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,†]

[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

[†]Corresponding Author: pranam@upenn.edu
Code: https://huggingface.co/ChatterjeeLab/AlloGen

Abstract

Protein binder design commonly optimizes compatibility with a single receptor conformation, leaving preference over alternative functional states unresolved. To address this gap, we introduce **AlloGen**, a modular method that decouples backbone generation from state-selectivity scoring. We train its SE(3)-invariant interface graph transformer, Q_θ, through interface-quality regression followed by paired conformational discrimination. By comparing scores across states, we rerank candidates and guide frozen backbone generators without retraining. On a benchmark of 65 targets spanning 15 protein families, Q_θ achieved a mean Spearman correlation of 0.520 on eight held-out targets. All 15 evaluated generator–guidance combinations achieved positive predicted selectivity averaged over these targets, and RFdiffusion with Langevin refinement reached $\bar{S}_{\text{cons}} = 0.677$ on calmodulin (CaM). To test the predicted preference experimentally, we measured binding by bio-layer interferometry. Five designed peptides bound holo CaM, with the strongest measured K_D of 46.6 nM, and none showed detectable apo CaM binding under the tested conditions. Together, these results support computational prioritization of conformation-selective binders.

1 Introduction

Protein conformational transitions govern signaling, catalysis, and regulation across diverse protein families [15, 26, 39, 44]. In kinases, nuclear receptors, and GPCRs, these transitions connect conformations with distinct biological functions. Binder design therefore requires control over conformational preference as well as binding affinity [9, 22, 23, 40]. This capability is relevant to allosteric drug design, conformational biosensors, and synthetic biology switches [22, 24].

Generative models support protein binder design at both sequence and structural levels. At the sequence level, masked language models generate peptide binders conditioned on target sequence [6], contrastive language models enable de novo peptide design for conformationally diverse targets [3], and multi-objective discrete diffusion supports therapeutic peptide generation [4, 5, 35, 41]. Structure-conditioned methods extend these capabilities to binder geometry. RFdiffusion generates functional protein binders [43], PXDesign supports fast, modular binder design [29], and Proteina-ComplexA enables atomistic design through generative pretraining and test-time compute [12]. BoltzGen supports binder design across protein families [33], and BindCraft combines structure prediction with sequence optimization to produce experimentally validated binders [28]. In standard single-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

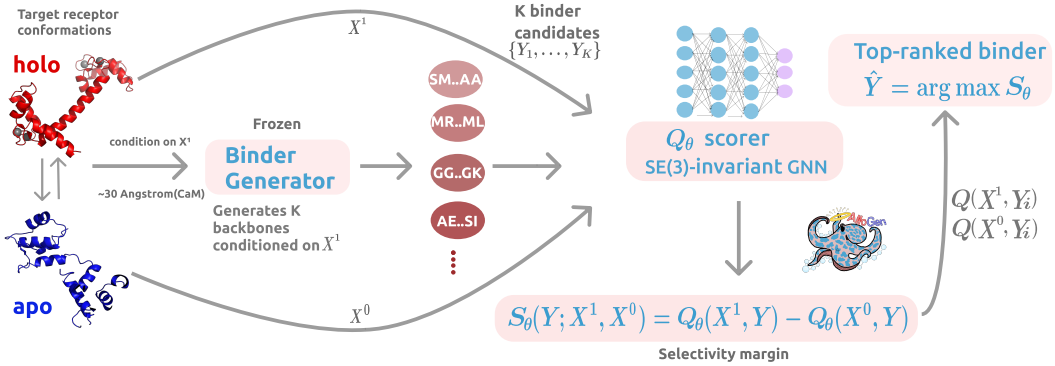


Figure 1: **AlloGen pipeline.** (*Left*) We use two receptor conformations: the goal state X^1 (holo, red) and the undesired state X^0 (apo, blue). (*Centre*) A frozen pretrained generator produces K binder backbone candidates conditioned on X^1. (*Right*) We score each candidate in both conformational frames with Q_θ and select the top candidate $\hat{Y}$ by the selectivity gap $\Delta Q = Q_\theta(X^1, Y) - Q_\theta(X^0, Y)$ through best-of-K reranking. DockQ regression followed by contrastive fine-tuning trains Q_θ independently of the backbone generator, enabling reuse across generators without retraining.

state settings, these workflows optimize compatibility with one receptor conformation, although a resulting binder may also bind an alternative state. State-selective design therefore requires an explicit comparison across conformations, which single-state interface scores can assess only through separate evaluations.

To address this gap, we introduce **AlloGen**, which decouples binder generation from selectivity evaluation (Figure 1). We train its SE(3)-invariant interface graph transformer, Q_θ, on paired receptor conformations to score each candidate in both states. The difference between these scores then guides candidate reranking, gradient-based refinement, and resampling with frozen generators [11, 14]. This separation allows us to apply the same scorer across generator architectures without retraining.

Our contributions are as follows:

1. **Problem formulation, benchmark, and scorer.** We develop a paired-state scoring task with: **(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 for interface-quality regression followed by conformational discrimination. On eight held-out targets, Q_θ achieved $\bar{\rho} = 0.520$, exceeding the evaluated contact-based proxies.
2. **Scoring controls and experimental validation.** We assess target specificity and apo rejection in the main text (Section 5.3), then compare selectivity scores with four Q_θ-independent metrics in the appendix: ProteinMPNN log-likelihood, Boltz-2 interface predicted TM-score (ipTM), AlphaFold 3 ipTM, and Rosetta InterfaceAnalyzer. BLI measurements on CaM directly test state-dependent binding in selected designs.
3. **End-to-end selectivity-guided design.** All 15 generator–guidance combinations achieved positive mean predicted selectivity over the eight test targets. TDS and SMC had the highest cross-target means for each generator. In a CaM case study, RFdiffusion combined with Langevin refinement reached $\bar{S}_{\text{cons}} = +0.677$ at 88% design success rate, and best-of-K reranking reached $\bar{S} = +0.885$ at $K = 10$. Five selected peptides showed measurable holo CaM binding with no detectable apo binding (Section 5.5).

2 Related Work

Structure-Based Binder Design. Structure-based generative models design binders conditioned on target geometry. RFdiffusion [43] generates binder backbones conditioned on a fixed target surface via backbone diffusion, followed by sequence design with ProteinMPNN [10]. PPDiff [32] and PXDesign [29] extend this to joint sequence-structure generation, while Proteina-ComplexA [12] 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 [3–8, 35–37, 41]. Standard structure-conditioned design settings optimize compatibility with one receptor conformation. AlloGen adds an explicit two-state score for reranking or guiding their backbone outputs.

Protein–Protein Interface Scoring. Physics-based methods such as FoldX [30] and PRODIGY [48] estimate binding free energies, and inverse folding models such as ProteinMPNN [10] and ESM-IF [17] serve as zero-shot affinity proxies. GNN-based scorers including GNN-DOVE [42] and DeepRank-GNN-esm [47] learn interface quality from structural features for docking re-ranking. Comparing such scores across receptor states assesses conformational preference. To train explicitly for this comparison, we combine interface-quality regression with paired contrastive fine-tuning of Q_θ.

Conformational Selectivity: Sampling, Design, and Negative Design. Conformation-selective design requires structural characterization of alternative states and binders that discriminate between them. On the characterization side, AlphaFlow [18] uses flow matching to generate structural ensembles, AFsample2 [21] diversifies predictions via MSA subsampling, and AlphaFold 3 [1] extends structure prediction to biomolecular complexes. Designing binders that distinguish these states further requires objectives for molecular recognition. In this setting, negative design has a long history in computational protein engineering [13, 34], and multi-state design in Rosetta [16] optimizes sequences across multiple conformations. Extending these sequence-level objectives to backbone generation requires a differentiable score for receptor–binder geometry. Contrastive objectives inspired by InfoNCE [27] have been applied to molecular representation learning, and classifier guidance [11] and TDS [46] support gradient-based steering and particle reweighting, respectively. Concurrent work by Xue et al. [49] targets state-specific peptide design for GPCRs via a conformation-specific folding filter, and ProDiT [19] introduces multi-state structure diffusion. AlloGen combines paired contrastive scoring with guidance of three generators and evaluates transfer on eight held-out targets from a 65-target benchmark spanning 15 families.

3 Preliminaries

3.1 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 4.3 enforces this conditional behaviour through paired contrastive supervision. 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 4.4 and $\mathcal{N} = \{X^0\}$ is the set of undesired conformations.

3.2 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 [20]. 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, are SE(3)-invariant.

3.3 DockQ as Interface Quality Proxy

DockQ [2] 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 use this geometric measure to supervise Phase 1 training (Section 4.3), initializing the scorer to assess receptor–binder fit before paired conformational discrimination in Phase 2. A DockQ value > 0.23 is the conventional threshold for an acceptable docking model [2].

4 Method

4.1 Interface Graph Construction

To score local receptor–binder geometry, we construct a sparse interface graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for each complex (X, Y). The node set $\mathcal{V}$ contains residues from both X and Y with at least one inter-chain C_α contact within an 8 Å cutoff, 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 [25] 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.

4.2 State-Selectivity Scorer Q_θ

Using these invariant features, we assess receptor–binder fit with a dense edge-biased graph transformer [38], Q_θ. Comparing its outputs across conformations yields the selectivity score. 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 $\mathbf{h}_{\mathrm{pool}}^{(L)} = [\mathrm{mean}_i\, \mathbf{h}_i^{(L)};\, \max_i \mathbf{h}_i^{(L)}]$ and passed through an MLP with sigmoid activation to produce:

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

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 A.

4.3 Two-Phase Training of Q_θ

To learn both interface quality and conformational preference, we train Q_θ in two phases. We first establish a representation of receptor–binder geometry, then optimize paired state discrimination.

Phase 1: Interface Quality Regression. We first train Q_θ to predict interface quality as measured by DockQ [2]:

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

The training data include native holo complexes, apo mismatches, rigid-body decoys, and hard negatives described in Section 5.1. Together, these examples initialize the scorer to distinguish compatible interfaces from mismatches before Phase 2.

Phase 2: Selectivity Fine-Tuning. Building on this initialization, we fine-tune Q_θ to distinguish conformational preference from target-specific biases. For paired triplets (X^+, X^-, Y) where $X^+ = X^1$ and $X^- = X^0$ for the same binder Y, we minimize a multi-negative InfoNCE loss [27] 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 apo term trains conformational discrimination, while the cross-target term discourages high scores based on receptor identity alone.

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

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

where τ is a temperature hyperparameter reported in Appendix A. The regression loss $\mathcal{L}_q$ is dropped in Phase 2 to avoid conflicting gradient signals; the curriculum comparison is reported in Figure 2.

Backbone-geometry augmentation. Training complexes have known sequences, but generated binder backbones lack sequence identity and sidechain coordinates. To train Q_θ for this backbone-only inference setting, we mask binder-side sequence features with probability p_{drop}:

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

where $\mathbf{0}_{\mathrm{AA}}$, $\mathbf{0}_\chi$, and $\mathbf{0}_{\mathrm{ESM}}$ replace the amino acid identity, sidechain torsion angles, and ESM-2 embeddings with zeros, respectively. All backbone torsions and edge features are preserved, so scoring can use backbone geometry when binder sequence features are unavailable. The same masking decision is applied consistently to both X^+ and X^- within each training pair.

4.4 Selectivity-Guided Binder Design

To use the trained scorer for design, we sample K binder candidates $Y^{(1)}, \ldots, Y^{(K)}$ from the frozen generator $p_\psi(\cdot \mid X^1)$ and evaluate each against X^1 and X^0 by rigid placement onto their aligned structures. We define the *selectivity margin* as:

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

where $\mathcal{N} = \{X^0\}$ is the set of undesired conformations. S_θ is positive when the holo score exceeds the apo score, zero when the two scores are equal, and negative when the apo score is higher. 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 [10] on the selected backbone.

Gradient-based guidance. Beyond reranking, we use four guidance strategies to update coordinates or reweight candidates with the trained scorer. *Langevin backbone refinement* [31, 45] 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; this deterministic variant uses no injected Langevin noise and operates on

fully denoised backbones. *Classifier guidance* [11] uses holo-score gradients during denoising to favor holo compatibility; gradient reliability decreases at high noise levels. *Twisted diffusion sampling* (TDS) [46] reweights diffusion particles using holo scores without gradient-based coordinate updates. *Sequential Monte Carlo* (SMC) [46] iteratively resamples complete trajectories by S_θ across multiple generation rounds, enriching the candidate pool for predicted selectivity. Detailed formulations and computational cost are provided in Appendix A.3 and E.

5 Experiments

We first tested interface-quality ranking and conformational discrimination on held-out targets (Sections 5.2–5.3), then evaluated the scorer for binder design (Section 5.4) and measured state-dependent binding in selected CaM designs (Section 5.5).

5.1 Experimental Setup

Target selection. We curated 65 two-state proteins spanning 15 protein families and diverse conformational mechanisms (Table 6). We required: (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 the remaining 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 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 6). 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, excluding test-target generator decoys from training. 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 with DockQ and the bounded score gap $S(Y) = Q_\theta(X^1, Y) - Q_\theta(X^0, Y)$. Tables and plots use this bounded gap; S_θ in Eq. (8) denotes the distinct logit-space objective. Consensus selectivity $\bar{S}_{\text{cons}}$ averages the bounded gap across designs and three independently trained checkpoints. We use ProteinMPNN ΔNLL, Boltz-2 ΔipTM, and AlphaFold 3 ΔipTM as independent computational measures of state preference. Implementation details are in Appendix A.

5.2 Scorer Performance

To test transfer to unseen targets, we evaluated Q_θ on the eight held-out test targets. Q_θ attained $\bar{\rho}$=$\mathbf{0.520 \pm 0.010}$ (3 random seeds for training), with positive correlation on all 8 OOD targets and 4 of 8 exceeding ρ=0.5 (Figure 2).

Table 1: 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	$\mathbf{0.469} \pm .034$
BCL-2	132	$+0.160$	-0.050	-0.082	-0.039	$\mathbf{0.667} \pm .014$
CaM	96	$+0.199$	-0.316	$+0.014$	$+0.059$	$\mathbf{0.474} \pm .061$
ERα	72	$+0.551$	-0.292	-0.036	$+0.086$	$\mathbf{0.664} \pm .013$
Integrin	60	-0.019	-0.195	$+0.053$	$+0.076$	$\mathbf{0.366} \pm .085$
MDM2	143	$+0.163$	-0.050	-0.130	-0.037	$\mathbf{0.589} \pm .036$
PAI-1	156	$+0.058$	$+0.062$	-0.146	-0.043	$\mathbf{0.421} \pm .045$
Ran	2268	$+0.264$	-0.270	-0.138	-0.018	$\mathbf{0.511} \pm .065$
Mean		$+0.143$	-0.154	-0.070	-0.005	$\mathbf{0.520} \pm .010$

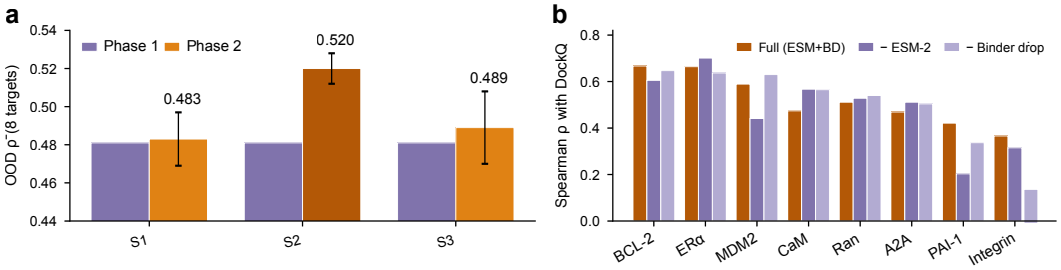


Figure 2: Q_θ **selectivity performance. (a)** Q_θ scoring performance ablation by data augmentation strategies and two phases; **(b)** Q_θ scoring performance ablation per target by different features.

Training curriculum, data augmentation, and feature design. We then used ablations to assess the contributions of training, augmentation, and features (Figure 2). *(i) Two-phase curriculum.* Phase 2 contrastive fine-tuning improved over Phase 1 regression across all augmentation configurations (Figure 2a). *(ii) Data augmentation.* The full Augmented configuration yielded the largest single gain ($\Delta\bar{\rho}$=$+0.037$), driven by GenDecoys, whose synthetic binder geometries broadened interface coverage beyond rigid-body and FastRelax decoys. *(iii) Feature design.* In per-target ablations (Figure 2b), removing ESM-2 reduced $\bar{\rho}$ across most targets, and disabling binder dropout further reduced performance, with the largest drops on harder OOD targets (Integrin, PAI-1, A_{2A}). The combined Full configuration (ESM + BD) performed best on all 8 OOD targets. InfoNCE batch-size ablations are in Appendix C, Table 9.

5.3 Conformational Preference of Q_θ

To extend the ranking evaluation to state preference, we compared Q_θ with contact-based proxies and assessed target specificity and holo–apo score differences.

Comparison against energy-based baselines. We compared Q_θ with three contact-based proxies on the eight test targets: PRODIGY [48], interface residue count, and cross-chain edge density (Table 1). Their mean correlations with DockQ were $+0.143$, -0.154, and -0.070, respectively, compared with 0.520 for Q_θ. PRODIGY retained positive correlation on several targets, including ERα (0.551). Across the benchmark, Q_θ ranked interface quality more accurately than these proxies.

Cross-target specificity. To test whether high scores reflected target-specific compatibility, we scored each target's 50 vanilla binders against all eight test receptors (Figure 3a). The ratio of diagonal to off-diagonal scores was 19.8, supporting target-specific compatibility. We next tested whether this compatibility differed between conformations of the same target.

Per-target apo rejection. For this within-target comparison, we scored 50 vanilla designs per target against both holo and apo structures (Figure 3b). Single-seed scoring favored holo on seven of eight targets; MDM2 and BCL-2 had positive gaps for all designs (Appendix C.4). The three-seed

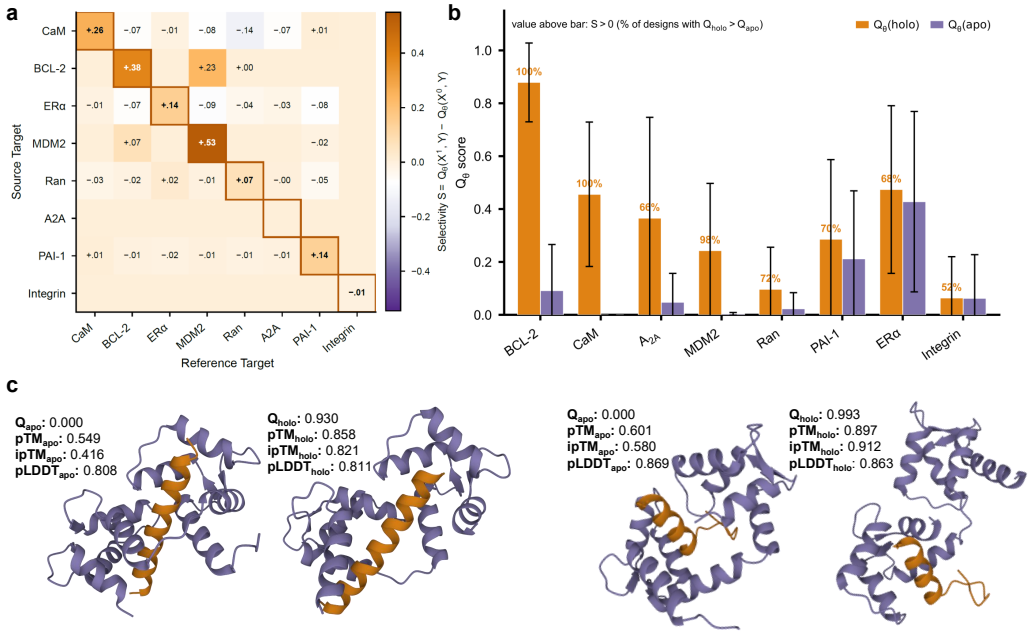


Figure 3: Q_θ **conformational selectivity and CaM selectivity design. (a)** Cross target selectivity matrix between Source Target (binders generated for) and the Reference Target (scored against). Zero values (0.00) are omitted. **(b)** Holo vs. apo Q_θ scores for 50 vanilla designs per target; **(c)** 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).

ensemble yielded 100% holo preference for CaM and 98% for MDM2. Integrin remained difficult, with an ensemble mean gap of $+0.001$ and 52% positive gaps.

5.4 End-to-End Binder Design

To test how these scoring properties translated to design, we combined three generators, RFdiffusion [43], PXDesign [29], and Proteina-ComplexA [12], with vanilla generation and four guidance strategies: classifier guidance, SMC, TDS, and gradient ascent (Langevin) refinement. We evaluated the resulting 15 generator$\times$guidance combinations on the 8 held-out OOD test targets. Each design was scored by an ensemble of three independently trained Augmented-S2 Q_θ checkpoints.

Generator $\times$ guidance benchmark across OOD targets. Cross-target means favored TDS and SMC for each generator (Table 2). Langevin refinement improved RFdiffusion (0.249 to 0.314) and Proteina-ComplexA (0.213 to 0.267), but reduced PXDesign selectivity (0.166 to 0.124). BCL-2 had the highest target mean (0.781), whereas Integrin, ERα, and PAI-1 had lower means. ERα combined strong scoring correlation (0.664) with weak design selectivity (0.043), suggesting that candidate generation also constrained performance. Full target-level results are in Table 2. Candidate counts and sampling costs differed across methods (Appendix D). Target means across all 15 combinations were CaM 0.446, BCL-2 0.781, ERα 0.043, A_{2A} 0.242, MDM2 0.502, PAI-1 0.127, Ran 0.263, and Integrin 0.038 (grand mean 0.305).

End-to-end CaM-selective design. For a detailed assessment of the complete workflow, we focused on CaM, whose Ca^{2+}-dependent rearrangement exposes a peptide-binding cleft. All 15 combinations had positive mean consensus selectivity on CaM. RFdiffusion with Langevin refinement reached $\bar{S}_{\text{cons}} = 0.677$ with an 88% joint computational success rate, compared with 100% success for vanilla RFdiffusion (Figure 4, Table 13). Gradient-noise analysis was consistent with reduced reliability of classifier guidance at high noise (Appendix D.4).

With Q_θ-based reranking of a fixed vanilla CaM pool, best-of-5 and best-of-10 selection reached $\bar{S} = 0.787$ and 0.885, respectively (Table 20). Reranking adds scoring cost to an existing pool,

Table 2: Complete target-level consensus selectivity for all 15 generator–guidance combinations, grouped by generator. Mean averages the eight targets. Bold indicates the highest value within each generator and target, or in the Mean column.

Guidance	CaM	BCL-2	ERα	A_{2A}	MDM2	PAI-1	Ran	Integrin	Mean
RFdiffusion									
Vanilla	+0.455	+0.787	+0.046	+0.318	+0.238	+0.073	+0.074	+0.001	+0.249
Classifier	+0.427	+0.741	+0.054	+0.332	+0.271	+0.054	+0.091	+0.006	+0.247
Langevin	**+0.677**	+0.806	+0.106	+0.377	+0.366	+0.061	+0.108	+0.008	+0.314
SMC	+0.510	+0.841	**+0.325**	+0.760	+0.641	**+0.389**	+0.424	+0.013	**+0.488**
TDS	+0.367	**+0.880**	+0.050	**+0.924**	**+0.769**	+0.262	**+0.485**	**+0.079**	+0.477
PXDesign									
Vanilla	+0.517	+0.560	-0.031	+0.023	+0.208	-0.000	+0.033	+0.015	+0.166
Classifier	+0.521	+0.561	-0.029	+0.037	+0.227	-0.000	+0.056	+0.024	+0.175
Langevin	+0.022	+0.568	-0.027	+0.048	+0.262	-0.000	+0.084	+0.033	+0.124
SMC	**+0.545**	+0.868	**+0.117**	+0.377	+0.590	-0.000	+0.446	+0.027	+0.371
TDS	+0.514	**+0.969**	+0.029	**+0.445**	**+0.613**	-0.001	**+0.601**	**+0.185**	**+0.419**
Proteina-ComplexA									
Vanilla	+0.338	+0.774	-0.000	-0.006	+0.506	+0.053	+0.081	-0.041	+0.213
Classifier	+0.432	+0.805	-0.000	-0.004	+0.567	+0.064	+0.128	-0.002	+0.249
Langevin	+0.429	+0.826	-0.000	-0.003	+0.598	+0.069	+0.204	-0.011	+0.267
SMC	**+0.565**	+0.836	-0.000	-0.000	+0.794	+0.123	+0.469	+0.017	+0.350
TDS	+0.374	**+0.898**	+0.000	-0.000	**+0.883**	**+0.757**	**+0.662**	**+0.190**	**+0.470**

whereas refinement updates individual candidates. Two predicted CaM complexes had holo scores of 0.930 and 0.993 with near-zero apo scores (Figure 3c). To assess structural changes during refinement, we checked backbone geometry across eight targets. We detected no Ramachandran outliers after refinement and measured a mean bond-length deviation of 0.005 Å (Table 15). Changes in native-contact recovery ranged from -0.004 to 0.168 on the six targets with references (Table 21).

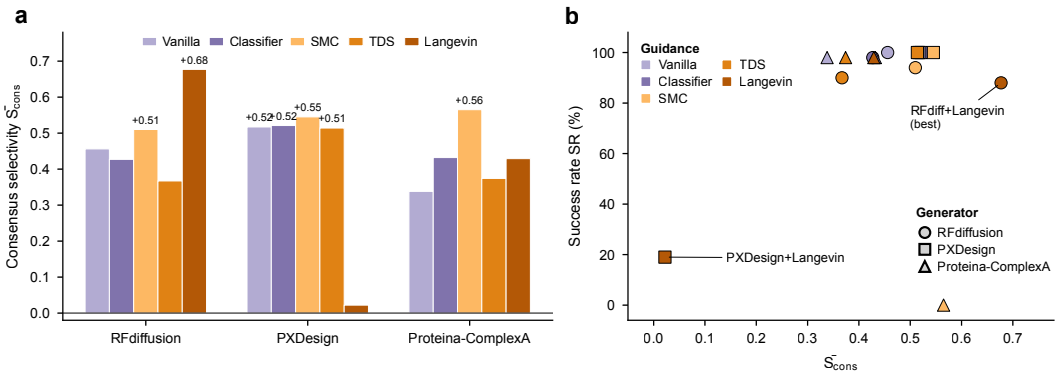


Figure 4: **Generation benchmark on CaM. (a)** Consensus selectivity $\bar{S}_{\text{cons}}$ across 15 generation $\times$ guidance approaches. **(b)** Selectivity vs. design success rate (designable $\times$ selective) across all generator $\times$ guidance combinations.

Independent computational cross-checks. To assess the designs beyond the optimization scorer, we compared Q_θ with independent computational metrics. Boltz-2 ΔipTM correlated with Q_θ selectivity on A_{2A} ($\rho = 0.500$) and CaM ($\rho = 0.349$), but correlations were weak on the other targets (Table 25). AlphaFold 3 favored holo on the training target ALK and held-out ERα, but favored apo on most BCL-2 designs (Table 26). Rosetta interface energies were most favorable for BCL-2 (-9.1 REU) and least favorable for Integrin (39.1 REU) in the separate panel reported in Table 27. ProteinMPNN gave positive mean ΔNLL on five of eight targets, and refinement reduced this metric on five targets despite increasing Q_θ (Table 23). Agreement varied across these computational

metrics, motivating direct binding measurements. Additional structural checks and failure analysis are in Appendix D.

5.5 Experimental Validation on Calmodulin

To test whether selected designs bound CaM preferentially under holo conditions, we evaluated ten designed peptides, comprising nine candidates and one lower-scoring negative control, together with an M13 positive control. Candidates came from RFdiffusion and Proteina-ComplexA variants and were selected using the predicted gap $\Delta_q = Q_\theta(X^1, Y) - Q_\theta(X^0, Y)$. BLI measurements compared $CaCl_2$-supplemented holo CaM with EGTA-treated apo CaM in duplicate (Figure 5; Appendix F).

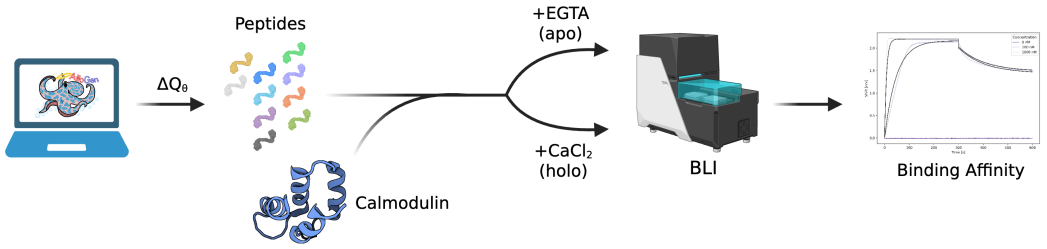


Figure 5: BLI workflow for the selected CaM peptide panel. Holo and apo assay conditions used $CaCl_2$ and EGTA, respectively. Measurements were performed in duplicate and fitted with a 1:1 binding model. The schematic summarizes the assay procedure.

Five designed peptides showed detectable holo CaM binding, including RFdiffusion designs 944 and 13 with fitted K_D values of 46.6 and 86.5 nM. The other positive designs had reported values of 413 nM, $> 1.00\,\mu$M, and $1.06\,\mu$M (Table 3). None of the designed peptides showed detectable apo CaM binding under the tested conditions. M13 bound holo CaM, and the designated negative control bound neither state. Four higher-scoring candidates also showed no detectable binding, including one with $\Delta_q = 0.993$. These measurements established state-dependent binding in selected CaM designs.

Table 3: BLI results for all ten designed peptides and M13. RF: RFdiffusion; Pro: Proteina-ComplexA; V: vanilla; Lg: Langevin; SM: SMC. The number is the original design or particle identifier. NB: no detectable binding under the tested conditions. M13 bound holo CaM without a reported fitted K_D.

Generator	Design	Holo K_D (nM)	Apo K_D	Δ_q
RF/V	944	46.6	NB	0.413
RF/V	13	86.5	NB	0.450
Pro/SM	0273	413	NB	0.932
RF/V	657	> 1000	NB	0.930
RF/Lg	25	1060	NB	0.930
Pro/V	0145	NB	NB	0.993
RF/Lg	330	NB	NB	0.905
Pro/V	0376	NB	NB	0.688
Pro/SM	0954	NB	NB	0.912
Pro/V	0105 (negative)	NB	NB	0.228
M13	Positive control	Bound	NB	–

6 Conclusion

We introduce AlloGen to connect differential state scoring with binder reranking and guidance across frozen generators. The benchmark supports transfer to held-out targets, with guidance gains that depend on the generator and target. In selected CaM designs, BLI measurements confirmed five peptides with detectable holo binding and no detectable apo binding under the tested conditions. Extending this approach to broader target validation, richer apo decoys, and score calibration will help establish its scope for conformation-selective binder design (Appendix G).

Acknowledgments and Disclosure of Funding

We thank the experimental team of the Chatterjee Laboratory for assistance with peptide design validation and helpful discussions throughout this project. We especially thank Adaptyv Bio for performing bio-layer interferometry measurements and assisting with experimental characterization of candidate binders. We thank Lauren Hong for providing the model logo.

This research was supported by a grant from the High-throughput Institute for Discovery (HIT-ID) at the University of Pennsylvania to the laboratory 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.

Competing Interests

P.C. is a co-founder of Gameto, Inc., UbiquiTx, Inc., AtomBioworks, Inc., and Kosha Sciences, Inc., and advises companies involved in biologics and therapeutic development. P.C.'s interests are reviewed and managed by the University of Pennsylvania in accordance with its conflict-of-interest policies. The remaining authors declare no competing interests.

References

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

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

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

[4] H. Cao, H. Shi, C. Wang, S. J. Pan, and P.-A. Heng. 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`.

[5] H. Cao, A. Pal, S. Tang, Y. Zhang, J. Zhang, P.-A. Heng, and P. Chatterjee. 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`.

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

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

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

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

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

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

[12] K. Didi, Z. Zhang, G. Zhou, D. Reidenbach, Z. Cao, S. Cha, T. Geffner, C. Dallago, J. Tang, M. M. Bronstein, M. Steinegger, E. Kucukbenli, A. Vahdat, and K. Kreis. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[26] R. Nussinov. Introduction to protein ensembles and allostery, 2016.

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

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

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

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

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

[32] Z. Song, T. Li, L. Li, and M. R. Min. 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.

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

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

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

[36] S. Tang, Y. Zhang, A. Tong, and P. Chatterjee. Gumbel-softmax score and flow matching for discrete biological sequence generation. In *ICLR 2025 GEM Workshop*, 2025.

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

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

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

[40] R. Vijayan, P. He, V. Modi, K. C. Duong-Ly, H. Ma, J. R. Peterson, R. L. Dunbrack Jr, and R. M. Levy. 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.

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

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

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

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

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

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

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

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

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

A Implementation Details

A.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 $\gamma = 1.5$.

A.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 4.

Table 4: 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.

A.3 Guidance Formulations

The strategies use either gradient updates of binder backbone coordinates $\mathbf{x} = \{\mathbf{p}_i\}_{i \in Y}$ or score-based resampling. The implementations below distinguish the holo-score objective used by classifier guidance and TDS from the two-state margin used by refinement and SMC.

Langevin backbone refinement. [31, 45] 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. [11] 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). [46] 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 $\mathrm{ESS} = (\sum_n w_n)^2 / \sum_n w_n^2$ drops below $N/2$.

Sequential Monte Carlo (SMC). [46] 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.

B 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 $\max(0, 1 - d_{\mathrm{RMSD}}/4)$, 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 6). 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.

Table 5: Dataset configurations and augmentation inventory. The target partition is 51/6/8 for training/validation/test in both configurations. The augmentation sources and configuration totals are listed separately.

Source	Baseline	Augmented
FastRelax negatives	959	959
Generator decoys	–	4,862
Cross-family negatives	–	254
Conformational decoys	–	1,075

Setting	Composition	Samples
S1 (Base)	Native complexes, rigid-body decoys, and FastRelax negatives	34,751
S2 (Full)	S1 plus generator, cross-family, and conformational decoys	39,867
S3 (Partial)	S1 plus cross-family and conformational decoys	36,080

Table 6: Training targets by family (51 total). CaM belongs to the test partition and ALK to training. Validation and test target identities are listed in Table 7.

Family (total targets)	Training targets
Kinase (9)	ABL1, **ALK**, Aurora A, BRAF, CDK2, EGFR, GRK2, SRC, SRC-SH2
Small GTPase (6)	Arf1, CDC42, KRAS, RhoA
Nuclear Receptor (5)	AR, GR, PPARγ, RXRα
GPCR/Ion Ch. (6)	CFTR, KcsA, nAChR
Phosphatase (4)	Calcineurin, PP2A, PTP1B, SHP2
Protease (6)	HIV-PR, MMP-2, M^{pro}, Thrombin, USP7
BCL-2 (2)	BCL-xL
14-3-3 (2)	–
Ca sensor (3)	S100B, Troponin C
G protein (2)	CheY, Gα_s
Chaperone/Enzyme (6)	ATCase, CypA, DHFR, Enolase, Hsp70, RNase A
Epigenetic (3)	BRD4, WDR5
Trafficking (5)	ABCB1, Importin-β, Maltose BP, p97, VHL
Cell adhesion (3)	CTLA-4, Integrin β_3
Signaling (3)	Gelsolin, Hemoglobin

Table 7: Held-out validation (6) and test (8) targets for the target-level split. The test partition contains CaM and excludes ALK.

Family	Validation	Test
Small GTPase (6)	Rac1	Ran
Nuclear Receptor (5)	–	ERα
GPCR/Ion Ch. (6)	β_2AR, μ-opioid	A_{2A}
Protease (6)	Caspase-3	–
BCL-2 (2)	–	BCL-2
14-3-3 (2)	14-3-3, 14-3-3σ	–
Ca sensor (3)	–	**CaM**
Epigenetic (3)	–	MDM2
Cell adhesion (3)	–	Integrin
Signaling (3)	–	PAI-1

C Extended Q_θ Results

All results in this section use the S2 configuration with the target split (Table 6), ESM-2 encoder, and binder-dropout rate p=0.3, consistent with the experimental setup in Section 5.2. The suffix S1/S2/S3 denotes the dataset augmentation configuration as defined in Table 5.

C.1 Per-Target Selectivity Performance

Phase 2 improved mean scoring correlation from 0.481 to 0.520 (Table 8). Gains were largest on Integrin (0.166) and A_{2A} (0.170), followed by MDM2 (0.037) and PAI-1 (0.027). Correlation decreased on BCL-2 (-0.049), ERα (-0.025), and CaM (-0.017), and was unchanged on Ran. The mean gain therefore reflected improvements on some targets and decreases on others.

Table 8: 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	$\mathbf{0.520} \pm 0.010$

C.2 Phase 2 Batch-Size Ablation

To assess sensitivity to the number of contrastive negatives, we varied the Phase 2 batch size; all three InfoNCE settings exceeded the Phase 1-only result. Batch size 256 had the highest mean correlation (0.530), followed by 512 (0.500) and 64 (0.496) (Table 9). Both negative diversity and optimization behavior varied with batch size, leaving their individual contributions unresolved.

Table 9: 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	**0.530**
InfoNCE, BS=512	0.500
InfoNCE, BS=64	0.496
Phase 1 only (MSE)	0.489

C.3 Scores Along Interpolated Conformations

To examine score variation between the endpoint states, we scored CaM binders against 11 interpolated receptor structures between apo and holo ($\tau = 0$ to 1). All ten vanilla designs had positive Spearman correlation between interpolation position and score, with mean $\rho = 0.518$ (Table 10). Langevin-refined designs had earlier score onset ($\tau_{50\%} = 0.4$–0.9, compared with 0.8–1.0). These correlations describe overall score trends along geometric interpolations; strict monotonicity and physical transition paths were not established.

Table 10: CaM interpolation analysis: 20 refined and 10 vanilla designs, each scored on 11 frames. The positive fraction refers to Spearman correlation with interpolation position, not strict monotonicity.

Set	Mean ρ	$\rho > 0$	Best ρ	$\tau_{50\%}$
Vanilla (10)	$+\mathbf{0.518}$	100% (10/10)	0.745	0.8–1.0
Langevin (20)	$+0.246$	80% (16/20)	0.891	0.4–0.9

C.4 Apo Rejection Analysis

Under single-seed scoring, seven of eight targets had positive mean holo–apo gaps, and all MDM2 and BCL-2 designs had positive gaps (Table 11). Integrin had a marginal mean apo preference ($\Delta Q = -0.009$) and 40% positive gaps. Under the three-seed ensemble, CaM had mean holo and apo scores of 0.456 and 0.000, with 100% positive gaps.

Table 11: 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%

C.5 Multi-Seed Scoring Robustness

In the additional CaM scoring comparison, vanilla and Langevin gaps varied little across three training seeds (Table 12). These seed identifiers and scores are reported separately from the consensus generation benchmark in Table 13.

Table 12: 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	$+0.255$	$+0.250$	$+0.263$	$\mathbf{+0.256}$	0.005

D Extended Design Results

D.1 Generation Benchmark Performance

We evaluated CaM design using RFdiffusion [43], PXDesign [29], and Proteina-ComplexA [12], each with vanilla sampling and four guidance strategies. Selectivity was evaluated with three independently trained scorer checkpoints (Table 13); structural and sequence metrics are listed in Table 14.

Table 13: CaM design selectivity and computational success rates. RF: RFdiffusion; PX: PXDesign; Pro: Proteina-ComplexA. All S values are bounded score gaps. Des: designable fraction; SR: reported fraction satisfying designability and positive selectivity. Structural metrics are listed in Table 14.

Method	N	$\bar{S}_{\text{cons}}$	Top-5 $\bar{S}$	$S > 0$	Des	SR
RF/Langevin	50	$\mathbf{+0.677}$	$\mathbf{0.993}$	100%	88%	**88%**
RF/SMC	64	$+0.510$	0.941	100%	94%	94%
RF/Vanilla	50	$+0.456$	0.932	100%	100%	100%
RF/Classifier	50	$+0.427$	0.891	100%	98%	98%
RF/TDS	40	$+0.367$	0.818	100%	90%	90%
PX/SMC	64	$+0.545$	0.851	100%	100%	100%
PX/Classifier	50	$+0.521$	0.891	100%	100%	100%
PX/Vanilla	50	$+0.517$	0.870	100%	100%	100%
PX/TDS	64	$+0.514$	0.851	100%	100%	100%
PX/Langevin	50	$+0.022$	0.424	24%	80%	19%
Pro/SMC	52	$\mathbf{+0.565}$	$\mathbf{0.992}$	**100%**	0%	0%
Pro/Classifier	52	$+0.432$	0.975	98%	100%	98%
Pro/Langevin	52	$+0.429$	0.975	98%	100%	98%
Pro/TDS	52	$+0.374$	0.981	98%	100%	98%
Pro/Vanilla	52	$+0.338$	0.942	98%	100%	98%

Table 14: Structural and sequence metrics for the CaM designs in Table 13. ipTM is from Boltz-2; scTM is structural self-consistency; NLL is the ProteinMPNN holo-state negative log-likelihood. Proteina-ComplexA generates full-atom complexes with co-designed sequences.

Method	ipTM	scTM	$\mathrm{NLL}_{\mathrm{holo}}$
RF/Langevin	0.742	0.806	**1.39**
RF/SMC	0.733	0.828	3.60
RF/Vanilla	**0.769**	0.820	3.59
RF/Classifier	0.715	0.826	3.54
RF/TDS	0.766	**0.835**	3.66
PX/SMC	0.710	0.698	3.71
PX/Classifier	0.757	0.684	3.74
PX/Vanilla	0.736	0.690	3.74
PX/TDS	0.762	0.690	3.74
PX/Langevin	0.639	0.664	1.58
Pro/SMC	0.420	0.340	1.38
Pro/Classifier	0.399	0.597	1.88
Pro/Langevin	0.381	0.596	1.88
Pro/TDS	0.376	0.592	1.86
Pro/Vanilla	0.387	0.601	1.80

D.2 Structural Validity of Langevin Refinement

To check whether refinement altered local backbone geometry, we compared 50 designs per target before and after refinement (Table 15). Both sets had zero detected Ramachandran outliers across eight targets. Mean ω deviation decreased from $4.2°$ to $2.3°$, while mean bond-length deviation increased from 0.000 to 0.005 Å. These measurements characterize local geometry; global stability and foldability require separate assessment.

Table 15: Backbone geometry for vanilla and Langevin-refined designs across 8 OOD targets. No Ramachandran outliers were detected in either set.

	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

D.3 Reranking vs. Gradient Refinement

We compared reranking and refinement on a CaM pool with mean vanilla gap 0.456 and 100% positive gaps (Table 16). Best-of-5 and best-of-10 selection reached 0.787 and 0.885, respectively, compared with 0.677 for Langevin refinement. All gaps were evaluated with the same Q_θ scorer used for selection or optimization.

Table 16: 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)	$\mathbf{+0.677}$	$\mathbf{100}\%$

D.4 Gradient Reliability Under Noise

To assess why guidance during denoising differed from refinement, we measured the cosine similarity between Q_θ gradients computed on clean and noise-perturbed backbones as a function of Gaussian noise scale σ. Cosine similarity dropped from 0.75 at $\sigma{=}0.1$ Å to 0.12 at $\sigma{=}0.5$ Å, and approached zero for $\sigma{\geq}2.0$ Å. Gradient reliability therefore degrades rapidly with noise, consistent with the advantage of refinement on denoised backbones over guidance at higher noise levels.

Table 17: Q_θ gradient cosine similarity under Gaussian backbone noise, averaged over 20 CaM designs.

σ (Å)	Norm	Cos	Margin
0.0	0.317	$\mathbf{1.000}$	$+0.359$
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	$+0.015$

D.5 Noise Schedule Alignment

To relate this noise sensitivity to denoising, we aligned the RFdiffusion noise schedule ($T{=}50$ denoising steps) with the gradient-reliability profile from Appendix D.4. Cosine similarity was 0.75 at $t{=}0$ and fell 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, once the trajectory enters the regime $\sigma{<}0.10$ Å. About 96% of steps in this 50-step diagnostic fell below a cosine similarity of 0.5; reliability under the main 200-step guidance configuration remains to be quantified.

Table 18: 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

D.6 Langevin Step-Size Sensitivity

To assess the balance between selectivity gains and backbone displacement, we swept the Langevin step size η on 50 CaM designs under ensemble scoring (Table 19). Selectivity increased from $\bar{S}{=}{+}0.506$ at $\eta{=}0.01$ to $\bar{S}{=}{+}0.579$ at $\eta{=}0.08$, then plateaued, while Cα RMSD from the vanilla backbone stayed below 0.14 Å throughout. At $\eta{=}0.04$, mean selectivity was 0.559 and RMSD was 0.10 Å. Generator-specific benchmark settings are listed in Appendix A.2.

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	$\mathbf{+0.579}$	$\mathbf{+0.146}$	$\mathbf{100\%}$	0.12
0.10	$+0.579$	$+0.146$	100%	0.14

D.7 Best-of-K Reranking

To quantify the effect of candidate-pool size under the selection scorer, we evaluated best-of-K selection from a fixed CaM pool using 10,000 bootstrap trials (Table 20). At $K = 5$, the mean selected gap was 0.787 and 94% of selections exceeded 0.5. At $K = 10$, these values reached 0.885 and 100%.

Table 20: 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	$\mathbf{+0.985}$	$\mathbf{100\%}$	$\mathbf{100\%}$

D.8 End-to-end Langevin Design Results

Structure-related metric analysis for Langevin refinement. We refined 50 vanilla designs per target for 100 steps (Table 21). Mean selectivity increased on seven of eight targets, with the largest gains on CaM (0.253) and MDM2 (0.181), and decreased on PAI-1 (-0.024). The fraction of retained designs with positive gaps was 100% for CaM and BCL-2 and at least 70% for seven targets. Across the eight targets, 390 of 400 designs were retained. Changes in native-contact recovery ranged from -0.004 to 0.168 on the six targets with contact references (Table 22).

Table 21: Refinement outcomes across the eight held-out targets. Positive-gap rate denotes the fraction of retained refined designs with $S > 0$, not the fraction improved over vanilla. Of 400 original designs, 390 were retained.

Target	N_{van}	N_{ref}	$S > 0$	$\bar{S}_{\text{van}}$	$\bar{S}_{\text{ref}}$	$\Delta\bar{S}$
CaM	50	50	100%	0.531	0.784	$+0.253$
ERα	50	50	74%	0.110	0.250	$+0.140$
BCL-2	50	50	100%	0.886	0.911	$+0.024$
MDM2	50	50	98%	0.216	0.397	$+0.181$
Ran	50	46	80%	0.119	0.249	$+0.130$
PAI-1	50	49	71%	0.188	0.164	-0.024
Integrin	50	48	58%	0.004	0.012	$+0.008$
A_{2A}	50	47	70%	0.363	0.412	$+0.049$

ProteinMPNN ΔNLL analysis for Langevin refinement. To test whether refinement also improved sequence-based state preference, we evaluated ProteinMPNN ΔNLL. Its mean was positive for five of eight vanilla target panels, with an all-target mean of 0.188 (Table 23). Langevin refinement

Table 22: Change in native-contact recovery after refinement, measured against the available crystal-contact reference. CaM and A_{2A} lack the short-peptide contact references used in this comparison.

Target	Δ fNAT
CaM	–
ERα	-0.004
BCL-2	$+0.168$
MDM2	$+0.084$
Ran	$+0.002$
PAI-1	-0.001
Integrin	$+0.012$
A_{2A}	–

reduced ΔNLL on five targets and reduced the all-target mean to 0.128. Thus, higher Q_θ scores did not imply higher sequence-recovery preference for holo. The paired tests assessed changes between vanilla and refined designs in this computational measure of state preference.

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. †Targets with negative mean vanilla ΔNLL. The cause of the disagreement with Q_θ is unresolved.

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	$+0.033$	$6\times10^{-5\,*}$
$A_{2A}^\dagger$	-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	–

Failure-mode analysis of negative-selectivity cases. We examined all negative-gap designs under the TS-S2 seed-1024 checkpoint in a pool of 482 CaM designs from nine generator–guidance configurations. After aligning each receptor to the 3CLN holo reference, we annotated five non-exclusive categories: length below 50 residues, binder centre-of-mass more than 25 Å from the designated groove (residues 84–88 and 124–125), fewer than five binder residues within 10 Å of the receptor, an inter-chain C_α distance below 2.5 Å, and apo score above 0.3. Ten designs (2.1%), all from PXDesign with Langevin refinement, had negative gaps (Table 24). All ten met the length criterion, four had misplaced interfaces, four had clashes, and two had insufficient interfaces. The length criterion applies to this generated-backbone analysis. Scorer errors and experimental state preference remain unresolved by these structural annotations.

Boltz-2 cross-check. We compared Boltz-2 ΔipTM with Q_θ gaps on 50 vanilla designs per target (Table 25). Mean ΔipTM was positive on four of eight targets. Design-level rank correlations were strongest on A_{2A} (0.500) and CaM (0.349), and weak or negative elsewhere. BCL-2 had a mean ΔipTM of -0.183 and only 24% positive differences, indicating disagreement between the two scoring methods.

Table 24: Failure-mode analysis of the 10 negative-selectivity designs out of 482 total designs from 9 guidance pipelines, scored under the TS-S2 seed-1024 Q_θ checkpoint. 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%

Table 25: 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%	$\mathbf{+0.500}$
CaM	$+0.032$	56%	$\mathbf{+0.349}$
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%	$+0.050$

AlphaFold-3 cross-check. We refolded a subset of designs with AlphaFold 3 in single-sequence mode (Table 26). All designs for the training target ALK and held-out ERα had positive ΔipTM in both vanilla and refined panels. Only 5% of the 19 BCL-2 designs had positive differences. The cause of this disagreement was not resolved by these measurements.

Table 26: 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%

D.9 Rosetta Interface Analysis

We computed Rosetta InterfaceAnalyzer metrics with repacked sidechains on 400 RFdiffusion designs, comprising 50 designs for each of eight targets (Table 27). This panel contains seven targets from the OOD test split plus the training target ALK, and excludes CaM. Mean interface energies were most favorable for BCL-2 (-9.1 REU) and least favorable for Ran (25.4 REU) and Integrin (39.1 REU). These energies characterize individual interfaces; a holo–apo energy comparison requires calculations in both states.

Table 27: Rosetta InterfaceAnalyzer metrics for 400 vanilla RFdiffusion designs (50 per target). The panel includes seven OOD targets and training target ALK; CaM is absent.

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

D.10 Statistical Analysis

Scoring results report means and standard deviations across three training seeds. Best-of-K analysis used 10,000 bootstrap trials. The p-values in Table 23 are reported for paired Wilcoxon signed-rank comparisons of vanilla and refined designs. These paired tests quantify changes between methods. BLI measurements were performed in duplicate and are reported descriptively, without significance tests or confidence intervals for K_D.

D.11 Additional Held-Out Target: Adenylate Kinase

To examine transfer beyond the benchmark, we evaluated adenylate kinase (AdK), which is outside the 65-target benchmark listed in Table 6. All reported combinations had positive mean Q_θ gaps (Table 28). TDS produced the largest mean gap for RFdiffusion (0.543) and Proteina-ComplexA (0.904), while SMC was highest for PXDesign (0.311). Boltz-2 holo–apo ipTM differences were mixed, so independent support for the predicted state preference remained inconclusive. Candidate counts vary by method. Structural and sequence metrics are listed in Table 29.

Table 28: Additional AdK design results, scored by the three-seed ensemble. RF, PX, and Pro denote RFdiffusion, PXDesign, and Proteina-ComplexA. Des and SR denote computational designability and joint success rates.

Method	N	$\bar{S}$	σ_S	$S > 0$	Des	SR
RF/TDS	50	**0.543**	**0.108**	100%	**84.0%**	**84.0%**
RF/SMC	18	0.522	0.160	100%	83.3%	83.3%
RF/Langevin	50	0.372	0.262	100%	82.0%	82.0%
RF/Classifier	50	0.298	0.248	100%	82.0%	82.0%
RF/Vanilla	50	0.260	0.253	100%	80.0%	80.0%
PX/SMC	32	**0.311**	0.321	87.5%	43.8%	40.6%
PX/TDS	50	0.305	**0.302**	**92.0%**	56.0%	**50.0%**
PX/Langevin	50	0.245	0.332	76.0%	**60.0%**	46.0%
PX/Vanilla	50	0.223	0.335	76.0%	58.0%	44.0%
PX/Classifier	50	0.124	0.341	74.0%	54.0%	42.0%
Pro/TDS	50	**0.904**	**0.130**	**100%**	60.0%	60.0%
Pro/SMC	50	0.878	0.143	**100%**	**66.0%**	**66.0%**
Pro/Langevin	50	0.481	0.463	82.0%	34.0%	26.0%
Pro/Classifier	50	0.420	0.444	80.0%	28.0%	20.0%
Pro/Vanilla	50	0.272	0.415	72.0%	34.0%	20.0%

Table 29: AdK structural and sequence metrics corresponding to Table 28. Both holo and apo ipTM are reported to preserve disagreements with the learned selectivity scores.

Method	ipTM holo	ipTM apo	scTM	NLL holo
RF/TDS	0.394	0.380	0.649	1.073
RF/SMC	0.368	0.405	0.660	**1.036**
RF/Langevin	0.395	**0.433**	0.643	1.049
RF/Classifier	0.410	0.415	**0.676**	1.044
RF/Vanilla	**0.414**	0.420	0.661	1.042
PX/SMC	0.298	0.314	0.454	1.262
PX/TDS	**0.352**	**0.361**	0.510	1.288
PX/Langevin	0.309	0.336	0.493	1.307
PX/Vanilla	0.326	0.313	**0.521**	1.318
PX/Classifier	0.348	0.328	0.477	**1.255**
Pro/TDS	0.419	**0.463**	0.545	1.130
Pro/SMC	**0.421**	0.423	**0.550**	1.126
Pro/Langevin	0.416	0.387	0.396	**1.093**
Pro/Classifier	0.377	0.388	0.400	1.101
Pro/Vanilla	0.410	0.395	0.400	1.096

E Efficiency Analysis

E.1 Compute Efficiency

Table 30: Computational cost on one A100 80GB GPU. RF, PX, and Pro denote RFdiffusion, PXDesign, and Proteina-ComplexA. GPU-hours correspond to the listed candidate counts.

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

E.2 Inference Speed

Q_θ achieved 2.98 ± 0.09 ms per-complex latency in single-sample mode on an A100 GPU. At batch size 16, amortized throughput reached 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. The per-method timings and their different candidate counts are listed in Table 30.

In the reported cost comparison, RFdiffusion with TDS had the highest $\bar{S}$/GPU-hour within RFdiffusion (0.281), and PXDesign with SMC had the highest ratio within PXDesign (0.116). RFdiffusion refinement increased time per design from 2.0 to 2.5 minutes. The ratios reflect the listed workloads, whose candidate counts differ across methods.

F Experimental Validation Details

F.1 Candidate Selection and Assay Protocol

We selected peptides from RFdiffusion and Proteina-ComplexA generator–guidance combinations using the predicted bounded gap $\Delta_q = Q_\theta(X^1, Y) - Q_\theta(X^0, Y)$. The measured panel in Table 3 contains nine candidates, one designated lower-scoring negative control, and an additional M13 positive control. Because we selected candidates by predicted preference, quantitative calibration requires a separately sampled set. The skeletal myosin light-chain kinase M13 peptide (CPC Scientific, SIGN-010) served as a Ca^{2+}-dependent binding control.

Calmodulin preparation. We used recombinant human calmodulin (MilliporeSigma, catalogue 208670). Holo and apo conditions were prepared by supplementation with $CaCl_2$ and EGTA, respectively, with the other assay conditions maintained identically.

Bio-layer interferometry. Adaptyv Bio measured binding on a Gator Bio Pro instrument. Designed peptides were immobilized through a Twin-Strep tag on compatible biosensors. After loading and baseline equilibration, sensors were exposed to CaM at 0, 100, and 1000 nM for association measurements, followed by transfer into assay buffer for dissociation measurements. Measurements were performed in duplicate (Figure 5).

Analysis and interpretation. Sensorgrams were reference-subtracted using control sensors and buffer-only wells. Association and dissociation traces were fitted to a 1:1 binding model to obtain rate constants and K_D. Concentration-dependent responses were classified as binding; responses below detection were recorded as NB. The reported $> 1.00\,\mu$M value is a bound, not a point estimate. The $1.06\,\mu$M fit is close to the highest tested concentration and should be interpreted accordingly. We report duplicate measurements without replicate-level uncertainty estimates, an independent preparation count, or a quantified detection limit. NB therefore does not establish zero affinity or a quantitative lower bound on the holo/apo selectivity ratio.

F.2 Complete Experimental Panel

Five designed peptides bound holo CaM, and none showed detectable apo binding under the tested conditions (Table 3). Four candidates with predicted gaps of 0.688–0.993 had no detectable binding to either state. This combination of binding and nonbinding candidates separates computational prioritization from experimental confirmation.

Table 31: Peptide sequences corresponding to Table 3. The sequences exclude assay tags.

Design	Sequence
RF/V 944	ATAAMIKTFQDVVVAAVREAREK
RF/V 13	SEAFARAAAVLAKARAAK
Pro/SM 0273	EGFKKLLKEALEIAK
RF/V 657	KVAEQAKQWILEMLAK
RF/Lg 25	EKLEALLREAGAARRAAKKAAEAA
Pro/V 0145	VDEDGDGKIDLPELSALLREKIK
RF/Lg 330	SELTKEILKKAMEMT
Pro/V 0376	AFGAEVKTPRTRFLDVLR
Pro/SM 0954	EEAARAAGLARLPRPLLLLQAL
Pro/V 0105	SEIAELLRRNPEGDPETLREALAA
M13	KRRWKKNFIAVSAANRFKKISSSGAL

The original identifiers combine the generator and scheme prefixes with `design_` or, for Pro/SM, `smc_particle_`, followed by the displayed identifier. The listed sequences are given in Table 31.

G Limitations and Discussion

AlloGen connects paired conformational scoring with binder generation, with computational transfer evaluated on a target-level split and state-dependent binding measured in selected CaM designs. Its use requires paired receptor structures, and transfer across sequence-clustered families remains to be tested. Backbone-only inference also differs from training on complete complexes, and limited apo-decoy coverage may constrain performance on subtle conformational changes. Richer apo conformations and broader target panels would help assess these dependencies.

Absolute Q_θ outputs are uncalibrated scores, and reranking and guidance are evaluated partly with the scorer used for optimization. Agreement across independent computational metrics was mixed, reinforcing the role of direct binding measurements. The CaM BLI panel established binding in selected designs, but duplicate measurements without a quantified detection limit or replicate-level uncertainty cannot quantify a holo/apo selectivity ratio. Additional targets, orthogonal assays, and calibration data are needed to establish quantitative selectivity and broader experimental transfer. With this validation, we envision using AlloGen to design and prioritize binders for defined receptor conformations across protein families.

Broader impact. State-selective binder design may support molecular probes and future therapeutic research. Therapeutic applications require further testing of cellular efficacy, delivery, and safety. Protein-design methods can also be misused to alter biological interactions, so downstream applications require context-specific assessment and experimental validation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction distinguish predicted selectivity from BLI measurements on CaM. The held-out scoring benchmark, generator comparisons, and complete experimental panel support the stated scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix G discusses paired-structure requirements, target-level splitting, uncalibrated scores, optimization with the evaluation scorer, mixed independent metrics, and the limited CaM assay panel.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper presents an empirical method and no formal theoretical results. The rigid-motion invariance of the geometric features is explained in Section 3.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [No]

Justification: The architecture, objectives, and computational settings are described in Sections 4 and A. The available assay description does not specify complete buffer and timing parameters, replicate-level data, or detection limits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The public project at `https://huggingface.co/ChatterjeeLab/AlloGen` includes code, model checkpoints, inference instructions, and a sample dataset. Availability of the complete benchmark and raw experimental records sufficient to reproduce every reported result has not been established.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The paper specifies the architecture, target split, learning rates, batch sizes, dropout, and guidance settings. The description does not fully specify optimizer selection, all model-selection decisions, or every assay setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Main scoring results include three-seed means and standard deviations, and reranking uses bootstrap trials. Design and BLI results do not uniformly include uncertainty estimates, and the paired-test description does not fully specify sidedness or multiple-comparison correction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix E reports A100 inference latency and generation costs by method. The total compute used for training, preliminary runs, and the full research project is not documented.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public protein structures and in vitro binding assays, without human participants or sensitive personal data. External methods are cited, and funding and competing interests are disclosed.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix G discusses potential applications and the need for downstream biological validation, along with possible misuse of protein-design methods.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifact is a specialized interface scorer and associated code, not a general-purpose generative foundation model. This work does not release a high-risk human dataset.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

Justification: Existing methods and structural data sources are cited. The project repository lists an Apache-2.0 license, but the manuscript does not enumerate the license and usage terms of every external asset.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The model, data-construction procedure, and guidance methods are described in the paper. The public project documents released code, checkpoints, and sample data, without establishing availability of the full experimental dataset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing or human participants. Experimental validation used recombinant calmodulin and designed peptides.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The study does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an original or non-standard component of the core method. ESM-2 is used as an optional pretrained protein-sequence feature encoder and is described and cited in the Method section. Writing and formatting assistance does not alter the methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.