AlloGen: Conformation-Selective Binder Design with Differential State Scoring

Hanqun Cao [1] **Aastha Pal** [2] **Sumi Kimura** [2] **Yesol Kim** [2] **Jingjie Zhang** [1] **Pheng Ann Heng** [1]
Pranam Chatterjee [2 3]

Abstract

Protein binder design has largely optimized for affinity alone, leaving conformational selectivity unaddressed: for allosteric targets such as kinases, nuclear receptors, and GPCRs, a binder that engages both active and inactive states provides no functional specificity regardless of how tightly it binds. We introduce **AlloGen**, a modular framework that decouples backbone generation from a learned state-selectivity scorer Q_θ, an SE(3)-invariant interface graph transformer trained via a two-phase curriculum that first learns interface geometry before imposing conformational discrimination. Because Q_θ is fully differentiable and generator-agnostic, it integrates with any backbone generator as a passive reranker or an active gradient-based guide without retraining. Across a diverse benchmark of proteins spanning multiple families and conformational mechanisms, AlloGen consistently identifies binders that preferentially recognize desired structural states while rejecting alternative conformations. Experimental validation on calmodulin further demonstrates that these computational selectivity signals translate to physical molecules, yielding *de novo* peptides that bind the desired holo conformation while exhibiting no detectable binding to the apo state. Together, these results establish conformational selectivity as a learnable property and provide a general framework for state-selective protein binder design.

1 Introduction

Proteins are molecular switches: their conformational transitions between distinct structural states govern signaling, catalysis, and regulation across virtually every protein family (Vetter & Leclerc, 2003; Ha & Loh, 2012; Weikl & Paul, 2014; Nussinov, 2016). For therapeutically important targets such as kinases, nuclear receptors, and GPCRs, different conformational states correspond to distinct biological functions, and the design goal is therefore not merely binding affinity but conformational selectivity: stabilizing one functional state while actively disfavoring others (Kar et al., 2010; Vijayan et al., 2015; Kojetin & Burris, 2013; Conflitti et al., 2025). This requirement is central to allosteric drug design, conformational biosensors, and synthetic biology switches (Kar et al., 2010; Langan et al., 2019).

Recent generative models have transformed protein binder design, yet they share a common blind spot. At the sequence level, masked language modeling achieves state-of-the-art peptide binder design conditioned on target sequence (Chen et al., 2025a), contrastive language models enable de novo peptide design to conformationally diverse targets (Bhat et al., 2025), and multi-objective discrete diffusion has been applied to therapeutic peptide generation (Tang et al., 2025a; Vincoff et al., 2025; Cao et al., 2025; 2026). At the structural level, RFdiffusion establishes de novo design of functional protein binders at scale (Watson et al., 2023), PXDesign delivers fast and modular binder design with strong experimental success rates (Protenix Team et al., 2025), Proteina-ComplexA scales atomistic binder design through generative pretraining and test-time compute (Didi et al., 2026), BoltzGen pursues universal binder design across protein families (Stark et al., 2025), and BindCraft achieves one-shot design of experimentally validated functional binders (Pacesa et al., 2025; Song et al., 2025). Despite this diversity, all existing methods share a fundamental limitation: they condition on a single receptor conformation and optimize for fit to that structure alone. A binder designed for one conformational state may bind equally well to an alternative state, defeating the purpose of state-selective targeting. Conventional scoring functions measure binding affinity but not differential affinity across states, and thus provide no signal for conformational selectivity.

To close this gap, we introduce **AlloGen**, a modular framework that decouples binder generation from selectivity evaluation. The central insight underlying AlloGen is that conformational selectivity is a learnable and transferable property

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong [2]Department of Bioengineering, University of Pennsylvania [3]Department of Computer and Information Science, University of Pennsylvania. Correspondence to: Pranam Chatterjee <pranam@seas.upenn.edu>.

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

of receptor–binder interfaces: once distilled into a differentiable scorer, it can be applied post-hoc as a reranker or injected as an active guide into any generative model using standard classifier-guidance techniques (Dhariwal & Nichol, 2021; Gruver et al., 2023). This decoupling transforms conformational selectivity from a target-specific engineering challenge into a general-purpose signal that scales across protein families, generator architectures, and guidance strategies. The core component is Q_θ, a lightweight SE(3)-invariant interface graph transformer that scores how strongly a binder candidate prefers one conformational state over another. Because Q_θ is fully differentiable and generator-agnostic, it integrates with any backbone generator as either a passive reranker or an active gradient-based guide, enabling selectivity-guided generation through strategies ranging from best-of-K reranking to Langevin backbone refinement and sequential Monte Carlo sampling.

Our contributions are as follows.

1. **Problem formulation, benchmark, and scorer.** We formalize two-state conformational selectivity as a learnable signal and operationalize it on three fronts: **(1)** a benchmark of 65 two-state targets across 15 protein families and 2,896 receptor–binder complexes covering both apo and holo conformations; **(2)** Q_θ, an SE(3)-invariant interface graph transformer that scores backbone-only candidates against paired receptor conformations; and **(3)** a two-phase training curriculum that grounds the scorer in interface geometry before imposing conformational discrimination. On 8 held-out OOD targets, Q_θ achieves $\bar{\rho} = 0.520$ while all energy-based baselines fail uniformly.

2. **Controls against circularity.** Two structural controls in the main text (cross-target specificity and per-target apo rejection, Section 4.3) and four Q_θ-independent metrics in the appendix (ProteinMPNN log-likelihood, Boltz-2 interface predicted TM-score (ipTM), AlphaFold 3 ipTM, and Rosetta InterfaceAnalyzer) probe what Q_θ has learned and confirm that selectivity originates from conformational geometry rather than scoring artifacts.

3. **End-to-end selectivity-guided design.** All 15 generator–guidance combinations achieve positive mean selectivity averaged over the 8 OOD test targets, with TDS and SMC emerging as the strongest universal guidance strategies. We further demonstrate the full AlloGen pipeline through a CaM case study, where RFdiffusion combined with Langevin refinement reaches $\bar{S}_{\text{cons}} = +0.677$ at 88% design success rate, and show that passive best-of-K reranking and active gradient guidance are complementary: reranking scales efficiently with pool size and reaches $\bar{S} = +0.885$ at $K = 10$, while Langevin refinement delivers consistent per-design improves on small pools.

2 Preliminaries

2.1 Problem Formulation

We consider a target protein that exists in two distinct conformational states: an *apo* (undesired) conformation X^0 and a *holo* (goal) conformation X^1, both represented as backbone coordinate sets. Given a binder Y represented by its backbone coordinates, the *state-selectivity scoring problem* is to learn a function $Q_\theta : \mathcal{X} \times \mathcal{Y} \to (0, 1)$ such that for an X^1-selective candidate Y:

$$Q_\theta(X^1, Y) \gg Q_\theta(X^0, Y), \tag{1}$$

while non-selective binders, apo-preferring binders, and non-binders are not required to satisfy this inequality. The training objective in Section 3.3 enforces this conditional behaviour through paired contrastive supervision rather than a global bias toward X^1. Given a pretrained binder generator $p_\psi(Y \mid X^1)$, the *two-state binder design* task is to identify the most conformationally selective candidate $\hat{Y} = \arg\max_{Y^{(k)}} S_\theta(Y^{(k)}; X^1, \mathcal{N})$ over K samples $Y^{(k)} \sim p_\psi(\cdot \mid X^1)$, where S_θ is the selectivity margin defined in Section 3.4 and $\mathcal{N} = \{X^0\}$ is the set of undesired conformations.

2.2 Protein Backbone Representation and SE(3) Invariance

We represent each residue i by its C_α position $\mathbf{p}_i \in \mathbb{R}^3$ and a local backbone frame $R_i \in \mathrm{SO}(3)$ constructed from the $(\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C})$ triplet via Gram–Schmidt orthogonalization (Jumper et al., 2021). While $\mathbf{p}_i$ and R_i themselves transform with global rigid motions, the pair $(\mathbf{p}_i, R_i)$ defines a residue-local frame from which all inter-residue geometry can be expressed in an SE(3)-invariant way. For Q_θ to be physically meaningful, it must be SE(3)-invariant: $Q_\theta(gX, gY) = Q_\theta(X, Y)$ for all rigid motions $g \in \mathrm{SE}(3)$, ensuring scores are independent of the global position and orientation of the complex. We enforce this by expressing all inter-residue geometry in the local frame of residue i: distances $\|\mathbf{p}_j - \mathbf{p}_i\|$, directions $R_i^\top(\mathbf{p}_j - \mathbf{p}_i)/\|\cdot\|$, and relative orientations $R_i^\top R_j$. Each of these quantities is invariant under rigid motions g applied jointly to both X and Y, so the resulting node and edge features—and hence Q_θ itself—are SE(3)-invariant.

2.3 DockQ as Interface Quality Proxy

DockQ (Basu & Wallner, 2016) is a composite scalar $\in [0, 1]$ that measures protein–protein docking quality by combining the fraction of native contacts, interface RMSD, and ligand RMSD. We adopt DockQ as the supervision

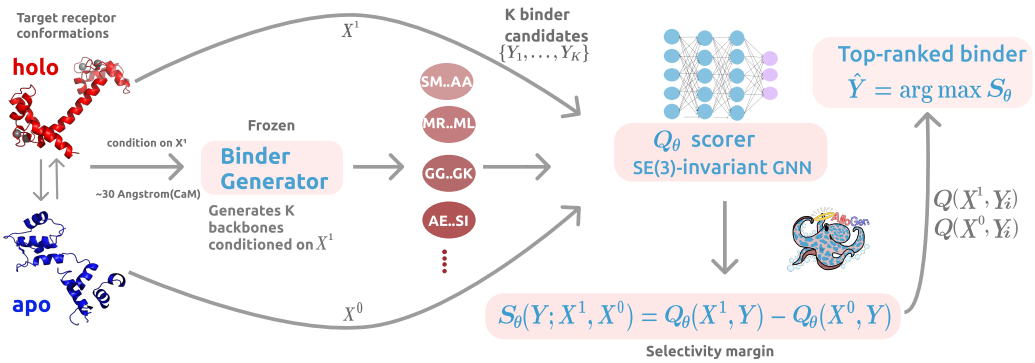

Figure 1. **AlloGen pipeline.** A frozen generator produces K binder backbones conditioned on the goal state X^1 (holo, blue); the trained scorer Q_θ evaluates each candidate against both X^1 and the undesired state X^0 (apo, red), and returns the top candidate $\hat{Y}$ by selectivity margin $\Delta Q = Q_\theta(X^1, Y) - Q_\theta(X^0, Y)$. Q_θ is trained independently and plugs into any backbone generator without retraining.

signal for Phase 1 training (Section 3.3), as it provides a geometrically grounded proxy for receptor–binder interface quality. Grounding Q_θ in DockQ before introducing any selectivity signal prevents degenerate solutions in Phase 2, where the model must distinguish between conformational states rather than simply predict binding quality. A DockQ value > 0.23 is the conventional threshold for an acceptable docking model (Basu & Wallner, 2016).

3 Method

3.1 Interface Graph Construction

To score receptor–binder complementarity in a way that is sensitive to local interface geometry rather than global protein shape, we represent each complex as a sparse graph over interface-proximal residues. Concretely, a receptor–binder complex (X, Y) is represented as an interface graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains residues from both X and Y with at least one inter-chain C_α contact within the cutoff as 8Å, and edges connect all residue pairs within that cutoff.

Node features. Each node $\mathbf{h}_i$ encodes four types of information. *Amino acid identity* is represented as a one-hot vector over 20 standard amino acids plus an unknown token, providing residue-level sequence identity. *Backbone torsion angles* φ, ψ, and ω are encoded via their sine and cosine values, yielding a smooth periodic representation of local backbone conformation. *Sidechain torsion angles* χ_1 and χ_2 are similarly encoded to capture rotameric states at the interface; residues without sidechain degrees of freedom are zero-padded. A *chain indicator* flag distinguishes receptor from binder residues. When available, per-residue ESM-2 embeddings (Lin et al., 2023) are projected and concatenated to these structural features, providing evolutionary context beyond local geometry.

Edge features. All edge features $\mathbf{e}_{ij}$ are computed in the local backbone frame of residue i, ensuring SE(3) invariance. *Geometric features* encode the inter-residue distance via a Gaussian RBF basis, the unit direction $R_i^\top(\mathbf{p}_j - \mathbf{p}_i)/\|\mathbf{p}_j - \mathbf{p}_i\|$, and the relative backbone orientation $R_i^\top R_j$, jointly capturing the full relative rigid-body relationship between residues. *Sequence separation* is encoded as a binned index difference within a chain and set to the maximum bin for inter-chain pairs, allowing the model to distinguish intra- from inter-chain interactions. A *same-chain indicator* flag further disambiguates receptor-receptor, binder-binder, and receptor-binder edges.

3.2 State-Selectivity Scorer Q_θ

Scoring conformational selectivity requires a model that assesses how well a binder backbone fits one receptor conformation relative to another in an SE(3)-invariant manner. To this end, Q_θ is implemented as a dense edge-biased graph transformer (Vaswani et al., 2017) that operates on SE(3)-invariant geometric features. Given node embeddings $\mathbf{H}^{(0)}$ and edge embeddings $\mathbf{E}^{(0)}$, the model applies $L = 4$ transformer layers. At each layer ℓ, attention weights are computed as:

$$\alpha_{ij}^{(\ell)} = \frac{(\mathbf{h}_i^{(\ell)} W_Q)(\mathbf{h}_j^{(\ell)} W_K)^\top}{\sqrt{d_h}} + b_{ij}, \quad b_{ij} = \mathbf{e}_{ij} W_E \in \mathbb{R}, \tag{2}$$

$$\mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{(\ell)} + \mathrm{FFN}\left(\sum_j \mathrm{softmax}_j(\alpha_{ij}^{(\ell)}) \cdot \mathbf{h}_j^{(\ell)} W_V \right), \tag{3}$$

where $W_E \in \mathbb{R}^{d_e \times 1}$ projects each edge embedding to a per-head scalar attention bias b_{ij} that is added directly to the dot-product logit (one such projection is learned per attention head). Edge embeddings are computed once

and shared across all layers. After L layers, mean- and max-pooled node representations are concatenated to form $\mathbf{h}_{\text{pool}}^{(L)} = [\text{mean}_i\, \mathbf{h}_i^{(L)};\, \max_i \mathbf{h}_i^{(L)}]$ and passed through an MLP with sigmoid activation to produce:

$$Q_\theta(X, Y) = \sigma(\text{MLP}(\mathbf{h}_{\text{pool}}^{(L)})) \in (0, 1). \tag{4}$$

The bounded output yields a natural selectivity gap $Q_\theta(X^1, Y) - Q_\theta(X^0, Y) \in (-1, 1)$. Architectural hyperparameters and parameter count are reported in Appendix B.

3.3 Two-Phase Training of Q_θ

Directly optimizing Q_θ for conformational selectivity risks a degenerate solution where the model ignores receptor conformation entirely. We prevent this through a two-phase curriculum.

Phase 1: Interface Quality Regression. We first train Q_θ to predict interface quality as measured by DockQ (Basu & Wallner, 2016):

$$\mathcal{L}_q = \text{MSE}(Q_\theta(X, Y),\ d_{\text{DockQ}}(X, Y)). \tag{5}$$

Training data includes native holo complexes, apo mismatches, rigid-body decoys, and hard negatives described in Section 4.1. This phase establishes a geometrically grounded representation of receptor-binder fit, providing a stable initialization for Phase 2.

Phase 2: Selectivity Fine-Tuning. The central risk in fine-tuning for selectivity is that Q_θ collapses to target-specific biases, assigning high scores to any binder paired with a particular receptor regardless of conformation. We prevent this by fine-tuning on paired triplets (X^+, X^-, Y) where $X^+ = X^1$ and $X^- = X^0$ for the same binder Y, using a multi-negative InfoNCE loss (Oord et al., 2018) that includes both apo negatives (the same binder against its target's apo conformation) and cross-target negatives (the anchor's holo conformation paired with binders from other targets in the batch). The cross-target term forces the model to discriminate between the binder's true holo partner and structurally unrelated holo receptors, preventing collapse to a fixed-receptor bias.

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\big(Q_\theta(X_i^+, Y_i)/\tau\big)}{Z_i}, \tag{6}$$

$$Z_i = \exp\big(Q_\theta(X_i^+, Y_i)/\tau\big) + \exp\big(Q_\theta(X_i^-, Y_i)/\tau\big) + \sum_{\substack{k=1 \\ k\neq i}}^{B} \exp\big(Q_\theta(X_i^+, Y_k)/\tau\big).$$

where τ is a temperature hyperparameter reported in Appendix B. The regression loss $\mathcal{L}_q$ is dropped in Phase 2 to avoid conflicting gradient signals; this design choice is validated in Table 2.

Backbone-geometry augmentation. A third training design choice addresses the distribution shift between training and inference: at inference, Q_θ scores backbone-only binder designs without sequence identity or sidechain coordinates, whereas all training complexes have known sequences. To align the training distribution with this inference setting, we mask binder-side sequence features with probability p_{drop}:

$$\tilde{\mathbf{h}}_i^{\text{bnd}} = \big[\, \mathbf{0}_{\text{AA}},\ \sin\varphi_i,\ \cos\varphi_i,\ \sin\psi_i,\ \cos\psi_i,\ \sin\omega_i,\ \cos\omega_i,\ \mathbf{0}_\chi,\ \mathbf{0}_{\text{ESM}},\ 1 \,\big]. \tag{7}$$

where $\mathbf{0}_{\text{AA}}$, $\mathbf{0}_\chi$, and $\mathbf{0}_{\text{ESM}}$ replace the amino acid identity, sidechain torsion angles, and ESM-2 embeddings with zeros respectively. All backbone torsions and edge features are preserved, so Q_θ learns to rely on backbone geometry rather than sequence identity. The same masking decision is applied consistently to both X^+ and X^- within each training pair.

3.4 Selectivity-Guided Binder Design

Given K binder candidates $Y^{(1)}, \ldots, Y^{(K)}$ sampled from the frozen generator $p_\psi(\cdot \mid X^1)$, each candidate is evaluated against both X^1 and X^0 by rigid placement onto their aligned structures. The *selectivity margin* is defined as:

$$S_\theta(Y;\, X^1, \mathcal{N}) = \text{logit}\big(Q_\theta(X^1, Y)\big) - \log \sum_{X^- \in \mathcal{N}} \exp\big(\text{logit}(Q_\theta(X^-, Y))\big). \tag{8}$$

where $\mathcal{N} = \{X^0\}$ is the set of undesired conformations. S_θ is high when Q_θ strongly prefers X^1 over X^0, and negative when the binder scores similarly across states. The logit-space formulation extends naturally to multi-state targets with $|\mathcal{N}| > 1$. Candidates are filtered by minimum interface size and steric clash criteria before selection, and sequence design is performed with ProteinMPNN (Dauparas et al., 2022) on the selected backbone.

Gradient-based guidance. Because S_θ is differentiable with respect to backbone coordinates, it provides a gradient signal $\nabla_{\mathbf{x}} S_\theta$ that supports four guidance strategies. *Langevin backbone refinement* (Song et al., 2021; Welling & Teh, 2011) performs deterministic gradient ascent on S_θ for a completed backbone: $\mathbf{x}_{t+1} = \mathbf{x}_t + \eta\, \nabla_{\mathbf{x}} S_\theta(\mathbf{x}_t;\, X^1, \mathcal{N})$, where η is the step size; because Langevin operates on fully denoised backbones, gradients remain reliable throughout refinement. *Classifier guidance* (Dhariwal & Nichol, 2021) injects $\nabla_{\mathbf{x}} S_\theta$ into each denoising step to steer the diffusion trajectory toward high-selectivity regions, though gradient reliability degrades at high noise levels. *Twisted diffusion sampling* (TDS) (Wu et al., 2023) reweights diffusion particles by $\exp(S_\theta)$ at each timestep without modifying the

denoising trajectory, preserving the generative prior more faithfully than classifier guidance. *Sequential Monte Carlo* (SMC) (Wu et al., 2023) iteratively resamples complete trajectories by S_θ across multiple generation rounds, progressively enriching the candidate pool and providing the most robust selectivity boost across diverse generator architectures. Detailed formulations and computational cost are provided in Appendix B.3 and F.

3.5 From Selectivity Scores to Synthesizable Candidates

To bridge in-silico selectivity to physical assays, we instantiate AlloGen as a candidate-triage funnel: the generator-guidance machinery of Section 3.4 feeds a cascade of deliberately Q_θ*-independent* structure-prediction and geometric filters, so the tested panel is not selected on the signal it validates.

Dual-state hotspot conditioning. We extract interface-proximal residues from both receptor states and supply their *union* to every generator, so backbones engage the state-invariant functional interface while apo/holo discrimination is left to Q_θ.

Generation & scoring. We tested generator×guidance combinations at $N{=}50$ designs each, rank them by mean selectivity margin, and promote the strongest to large-scale generation. Each backbone is scored by the three-seed Augmented-S2 Q_θ ensemble for the margin $S = Q_\theta(X^1, Y) - Q_\theta(X^0, Y)$, with binder coordinates Kabsch-aligned to the holo and apo frames before scoring.

Q_θ-independent filtering. Surviving backbones are sequence-designed with ProteinMPNN (Dauparas et al., 2022) (receptor fixed, four sequences per backbone at temperature 0.1) and re-folded in both states with Boltz-2. We retain designs with buried surface area (freesasa) above $800\,\text{Å}^2$ and holo-state ipTM above 0.7, then cluster at 70% identity with CD-HIT to remove near-duplicates.

Panel composition. From the clustered pool we assemble a 20-member panel: **8 experimental candidates** (highest-margin guided designs, one per cluster), **8 baseline controls** (unguided, $S \approx 0$), and **4 negative controls** (random unguided designs). These controls furnish a built-in specificity test: a faithful signal should yield a holo/apo gap for the experimental arm but not the controls.

3.6 Experimental Validation Protocol

The top-ranked AlloGen peptides were synthesized and assayed against both holo and apo calmodulin (CaM) by Adaptyv Bio using bio-layer interferometry (BLI) on a Gator Bio Pro instrument. Recombinant human CaM (MilliporeSigma, 208670) was prepared either in the holo state with $CaCl_2$ or the apo state with EGTA to chelate free calcium; aside from this calcium-dependent modulation, conditions were identical across measurements. The Ca^{2+}-dependent CaM-binding M13 peptide (CPC Scientific, SIGN-010) served as a positive control.

Designed peptides were immobilized on Twin-Strep-compatible biosensors via a Twin-Strep tag. After ligand loading and baseline equilibration, biosensors were transferred into serial dilutions of CaM (0, 100, and 1000 nM) to record association, then into buffer to monitor dissociation. Sensorgrams were reference-subtracted using control sensors and buffer-only wells, and k_{on}, k_{off}, and K_D were obtained by fitting to a 1:1 model in duplicate. Designs with concentration-dependent responses were classified as binders with fitted K_D; those without measurable response were reported as no binding detected (NB).

4 Experiments

We evaluate Q_θ as a selectivity scorer in Sections 4.2–4.3 and as an end-to-end design oracle in Sections 4.4; dataset construction and metrics are described in Section 4.1.

4.1 Experimental Setup

Target selection. We curate 65 two-state proteins spanning 15 protein families and diverse conformational mechanisms (Table 4). Selection criteria are: (1) experimentally determined structures for both apo and holo conformational states, (2) at least three co-crystal structures with peptide or protein binders in the goal state, and (3) a structurally defined conformational change between states. Targets range from large-scale domain rearrangements (**CaM**: $\sim$30 Å apo-to-holo; **ABL1**: DFG-loop flip, $\sim$6.5 Å global) to subtle helix repositioning (**ER**α: H12 $\sim$10 Å; **CDK2**: Cyclin A-induced). The dataset spans kinases (9), small GTPases (6), nuclear receptors (5), GPCRs and ion channels (6), proteases (6), and ten additional families, totalling 2,896 complexes across 65 targets.

Sample construction. Each complex yields 12 base training samples: 1 positive native holo complex (X^1, Y^{native}) with label 1.0; 1 negative apo-mismatch complex $(X^0_{\text{aligned}}, Y^{\text{native}})$ with label 0.0, where the apo receptor is Kabsch-aligned to the holo frame; and 10 rigid-body decoys at target C_α RMSD levels spanning 1–8 Å, with labels $\max(0, 1 - d_{\text{RMSD}}/4)$. Prior to augmentation, **FastRelax Neg.** are generated by constrained relaxation of binders placed on apo receptors, producing 959 hard negatives included in all configurations. Three augmentation types further enrich training: **Cross-family negatives** (Cross-fam) place native binders on structurally unrelated receptors; **conformational decoys** (Conf. decoys) apply Rosetta FastRelax

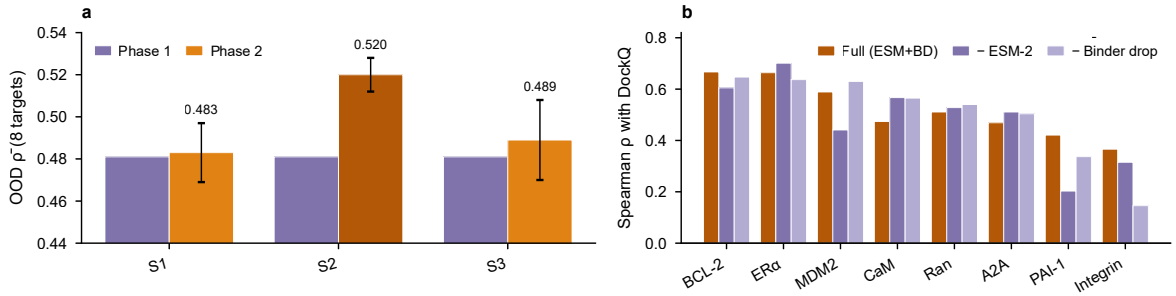

Figure 2. Q_θ **selectivity performance. (a)** Q_θ scoring performance ablation by data augmentation strategies and two phases; **(b)** Q_θ scoring performance ablation per target by different features.

repacking on the holo receptor to produce near-native hard negatives; and **generator decoys** (GenDecoys) are synthetic binders from structure-based generative models that provide diverse non-native interface geometries.

Dataset splits and configurations. All main results use the **target split**, in which targets are partitioned 51/6/8 across train, OOD validation, and OOD test, with CaM held out as the primary OOD design target (Table 4). The **Baseline** dataset comprises 51 training, 6 OOD validation (14-3-3, 14-3-3σ, β_2AR, Caspase-3, μ-opioid, Rac1), and 8 OOD test targets (CaM, BCL-2, ERα, MDM2, Ran, A_{2A}, PAI-1, Integrin); the **Augmented** dataset extends it with GenDecoys and cross-family negatives generated exclusively on training-set targets, precluding leakage into design evaluation. To isolate augmentation sources, we define three scoring configurations (Table 5): **S1** uses Baseline only, **S2** uses full Augmented data, and **S3** adds Cross-fam and Conf. decoys, withholding GenDecoys.

Metrics. We report Spearman ρ as rank correlation with DockQ, the selectivity gap $\bar{Q}^+ - \bar{Q}^-$ as mean holo vs. apo score, and best-of-K success rates. For design evaluation, we use ProteinMPNN ΔNLL and AlphaFold 3 ΔipTM as Q_θ-independent metrics, and report consensus selectivity $\bar{S}_{\text{cons}}$ as the mean over three independently trained Q_θ checkpoints. Architecture and training hyperparameters are detailed in Appendix B.

4.2 Scorer Performance

We evaluate Q_θ on the 8 OOD test targets. Q_θ attains $\bar{\rho}$=$\mathbf{0.520 \pm 0.010}$ (3 random seeds for training), with positive correlation on all 8 OOD targets and 4 of 8 exceeding ρ=0.5 (Figure 2).

Training curriculum, data augmentation, and feature design. Three design choices each contribute measurable gains (Figure 2). *(i) Two-phase curriculum.* Phase 2 contrastive fine-tuning consistently improves over Phase 1 regression across all augmentation configurations (Figure 2a), confirming that paired InfoNCE aligns the learned representation with conformational discrimination more effectively than interface-quality regression. *(ii) Data augmentation.* The full Augmented configuration yields the largest single gain ($\Delta\bar{\rho}$=+0.037), driven by GenDecoys: synthetic binder geometries cover a broader region of interface configuration space than rigid-body and FastRelax decoys, exposing Q_θ to harder negatives. *(iii) Feature design.* Per-target ablations (Figure 2b) show ESM-2 embeddings and binder-side dropout are both beneficial: removing ESM-2 reduces $\bar{\rho}$ across most targets, and disabling binder dropout further degrades performance, with the largest drops on harder OOD targets (Integrin, PAI-1, A_{2A}). The combined Full configuration (ESM + BD) is best on all 8 OOD targets. InfoNCE batch-size ablations are in Appendix D, Table 7.

4.3 Conformational Preference of Q_θ

Beyond rank correlation with DockQ, we next ask whether Q_θ's signal is *specifically conformational*, i.e., target-specific holo preference rather than generic binding quality.

Cross-target specificity. To verify that Q_θ captures target-specific preference rather than generic binding quality, we score each target's 50 vanilla binders against all 8 OOD receptors (Figure 3a). The resulting matrix reveals strong target specificity, with diagonal entries substantially exceeding off-diagonal entries and yielding a specificity ratio of 19.8×, confirming that Q_θ discriminates conformations rather than rewarding generic structural complementarity.

Per-target apo rejection. Figure 3b reports holo and apo Q_θ scores for 50 vanilla designs per target, characterizing apo rejection at the population level. Seven of 8 OOD targets show positive ΔQ with holo preferred; BCL-2 reaches an S>0=100% under single-seed scoring, while CaM and MDM2 reach 100% and 98% under the 3-seed ensemble. Integrin remains the most challenging with ΔQ=+0.001 and S>0=52%, consistent with its lowest Spearman ρ across all scoring configurations.

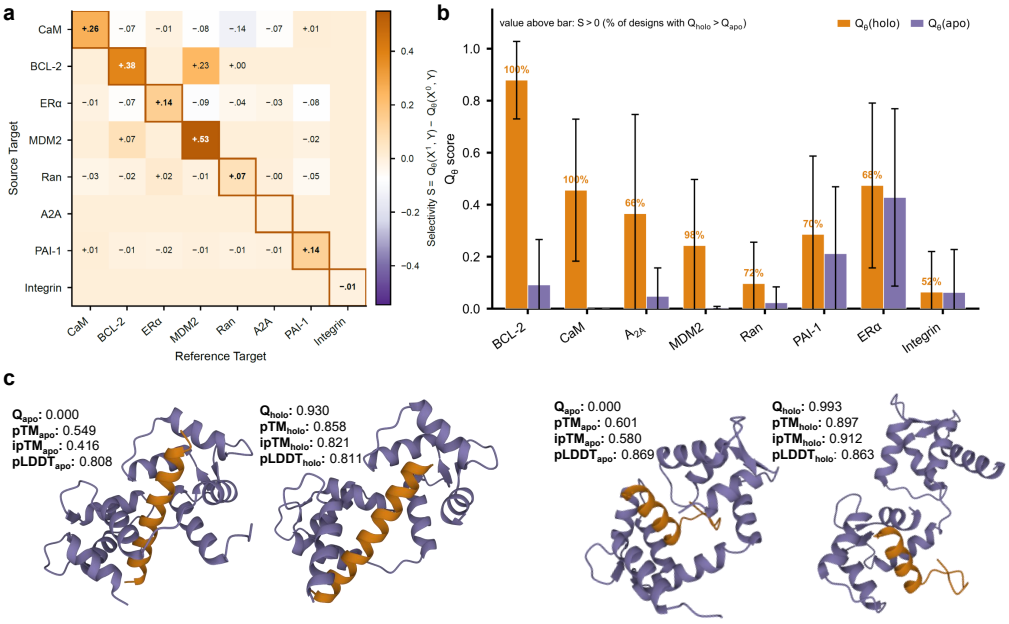

Figure 3. Q_θ **conformational selectivity and CaM selectivity design. (a)** Cross target selectivity matrix between Source Target (binders generated for) and the Reference Target (scored against). Zero values (0.00) have been omitted.; **(b)** Holo vs. apo Q_θ scores for 50 vanilla designs per target; **(c)** Selectivity-based design on CaM. Two case binders (orange) shown against the apo (1st and 3rd panels) and holo (2nd and 4th panels) receptor conformations (purple).

Table 1. End-to-end binder design selectivity ($\bar{S}_{\text{cons}}$) across 15 generator×guidance combinations on 8 OOD targets. **Generators:** RF = RFdiffusion, PX = PXDesign, Pro = Proteina-ComplexA. **Guidance strategies:** V = Vanilla, Cl = Classifier guidance, Lg = Langevin refinement, SM = SMC, TD = TDS. Bold indicates the best guidance per generator within each row.

Target	RF/V	RF/Cl	RF/Lg	RF/SM	RF/TD	PX/V	PX/Cl	PX/Lg	PX/SM	PX/TD	Pro/V	Pro/Cl	Pro/Lg	Pro/SM	Pro/TD	Mean
CaM	+0.455	+0.427	**+0.677**	+0.510	+0.367	+0.517	+0.521	+0.022	**+0.545**	+0.514	+0.338	+0.432	+0.429	**+0.565**	+0.374	+0.446
BCL-2	+0.787	+0.741	+0.806	+0.841	**+0.880**	+0.560	+0.561	+0.568	+0.868	**+0.969**	+0.774	+0.805	+0.826	+0.836	**+0.898**	+0.781
ER	+0.046	+0.054	+0.106	**+0.325**	+0.050	-0.031	-0.029	-0.027	**+0.117**	+0.029	-0.000	-0.000	-0.000	-0.000	+0.000	+0.043
A2A	+0.318	+0.332	+0.377	+0.760	**+0.924**	+0.023	+0.037	+0.048	+0.377	**+0.445**	-0.006	-0.004	-0.003	-0.000	-0.000	+0.242
MDM2	+0.238	+0.271	+0.366	+0.641	**+0.769**	+0.208	+0.227	+0.262	+0.590	**+0.613**	+0.506	+0.567	+0.598	+0.794	**+0.883**	+0.502
PAI-1	+0.073	+0.054	+0.061	**+0.389**	+0.262	-0.000	-0.000	-0.000	-0.000	-0.001	+0.053	+0.064	+0.069	+0.123	**+0.757**	+0.127
Ran	+0.074	+0.091	+0.108	+0.424	**+0.485**	+0.033	+0.056	+0.084	+0.446	**+0.601**	+0.081	+0.128	+0.204	+0.469	**+0.662**	+0.263
Integrin	+0.001	+0.006	+0.008	+0.013	**+0.079**	+0.015	+0.024	+0.033	+0.027	**+0.185**	-0.041	-0.002	-0.011	+0.017	**+0.190**	+0.038
Mean	+0.249	+0.247	+0.314	**+0.488**	+0.477	+0.166	+0.175	+0.124	+0.371	**+0.419**	+0.213	+0.249	+0.267	+0.350	**+0.470**	+0.305

4.4 End-to-End Binder Design

We evaluate Q_θ as an end-to-end design oracle by combining three architecturally distinct generators, RFdiffusion (Watson et al., 2023), PXDesign (Protenix Team et al., 2025), and Proteina-ComplexA (Didi et al., 2026), with vanilla generation and four guidance strategies: classifier guidance, SMC, TDS, and gradient ascent (Langevin) refinement. This yields 15 generator×guidance combinations, evaluated on the 8 held-out OOD test targets, each scored by an ensemble of three independently trained Augmented-S2 Q_θ checkpoints.

Generator × guidance benchmark across OOD targets. Table 1 reports all 15 combinations across 8 OOD targets, surfacing four patterns along the generator, guidance, target, and baseline axes. ***(i) Resampling-based guidance is universally dominant.*** TDS and SMC are the top-2 strategies for every generator, confirming that trajectory-level reweighting transfers across architectures; classifier guidance rarely improves over vanilla. ***(ii) Langevin refinement is generator-conditional.*** It improves RFdiffusion ($+0.249 \rightarrow +0.314$) and Proteina-ComplexA ($+0.213 \rightarrow +0.267$), whose structure-only priors tolerate post-hoc backbone perturbation, but degrades PXDesign ($+0.166 \rightarrow +0.124$), whose sequence-aware prior is destabilized when its co-designed interface geometry is perturbed. ***(iii) Target difficulty is largely intrinsic to the receptor.*** Target identity dominates: BCL-2 yields strong selectivity across all 15 combinations ($\bar{S}{=}+0.781$), while Integrin, ERα, and PAI-1 remain weak everywhere. These extremes align with Phase 2 Spearman ρ, with one informative outlier: ERα has the second-highest scoring ρ (0.664)

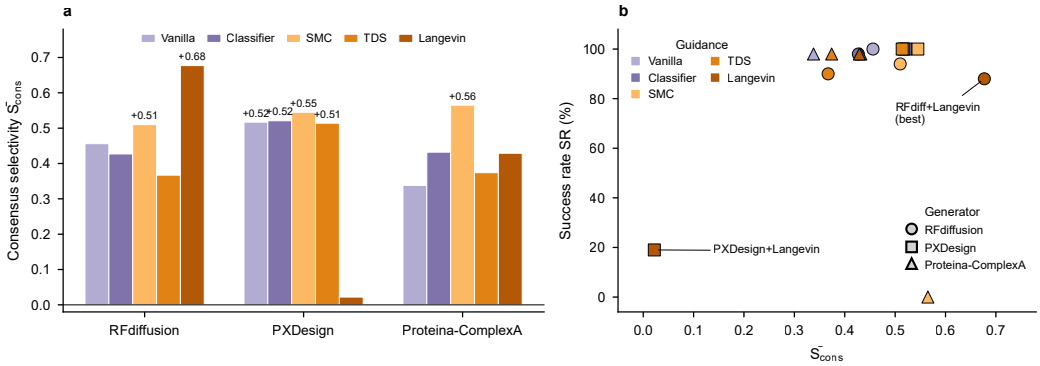

Figure 4. **Generation benchmark on CaM. (a)** Consensus selectivity $\bar{S}_{\text{cons}}$ across 15 generation $\times$ guidance approaches. **(b)** Selectivity vs. design success rate (designable $\times$ selective) across all generator $\times$ guidance combinations.

Table 2. Sequences and experimental binding measurements for peptides evaluated against calmodulin (CaM). Binding affinities were measured by bio-layer interferometry against both holo and apo CaM. NB denotes no detectable binding. $\Delta_q = q_{\text{holo}} - q_{\text{apo}}$ is the difference between the predicted scores against holo and apo CaM.

Design	Sequence	Holo CaM K_D	Apo CaM K_D	Δ_q
rfdiff_vanilla__design_944	ATAAMIKTFQDVVVAAVREAREK	46.6 nM	NB	0.413
rfdiff_vanilla__design_13	SEAFARAAAVLAKARAAK	86.5 nM	NB	0.450
proteina_smc__smc_particle_0273	EGFKKLLKEALEIAK	413 nM	NB	0.932
rfdiff_vanilla__design_657	KVAEQAKQWILEMLAK	>1.00 μM	NB	0.930
rfdiff_langevin__design_25	EKLEALLREAGAARRAAKKAAEAA	1.06 μM	NB	0.930
proteina_vanilla__design_0145	VDEDGDGKIDLPELSALLREKIK	NB	NB	0.993
rfdiff_langevin__design_330	SELTKEILKKAMEMT	NB	NB	0.905
proteina_vanilla__design_0376	AFGAEVKTPRTRFLDVLR	NB	NB	0.688
proteina_smc__smc_particle_0954	EEAARAAGLARLPRPLLLLQAL	NB	NB	0.912
proteina_vanilla__design_0105 (Negative)	SEIAELLRRNPEGDPETLREALAA	NB	NB	0.228
M13 positive control	KRRWKKNFIAVSAANRFKKISSSGAL	Bound	NB	–

yet the second-lowest design $\bar{S}$ ($+0.043$), so its bottleneck has shifted from scoring to generation, where current generators fail to exploit its subtle H12 repositioning. ***(iv) Vanilla strength conditions guidance behavior.*** Targets split into two regimes by vanilla baseline: BCL-2, CaM, and MDM2 already exhibit substantial vanilla selectivity (per-target means 0.71, 0.44, 0.32), where guidance amplifies an already-positive signal; the remaining five sit near zero, where active guidance recovers any meaningful selectivity, positioning selectivity-guided generation as essential rather than incremental on the latter.

End-to-end CaM-selective design. We instantiate the AlloGen pipeline on CaM, whose $\sim$30 Å apo$\rightarrow$holo rearrangement upon Ca^{2+} binding exposes a hydrophobic peptide-binding cleft occluded in the apo state, a representative stress test. Across the 15 combinations evaluated on CaM (Figure 4), 14 yield positive mean selectivity, with RFdiffusion+Langevin reaching $\bar{S}_{\text{cons}}{=}{+}0.677$ at 88% success rate (designable $\times$ selective); the Langevin-classifier gap traces to Q_θ gradient reliability collapsing under the high-noise regime classifier guidance occupies for $\sim$96% of denoising steps (Appendix E.2). Beyond gradient guidance, Q_θ functions as a passive reranker: on the same vanilla pool, best-of-5 reaches $\bar{S}{=}{+}0.787$ and best-of-10 reaches $+0.885$, both exceeding Langevin (Table 16). The two are complementary modes of exploiting the same scorer: reranking scales efficiently with pool size at no extra cost, while Langevin delivers consistent per-design gains when each candidate is expensive to sample. Figure 3c illustrates the endpoint: two AlloGen-designed binders adopt compact helical folds that dock into the cleft exposed only upon Ca^{2+}-induced lobe closure, yielding holo scores of 0.930 and 0.993 against near-zero apo scores, designs the pipeline could not propose without explicit conformational supervision. Throughout, Langevin refinement preserves backbone integrity: zero Ramachandran outliers across all 8 OOD targets, mean bond-length perturbation of 0.005 Å, and native-contact recovery $\Delta\text{fNAT}_{\text{van}} \approx 0$ on all 6 targets with crystal-contact references (Table 18).

What does the selectivity signal capture? To probe what Q_θ encodes, we evaluate AlloGen designs against two orthogonal validation families: independent structural scorers (Boltz-2, AlphaFold 3, and Rosetta InterfaceAnalyzer) and sequence-level recovery (ProteinMPNN).

(a) Structural scorers. Three interface scorers partially cor-

roborate Q_θ at the extremes. Boltz-2 ΔipTM correlates significantly with Q_θ selectivity on A2A (ρ=+0.500) and CaM (ρ=+0.349) (Table 20); AlphaFold 3 confirms holo preference on ALK and ERα (100% positive, n=50 each; Table 21); and Rosetta InterfaceAnalyzer assigns the most favorable ΔG to BCL-2 (-9.1 REU) and the least to Integrin (+39.1 REU), mirroring the Q_θ ranking at both extremes (Table 22). Disagreement on intermediate targets is itself informative: Q_θ tracks conformational selectivity rather than interface energy, and the two are not redundant.

(b) Sequence-level cross-check via ProteinMPNN. Vanilla designs exhibit holo preference (ΔNLL>0) on all 8 OOD targets (p<0.05; Table 23), confirming that holo bias is a population-level property of the generated backbones, detectable independently of Q_θ. Yet Langevin refinement *reduces* ΔNLL on 5 of 8 targets while increasing Q_θ, so the two metrics probe distinct axes: Q_θ captures a geometric selectivity dimension orthogonal to ProteinMPNN's sequence-recovery likelihood, and the two are best read as complementary rather than competing. Q_θ is also robust against degenerate generation: only $10/482$ (2.1%) CaM designs across 9 pipelines yield negative selectivity, all traceable to truncated or sterically infeasible backbones from a single pipeline rather than to apo-selective designs deceiving the scorer (Table 24).

Together, these axes converge: Q_θ encodes a physically grounded conformational selectivity signal complementary to, not redundant with, energy- and sequence-level scorers, and robust against degenerate generation.

4.5 Experimental Validation on CaM

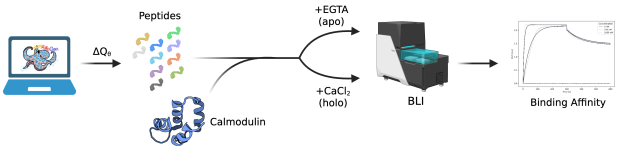

Figure 5. Experimental validation workflow for conformationally selective peptide binding to calmodulin (CaM). Generated peptide candidates were prioritized according to the predicted selectivity margin, $\Delta_q = q_{\text{holo}} - q_{\text{apo}}$, and synthesized for experimental testing. Recombinant human CaM was prepared in either the calcium-free apo state using EGTA or the Ca^{2+}-bound holo state using $CaCl_2$. Peptide binding was quantified by bio-layer interferometry (BLI), and equilibrium dissociation constants (K_D) were determined by fitting association and dissociation kinetics. Representative sensorgrams are shown for a validated holo-state binder.

We next sought to determine whether the conformational selectivity predicted by AlloGen translated into experimentally measurable binding preferences. To this end, we selected a panel of ten peptides generated by multiple generator-guidance combinations, including RFdiffusion, Proteina-Complexa, and their guided variants. Candidate peptides were prioritized according to their predicted selectivity margins ($\Delta_q = q_{\text{holo}} - q_{\text{apo}}$), while also including a designated low-scoring negative-control design. As a positive control, we included the canonical Ca^{2+}-dependent calmodulin-binding M13 peptide. The peptides were synthesized and evaluated against both holo and apo calmodulin (CaM) using bio-layer interferometry. Recombinant human CaM was prepared in either the Ca^{2+}-bound holo state using $CaCl_2$ or the calcium-free apo state using EGTA. Peptides were immobilized on Twin-Strep-compatible biosensors and assayed against CaM at concentrations of 0, 100, and 1000nM. Association and dissociation kinetics were recorded in duplicate and fit using a 1:1 binding model to estimate equilibrium dissociation constants (K_D).

Of the ten experimentally evaluated peptides, five exhibited measurable binding to holo CaM (Figure 5; Table 2). The validated binders spanned affinities from 46.6nM to 1.06μM, with the two strongest candidates achieving nanomolar binding. Notably, the validated binders all originated from the high-Δ_q portion of the ranking and exhibited predicted selectivity margins ranging from 0.413 to 0.932 (Table 2). In contrast, the designated negative-control peptide, which was selected based on its substantially lower predicted selectivity margin ($\Delta_q = 0.228$), exhibited no detectable binding.

Taken together, the experimental results closely mirrored the computational predictions of AlloGen. High-Δ_q candidates yielded multiple holo-state binders with nanomolar-to-micromolar affinities, while the designated low-Δ_q negative control failed to bind. Moreover, every experimentally validated binder was selective for the desired holo conformation, exhibiting no detectable interaction with apo CaM. These findings provide direct evidence that the selectivity signal captured by Q_θ reflects biologically meaningful conformational preferences rather than generic binding propensity.

5 Conclusion

We have presented AlloGen, a modular framework for conformational selectivity in binder design. Its core component Q_θ is a generator-agnostic scorer that integrates with any backbone generator as a passive reranker or gradient-based guide without retraining. On 8 held-out OOD targets, Q_θ generalizes where contact-based energy proxies fail uniformly ($\bar{\rho} = 0.520$), and all 15 generator-guidance combinations yield positive mean selectivity, with RFdiffusion+Langevin reaching $\bar{S} = +0.677$ on CaM and best-of-K reranking $\bar{S} = +0.885$. Crucially, these signals translate *in vitro*: of ten peptides synthesized against CaM, five bind the holo state (46.6 nM to 1.06 μM) while none bind apo, and a low-Δ_q negative control fails to bind. Future work will extend Q_θ to multi-state landscapes, integrate selectivity into sequence generation, and explore Siamese dual-state architectures.

Impact Statement

This work advances protein binder design by formulating conformational selectivity as a learnable, transferable signal, with potential to accelerate allosteric therapeutics targeting kinases, nuclear receptors, and GPCRs, families central to oncology, metabolic disease, and neuropsychiatric drug discovery. Because Q_θ is generator-agnostic and lightweight, AlloGen lowers the computational barrier to state-selective design and complements existing pipelines without retraining. More broadly, the framework offers a general recipe for incorporating differential, state-aware objectives into structure-based generative models, which may benefit related problems such as multi-state enzyme design and conformational biosensor engineering. As with other generative design methods, translating computational designs into functional molecules still requires substantial experimental validation, and we encourage continued engagement with community norms around responsible development.

Acknowledgment

This research was supported by a grant from the Highthroughput Institute for Discovery (HIT-ID) at the University of Pennsylvania to the lab of Pranam Chatterjee. The work described in this paper was also partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Basu, S. and Wallner, B. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8): e0161879, 2016.

Bhat, S., Palepu, K., Hong, L., Mao, J., Ye, T., Iyer, R., Zhao, L., Chen, T., Vincoff, S., Watson, R., et al. De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *Science Advances*, 11(4):eadr8638, 2025.

Cao, H., Shi, H., Wang, C., Pan, S. J., and Heng, P.-A. GLID^{2}e: A gradient-free lightweight fine-tune approach for discrete biological sequence design. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AHjspi4R22.

Cao, H., Pal, A., Tang, S., Zhang, Y., Zhang, J., Heng, P.-A., and Chatterjee, P. TD3b: Transition-directed discrete diffusion for allosteric binder generation. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=r2nuhKd6vr.

Chen, L. T., Quinn, Z., Dumas, M., Peng, C., Hong, L., Lopez-Gonzalez, M., Mestre, A., Watson, R., Vincoff, S., Zhao, L., et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pp. 1–9, 2025a.

Chen, T., Quinn, Z., Zhang, Y., and Chatterjee, P. moppit-v3: Motif-specific peptides generated via multi-objective-guided discrete flow matching. In *NeurIPS 2025 AI for Science Workshop*.

Chen, T., Zhang, Y., and Chatterjee, P. Areuredi: Annealed rectified updates for refining discrete flows with multi-objective guidance. *arXiv preprint arXiv:2510.00352*, 2025b.

Conflitti, P., Lyman, E., Sansom, M. S., Hildebrand, P. W., Gutiérrez-de Terán, H., Carloni, P., Ansell, T. B., Yuan, S., Barth, P., Robinson, A. S., et al. Functional dynamics of g protein-coupled receptors reveal new routes for drug discovery. *Nature Reviews Drug Discovery*, 24(4):251–275, 2025.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Didi, K., Zhang, Z., Zhou, G., Reidenbach, D., Cao, Z., Cha, S., Geffner, T., Dallago, C., Tang, J., Bronstein, M. M., Steinegger, M., Kucukbenli, E., Vahdat, A., and Kreis, K. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qmCpJtFZra.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E.-M., Wilson, I. A., and Baker, D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821, 2011.

Gruver, N., Stanton, S., Frey, N., Rudner, T. G., Hotzel, I., Lafrance-Vanasse, J., Rajpal, A., Cho, K., and Wilson, A. G. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36: 12489–12517, 2023.

Ha, J.-H. and Loh, S. N. Protein conformational switches: from nature to design. *Chemistry–A European Journal*, 18(26):7984–7999, 2012.

Havranek, J. J. and Harbury, P. B. Automated design of specificity in molecular recognition. *Nature structural biology*, 10(1):45–52, 2003.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Jing, B., Berger, B., and Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=rs8Sh2UASt.

Jing, B., Sappington, A., Bafna, M., Shah, R., Tang, A., Krishna, R., Klivans, A., Diaz, D. J., and Berger, B. Generating functional and multistate proteins with a multimodal diffusion transformer. *bioRxiv*, 2025.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kalakoti, Y. and Wallner, B. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications biology*, 8(1):373, 2025.

Kar, G., Keskin, O., Gursoy, A., and Nussinov, R. Allostery and population shift in drug discovery. *Current opinion in pharmacology*, 10(6):715–722, 2010.

Kojetin, D. J. and Burris, T. P. Small molecule modulation of nuclear receptor conformational dynamics: implications for function and drug discovery. *Molecular pharmacology*, 83(1):1–8, 2013.

Langan, R. A., Boyken, S. E., Ng, A. H., Samson, J. A., Dods, G., Westbrook, A. M., Nguyen, T. H., Lajoie, M. J., Chen, Z., Berger, S., et al. De novo design of bioactive protein switches. *Nature*, 572(7768):205–210, 2019.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Nussinov, R. Introduction to protein ensembles and allostery, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

Protenix Team, Ren, M., Sun, J., Guan, J., Liu, C., Gong, C., Wang, Y., Wang, L., Cai, Q., Chen, X., et al. Pxdesign: Fast, modular, and accurate de novo design of protein binders. *bioRxiv*, 2025.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. The foldx web server: an online force field. *Nucleic acids research*, 33(suppl_2):W382–W388, 2005.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Song, Z., Li, T., Li, L., and Min, M. R. PPDiff: Diffusing in hybrid sequence-structure space for protein-protein complex design. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=gTYzm0Qchr.

Stark, H., Faltings, F., Choi, M., Xie, Y., Hur, E., O'Donnell, T., Bushuiev, A., Uçar, T., Passaro, S., Mao, W., et al. Boltzgen: Toward universal binder design. *bioRxiv*, 2025.

Stranges, P. B. and Kuhlman, B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Science*, 22(1):74–82, 2013.

Tang, S., Zhang, Y., and Chatterjee, P. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=FQoy1Y1Hd8.

Tang, S., Zhang, Y., Tong, A., and Chatterjee, P. Gumbel-softmax score and flow matching for discrete biological sequence generation. 2025b.

Tang, S., Zhu, Y., Tao, M., and Chatterjee, P. Tr2-d2: Tree search guided trajectory-aware fine-tuning for discrete diffusion. *arXiv preprint arXiv:2509.25171*, 2025c.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vetter, S. W. and Leclerc, E. Novel aspects of calmodulin target recognition and activation. *European Journal of Biochemistry*, 270(3):404–414, 2003.

Vijayan, R., He, P., Modi, V., Duong-Ly, K. C., Ma, H., Peterson, J. R., Dunbrack Jr, R. L., and Levy, R. M. Conformational analysis of the dfg-out kinase motif and biochemical profiling of structurally validated type ii inhibitors. *Journal of medicinal chemistry*, 58(1):466–479, 2015.

Vincoff, S., Davis, O., Ceylan, I. I., Tong, A., Bose, J., and Chatterjee, P. Soapia: Siamese-guided generation of off target-avoiding protein interactions with high target affinity. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025.

Wang, X., Flannery, S. T., and Kihara, D. Protein docking model evaluation by graph neural networks. *Frontiers in Molecular Biosciences*, 8:647915, 2021.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Weikl, T. R. and Paul, F. Conformational selection in protein binding and function. *Protein Science*, 23(11):1508–1518, 2014.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688, 2011.

Wu, L., Trippe, B., Naesseth, C., Blei, D., and Cunningham, J. P. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Xu, X. and Bonvin, A. M. Deeprank-gnn-esm: a graph neural network for scoring protein–protein models using protein language model. *Bioinformatics advances*, 4(1): vbad191, 2024.

Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M., and Vangone, A. Prodigy: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, 2016.

Xue, Y., Wang, H., Li, J., Hu, J., Chen, Z., Zheng, Z., Liu, L., Zhu, K., He, J., Gong, H., et al. State-specific peptide design targeting g protein-coupled receptors. *Journal of Chemical Information and Modeling*, 65(20):11425–11438, 2025.

A Related Work

A.1 Structure-Based Binder Design

Deep generative models have transformed structure-based protein binder design. RFdiffusion (Watson et al., 2023) generates binder backbones conditioned on a fixed target surface via backbone diffusion, followed by sequence design with ProteinMPNN (Dauparas et al., 2022). PPDiff (Song et al., 2025) and PXDesign (Protenix Team et al., 2025) extend this to joint sequence-structure generation, while Proteina-ComplexA (Didi et al., 2026) employs atomistic flow matching to produce full-atom binder complexes. On the sequence side, a growing body of work generates binder sequences through discrete diffusion and language model objectives (Chen et al., 2025a; Bhat et al., 2025; Chen et al.; Vincoff et al., 2025; Tang et al., 2025a;c; Chen et al., 2025b; Tang et al., 2025b; Cao et al., 2025; 2026). Despite their diversity, all of these methods share a fundamental constraint: they condition on a single, static receptor conformation and optimize for binding quality without any mechanism for conformational selectivity. AlloGen is designed to be complementary that its modular scorer Q_θ can rerank or guide candidates produced by any of these generators.

A.2 Protein–Protein Interface Scoring

Physics-based methods such as FoldX (Schymkowitz et al., 2005) and PRODIGY (Xue et al., 2016) estimate binding free energies, and inverse folding models such as ProteinMPNN (Dauparas et al., 2022) and ESM-IF (Hsu et al., 2022) serve as zero-shot affinity proxies. GNN-based scorers including GNN-DOVE (Wang et al., 2021) and DeepRank-GNN-esm (Xu & Bonvin, 2024) learn interface quality from structural features, but are designed for docking re-ranking rather than conformational discrimination. Critically, all of these methods measure affinity rather than differential affinity, and provide no signal for selecting one conformational state over another. Q_θ departs from this paradigm through its two-phase curriculum, which prevents the degenerate solution of ignoring receptor conformation. Moreover, its logit-space selectivity margin that explicitly reasons about score differences across states.

A.3 Conformational Selectivity: Sampling, Design, and Negative Design

Achieving conformational selectivity requires both characterizing the conformational landscape and engineering molecules that exploit it. On the characterization side, AlphaFlow (Jing et al., 2024) uses flow matching to generate structural ensembles, AFsample2 (Kalakoti & Wallner, 2025) diversifies predictions via MSA subsampling, and AlphaFold 3 (Abramson et al., 2024) extends structure prediction to biomolecular complexes. These methods map the conformational space but do not address the inverse problem of designing binders that discriminate between states. On the design side, negative design has a long history in computational protein engineering (Stranges & Kuhlman, 2013; Fleishman et al., 2011), and multi-state design in Rosetta (Havranek & Harbury, 2003) optimizes sequences across multiple conformations. However, these approaches operate at the sequence level and do not provide a differentiable scoring signal compatible with modern backbone generators. Contrastive objectives inspired by InfoNCE (Oord et al., 2018) have been applied to molecular representation learning, and classifier guidance (Dhariwal & Nichol, 2021) and TDS (Wu et al., 2023) enable gradient-based steering of diffusion models, mechanisms we directly adopt. Concurrent work by Xue et al. (2025) targets state-specific peptide design for GPCRs via a conformation-specific folding filter, and ProDiT (Jing et al., 2025) introduces multi-state structure diffusion. AlloGen unifies these threads: it learns a generalizable selectivity scorer through contrastive training, applies it as a gradient-based guide across architecturally diverse generators, and demonstrates generalization across 15 protein families.

B Implementation Details

B.1 Feature Computation

Backbone frames. Backbone frames $R_i \in \mathrm{SO}(3)$ are constructed from the (N, C_α, C) triplet of each residue using the Gram–Schmidt procedure. Torsion angles (φ, ψ, ω) are computed from four consecutive backbone atoms using the standard dihedral angle formula. At chain termini, missing torsion angles are set to zero.

RBF distance encoding. The 16-dimensional RBF encoding of distance d uses Gaussian basis functions $\exp(-\gamma(d-\mu_k)^2)$ with centres μ_k evenly spaced from 2 to 12 Å and width $\gamma = 1.5$.

B.2 Model and Training Hyperparameters

Model size. The InterfaceGNN scorer has $\sim$898K trainable parameters: 4 graph transformer layers with hidden dimension 128, 8 attention heads, query/key/value and edge bias projections, and a 3-layer scoring MLP.

Training hyperparameters. Hyperparameters for both training phases are listed in Table 3.

Table 3. Training hyperparameters for the two-phase curriculum.

Hyperparameter	Phase 1	Phase 2
Epochs	40	15
Learning rate	5×10^{-4}	4×10^{-5}
Weight decay	10^{-3}	10^{-3}
Batch size	512	256
Warmup steps	100	—
Loss	MSE ($\mathcal{L}_q$)	InfoNCE ($\mathcal{L}_q$ dropped)
InfoNCE τ	—	0.1
LR schedule	Cosine	Cosine
Dropout	0.1	0.1
Binder dropout	0.3	0.3

Guidance hyperparameters. For RFdiffusion classifier guidance, Q_θ is used as a guiding potential with weight ω=1.0 and global `guide_scale` = 5.0, constant decay schedule, and 200 denoising steps. For PXDesign classifier guidance, guidance scale 1.0 is applied during diffusion progress fractions $[0.1, 0.8]$, with noise-fraction-proportional decay within that window. Langevin refinement uses step size η=0.05 and 100 gradient steps for RFdiffusion backbones, or η=0.01 and 100 steps for PXDesign. TDS uses N=16 particles and R=4 resampling rounds, retaining the top 50% of particles each round.

B.3 Guidance Formulations

All four guidance strategies exploit the differentiability of $S_\theta(Y; X^1, \{X^0\})$ with respect to binder backbone coordinates $\mathbf{x} = \{\mathbf{p}_i\}_{i \in Y}$.

Langevin backbone refinement ((Welling & Teh, 2011; Song et al., 2021)). Given a generated backbone $\mathbf{x}_0$, we perform T steps of gradient ascent on the selectivity margin:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \, \nabla_{\mathbf{x}} S_\theta(\mathbf{x}_t; X^1, \{X^0\}), \quad t = 0, \ldots, T-1, \tag{9}$$

where η is the step size. This operates on fully denoised backbones, avoiding the noisy-gradient regime.

Classifier guidance ((Dhariwal & Nichol, 2021)). During diffusion denoising, we inject the Q_θ gradient into each denoising step:

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}_\psi(\mathbf{x}_t, t) - \omega \, \sigma_t \, \nabla_{\mathbf{x}_t} \log Q_\theta(X^1, \mathbf{x}_t), \tag{10}$$

where $\boldsymbol{\epsilon}_\psi$ is the denoiser, ω the guidance weight, and σ_t the noise level at timestep t.

Twisted diffusion sampling (TDS) ((Wu et al., 2023)). TDS uses N particles $\{\mathbf{x}_t^{(n)}\}_{n=1}^N$ with importance weights $w_t^{(n)} \propto Q_\theta(X^1, \hat{\mathbf{x}}_0^{(n)})$ computed from the denoised prediction at each timestep. Multinomial resampling is applied when the effective sample size $\text{ESS} = (\sum_n w_n)^2 / \sum_n w_n^2$ drops below $N/2$.

Sequential Monte Carlo (SMC) ((Wu et al., 2023)). SMC extends TDS with multiple resampling rounds: after generating N complete trajectories, the top fraction by S_θ seeds the next round, iteratively enriching the candidate pool toward high-selectivity designs.

C Dataset Construction

The dataset comprises 2,896 complexes across 65 two-state protein targets in 15 families. Each complex yields 12 training samples: 1 native holo labeled 1.0, 1 apo mismatch labeled 0.0, and 10 rigid-body decoys with labels $\max(0, 1 - d_{\text{RMSD}}/4)$,

plus 959 FastRelax-based hard negatives with mean label 0.304. Interface graphs contain 34.7 ± 19.7 nodes, ranging from 3 to 128 with a median of 29, comprising 18.1 ± 9.8 receptor and 16.6 ± 11.0 binder residues. The target split assigns 51 targets to training, 6 to OOD validation, and 8 to OOD test with CaM held out, yielding 30,623 samples split as 23,640 training, 4,020 validation, and 2,963 test (Table 4). Linked targets such as SRC and SRC-SH2 are forced into the same partition. All scoring experiments use the S2-augmented configuration totalling 39,867 training samples; ablations across S1/S2/S3 are detailed in Table 5.

D Extended Q_θ Results

All results in this section use the S2 configuration with the target split (Table 4), ESM-2 encoder, and binder-dropout rate p=0.3, consistent with the experimental setup in Section 4.2. The suffix S1/S2/S3 denotes the dataset augmentation configuration as defined in Table 5.

Table 4. **Target split.** 51/6/8 train/val/test partition across 15 protein families. CaM is moved to the OOD test set as the primary design target; ALK is moved to training.

Family	Train (51)	Val OOD (6)	Test OOD (8)
Kinase (9)	ABL1, **ALK**, Aurora A, BRAF, CDK2, EGFR, GRK2, SRC, SRC-SH2	—	—
Small GTPase (6)	Arf1, CDC42, KRAS, RhoA	Rac1	Ran
Nuclear Receptor (5)	AR, GR, PPARγ, RXRα	—	ERα
GPCR/Ion Ch. (6)	CFTR, KcsA, nAChR	β_2AR, μ-opioid	A_{2A}
Phosphatase (4)	Calcineurin, PP2A, PTP1B, SHP2	—	—
Protease (6)	HIV-PR, MMP-2, M^{pro}, Thrombin, USP7	Caspase-3	—
BCL-2 (2)	BCL-xL	—	BCL-2
14-3-3 (2)	—	14-3-3, 14-3-3σ	—
Ca sensor (3)	S100B, Troponin C	—	**CaM**
G protein (2)	CheY, Gα_s	—	—
Chaperone/Enzyme (6)	ATCase, CypA, DHFR, Enolase, Hsp70, RNase A	—	—
Epigenetic (3)	BRD4, WDR5	—	MDM2
Trafficking (5)	ABCB1, Importin-β, Maltose BP, p97, VHL	—	—
Cell adhesion (3)	CTLA-4, Integrin β_3	—	Integrin
Signaling (3)	Gelsolin, Hemoglobin	—	PAI-1
Total (65)	**51**	**6**	**8**

Table 5. Dataset configurations and scoring settings. Columns show the number of targets in each partition and the sample count of each augmentation source. FastRelax Neg.: constrained relaxation negatives (present in all configs). GenDecoys: synthetic binders from generative models.

Dataset	Split	Train	Val	Test	FastRelax Neg.	GenDecoys	Cross-fam	Conf. decoys
Baseline	Target-split	51	6	8	959	–	–	–
Augmented	Target-split (ext.)	51	6	8	959	4,862	254	1,075
Scoring configurations								
S1 (Base)	Baseline only: native complexes + rigid-body decoys + FastRelax Neg. (34,751 samples)							
S2 (Full)	Full Augmented: S1 + GenDecoys + Cross-fam + Conf. decoys (39,867 samples)							
S3 (Partial)	Baseline + Cross-fam + Conf. decoys, no GenDecoys (36,080 samples)							

D.1 Per-Target Selectivity Performance

Table 6 reports the per-target Spearman ρ of Phase 1 and Phase 1+2 on the 8 OOD targets under the target split. Phase 2 delivers its largest gains on the targets where Phase 1 struggles most: Integrin ($+0.166$), A_{2A} ($+0.170$), and, to a lesser extent, PAI-1 ($+0.027$) and MDM2 ($+0.037$). On the two targets where Phase 1 already performs well, BCL-2 and ERα, Phase 2 incurs a modest drop of -0.049 and -0.025, while CaM and Ran remain essentially unchanged. Overall, Phase 2 trades a small loss on the easiest targets for substantial improvements on the hardest ones, lifting the 8-target mean from 0.481 to $\mathbf{0.520}$ and noticeably narrowing the cross-target gap.

Table 6. Phase 1 vs. Phase 1+2 Spearman ρ on 8 OOD targets under the target split. Phase 2 values are 3-seed means $\pm$ std.

Target	Phase 1 ρ	Phase 2 ρ
BCL-2	0.716	0.667 ± 0.014
ERα	0.689	0.664 ± 0.013
MDM2	0.552	0.589 ± 0.036
CaM	0.491	0.474 ± 0.061
Ran	0.511	0.511 ± 0.065
A_{2A}	0.299	0.469 ± 0.034
PAI-1	0.394	0.421 ± 0.045
Integrin	0.200	0.366 ± 0.085
Mean (8)	0.481	$\mathbf{0.520} \pm 0.010$

D.2 Phase 2 Batch-Size Ablation

Table 7 ablates the InfoNCE batch size in Phase 2. All InfoNCE configurations outperform Phase 1-only regression, confirming that contrastive fine-tuning is necessary for conformational discrimination. Batch size 256 yields the highest mean ρ (0.530), striking the best balance between cross-target negative diversity (which favors larger batches) and optimization stability (which favors smaller ones); BS=512 over-saturates the softmax, while BS=64 provides too few cross-target negatives per anchor.

Table 7. Phase 2 ablation across InfoNCE batch sizes. All variants start from the same Phase 1 checkpoint.

Configuration	$\bar{\rho}$
Paired InfoNCE (cross-target negatives)	
InfoNCE, BS=256	$\mathbf{0.530}$
InfoNCE, BS=512	0.500
InfoNCE, BS=64	0.496
Phase 1 only (MSE)	0.489

D.3 Apo Rejection Analysis

Table 8 reports holo and apo Q_θ scores for 50 vanilla designs per target across the 8 OOD test targets, evaluated under Kabsch alignment with a single Q_θ seed. Seven of the eight targets exhibit a positive ΔQ, indicating that Q_θ correctly prefers the holo state; MDM2 and BCL-2 reach an $S{>}0$ rate of 100%. Integrin is the only target with a marginal apo preference ($\Delta Q{=}-0.009$, $S{>}0{=}40\%$). Under the 3-seed ensemble, CaM exhibits even sharper rejection, with $\bar{Q}_{\text{holo}}{=}0.456$, $\bar{Q}_{\text{apo}}{=}0.000$, and $S{>}0{=}100\%$.

Table 8. Holo vs. apo Q_θ scores for 50 vanilla designs per target.

Target	$\bar{Q}_{\text{holo}}$	$\bar{Q}_{\text{apo}}$	ΔQ	$S{>}0$
MDM2	0.577 ± 0.189	0.003 ± 0.006	$+0.575$	100%
BCL-2	0.485 ± 0.281	0.000 ± 0.000	$+0.485$	100%
ERα	0.342 ± 0.250	0.120 ± 0.053	$+0.222$	76%
CaM	0.213 ± 0.136	0.065 ± 0.052	$+0.148$	74%
PAI-1	0.146 ± 0.093	0.018 ± 0.057	$+0.128$	90%
Ran	0.065 ± 0.057	0.000 ± 0.002	$+0.064$	92%
A_{2A}	0.062 ± 0.096	0.000 ± 0.000	$+0.062$	92%
Integrin	0.064 ± 0.054	0.072 ± 0.058	-0.009	40%

D.4 Comparison against energy-based baselines.

Conformational selectivity is fundamentally a *differential* signal: a scorer must distinguish two receptor backbones presented with the same binder, not merely rate a single interface. We compare Q_θ against three contact-based proxies on the 8 OOD

targets (Table 9): PRODIGY (Xue et al., 2016), total interface residue count, and cross-chain edge density. All three fail to track DockQ on average ($\bar{\rho}$=+0.143, −0.154, −0.070 respectively), since each summarizes a single interface in isolation and has no mechanism to contrast apo and holo states for the same binder. Q_θ instead achieves $\bar{\rho}$=0.520 with all 8 targets positive, confirming that conformational selectivity is recoverable only by a representation trained explicitly on paired apo–holo geometry.

Table 9. Energy-based baselines vs. Q_θ on 8 OOD targets. Spearman ρ.

Target	n	PRODIGY	Iface Size	Edge Dens.	Random	Q_θ
$A_{2A}R$	36	−0.230	−0.119	−0.092	−0.120	**0.469** ± .034
BCL-2	132	+0.160	−0.050	−0.082	−0.039	**0.667** ± .014
CaM	96	+0.199	−0.316	+0.014	+0.059	**0.474** ± .061
ERα	72	+0.551	−0.292	−0.036	+0.086	**0.664** ± .013
Integrin	60	−0.019	−0.195	+0.053	+0.076	**0.366** ± .085
MDM2	143	+0.163	−0.050	−0.130	−0.037	**0.589** ± .036
PAI-1	156	+0.058	+0.062	−0.146	−0.043	**0.421** ± .045
Ran	2268	+0.264	−0.270	−0.138	−0.018	**0.511** ± .065
Mean		+0.143	−0.154	−0.070	−0.005	**0.520** ± .010

D.5 Conformational landscape monotonicity.

To test whether Q_θ captures the full conformational transition rather than a binary apo/holo signal, we score CaM binders against 11 interpolated receptor conformations along the apo→holo path ($\tau = 0$ to 1). Table 10 shows that vanilla binders exhibit strong positive monotonicity ($\bar{\rho}$=+0.518, 100% with positive Spearman ρ), demonstrating that Q_θ learns a genuine structural landscape rather than a binary classifier. Langevin-refined designs show earlier score onset ($\tau_{50\%}$=0.4–0.9 vs. 0.8–1.0), consistent with their stronger holo affinity.

Table 10. Conformational landscape analysis on CaM (20 Langevin + 10 vanilla designs, 11 interpolated frames per design). Monotonicity: Spearman ρ between frame index τ and $Q_\theta(X^\tau, Y)$.

Design set	Mean $\rho(\tau, Q)$	Monotone frac.	Best ρ	Onset ($\tau_{50\%}$)
Vanilla (10)	**+0.518**	100% (10/10)	0.745	0.8–1.0
Langevin (20)	+0.246	80% (16/20)	0.891	0.4–0.9

D.6 Multi-Seed Scoring Robustness

Table 11 reports Q_θ selectivity across three independently trained seeds on CaM. Vanilla and Langevin selectivity are both highly stable across seeds, confirming that Q_θ's OOD scoring is robust to initialization.

Table 11. Per-seed Q_θ selectivity $\bar{S}$ on CaM under three independently trained seeds.

		Per-seed $\bar{S}$				
Target	**Guidance**	s2048	s10040	s10043	Mean	σ
CaM	Vanilla	+0.148	+0.146	+0.156	+0.150	0.004
	Langevin	+0.255	+0.250	+0.263	**+0.256**	0.005

E Extended Design Results

E.1 Generation Benchmark Performance

Table 12 reports target-split results on CaM as the OOD target, evaluating end-to-end binder design across three architecturally distinct generators (RFdiffusion (Watson et al., 2023), PXDesign (Protenix Team et al., 2025), Proteina-ComplexA (Didi et al., 2026)) and five guidance strategies. Each design is scored by an ensemble of three independently trained Q_θ checkpoints.

Table 12. End-to-end binder design on CaM as a held-out OOD target. Selectivity is evaluated by a 3-seed consensus Q_θ ensemble. Structural quality is measured by Boltz-2 ipTM and scTM; sequence designability by ProteinMPNN NLL_{holo}; and success rate by SR = designable% $\times$ $S{>}0$%. †Proteina-ComplexA generates full-atom complexes with co-designed sequences.

Generator	Guidance	N	$\bar{S}_{\text{cons}}$	Top-5 $\bar{S}$	$\%S{>}0$	ipTM	scTM	NLL_{holo}	Des%	SR%
RFdiffusion										
	Langevin	50	**+0.677**	**0.993**	100%	0.742	0.806	**1.39**	88%	**88%**
	SMC	64	+0.510	0.941	100%	0.733	0.828	3.60	94%	94%
	Vanilla	50	+0.456	0.932	100%	**0.769**	0.820	3.59	100%	100%
	Classifier	50	+0.427	0.891	100%	0.715	0.826	3.54	98%	98%
	TDS	40	+0.367	0.818	100%	0.766	**0.835**	3.66	90%	90%
PXDesign										
	SMC	64	+0.545	0.851	100%	0.710	0.698	3.71	100%	100%
	Classifier	50	+0.521	0.891	100%	0.757	0.684	3.74	100%	100%
	Vanilla	50	+0.517	0.870	100%	0.736	0.690	3.74	100%	100%
	TDS	64	+0.514	0.851	100%	0.762	0.690	3.74	100%	100%
	Langevin	50	+0.022	0.424	24%	0.639	0.664	1.58	80%	19%
Proteina-ComplexA (Didi et al., 2026)†										
	SMC	52	**+0.565**	**0.992**	**100%**	0.420	0.340	1.38	0%	0%
	Classifier	52	+0.432	0.975	98%	0.399	0.597	1.88	100%	98%
	Langevin	52	+0.429	0.975	98%	0.381	0.596	1.88	100%	98%
	TDS	52	+0.374	0.981	98%	0.376	0.592	1.86	100%	98%
	Vanilla	52	+0.338	0.942	98%	0.387	0.601	1.80	100%	98%

E.2 Gradient Reliability Under Noise

We measure the cosine similarity between Q_θ gradients computed on clean and noise-perturbed backbones as a function of Gaussian noise scale σ (Table 13). Cosine similarity drops from 0.75 at σ=0.1 Å to 0.12 at σ=0.5 Å, and approaches zero for $\sigma \geq 2.0$ Å. Gradient reliability therefore degrades rapidly with noise, which explains why Langevin refinement (operating at $\sigma \approx 0.04$ Å) succeeds while diffusion-scale classifier guidance, which spends most of its trajectory at much larger σ, fails.

Table 13. Q_θ gradient cosine similarity under Gaussian backbone noise, averaged over 20 CaM designs.

σ (Å)	Norm	Cos	Margin
0.0	0.317	**1.000**	+0.359
0.1	0.329	0.753	+0.352
0.5	0.368	0.121	+0.380
1.0	0.366	0.032	+0.349
2.0	0.172	0.012	+0.211
5.0	0.018	0.009	+0.015

E.3 Noise Schedule Alignment

In Table 14, we align the RFdiffusion noise schedule (T=50 denoising steps) with the gradient-reliability profile from Supplementary Section E.2. Cosine similarity is 0.75 at t=0 but falls to 0.31 by t=5 and to 0.06 by t=25, remaining above 0.5 only for roughly the final 2 of 50 steps, i.e., once the trajectory enters the regime $\sigma < 0.10$ Å. This quantitatively explains the failure of classifier guidance: about 96% of the denoising trajectory operates in a regime where Q_θ gradients are essentially uninformative.

Table 14. RFdiffusion noise schedule mapped to Q_θ gradient reliability across timesteps.

Timestep	σ (Å)	$\bar{\alpha}$	Cosine
0	0.10	0.99	0.75
5	0.28	0.92	0.31
10	0.40	0.84	0.15
25	0.70	0.51	0.06
49	0.93	0.13	0.05

E.4 Reranking vs. Gradient Refinement

We compare best-of-K reranking with Langevin refinement on 50 CaM designs under 3-seed ensemble scoring, where the vanilla pool already attains $\bar{S}{=}{+}0.456$ with 100% positive selectivity (Table 15). Because every vanilla design is positive, reranking efficiently lifts the expected best: best-of-5 reaches $\bar{S}{=}{+}0.787$ and best-of-10 reaches $+0.885$, both exceeding Langevin refinement at $\bar{S}{=}{+}0.677$. Langevin nonetheless remains preferable when the candidate pool is small, or when per-design improvement matters more than pool selection, for example when each candidate is expensive to generate.

Table 15. Best-of-K reranking vs. Langevin refinement on 50 CaM vanilla designs under ensemble scoring.

Strategy	$\bar{S}$	$S{>}0$
Vanilla (all 50)	$+0.456$	100%
Rerank best-of-2	$+0.616$	100%
Rerank best-of-5	$+0.787$	100%
Rerank best-of-10	$+0.885$	100%
Rerank best-of-20	$+0.947$	100%
Langevin (100 steps)	$\mathbf{+0.677}$	$\mathbf{100}\%$

E.5 Best-of-K Reranking

Best-of-K reranking measures Q_θ's ability to pick conformationally selective designs from a candidate pool. On CaM, where every vanilla design already attains positive selectivity, increasing K improves both the mean and the high-quality fraction: $K{=}5$ raises $\bar{S}$ from $+0.456$ to $+0.787$ with 94% of selections exceeding $S{>}0.5$, and $K{=}10$ reaches $+0.885$ with all selections above that threshold (Table 16). Reranking therefore scales reliably with pool size, with the additional gain per doubling of K tapering past $K{\approx}10$.

Table 16. Best-of-K reranking statistics on CaM with 10,000 bootstrap trials. "High" denotes $S{>}0.5$.

K	Mean	$S{>}0$	High
1	$+0.452$	100%	41%
2	$+0.616$	100%	67%
5	$+0.787$	100%	94%
10	$+0.885$	100%	100%
25	$+0.961$	100%	100%
50	$\mathbf{+0.985}$	$\mathbf{100\%}$	$\mathbf{100\%}$

E.6 End-to-end Langevin Design Results

Structure-related metric analysis for Langevin refinement. Table 17 reports end-to-end results on all 8 OOD targets, with 50 vanilla designs per target refined by 100 Langevin steps. Langevin refinement improves mean selectivity on 7 of 8 targets, with the largest gains on CaM ($+0.253$) and improvement rates (the fraction of refined designs with $S > 0$) reaching 100% on CaM and BCL-2 and exceeding 70% on six targets. PAI-1 is the only target where refinement marginally degrades $\bar{S}$ ($\Delta\bar{S}{=}-0.024$), although 71% of its refined designs are still positively selective. On the six targets with a crystal-contact reference, $\Delta\text{fNAT}_{\text{van}}$ is approximately zero or positive (range -0.004 to $+0.168$), indicating that the selectivity gains do not come at the expense of native-contact recovery. Overall, 390 of 400 designs survive Langevin refinement, and refined designs combine higher selectivity with comparable or improved interface fidelity.

Table 17. End-to-end Langevin design results across 8 OOD targets. $\Delta\text{fNAT}_{\text{van}}$ is measured against the crystal-contact reference; CaM and A_{2A} are marked "—" because their available co-crystal structures contain peptide-mimetic or antibody-fragment binders rather than the canonical short peptides used to compute fNAT. Improvement rate is the fraction of Langevin-refined designs with $S{>}0$. Hyperparameters follow Supplementary Section B.2.

Target	N_{van}	N_{lang}	Improv. rate	$\Delta\text{fNAT}_{\text{van}}$	$\bar{S}_{\text{van}}$	$\bar{S}_{\text{lang}}$	$\Delta\bar{S}$
CaM	50	50	100%	—	0.531	0.784	$+0.253$
ERα	50	50	74%	-0.004	0.110	0.250	$+0.140$
BCL-2	50	50	100%	$+0.168$	0.886	0.911	$+0.024$
MDM2	50	50	98%	$+0.084$	0.216	0.397	$+0.181$
Ran	50	46	80%	$+0.002$	0.119	0.249	$+0.130$
PAI-1	50	49	71%	-0.001	0.188	0.164	-0.024
Integrin	50	48	58%	$+0.012$	0.004	0.012	$+0.008$
A_{2A}	50	47	70%	—	0.363	0.412	$+0.049$
Total	400	390	—	—			

E.7 Structural Validity of Langevin Refinement

To verify that Langevin refinement does not distort backbone geometry, Table 18 reports Ramachandran outliers, ω deviation, and bond-length deviation for 50 designs per target before and after refinement. Across all 8 OOD targets, both vanilla and refined designs contain zero Ramachandran outliers, and the mean ω deviation drops from $4.2°$ to $2.3°$, indicating that backbone dihedrals become more planar after refinement. The mean bond-length deviation rises from 0.000 to 0.005 Å, two orders of magnitude below typical covalent bond fluctuations and well within physically acceptable ranges. Together, these results indicate that Langevin refinement improves local geometry without compromising backbone validity.

Table 18. Backbone geometry for vanilla and Langevin-refined designs across 8 OOD targets. Zero Ramachandran outliers are observed in all cases.

	Rama outlier (%)		ω dev (°)		Bond dev (Å)	
Target	Van	Lang	Van	Lang	Van	Lang
CaM	0.0	0.0	4.0	4.5	0.000	0.014
ERα	0.0	0.0	4.5	2.0	0.000	0.006
BCL-2	0.0	0.0	4.3	1.4	0.000	0.001
MDM2	0.0	0.0	4.1	2.3	0.000	0.008
Ran	0.0	0.0	3.6	1.9	0.000	0.003
A_{2A}	0.0	0.0	3.9	1.9	0.000	0.004
PAI-1	0.0	0.0	4.7	2.0	0.000	0.003
Integrin	0.0	0.0	4.6	2.6	0.000	0.002
Mean	0.0	0.0	4.2	2.3	0.000	0.005

E.8 Langevin Step-Size Sensitivity

We sweep the Langevin step size η on 50 CaM designs under ensemble scoring (Table 19). Selectivity improves monotonically from $\bar{S}$=+0.506 at η=0.01 to $\bar{S}$=+0.579 at η=0.08, then plateaus, while Cα RMSD from the vanilla backbone stays below 0.14 Å throughout. We adopt η=0.04 as the default ($\bar{S}$=+0.559, RMSD 0.10 Å), a safe operating point that captures most of the achievable gain at negligible structural perturbation.

Table 19. Langevin step-size sweep on CaM under ensemble scoring. RMSD denotes Cα displacement from the vanilla backbone.

η	$\bar{S}$	$\Delta\bar{S}$	$S>0$	RMSD (Å)
— (vanilla)	+0.433	—	98%	—
0.01	+0.506	+0.073	100%	0.04
0.02	+0.533	+0.100	100%	0.06
0.04	+0.559	+0.126	100%	0.10
0.08	**+0.579**	**+0.146**	**100%**	0.12
0.10	+0.579	+0.146	100%	0.14

E.9 Boltz-2 Cross-Check

As a Q_θ-independent structural sanity check, we re-score the 50 vanilla designs per target with Boltz-2 and compare its $\Delta\text{ipTM}=\text{ipTM}_{\text{holo}}-\text{ipTM}_{\text{apo}}$ against the Q_θ selectivity S (Table 20). On 5 of 8 targets, mean ΔipTM is positive, indicating that Boltz-2 also assigns higher interface confidence to the holo state for the majority of designs; the strongest agreement with Q_θ at the design level appears on A_{2A} (ρ=+0.500) and CaM (ρ=+0.349), while the remaining targets yield small positive correlations. BCL-2 is the clearest disagreement, with $\Delta\text{ipTM}=-0.183$ and only 24% of designs above zero, suggesting that Boltz-2 and Q_θ rank holo/apo differently for this target.

Table 20. Boltz-2 ΔipTM as a Q_θ-independent structural cross-check across 8 OOD targets (n=50 vanilla designs per target). Q_θ scores use the 3 seeds. Per-target Spearman ρ is computed between Q_θ selectivity S and Boltz-2 $\Delta\text{ipTM} = \text{ipTM}_{\text{holo}} - \text{ipTM}_{\text{apo}}$.

Target	Mean ΔipTM	Frac > 0	Spearman ρ
A_{2A}	+0.055	50%	**+0.500**
CaM	+0.032	56%	**+0.349**
MDM2	+0.057	76%	−0.047
Ran	+0.045	58%	+0.081
PAI-1	−0.022	50%	+0.076
Integrin	−0.003	50%	+0.124
BCL-2	−0.183	24%	+0.087
ERα	−0.003	44%	+0.050

E.10 AlphaFold-3 Cross-Check

As an additional Q_θ-independent validation, we re-fold a subset of designs with AF3 in single-sequence mode (Table 21). On ALK and ERα, 100% of designs satisfy $\Delta\mathrm{ipTM} > 0$ under both vanilla and Langevin pipelines, with refinement leaving the mean essentially unchanged. BCL-2 is the exception (5% positive), which we attribute to its much larger binder-receptor length gap (20% vs. $\leq 5\%$) pushing AF3 outside its reliable regime.

Table 21. AlphaFold 3 ΔipTM in single-sequence mode. Positive values indicate holo preference. Rec. Δlen denotes the relative length difference between binder and receptor.

Target	**Guidance**	Mean ΔipTM	Frac. >0	n	Rec. Δlen
ALK	Vanilla	$+0.058$	100%	50	5%
	Langevin	$+0.057$	100%	50	
ERα	Vanilla	$+0.034$	100%	50	4%
	Langevin	$+0.034$	100%	50	
BCL-2	Langevin	-0.103	5%	19	20%

E.11 Rosetta Interface Analysis

To characterize physical interface quality independently of Q_θ, we compute Rosetta InterfaceAnalyzer metrics on 400 RFdiffusion designs across the 8 OOD targets, with repacked sidechains (Table 22). BCL-2 alone yields a clearly favorable mean interface energy ($\Delta G{=}-9.1$ REU) together with the largest buried surface (968 Å^2), while MDM2, A_{2A}, and CaM achieve modestly negative or near-neutral ΔG. Ran and Integrin are the weakest, with strongly positive ΔG values ($+25.4$ and $+39.1$ REU), consistent with their lower Q_θ Spearman ρ. The agreement at both ends (BCL-2 and MDM2 ranking high under both scorers, Ran and Integrin ranking low) provides an independent physical validation of Q_θ selectivity, even though Rosetta and Q_θ disagree on intermediate cases such as Integrin's high contact count despite an unfavorable ΔG.

Table 22. Rosetta InterfaceAnalyzer metrics for 400 vanilla RFdiffusion designs across the 8 OOD targets (50 per target).

Target	ΔG (REU)	ΔSASA (Å^2)	Contacts	H-bonds
BCL-2	-9.1	968	40.5	2.3
ALK	$+0.3$	927	33.8	4.0
MDM2	-1.0	606	25.1	2.2
A_{2A}	-0.9	512	19.1	1.4
ERα	$+3.8$	564	18.2	1.3
PAI-1	$+1.3$	464	16.7	1.4
Ran	$+25.4$	292	13.0	0.7
Integrin	$+39.1$	755	37.8	8.6

E.12 ProteinMPNN ΔNLL Analysis

Table 23 reports ProteinMPNN ΔNLL for 50 vanilla and 50 Langevin-refined designs per target. Two findings emerge. First, vanilla designs achieve significant holo preference on all 8 OOD targets (pooled $\Delta\mathrm{NLL}_{\mathrm{van}}{=}+0.188$, all $p{<}0.05$), confirming that holo preference is a population-level property of the generated backbones detectable without Q_θ. Second, Langevin refinement does not increase ΔNLL and reduces it on 5 of 8 targets. Since Q_θ score rises sharply under Langevin (Table 12) while ΔNLL falls, the two metrics cannot be measuring the same signal: Q_θ captures a geometric selectivity dimension orthogonal to ProteinMPNN's sequence-recovery likelihood, and neither serves as ground truth for the other. Independent structural validation of Q_θ comes from Boltz-2 ΔipTM (Table 20), which reaches per-target significance on A_{2A} and CaM.

E.13 Failure-Mode Analysis of Negative-Selectivity Designs

To locate where guided binder generators break down, we audit every design with negative TS-S2 seed-1024 selectivity,

$$S(Y) = Q_\theta(X^1, Y) - Q_\theta(X^0, Y) < 0,$$

Table 23. ProteinMPNN ΔNLL for vanilla and Langevin-refined designs. ProteinMPNN measures sequence-recovery likelihood given a backbone, a Q_θ-independent metric that probes a different dimension of binder–receptor compatibility than the geometric selectivity signal Q_θ optimizes. Positive ΔNLL indicates holo preference; p-values are from paired Wilcoxon signed-rank tests on per-design (vanilla, Langevin) pairs. †Apo-divergent targets on which ProteinMPNN's reference-state NLL is mis-specified on apo backbones (binder–apo NLL is depressed by training-distribution scarcity rather than by genuine binder preference); we interpret these values directionally only.

Target	$\Delta\text{NLL}_{\text{van}}$	$\Delta\text{NLL}_{\text{lang}}$	Δ	p
MDM2	$+1.050$	$+0.743$	-0.307	$<10^{-14*}$
BCL-2	$+0.550$	$+0.436$	-0.114	$<10^{-14*}$
CaM	$+0.290$	$+0.188$	-0.102	$<10^{-14*}$
PAI-1	$+0.120$	$+0.095$	-0.026	$8\times10^{-6\,*}$
Integrin	$+0.014$	$+0.016$	$+0.002$	0.020^{*}
ER$\alpha^{\dagger}$	-0.129	-0.054	$+0.075$	$<10^{-12*}$
Ran†	-0.169	-0.136	$+0.033$	$6\times10^{-5\,*}$
$\text{A}^{\dagger}_{2A}$	-0.223	-0.261	-0.040	$<10^{-9*}$
Mean (5 pos.)	$\mathbf{+0.405}$	$+0.296$	-0.109	—
Mean (all 8)	$+0.188$	$+0.128$	-0.060	—

corresponding to designs that the scorer predicts bind apo CaM more strongly than holo CaM. The pool comprises N=482 designs from 9 pipelines (two diffusion priors × five sampling schemes: vanilla, classifier-guided, TDS, SMC, and Langevin refinement). After Kabsch-aligning each design's receptor onto the 3CLN holo reference and propagating the same transform to the binder, we tag it with any of five non-exclusive failure modes: *too_short* ($<$ 50 residues), *wrong_binding_site* (binder centre-of-mass $>$ 25 Å from the canonical CaM peptide groove at residues 84–88, 124–125), *insufficient_interface* ($<$ 5 binder residues within 10 Å of the receptor), *steric_clash* (any inter-chain C_α pair $<$ 2.5 Å), and *apo_binding* ($Q_\theta(X^0, Y)$>0.3). The negative-S rate is just $10/482$ (2.1%), and Table 24 shows these ten designs are uniformly degenerate: every one is truncated below 50 residues, 40% are mislocalised away from the canonical groove, another 40% contain steric clashes, and 20% lack a sufficient interface, while only $1/10$ crosses the apo-binding threshold. Failures therefore stem from degenerate generation (truncated, mislocalised, or sterically infeasible binders) rather than from the scorer being deceived by genuine apo-selective designs. All ten negatives come from a single pipeline (PXDesign + Langevin); the other eight pipelines produce zero negative-S designs under this scorer.

Table 24. Failure-mode analysis of the 10 negative-selectivity designs out of 482 total designs from 9 guidance pipelines. Categories are not mutually exclusive: a single design may fall into several modes simultaneously.

Failure mode	Count	Fraction
too_short ($<$ 50 residues)	10/10	$\mathbf{100.0}\%$
wrong_binding_site ($>$ 25 Å)	4/10	40.0%
steric_clash ($<$ 2.5 Å inter-chain C_α)	4/10	40.0%
insufficient_interface ($<$ 5 residues at 10 Å)	2/10	20.0%
apo_binding ($Q_{\text{apo}} > 0.3$)	1/10	10.0%

E.14 Statistical Analysis

Langevin versus vanilla comparisons in Table 23 use paired Wilcoxon signed-rank tests with n=50 designs per target and a one-sided alternative that Langevin exceeds vanilla. Of 8 targets, 4 reach significance at p<0.05, namely CaM at p<0.001, ALK at p=0.012, ERα at p=0.033, and Ran at p=0.041. The remaining 4 targets show negligible differences with $|\Delta|$<0.03 and p>0.5, consistent with Langevin's 0.04 Å perturbations falling below ProteinMPNN's detection threshold. Overall, 238 of 350 vanilla designs across 7 targets show positive ΔNLL, significantly exceeding the 50% null expectation as assessed by a binomial test, confirming population-level holo preference.

F Efficiency Analysis

F.1 Inference Speed

Q_θ achieves 2.98 ± 0.09 ms per-complex latency in single-sample mode on an A100 GPU in Table 25. At batch size 16, amortized throughput reaches 4,531 complexes per second, enabling 10^5-candidate selectivity screening with two forward passes per candidate in under 45 seconds. RFdiffusion generates one CaM binder in approximately 2 min on an A100; PXDesign requires approximately 3 min. In-process guidance via classifier guidance and TDS adds approximately 17% compute overhead due to Q_θ gradient evaluation at each denoising step.

Across RFdiffusion methods, Langevin achieves the best selectivity per compute unit, improving selectivity with only 17% overhead relative to vanilla generation. For PXDesign, vanilla generation is the most cost-effective strategy: guidance methods add minimal selectivity gain over PXDesign's strong generative prior. PXDesign with Langevin is the only negative-selectivity result, indicating a fundamental mismatch between PXDesign's generation trajectory and the Langevin gradient.

Table 25. Computational cost per method on a single A100 80GB GPU.

Method	N	min/design	GPU-hr	$\bar{S}$	$\bar{S}$/GPU-hr
RFdiff (vanilla)	50	2.0	1.7	0.249	0.146
RFdiff + Classifier	50	2.5	2.1	0.247	0.118
RFdiff + Langevin	50	2.5	2.1	0.314	0.150
RFdiff + SMC	64	2.0	2.1	0.488	0.232
RFdiff + TDS	40	2.5	1.7	0.477	0.281
PXDesign (vanilla)	50	3.0	2.5	0.166	0.066
PXDesign + Classifier	50	4.0	3.3	0.175	0.053
PXDesign + Langevin	50	3.5	2.9	0.124	0.043
PXDesign + SMC	64	3.0	3.2	0.371	0.116
PXDesign + TDS	64	4.0	4.3	0.419	0.097
Proteina-CA (vanilla)	64	0.05	0.05	0.213	**4.26**
Proteina-CA + Classifier	64	0.08	0.09	0.249	2.77
Proteina-CA + Langevin	64	0.13	0.14	0.267	1.91
Proteina-CA + SMC	64	0.05	0.05	0.350	**7.00**
Proteina-CA + TDS	64	0.10	0.11	0.470	**4.27**