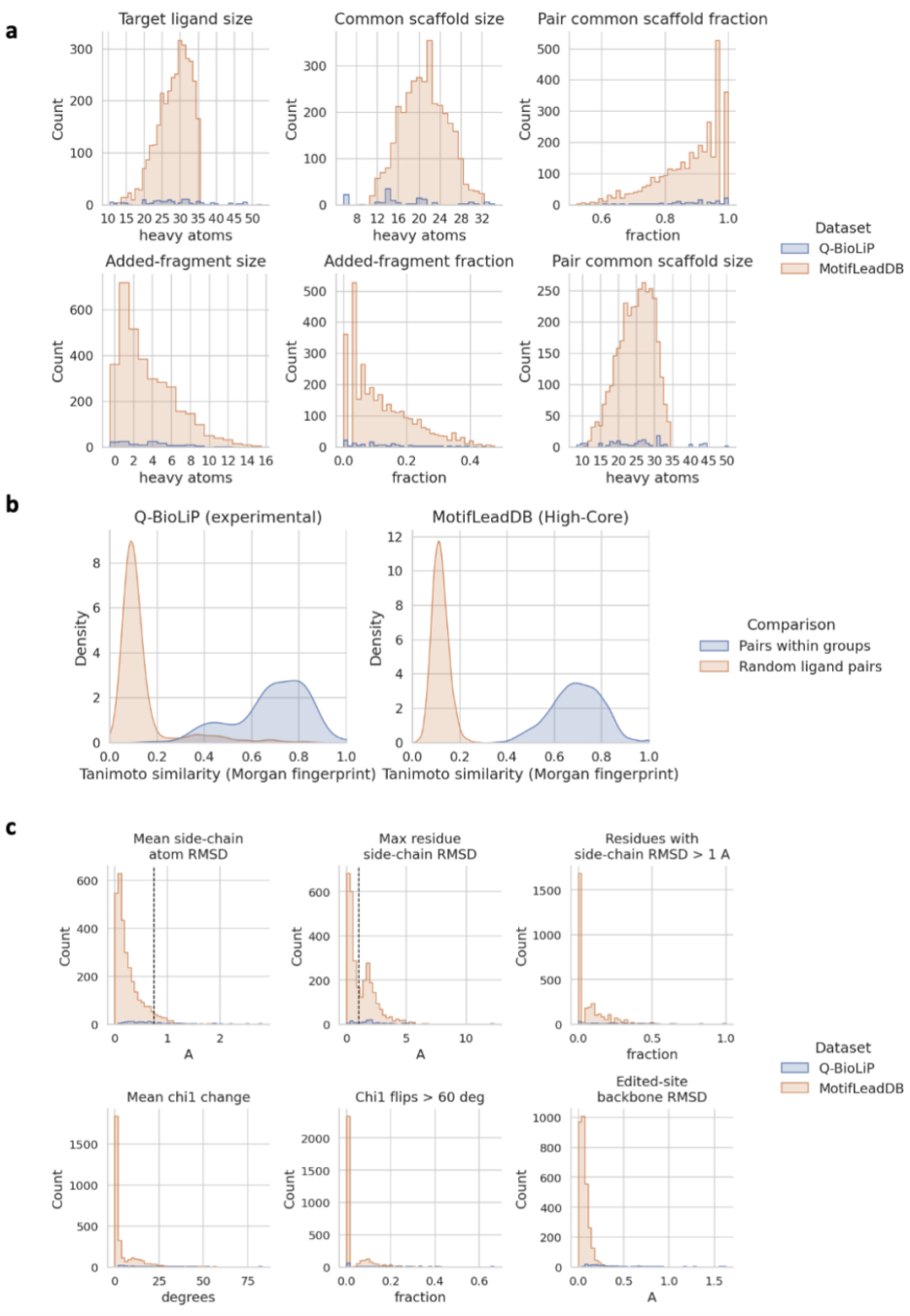

Figure 4: Characterization of the scaffold-preserved benchmark. The figure summarizes ligand size, common-scaffold size, pair common-scaffold fraction, added-fragment size, within-group ligand similarity, and local pocket side-chain rearrangements. Q-BioLiP is the experimentally determined benchmark split and MotifLeadDB is the high-confidence supplementary split.

D Ligand Ranking Score and Protocol

This section gives the full definition of the ligand-level score used for exploratory ligand prioritization. The main text gives the conceptual motivation. Here we provide the exact aggregation procedure, normalization, and component weights used in Figure 5.

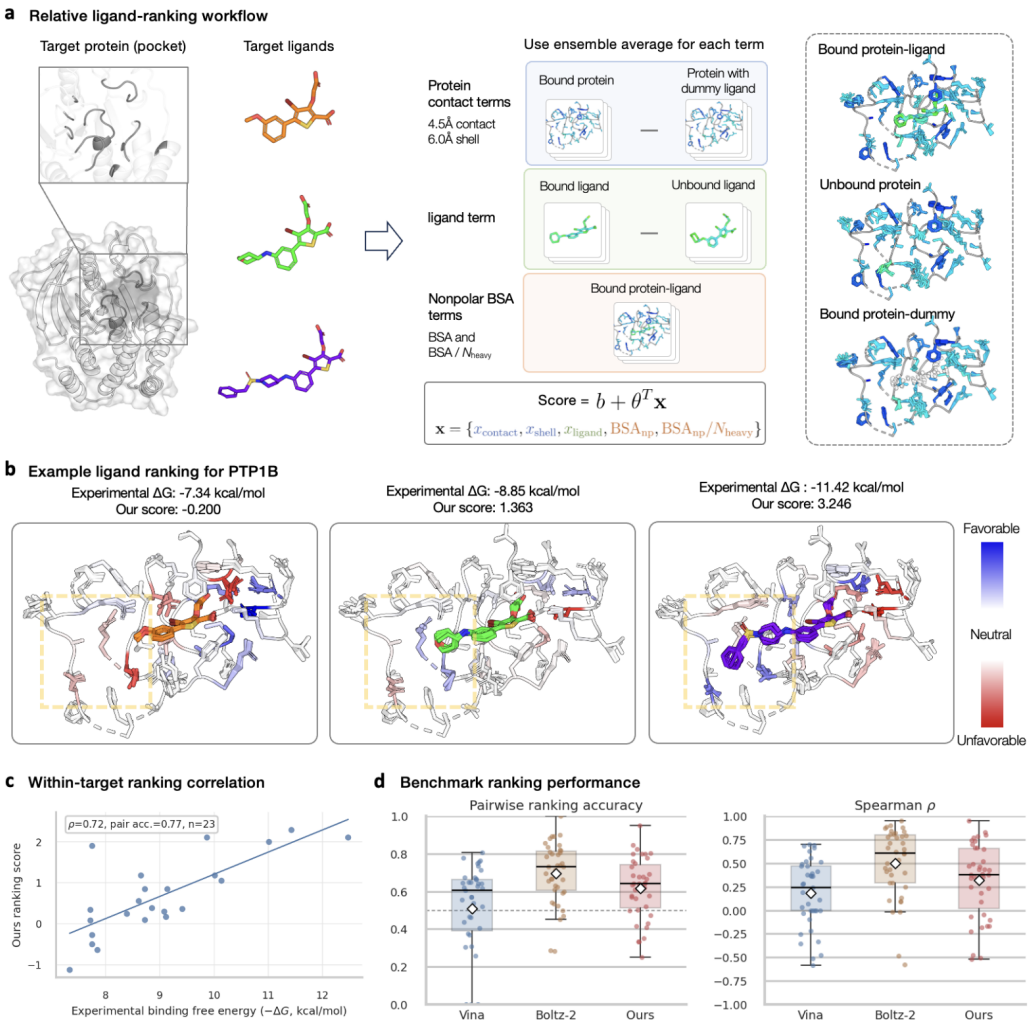

Figure 5: Downstream usage of local ensemble features for ligand prioritization across congeneric series. (a) Workflow for comparing ligands within a congeneric series. Sampled bound, unbound, and dummy-reference ensembles are used to extract protein contact-shell, ligand, and nonpolar buried-surface-area features, combined by a fitted linear ranking model. (b) PTP1B ligands with experimental binding free energies and predicted scores. Token-wise attributions highlight favorable and unfavorable contributions. (c) Experimental binding free energy and the proposed score for the PTP1B example series. (d) Ranking benchmark across the test target series, comparing AutoDock Vina docking scores, Boltz-2 predicted affinity, and our score by pairwise ranking accuracy and Spearman correlation.

D.1 Algorithm for exploratory ligand prioritization

The ranking protocol uses the conformer sampler as a local evaluator rather than as a molecule generator. For each ligand in a matched series, we generate an ensemble of local protein–ligand conformations from the same structural protocol and compute token-wise log-probability-derived scores along the sampling trajectory. The final ligand score is obtained by comparing the bound ensemble with matched reference conditions and aggregating the resulting token-level differences over conformers.

For the FEP relative-ranking benchmark, we use congeneric target series from the full benchmark dataset (Ross et al., 2023). We divide the target series into disjoint training and test sets based on

protein-sequence clusters, keeping related target series in the same split. The training split is used to fit the ranking model, whereas the test split is used only for evaluation. Experimental binding strength is defined as the negative of the reported affinity value, so that larger values indicate stronger binding. Since the analysis evaluates rank order within each target series rather than absolute affinity calibration across targets, we apply two target-level eligibility filters before computing ranking metrics. A series must contain at least five ligands and at least two distinct experimental affinity values. These filters remove very small or experimentally tied series for which target-level rank correlation and pairwise ranking accuracy are not informative.

Algorithm 1 Relative ligand ranking from conformer ensembles

Input. protein P, ligands $\{L_i\}_{i=1}^{N}$, initial complexes, conformers per state K, temperature β, fitted weights $\mathbf{w}$, bias b
Output. within-series ranking of $\{L_i\}$
for each ligand L_i **do**
  Generate matched bound, unbound-reference, and dummy-reference conformer sets with identical sampling settings.
  Compute raw node scores s_n and stabilized token scores using Eq. (24).
  Compute bound-state conformer weights $\pi_k \propto \exp(\beta\, S_{\mathrm{bound}}(k))$ using Eq. (28).
  Compute protein contact and shell terms from the π_k-weighted bound–dummy difference using Eq. (29).
  Compute the ligand term from the π_k-weighted bound–unbound difference using Eq. (30).
  Compute nonpolar buried surface area $\mathrm{BSA}_{\mathrm{np}}$ and $\mathrm{BSA}_{\mathrm{np}}$ per ligand heavy atom.
  Form the feature vector $\mathbf{x}_i = [\,\text{contact},\ \text{shell},\ \text{ligand},\ \mathrm{BSA}_{\mathrm{np}},\ \mathrm{BSA}_{\mathrm{np}}/\text{heavy atom}\,]$.
  Compute $\mathrm{Score}_i = \mathbf{w}^{\top}\mathbf{x}_i + b$ using Eq. (31).
end for
Rank ligands within the same target series by Score_i.

The exact component definitions are given below. In the main experiments we use the fitted linear score as the primary ligand-ranking score, while total log-probability, ligand-only log-probability, and alternative normalizations are treated as ablation variants.

Reference-state construction. All three states use identical sampling settings and share the same target-level pocket crop, defined as the union of protein residues with any atom within 8.0 Å of any ligand in the series. Matched conformers use the same initial prior noise across states. For the *unbound* reference, the ligand and its coarse anchor nodes are translated out of the pocket along the pocket-to-ligand axis by the protein radius plus the ligand radius plus 15 Å. The protein pocket and co-bound molecules and ions remain in the conditioning graph, thereby removing protein–ligand contacts while preserving the ligand-internal anchor geometry.

The *dummy-ligand* reference uses the sampled bound geometry, bond connectivity, and coarse anchors, but replaces each ligand atom with an atomic-number-zero placeholder encoded as the unknown-element category. Ligand coordinates are pinned to the sampled bound geometry through endpoint replacement during sampling, while the pocket side chains remain free to evolve. The dummy state is used only through the bound–dummy difference on the matched geometry. Protein-side differences are averaged over the side-chain heavy atoms of each residue. Contact and shell terms include residues within 4.5 and 6.0 Å, respectively, of a bound ligand heavy atom in at least 25% of conformers.

D.2 Feature aggregation and fitted score

Let $k = 1, \ldots, K$ index conformers in an ensemble, and let $r \in \mathcal{R}$ denote protein contact residues (defined consistently across bound/unbound/dummy states). All token values in this section are the stabilized scores of Eq. (24). We write $\widetilde{s}_u^{\mathrm{prot}}(r,k)$ for the protein score assigned to residue r under state $u \in \{\mathrm{bound}, \mathrm{unbound}, \mathrm{dummy}\}$ at conformer k, and $\widetilde{s}_u^{\mathrm{lig}}(k)$ for the ligand-side score aggregated over ligand tokens.

Table 6: Fitted ligand-ranking model coefficients.

Feature	Coef.	Interpretation
Protein contact	-0.0027	4.5 Å protein-side chemistry signal (bound–dummy).
Protein shell	0.161	6.0 Å protein-side chemistry signal (bound–dummy).
Ligand term	0.224	Element-normalized ligand-side binding-environment signal (bound–unbound).
Nonpolar BSA	0.0102	Total nonpolar buried surface area.
Nonpolar BSA/heavy atom	-0.286	Nonpolar buried surface area normalized by ligand size.

We aggregate conformers using softmax weights derived from the bound-state conformer score $S_{\mathrm{bound}}(k)$.

$$\pi_k \;=\; \frac{\exp\big(\beta\, S_{\mathrm{bound}}(k)\big)}{\sum_{j=1}^{K}\exp\big(\beta\, S_{\mathrm{bound}}(j)\big)}, \qquad \langle f(k)\rangle_k \;=\; \sum_{k=1}^{K}\pi_k\, f(k). \tag{28}$$

We use $\beta = 0.5$ in all experiments. Angle brackets $\langle\cdot\rangle_r$ denote averaging over residues in $\mathcal{R}$.

We define the chemistry-specific component on the protein side as

$$\Delta S_{\mathrm{chem}} = \Big\langle \Big\langle \widetilde{s}^{\,\mathrm{prot}}_{\mathrm{bound}}(r,k) - \widetilde{s}^{\,\mathrm{prot}}_{\mathrm{dummy}}(r,k) \Big\rangle_r \Big\rangle_k , \tag{29}$$

The ligand-side component is defined as

$$\Delta S_{\mathrm{lig}} = \Big\langle \widetilde{s}^{\,\mathrm{lig}}_{\mathrm{bound}}(k) - \widetilde{s}^{\,\mathrm{lig}}_{\mathrm{unbound}}(k) \Big\rangle_k . \tag{30}$$

Here, ΔS_{chem} measures the protein-side plausibility gained from ligand-specific chemistry beyond matched ligand-shaped occupancy. ΔS_{lig} measures how the pocket changes ligand-side plausibility relative to the matched unbound state.

The final score used in Figure 5 is a fitted linear ranking model,

$$\boxed{\mathrm{Score} = \mathbf{w}^{\top}\mathbf{x} + b} \tag{31}$$

where $\mathbf{x}$ contains five features, namely a protein contact term, a protein shell term, an element-normalized ligand term, nonpolar buried surface area ($\mathrm{BSA}_{\mathrm{np}}$), and nonpolar buried surface area per ligand heavy atom. The protein contact and shell terms use ΔS_{chem} at 4.5 Å and 6.0 Å, respectively, and the ligand term uses ΔS_{lig}. Both BSA terms are averaged over the generated bound-state conformers using the same weights π_k, and are therefore ensemble-derived structural features rather than static ligand descriptors.

The coefficients are fitted by L2-regularized pairwise logistic regression ($\alpha = 1.0$) using ligand pairs from the same target series. Each training example is the feature difference $\mathbf{x}_a - \mathbf{x}_b$, labeled by which ligand has the stronger experimental binding measurement. Feature standardization and ligand normalization are fitted only on the training split. The fit uses 557 ligands from 53 target series, and evaluation uses 782 ligands from 39 held-out target series. Spearman correlation and pairwise accuracy are computed within each target series and then summarized across eligible targets.

To quantify this contribution, we refit the same pairwise logistic model using restricted feature subsets while keeping the train/test split and target-local pair construction fixed (Table 7). The BSA features alone provide a substantial ranking signal, while the full feature set modestly improves both target-averaged metrics. Thus, the protein- and ligand-side plausibility terms provide complementary information beyond packing descriptors.

E Ablation Studies

E.1 Backbone architecture ablation

The implementation supports several equivariant backbone choices. The main model uses a cross-product GVP-GNN because GVP message passing explicitly represents scalar and vector features (Jing et al., 2021), while the final cross-product coordinate update captures local oriented

Table 7: Feature-subset ablation for the fitted ligand-ranking score on the FEP test target series. Both variants are fitted using the same training split; metrics are computed within each target series and then averaged over targets.

FEATURE SET	TARGET-MEAN SPEARMAN	TARGET-MEAN PAIRWISE ACC.
FULL MODEL	0.317	0.619
BSA ONLY	0.280	0.607

Table 8: Validation ablations for the conformer sampler. Values are mean ± standard deviation over validation targets. The screening ablations were trained for approximately 30 epochs and evaluated using the best checkpoint in that window. The full model row is not part of the controlled ablation screen and is included only as a reference for the fully trained production model.

Group	Variant	lDDT ↑	lDDT-PLI ↑	Chi1 ↑	Chi2 ↑	CB RMSD ↓	Clash rate ↓
Backbone	EGNN	0.847±0.045	0.760±0.090	0.750±0.089	0.632±0.094	0.103±0.048	0.521±0.999
Backbone	GVP	0.865±0.043	0.781±0.094	0.794±0.090	0.684±0.101	0.112±0.102	0.528±1.072
Backbone	GVP + cross	0.856±0.043	0.781±0.090	0.776±0.095	0.636±0.097	0.139±0.034	0.500±1.014
Time distribution	Uniform	0.828±0.042	0.734±0.090	0.695±0.091	0.582±0.101	0.153±0.049	0.646±1.129
Time distribution	Mixture of uniform and beta	0.770±0.044	0.645±0.059	0.515±0.073	0.472±0.076	0.270±0.099	0.707±0.936
Time distribution	Logit-normal	0.856±0.043	0.781±0.090	0.776±0.095	0.636±0.097	0.139±0.034	0.500±1.014
Reference	Full model	0.883±0.042	0.792±0.094	0.830±0.091	0.723±0.111	0.089±0.029	0.591±1.208

geometry that is useful for side-chain and ligand placement. The network builds a joint graph over ligand atoms, local protein atoms, coarse ligand anchor nodes, and optional co-bound molecules and ions. Protein–ligand and intra-pocket edges are formed using distance cutoffs, and distance features are embedded with radial basis functions. Self-conditioning and optional input-coordinate features provide additional coordinate-derived context.

Let $(\mathbf{h}, \mathbf{z})$ denote the scalar node states and vector-derived coordinates returned by the GVP stack, and let $\mathbf{a}_{ij}$ denote the scalar edge features for edge (i, j). The final coordinate update used by the default backbone is

$$\begin{aligned}
\mathbf{r}_{ij} &= \mathbf{z}_i - \mathbf{z}_j, \\
\mathbf{q}_{ij} &= (\mathbf{z}_i - \bar{\mathbf{z}}_{b_i}) \times (\mathbf{z}_j - \bar{\mathbf{z}}_{b_j}), \\
\alpha_{ij} &= R \tanh\big(\phi_x^d([\mathbf{h}_i, \mathbf{h}_j, \mathbf{a}_{ij}])\big), \\
\beta_{ij} &= R \tanh\big(\phi_x^\times([\mathbf{h}_i, \mathbf{h}_j, \mathbf{a}_{ij}])\big), \\
\widehat{\mathbf{x}}_{1,i} &= \mathbf{z}_i + \mathbf{c}_i + \frac{1}{C} \sum_{j \in \mathcal{N}(i)} \left[\mathrm{norm}(\mathbf{r}_{ij})\alpha_{ij} + \mathrm{norm}(\mathbf{q}_{ij})\beta_{ij}\right].
\end{aligned} \tag{32}$$

Here $\mathrm{norm}(\mathbf{u}) = \mathbf{u}/(\|\mathbf{u}\|_2 + \epsilon)$, $\bar{\mathbf{z}}_b$ is the coordinate mean for graph b, C is the aggregation normalization constant, R is the coordinate range, and $\mathbf{c}_i$ restores the local anchor offset when coordinates are represented relative to anchors.

The architecture ablation compares this default cross-product GVP-GNN against an EGNN backbone (Satorras et al., 2021) and a plain GVP-GNN without the final cross-product coordinate update. Screening ablations were trained for approximately 30 epochs and evaluated using the best validation checkpoint within that window; validation metrics had largely saturated by then.

In the backbone comparison, the plain GVP backbone outperformed EGNN on most side-chain and contact-shell metrics. The plain GVP and cross-product GVP variants were close in lDDT-PLI, with the plain GVP giving slightly better side-chain torsion and CB RMSD metrics in this short ablation. We nevertheless use the cross-product update in the final model because the cross-product term provides an SE(3)-equivariant oriented coordinate update, which can represent handed local geometry. Adding this final cross-product update is also lighter than adding an additional full GVP message-passing layer.

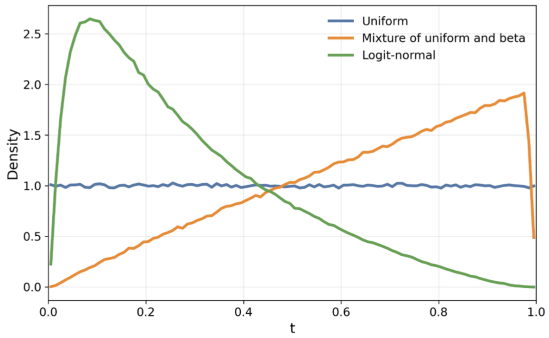

Figure 6: Training time distributions used in the ablation. The logit-normal distribution places more probability mass on noisier states, whereas the mixture of uniform and beta places more mass near the denoised endpoint.

E.2 Impact of time distribution

Training samples a continuous time t for each complex and learns to predict the denoised endpoint x_1 from the interpolated state x_t. The controlled ablation compares uniform sampling, the Proteina-style mixture of uniform and beta distributions (Geffner et al., 2025), and the logit-normal default used by the final model. These distributions place different amounts of training probability near the noisy and denoised ends of the path (Figure 6). Sampling at inference uses the Euler integration protocol described in Section B.3.

Among the tested settings, the time distribution had the largest effect on validation quality. The logit-normal distribution substantially improved lDDT, lDDT-PLI, side-chain torsion accuracy, CB RMSD, and clash rate relative to both uniform sampling and the mixture of uniform and beta (Table 8). This suggests that, for local protein–ligand conformer refinement, allocating more training probability to noisier states is critical for learning a robust denoising field.

F Additional Experimental Results

F.1 MotifLeadDB scaffold-preserved benchmark

We repeat the main-text Q-BioLiP scaffold-preserved analyses on MotifLeadDB High-Core using the same construction and evaluation protocol described in Section C.2. The all-sample structural metrics on the full benchmark and induced-fit-required subset are reported in Table 9, and the score-based conformer-selection analysis is shown in Figure 7b. Together, these results show that the Q-BioLiP trends broadly persist on this larger congeneric benchmark.

Table 9: Scaffold-preserved congeneric cross-docking on the full MotifLeadDB benchmark and an induced-fit-required subset. The subset uses the same target-structure-defined criteria as Table 2 and contains the 60 tasks (32 groups) evaluable by every method. Metrics use the same all-sample and congeneric-group-balanced aggregation as Table 2.

Method	Scaffold restraint	Pocket side chains	Scaffold RMSD ↓	Scaffold IFP ↑	Frag. RMSD ↓	Frag. IFP ↑	Frag. lDDT-PLI ↑
Full MotifLeadDB benchmark							
AutoDock Vina†	none	fixed	5.44 ± 1.28	0.62 ± 0.12	6.92 ± 2.13	0.24 ± 0.14	0.41 ± 0.12
rDock†	hard	fixed	1.59 ± 0.70	0.84 ± 0.08	2.99 ± 1.51	0.50 ± 0.21	0.67 ± 0.13
SigmaDock‡	none	fixed	3.52 ± 1.79	0.77 ± 0.10	5.48 ± 2.95	0.34 ± 0.22	0.52 ± 0.18
GALigandDock†	none	flexible	6.16 ± 1.27	0.58 ± 0.10	7.65 ± 2.00	0.20 ± 0.10	0.37 ± 0.09
PLACER‡	none	flexible	3.33 ± 1.02	0.74 ± 0.09	4.93 ± 1.79	0.34 ± 0.17	0.54 ± 0.12
Ours‡	soft (learned)	fixed	**1.29 ± 0.55**	**0.89 ± 0.07**	**2.50 ± 1.26**	**0.55 ± 0.22**	**0.73 ± 0.12**
		flexible	1.33 ± 0.55	0.86 ± 0.07	2.56 ± 1.27	0.53 ± 0.21	0.70 ± 0.12
Induced-fit-required subset (60 tasks; 32 groups)							
AutoDock Vina†	none	fixed	5.51 ± 1.11	0.61 ± 0.12	7.61 ± 2.29	0.19 ± 0.11	0.36 ± 0.09
rDock†	hard	fixed	1.81 ± 0.96	0.81 ± 0.10	3.84 ± 2.16	0.43 ± 0.24	0.58 ± 0.15
SigmaDock‡	none	fixed	4.46 ± 1.47	0.72 ± 0.11	7.25 ± 2.93	0.23 ± 0.19	0.40 ± 0.16
GALigandDock†	none	flexible	6.15 ± 1.10	0.57 ± 0.10	8.27 ± 2.01	0.16 ± 0.08	0.33 ± 0.08
PLACER‡	none	flexible	3.42 ± 0.94	0.72 ± 0.07	5.52 ± 2.14	0.29 ± 0.15	0.49 ± 0.14
Ours‡	soft (learned)	fixed	**1.51 ± 0.85**	**0.86 ± 0.09**	3.25 ± 1.62	0.43 ± 0.28	0.62 ± 0.14
		flexible	1.55 ± 0.84	0.83 ± 0.09	**3.23 ± 1.63**	**0.43 ± 0.25**	**0.62 ± 0.14**

†Classical method. ‡Machine-learning-based method.

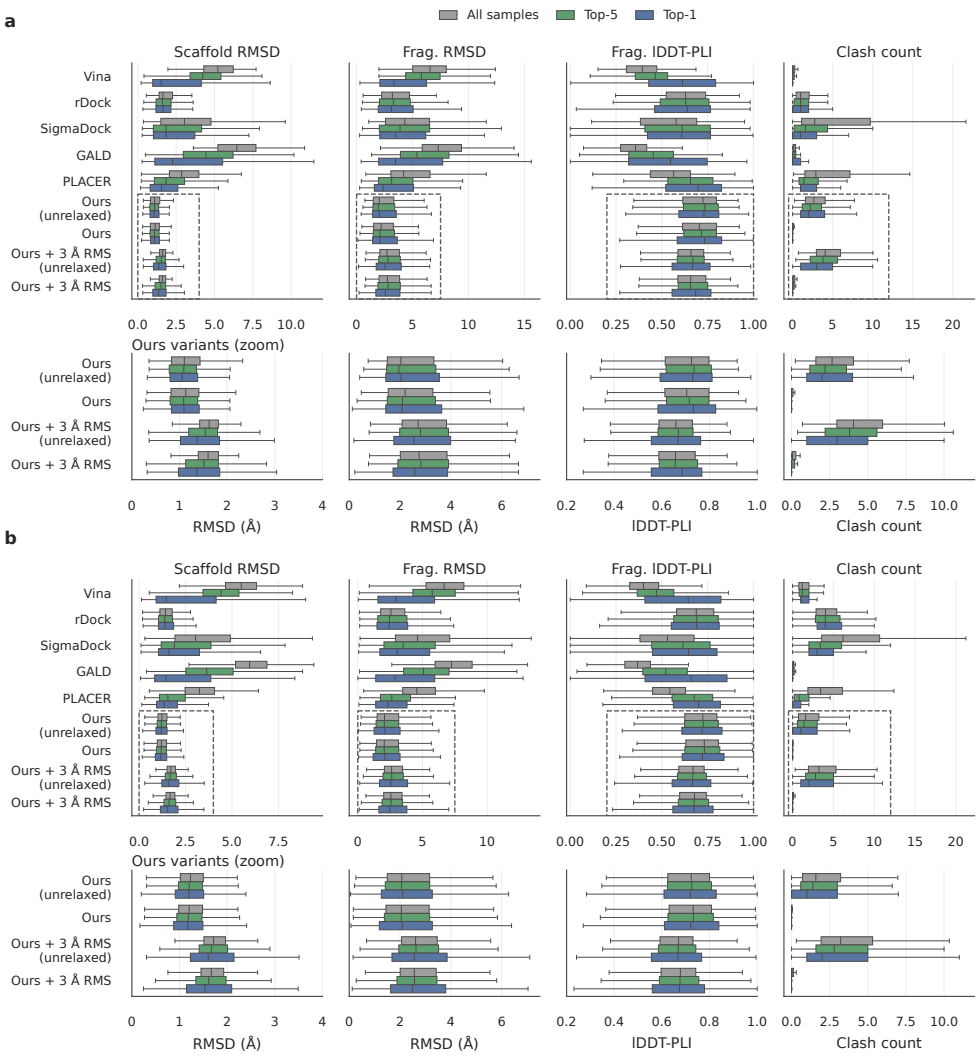

Figure 7: Score-based conformer selection on scaffold-preserved congeneric cross-docking benchmarks. (a) Experimentally determined Q-BioLiP benchmark. (b) MotifLeadDB benchmark. Gray, green, and blue boxes show all samples, top-5, and top-1 conformers selected by each method's scoring function, respectively; the lower row of each panel enlarges our model variants. Per-method selection criteria are defined in Section C.3.

F.2 Effect of side-chain flexibility in congeneric cross-docking

To isolate the effect of side-chain flexibility, we compare the standard joint ligand–side-chain sampler with an otherwise matched variant in which pocket side chains remain fixed. Across the paired full sets (129 Q-BioLiP and 3373 MotifLeadDB tasks), fixing side chains slightly improves the mean structure metrics (Table 2 and Supplementary Table 9). This suggests that many source pockets are already compatible with the target analog and that additional side-chain flexibility can introduce sampling variance when pocket adaptation is not required.

On the induced-fit-required subset defined in Table 2, flexible sampling improves four of five structure metrics for the 13 common Q-BioLiP tasks, whereas differences remain small or mixed for the 60 common MotifLeadDB tasks (Supplementary Table 9).

Table 10: Sampling and reference-free ranking of near-native substituted-fragment poses on the induced-fit-required Q-BioLiP subset. Sampling success and ranking success@K report recovery of a pose with fragment RMSD ≤ 2 Å in the available ensemble and top K poses, respectively. Values are percentages averaged across 11 congeneric groups (13 tasks).

Method	Sampling success	Ranking success@1	Ranking success@5
AutoDock Vina	22.73	0.00	13.64
rDock	63.64	4.55	9.09
SigmaDock	45.45	18.18	22.73
GALigandDock	59.09	13.64	27.27
PLACER	45.45	18.18	36.36
Ours (side-chain fixed)	81.82	**54.55**	**72.73**
Ours	**90.91**	**54.55**	**72.73**

For near-native fragment poses (fragment RMSD ≤ 2 Å), flexibility increases Q-BioLiP sampling success from 81.82% to 90.91%, while ranking success@1 and success@5 remain unchanged. Thus, broader sampling support does not automatically translate into better top-ranked recovery (Table 10).

Figure 8 provides qualitative examples of ligand-induced side-chain adaptation in the scaffold-preserved fragment substitution setting. Jointly sampling the pocket side chains relieves clashes and forms local interactions that fixed-receptor placement of the edited ligand would miss.

F.3 Sampling diversity control

Table 11 evaluates the two sampling-time diversity controls described in Section B.1 on 96 PLINDER targets. Increasing either the coarse-anchor perturbation or the ligand-prior standard deviation broadens the generated ensemble. Relative to the default setting, the strongest combined perturbation increases diversity from 1.28 to 1.83 Å, while lDDT-PLI decreases from 0.81 to 0.70, IFP similarity from 0.65 to 0.52, and PB-validity from 95.8% to 86.1%. Side-chain χ_1/χ_2 accuracy is largely retained, indicating that these controls primarily trade native-contact recovery for broader ligand sampling.

Table 11: Sampling diversity control through coarse-anchor perturbation and the ligand-prior standard-deviation multiplier. Values are target-level mean $\pm$ standard deviation.

Coarse perturb RMS (Å)	Ligand-prior σ multiplier**	PB-valid (%)* ↑	Clash ↓	lDDT-PLI ↑	IFP sim. ↑	Diversity ↑	χ_1/χ_2 acc. ↑
0.0	1.0	95.8 ± 18.9	0.01 ± 0.04	0.81 ± 0.09	0.65 ± 0.13	1.28 ± 0.30	0.85 ± 0.08 / 0.74 ± 0.11
	1.5	92.6 ± 24.7	0.01 ± 0.04	0.80 ± 0.09	0.63 ± 0.12	1.35 ± 0.31	0.85 ± 0.08 / 0.74 ± 0.11
	2.0	91.8 ± 26.1	0.01 ± 0.04	0.78 ± 0.09	0.60 ± 0.12	1.43 ± 0.31	0.85 ± 0.07 / 0.74 ± 0.11
	2.5	88.3 ± 30.2	0.01 ± 0.04	0.75 ± 0.08	0.57 ± 0.11	1.53 ± 0.32	0.84 ± 0.07 / 0.74 ± 0.11
1.0	1.0	95.8 ± 19.0	0.01 ± 0.04	0.81 ± 0.09	0.64 ± 0.13	1.32 ± 0.32	0.85 ± 0.08 / 0.74 ± 0.11
	1.5	92.6 ± 24.7	0.01 ± 0.04	0.79 ± 0.09	0.62 ± 0.12	1.39 ± 0.33	0.85 ± 0.08 / 0.74 ± 0.11
	2.0	90.7 ± 27.5	0.01 ± 0.04	0.77 ± 0.09	0.59 ± 0.11	1.48 ± 0.33	0.84 ± 0.08 / 0.74 ± 0.11
	2.5	88.5 ± 30.2	0.01 ± 0.04	0.75 ± 0.08	0.56 ± 0.10	1.58 ± 0.34	0.84 ± 0.07 / 0.74 ± 0.11
2.0	1.0	92.8 ± 24.5	0.01 ± 0.03	0.79 ± 0.09	0.62 ± 0.12	1.42 ± 0.33	0.84 ± 0.08 / 0.74 ± 0.11
	1.5	91.7 ± 26.2	0.01 ± 0.03	0.77 ± 0.09	0.60 ± 0.12	1.48 ± 0.34	0.84 ± 0.08 / 0.74 ± 0.11
	2.0	90.7 ± 27.5	0.01 ± 0.03	0.76 ± 0.08	0.58 ± 0.11	1.56 ± 0.34	0.84 ± 0.08 / 0.74 ± 0.11
	2.5	88.0 ± 30.2	0.01 ± 0.04	0.73 ± 0.08	0.55 ± 0.10	1.66 ± 0.35	0.84 ± 0.08 / 0.74 ± 0.11
3.0	1.0	93.8 ± 22.4	0.01 ± 0.03	0.75 ± 0.08	0.58 ± 0.11	1.62 ± 0.35	0.84 ± 0.08 / 0.74 ± 0.11
	1.5	91.8 ± 26.0	0.01 ± 0.04	0.74 ± 0.08	0.56 ± 0.10	1.67 ± 0.35	0.84 ± 0.08 / 0.74 ± 0.11
	2.0	89.2 ± 29.0	0.01 ± 0.04	0.72 ± 0.08	0.54 ± 0.10	1.73 ± 0.36	0.84 ± 0.08 / 0.73 ± 0.11
	2.5	86.1 ± 32.5	0.01 ± 0.03	0.70 ± 0.07	0.52 ± 0.09	1.83 ± 0.36	0.83 ± 0.07 / 0.73 ± 0.11

*Mean percentage of generated poses per target that pass all PoseBusters `mol`-mode checks.
**The baseline ligand-prior per-coordinate standard deviation is $\sigma_0 = 0.2$ Å; the effective standard deviation equals 0.2 Å times the listed multiplier.

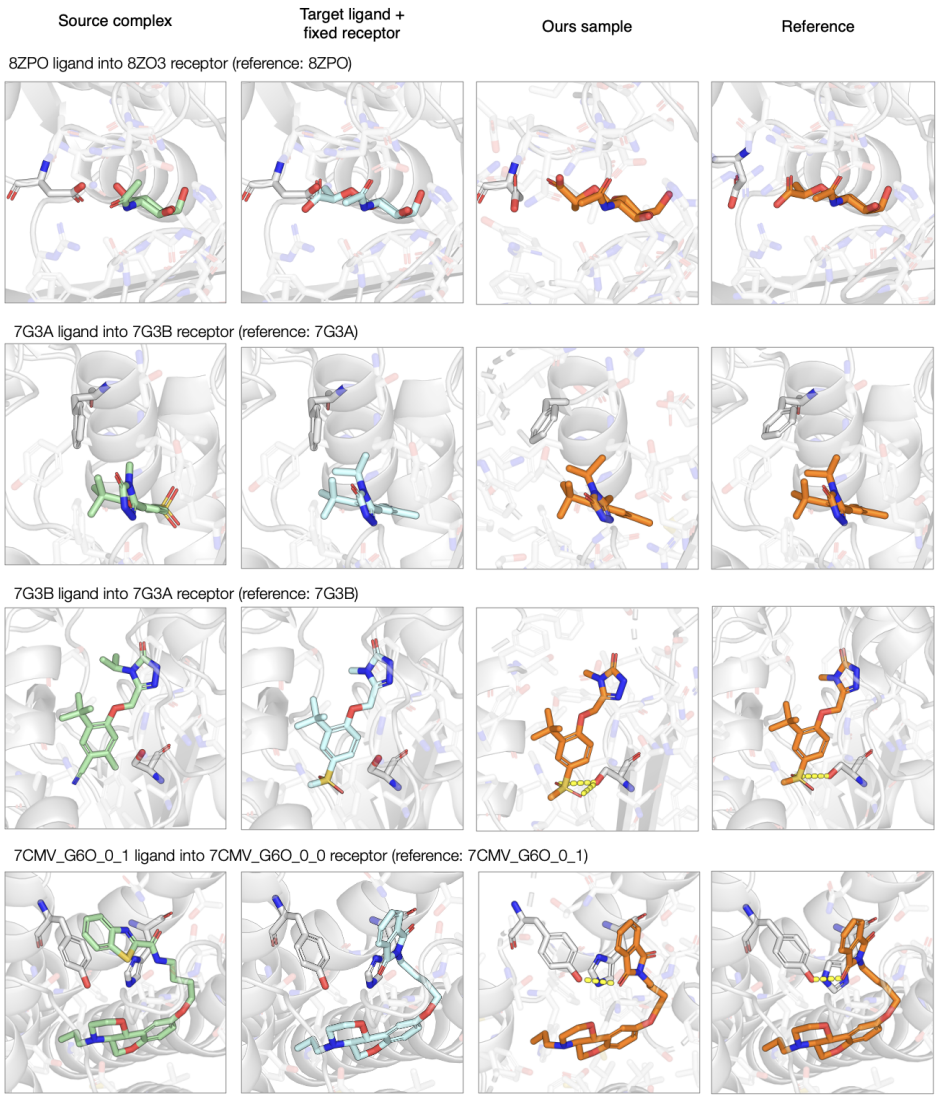

Figure 8: Ligand-induced side-chain adaptation in scaffold-preserved fragment substitution. Each row shows a directed congeneric cross-docking task in which the target ligand is placed into the holo receptor associated with a source ligand. Row labels follow "target ligand into source receptor," with the experimental target complex indicated as the reference. Columns show the source complex, the target ligand in the fixed source receptor, a generated conformer with jointly sampled ligand and side chains, and the experimental target reference. Compared with fixed-receptor placement, joint sampling relieves steric clashes and recovers local interactions, including the hydrogen bonds shown as dashed lines, toward the target pocket arrangement.

F.4 PLACER ROBUSTNESS AND PB-VALIDITY ANALYSIS

F.4.1 PLACER INPUT PERTURBATION CONTROL

PLACER is evaluated both from the benchmark input complex and from a perturbed input complex. The perturbed-input control tests whether PLACER's local ensemble generation is sensitive to small changes in the starting ligand and nearby side-chain coordinates, rather than only reconstructing the provided pose. For this control, pocket side chains within 8 Å of the ligand were initialized with random side-chain χ angles, and ligand rotatable bonds were perturbed with Gaussian torsion noise

Table 12: PLACER input perturbation control. Values use the same metrics as Table 1. Validity* is PoseBusters PB-valid as in Table 1.

Metric	PLACER	PLACER perturb	Δ perturb
Validity* ↑	53.2	52.2	−0.9
Clash ↓	0.0306	0.0339	+0.0033
lDDT-PLI ↑	0.6867	0.6831	-0.0036
IFP sim. ↑	0.5084	0.5092	+0.0008
Diversity ↑	1.8605	1.8885	+0.0280
χ_1 acc. ↑	0.9162	0.9142	-0.0020
χ_2 acc. ↑	0.8470	0.8448	-0.0022

Table 13: PoseBusters failure modes for PLACER outputs. Target-level PB-valid is the mean percentage of PB-valid poses per target. Failure-mode columns report the percentage of PB-invalid poses that failed each PoseBusters `mol` check. A pose may fail multiple checks.

Method	Targets	Poses	PB-valid target mean	All-valid targets	Partial-invalid targets	All-invalid targets	Bond angles	Bond lengths	Internal clash	Internal energy
PLACER	824	26338	53.2	6.6	86.8	6.7	80.9	56.0	47.0	27.5
PLACER perturb	817	26114	52.2	6.6	86.2	7.2	82.4	58.9	47.5	25.2

(standard deviation 20°, clipped to 45°). These settings are intended as a moderate local perturbation, not a global redocking or binding-mode disruption.

The resulting metrics are nearly unchanged from the unperturbed PLACER run. We therefore interpret this control as evidence that the main PLACER result is not an artifact of the particular small local perturbations used in our benchmark, rather than as evidence that PLACER has been stress-tested under large binding-mode errors.

F.4.2 PLACER PB-VALID FAILURE MODES

The low PB-valid rate for PLACER is analyzed separately from the aggregate benchmark table because RDKit sanitization remains high while PoseBusters identifies local three-dimensional geometry failures. We reran PoseBusters `mol`-mode checks on all 26,338 PLACER and 26,114 perturbed-input PLACER poses (824 and 817 targets, respectively) and summarized which checks are responsible for PB-invalid poses.

Across both PLACER settings, only ≈6.6% of targets have all poses PB-valid, while ≈87% are partially invalid and ≈7% are fully PB-invalid at the target level. At the pose level, ≈47% of PLACER poses fail at least one check. Among PB-invalid poses, failures are dominated by local three-dimensional geometry rather than graph-level chemistry, including implausible bond angles (81–82% of invalid poses), bond lengths (56–59%), internal steric clashes (47–48%), and high internal energy (25–27%). RDKit sanitization and InChI-conversion failures account for less than 1% of invalid poses, indicating that PLACER often preserves a valid molecular graph while producing strained or distorted coordinates. This pattern is consistent with targets where our sampler remains PB-valid while PLACER does not (Figure 9).

Supplementary Figure 9 illustrates these failures directly. Representative ligands show the dominant local geometry artifacts flagged by PoseBusters, and the corresponding failure-mode summary confirms that the same bond-angle, bond-length, and internal-clash checks dominate both the default and perturbed-input PLACER settings. Because PoseBusters checks are not mutually exclusive, the plotted percentages report the fraction of PB-invalid poses failing each check and do not sum to 100%.

F.5 INTERPRETATION OF LOG-PROBABILITY

The flow formulation assigns token-wise log-density attributions to the generated structures. For numerical stability, the practical scores in this section use the clipped and rescaled token values from

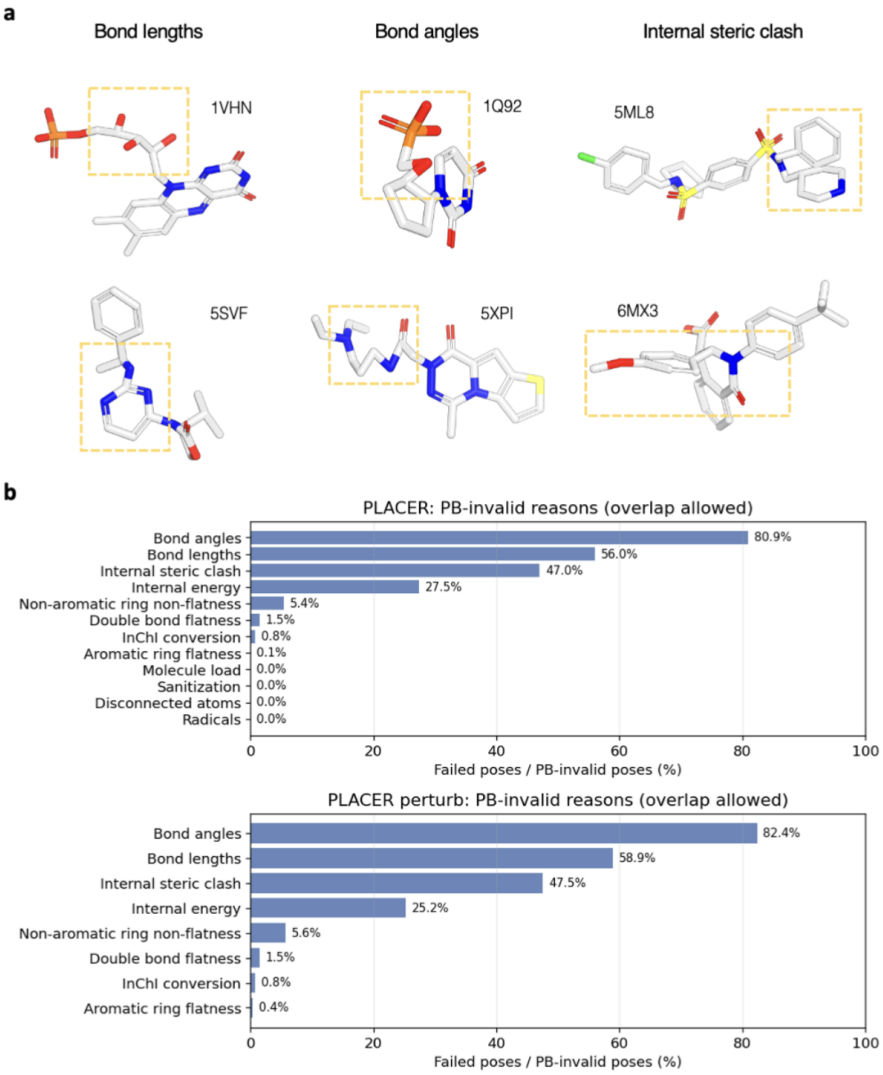

Figure 9: PLACER PB-invalid ligand-geometry failure modes. **a**, Representative PLACER outputs grouped by the dominant PoseBusters `mol`-mode failure, including implausible bond lengths, implausible bond angles, or internal steric clashes. The examples illustrate that the molecular graphs can remain chemically recognizable while the generated three-dimensional coordinates contain severe local distortions. **b**, PoseBusters failure reasons among PB-invalid PLACER poses for the default and perturbed-input settings. Percentages are computed over PB-invalid poses, and checks are overlapping, so one pose can contribute to multiple failure categories.

Eq. (24). We refer to these values below as token-wise plausibility scores rather than calibrated physical energies. Ligand atoms receive ligand-side values, and protein side chains receive residue-level values. These signals can then be aggregated over different structural regions. In the analyses below, we compare total complex, ligand-only, protein-side, and contact-focused score summaries.

For the structure-level conformer-selection analysis in Supplementary Figure 7a, we use `logp_contact` as defined in Section 3.2. This score combines ligand-side and contact-residue protein-side scores using a fixed target-specific contact-residue set. Conformers are ranked within each target by this scalar and used to define the `logp_contact` top-ranked subsets in Supplementary Figure 7a. For the PLINDER percentile analysis in Figure 3b, conformers are instead assigned to within-target score-percentile bins. Structural metrics are centered by their target-level means

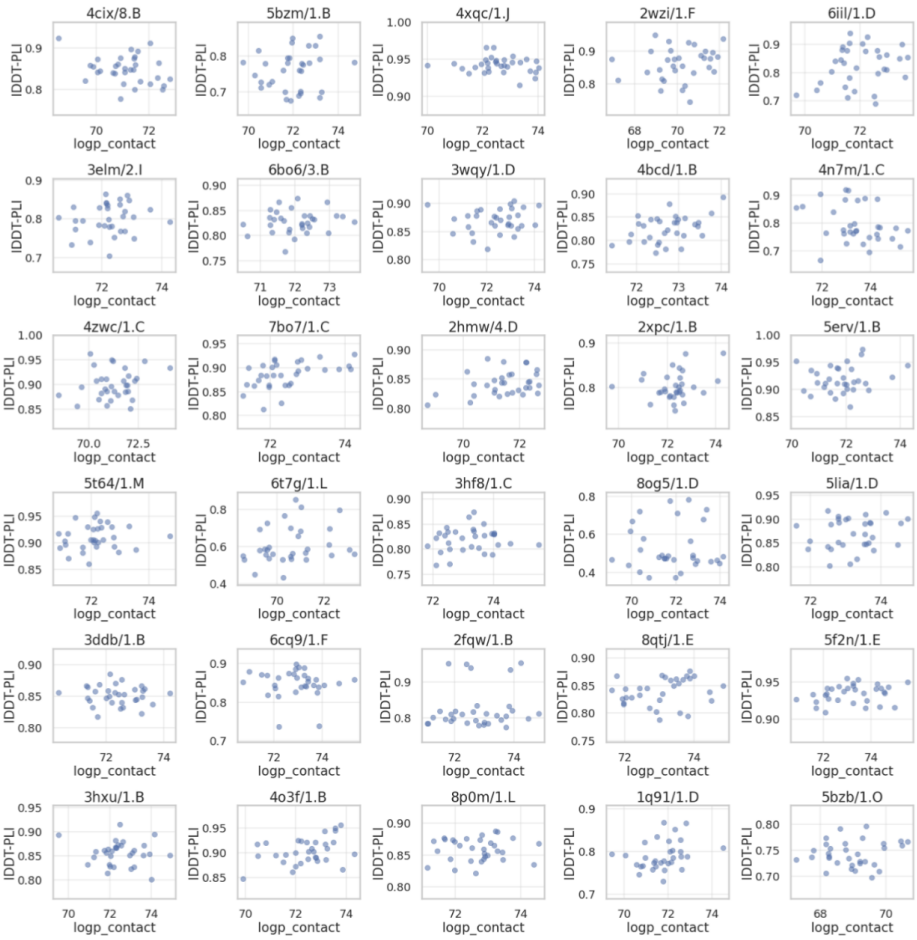

Figure 10: Relationship between contact log-probability and reference interaction quality on PLINDER testset conformer ensembles. Each panel shows one randomly sampled PLINDER testset target. Points are generated conformers. The orange point is the top-ranked conformer selected by `logp_contact`. The y-axis reports lDDT-PLI computed against the experimental complex. Across many targets, higher contact log-probability tends to coincide with higher lDDT-PLI, supporting its use as a reference-free conformer-ranking signal. The relationship is intentionally not perfect. lDDT-PLI measures local protein–ligand distance recovery, whereas the learned log-probability is also affected by ligand geometry, steric strain, side-chain plausibility, and local clashes.

before aggregation, so the plotted changes measure within-target conformer ordering rather than between-target difficulty.

This score is meant to distinguish ligand-intrinsic plausibility from contact-focused protein–ligand interaction plausibility. A ligand conformer can have a high ligand-side score because its internal geometry is reasonable even if it makes weak or strained pocket contacts. The primary structure-level score in the main text is therefore `logp_contact`, not the raw total log-probability.

Figure 11 complements the target-centered analysis in Figure 3b by reporting the absolute metric values in the same within-target score-percentile bins. This view shows the structural quality represented by each bin on its original scale while preserving the same comparison between `logp_contact` and the raw joint log-density. Across the evaluated PLINDER ensembles, the absolute lDDT-PLI, pocket lDDT, and IFP similarity increase toward higher `logp_contact` percentiles, whereas ligand RMSD, clash count, and clash rate decrease. Thus, the centered trends in the main text correspond to direct improvements in the underlying conformer-quality metrics.

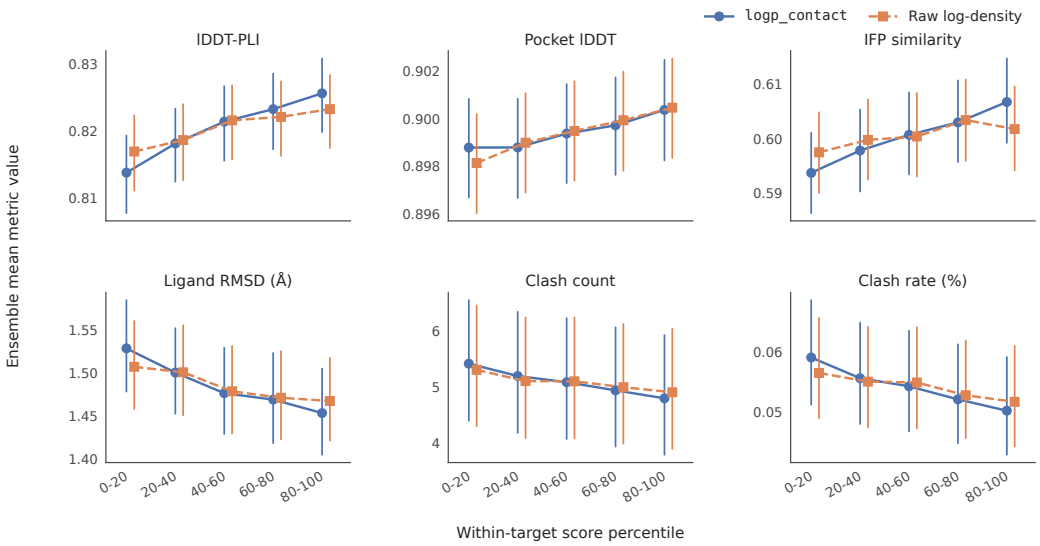

Figure 11: Absolute conformer-quality metrics across within-target score-percentile bins on the PLINDER test set. Reference-based metrics (lDDT-PLI, pocket lDDT, and IFP similarity) are computed against the experimental complex; ligand RMSD, clash count, and clash rate measure conformer geometry and physical validity. Blue denotes the contact-focused score `logp_contact`, and orange denotes the raw joint log-density. Error bars are 95% bootstrap confidence intervals across target-level bin means.

F.5.1 Token-wise log-probability examples by interaction type

Figure 12 gives qualitative PLINDER testset examples of how token-wise log-probability behaves across generated conformers. Each row is one target illustrating a different local interaction type, and within each row we compare a high-scoring and a low-scoring conformer from the same generated ensemble against the experimental reference. High-scoring conformers preserve the reference hydrogen bond or hydrophobic packing contact, whereas low-scoring conformers lose the contact, introduce a steric clash, or adopt a strained ligand geometry. These examples are the qualitative counterpart to the quantitative score–quality relationship in Figures 10 and 11, showing that the score rises when the decomposed tokens correspond to physically reasonable local interactions and falls under clashes or strain.

F.6 Within-group affinity–structure relationships

As an additional analysis, we ask whether generated-ensemble structural quality and sampling diversity are associated with experimental binding affinity within matched local contexts. Structural quality is measured by lDDT-PLI and IFP similarity, with fragment lDDT-PLI included when a substituted fragment is defined, while conformational diversity measures the spread of the sampled ensemble. This analysis does not use the fitted ligand-ranking model of Section 4.4; it directly compares experimental affinity with these structural properties. To remove between-target differences, both affinity and each metric are centered within the corresponding protein–site or congeneric group before computing Spearman correlation. Across the two benchmarks, higher relative affinity generally follows better structural recovery and lower ensemble diversity, although the uncertainty of individual associations varies across metrics and subsets.

On the PLINDER test set, ligands are grouped by protein and binding site, and the reported $\mathrm{p}K_i$ or $\mathrm{p}K_d$ value is compared with lDDT-PLI, IFP similarity, and conformational diversity (Figure 13a). Relative affinity shows positive associations with lDDT-PLI ($\rho = 0.53$) and IFP similarity ($\rho = 0.38$), and a weak negative association with conformational diversity ($\rho = -0.21$). However, this analysis contains only 25 ligands across 12 groups, combines reported $\mathrm{p}K_i$ and $\mathrm{p}K_d$ endpoints, and spans

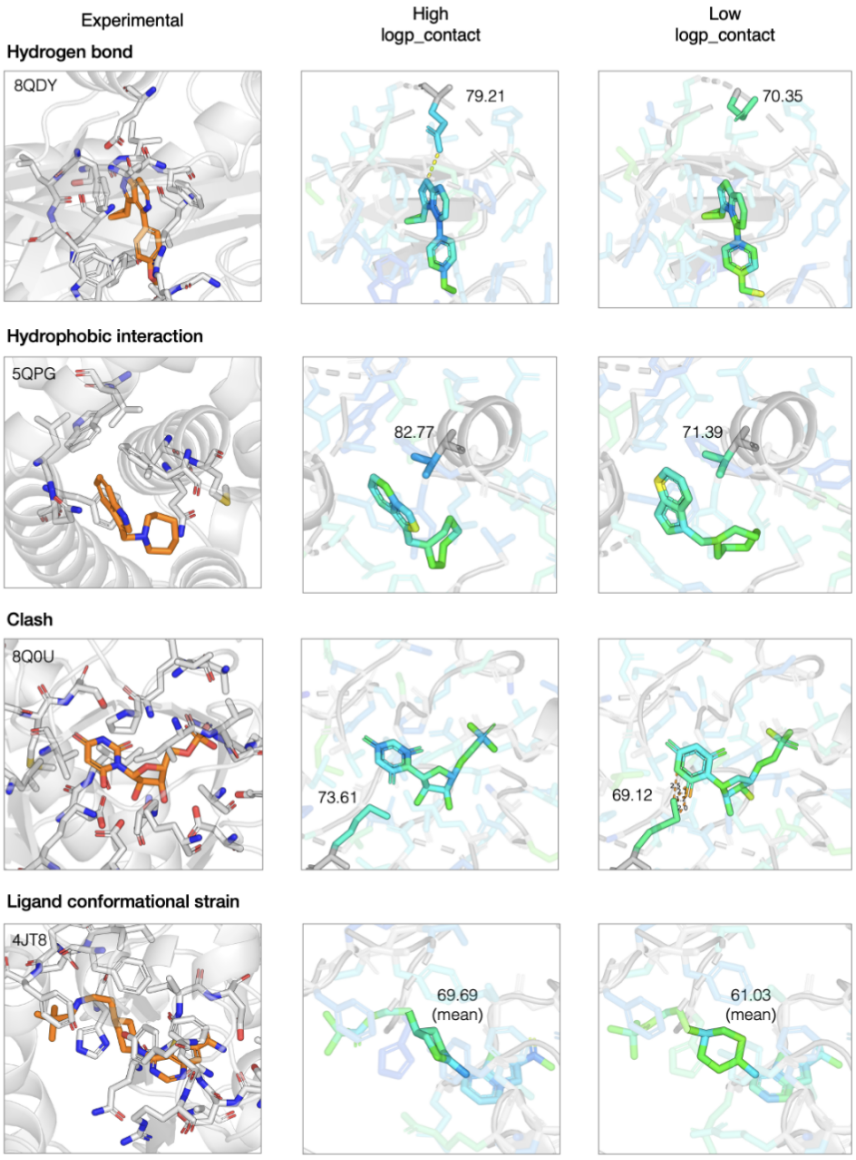

Figure 12: Qualitative examples of token-wise log-probability across generated conformers on PLINDER testset targets. Each row illustrates one local interaction type with a representative target (hydrogen bond, 8QDY, hydrophobic interaction, 5QPG, steric clash, 8Q0U, ligand conformational strain, 4JT8). The left column shows the experimental reference complex, and the middle and right columns show a high-scoring and a low-scoring generated conformer from the same ensemble. Ligand atoms and pocket side chains are colored by their scaled token-wise log-probability, and the annotated value is the displayed conformer's contact score `logp_contact`, where higher indicates a more locally plausible structure. High-scoring conformers reproduce the reference hydrogen bond or hydrophobic packing, whereas low-scoring conformers lose the contact, introduce a steric clash, or place the ligand in a strained conformation.

heterogeneous protein–site contexts that are difficult to control. We therefore repeat the analysis on the larger MotifLeadDB congeneric dataset using only $\mathrm{p}K_d$, which provides a more controlled within-series comparison.

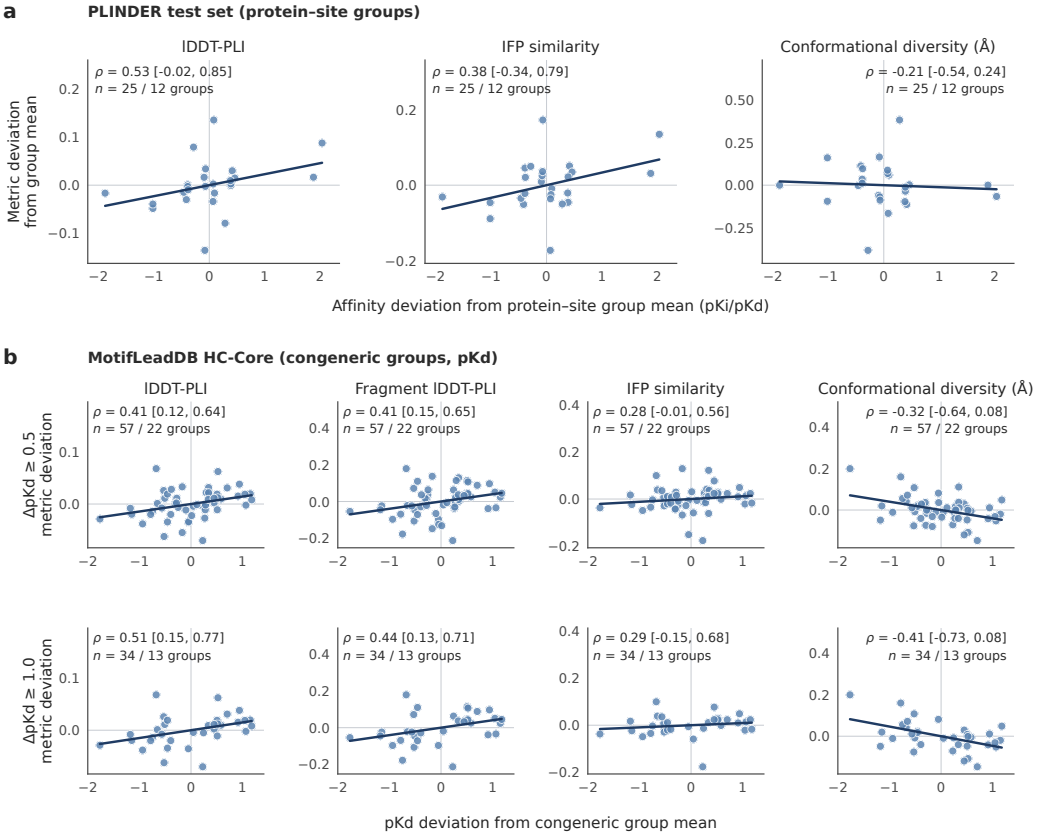

Figure 13: Within-group relationships between experimental affinity and generated-ensemble structural metrics. (a) PLINDER test-set ligands grouped by protein and binding site. The reported $\mathrm{p}K_i$ or $\mathrm{p}K_d$ and each metric are centered within the corresponding protein–site group. (b) MotifLeadDB HC-Core ligands grouped by congeneric series. Rows retain groups with minimum within-group $\mathrm{p}K_d$ ranges of 0.5 and 1.0 log units. Columns report lDDT-PLI, fragment lDDT-PLI where a substituted fragment is defined, IFP similarity, and conformational diversity. Each panel reports Spearman's ρ with a group-bootstrap 95% confidence interval, the number of ligands n, and the number of retained groups. Lines are linear trend guides rather than fitted affinity predictors.

For MotifLeadDB, we analyze $\mathrm{p}K_d$ within congeneric groups and retain groups whose observed $\mathrm{p}K_d$ range is at least 0.5 or 1.0 log units, corresponding to approximately 3.2-fold or 10-fold differences in K_d, respectively (Figure 13b). Higher relative potency is associated with higher lDDT-PLI, fragment lDDT-PLI, and IFP similarity and with lower conformational diversity. Together, the consistent directions across both thresholds link stronger binding within a congeneric series to better recovered local geometry and interactions and to a more focused conformational ensemble.

F.7 Ligand prioritization across the FEP benchmark

Figure 14 shows, for every test series, the per-target relationship between the fitted ranking score and experimental relative binding free energy ($-\Delta G$), with the within-target Spearman ρ and the number of ligands n annotated on each panel. The ranking model is fit only on the training split, so we report the full set of 39 test targets rather than a selected subset, and these targets are disjoint from the series used to fit the score. Across these test targets, the score is positively associated with experimental affinity for most series. The median within-target ρ is 0.38, $29/39$ targets have $\rho > 0$, and the pooled within-target pairwise ranking accuracy is 0.64, consistent with a useful but imperfect prioritization signal. A subset of series shows weak or negative correlation ($10/39$ targets with $\rho < 0$, of which 3 have $\rho \leq -0.3$). This is expected. The score is an exploratory, non-calibrated structural quantity

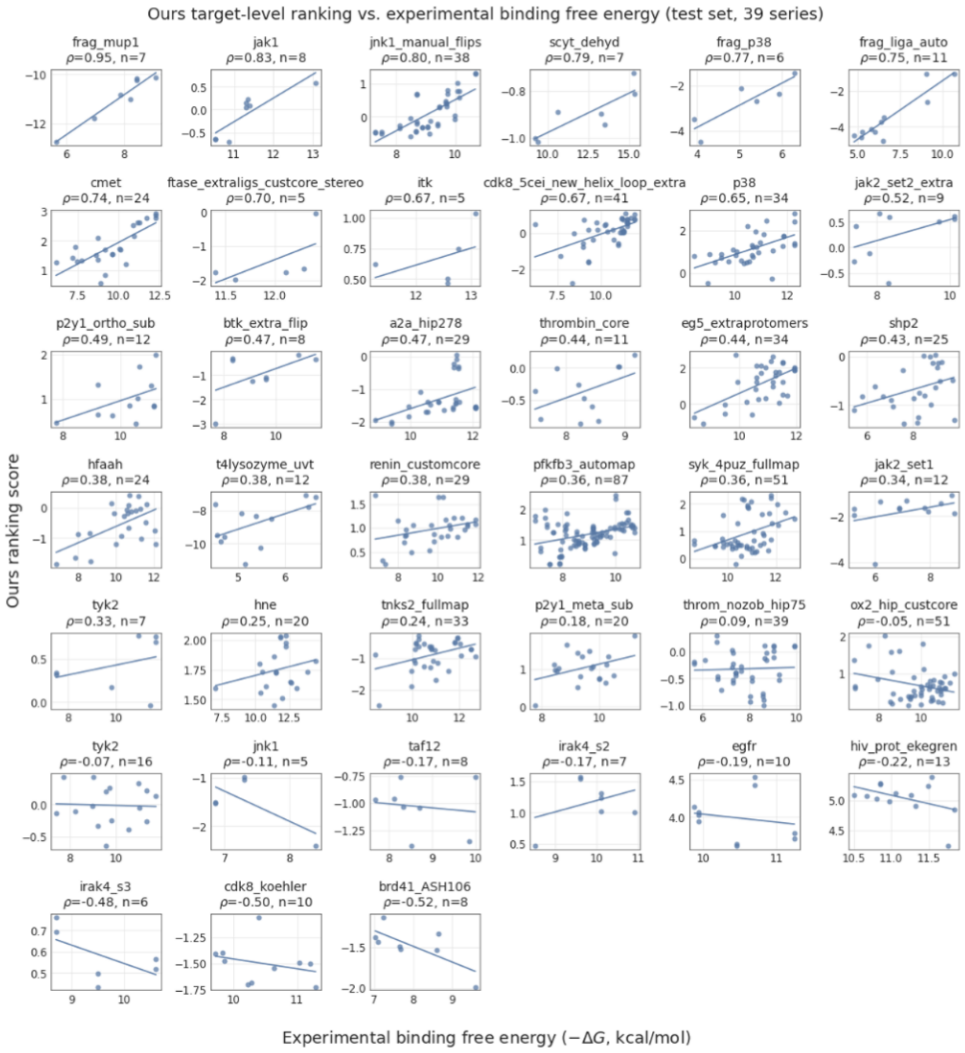

Figure 14: Per-target ranking versus experimental binding free energy across the 39 test FEP congeneric series. These series are disjoint from the training split used to fit the score. Each panel plots the fitted "Ours" ranking score (y) against experimental binding free energy ($-\Delta G$, kcal/mol, larger = stronger binding) for one target series, annotated with the within-target Spearman ρ and the number of ligands n. A positive ρ indicates that the score increases with experimental binding strength. All test series are shown, including weak- and negative-correlation series.

rather than a free-energy estimate, and several of these series involve effects the local model does not capture, such as binding-mode switches, water- or metal-mediated interactions, or large scaffold changes. Per-target ρ is also unstable for small series. Nine of 39 targets have $n \leq 7$ ligands, so individual low- or negative-ρ panels in this regime should not be over-interpreted.