PROTEÍNA-REFORZA: REINFORCEMENT LEARNING PROTEIN BINDER DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow- and diffusion-based generative models have transformed *de novo* protein binder design, but they are typically trained to reproduce molecular data distributions without explicit alignment to design objectives. Here, we introduce *Proteína-Reforza*, a reinforcement learning (RL) approach that post-trains Proteína-Complexa to improve protein binder design. We adapt the Advantage Weighted Matching RL framework to Proteína-Complexa's partially latent representation for atomistic joint sequence–structure generation. Rewards include structure prediction confidence scores, physics-based Rosetta energies, affinity predictions, as well as a novel diversity reward, and we additionally train reward models on experimental binding data to enable RL guided by experimental outcomes. We apply Proteína-Reforza to both minibinder and cyclic peptide design, for both protein and ligand targets, and demonstrate general model alignment as well as target-specific adaptation for challenging design tasks. Proteína-Reforza achieves state-of-the-art performance on established *in silico* binder design benchmarks, substantially outperforming existing methods, and alleviating the need for extensive post-hoc filtering. Further analyses and ablations reveal how RL shapes the generated binders and support their structural and biophysical plausibility, without noticeable reward hacking. Experimental evaluation of generated minibinders and cyclic peptides confirms that these computational gains translate into improved binding success in the lab.

1 INTRODUCTION

De novo protein binder design is a critical task in therapeutics, diagnostics and biological research. Modern flow-based generative models such as Proteína-Complexa condition on targets and codesign binder sequences and fully atomistic structures (Didi et al., 2026b). However, these generators are usually pretrained to reproduce molecular data distributions without explicitly optimizing design objectives. Successful design thus often requires extensive sampling and filtering. Inference-time search can help but adds computation without updating the model (Didi et al., 2026a). Hence, incorporating design objectives directly into training could increase the yield of promising candidates and reduce this overhead. Meanwhile, reinforcement learning (RL) aligns language and visual generative models with desired outcomes (Ouyang et al., 2022; Black et al., 2024). By favouring high-reward outputs, it incorporates task objectives into the learned distribution. For protein binder design, this offers a route to improving generation using computational and experimental feedback while retaining efficient sampling.

Here, we introduce *Proteína-Reforza*, an RL approach for post-training Proteína-Complexa (Fig. 1). We adapt Advantage Weighted Matching (AWM) (Xue et al., 2026) to the model's partially latent representation, which combines explicit backbone coordinates with continuous per-residue latent variables encoding amino acid identity and atomic structure (Geffner et al., 2026). AWM uses reward-dependent advantages to weight updates on generated samples, connecting RL post-training to the flow matching pretraining objective. Our framework supports both general model alignment, for which we construct suitable protein and small molecule target sets for RL training, as well as target-specific model adaptation for challenging problems. To the best of our knowledge, Proteína-Reforza represents the first principled application of RL to modern fully atomistic protein binder design models.

A central component is the choice of rewards. We first use structure prediction confidence scores, widely employed to assess candidate binder–target complexes (Watson et al., 2023; Pacesa et al., 2025). This allows the generator to learn directly from criteria otherwise applied during downstream filtering. We also explore alignment with physics-based objectives. All molecular interactions are governed

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

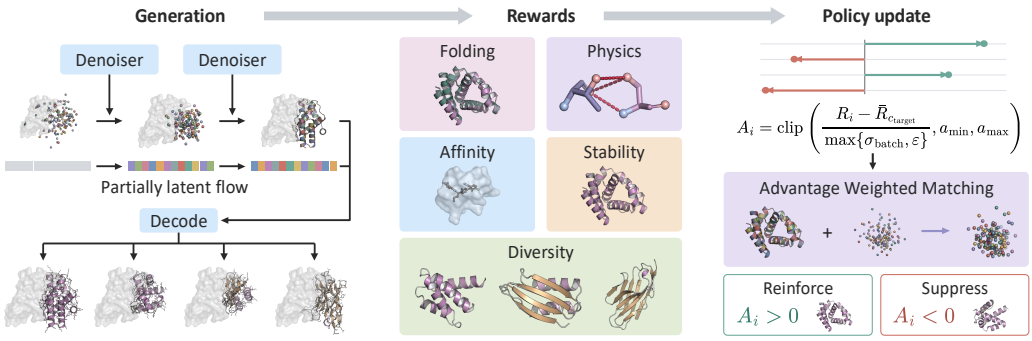


Figure 1: **Proteína-Reforza** uses Advantage Weighted Matching (Xue et al., 2026) adapted to Proteína-Complexa's partially latent representation (Didi et al., 2026b) for reinforcement learning: *(left, rollout)* Partially latent candidates are generated and decoded into groups of binders. *(middle, scoring)* The proteins are scored via reward functions. *(right, update)* The model is updated; good candidates are reinforced, poor ones suppressed.

by physical laws, and structural plausibility alone does not guarantee energetically favorable binding. We therefore incorporate Rosetta force field-based binding energy scores to provide feedback on the energetics of generated complexes (Alford et al., 2017). Furthermore, we leverage binding affinity predictors (Shenoy et al., 2026) as rewards to score ligand-protein complexes. These rewards complement learned confidence estimates, bringing physical considerations into generative post-training.

Beyond computational scoring, experimental measurements determine whether designed proteins actually bind. Inspired by the use of learned reward models in reinforcement learning from human feedback (Ouyang et al., 2022; Lambert, 2026), we train reward models on experimental binding data and use their predictions to guide RL. This transfers information from measured binding outcomes into the generative model through a learned surrogate, allowing experimental data to directly influence generation. In protein design, it is particularly critical for alignment to preserve access to multiple candidate design solutions. Therefore, we also introduce a novel diversity reward that can optionally be incorporated to encourage diverse candidate generation alongside the optimization of design quality. We further extend the generation framework to cyclic peptides and apply RL in this setting, broadening the scope of Proteína-Reforza beyond minibinders to a distinct class of constrained molecular designs.

Across established *in silico* binder design benchmarks, Proteína-Reforza achieves state-of-the-art performance by a large margin, outperforming existing methods and its unaligned base model. We demonstrate substantial improvements across minibinder and cyclic peptide design tasks for both ligand and protein targets (see Fig. 2), as well as both general model alignment and adaptation to individual challenging targets. These gains increase the fraction of promising candidates output by the model and alleviate reliance on extensive post-hoc filtering. Extensive analyses and ablations examine how RL changes the generated binders, support their structural and biophysical plausibility, and reveal no evidence of systematic reward hacking. Finally, experimental wet lab evaluation of generated minibinders and cyclic peptides, including designs against therapeutically relevant targets, confirms that the computational improvements translate into increased binding success in the laboratory. Our results establish RL post-training as a practical route to molecular generative models that learn from structure prediction confidence, physics and experimental feedback to produce functional binders more reliably.

Key contributions. *(a)* We adapt AWM to Proteína-Complexa and introduce Proteína-Reforza, the first systematically post-trained atomistic generative model for binder design. *(b)* We demonstrate RL with structure prediction confidence rewards, physics-based force field rewards, affinity prediction rewards, as well as diversity objectives, for multiple design tasks in both general and target-specific manner. *(c)* We train a new reward model on experimental binding data and showcase RL guided by wet lab feedback. *(d)* We extend Proteína-Reforza to a new molecular modality not present in the base model, cyclic peptides, and also perform successful RL alignment. *(e)* Experimental validation confirms that Proteína-Reforza's state-of-the-art *in silico* performance translates to enhanced success in the lab.

2 Background and Related Work

Proteína-Complexa. We build on Proteína-Complexa (Didi et al., 2026b), a target-conditioned extension of La-Proteína (Geffner et al., 2026) for fully atomistic binder design. Its partially latent representation combines explicit binder alpha carbon coordinates $\mathbf{x}^{C_\alpha}$ with per-residue latents $\mathbf{z}$ encoding amino acid identities $\mathbf{s}$ and remaining atomic coordinates $\mathbf{x}^{\neg C_\alpha}$. This representation is learned via a variational autoencoder (VAE) (Kingma & Welling, 2014), whose encoder $\mathcal{E}$ provides

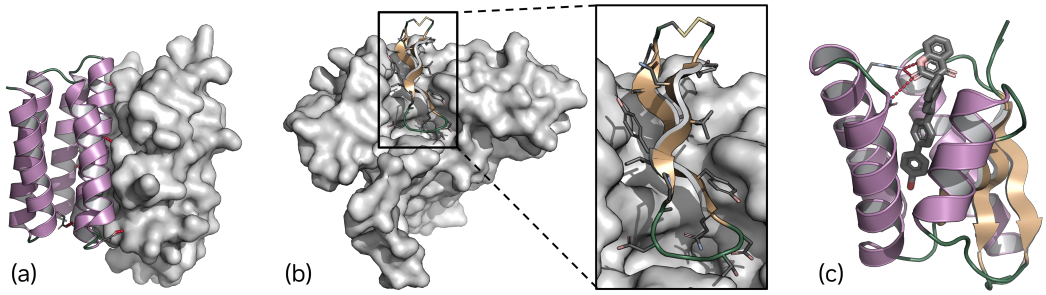


Figure 2: **Binders** generated by Proteína-Reforza. *(a)* Protein binder to PD-L1 target, experimentally tested (64.7 nM, see Sec. 5). *(b)* Disulfide cyclic peptide binder to TROP2 target. *(c)* Protein binder to UA7 ligand target.

training latents $\mathbf{z} = \mathcal{E}(\mathbf{x}^{C_\alpha}, \mathbf{x}^{\neg C_\alpha}, \mathbf{s})$ from clean binder inputs, and whose decoder $\mathcal{D}(\mathbf{x}^{C_\alpha}, \mathbf{z})$ reconstructs the binder sequence and Atom37 coordinates. With separate coordinate and latent interpolation times $t_x, t_z \in [0, 1]$, flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023) learns coupled velocity fields $\mathbf{v}_x^\theta$ and $\mathbf{v}_z^\theta$ for a target-conditioned probability path $p_{t_x,t_z}(\mathbf{x}_{t_x}^{C_\alpha}, \mathbf{z}_{t_z} | \mathbf{c}^{\text{target}})$ connecting a tractable noise prior to the encoded binder data distribution. Here, $\mathbf{c}^{\text{target}}$ comprises target Atom37 coordinates, amino acid identities, and binary hotspot indicators. Conditional flow matching (CFM) trains these velocity fields against tractable endpoint-conditioned velocities. Specifically, Proteína-Complexa uses rectified flow interpolants $\mathbf{x}_{t_x}^{C_\alpha} = t_x \mathbf{x}^{C_\alpha} + (1 - t_x)(\mathbf{x}_0^{C_\alpha} + \vec{d}\,\mathbf{1}_N)$ and $\mathbf{z}_{t_z} = t_z \mathbf{z} + (1 - t_z)\mathbf{z}_0$ with the noise distributions $p_0^{C_\alpha} = \mathcal{N}(\mathbf{x}_0^{C_\alpha} | \mathbf{0}, \boldsymbol{I})$, $p_0^{\mathbf{z}} = \mathcal{N}(\mathbf{z}_0 | \mathbf{0}, \boldsymbol{I})$ and the random global translation $\vec{d} \sim p_0^{\vec{d}} = \mathcal{N}(\vec{d} | \vec{0}, c_d^2\, I_{xyz})$ broadcast over the N binder residues (c_d=0.2nm), which helps the model learn global binder positioning. Writing $\mathbf{h} = (\mathbf{x}_{t_x}^{C_\alpha}, \mathbf{z}_{t_z}, \mathbf{c}^{\text{target}}, t_x, t_z)$, the objective then is $\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}\big[||\mathbf{v}_z^\theta(\mathbf{h}) - (\mathbf{z} - \mathbf{z}_0)||^2 + ||\mathbf{v}_x^\theta(\mathbf{h}) - (\mathbf{x}^{C_\alpha} - [\mathbf{x}_0^{C_\alpha} + \vec{d}\,\mathbf{1}_N])||^2\big]$. The transformer neural network jointly processes clean conditioning tokens and noisy binder tokens. See Didi et al. (2026b) for further architecture and training details.

Although sampling can use low-temperature stochastic differential equations, here we adopt the more efficient temporal score rescaling technique (Xu et al., 2026) with its adaptation to the partially latent framework introduced by Anonymous (2026). Writing $\mathbf{y}_x = \mathbf{x}_{t_x}^{C_\alpha}$ and $\mathbf{y}_z = \mathbf{z}_{t_z}$, the rescaled velocities are $\widetilde{\mathbf{v}}_a^\theta(\mathbf{h}; \gamma_a) = (1 - \gamma_a(t_a))\mathbf{y}_a / t_a + \gamma_a(t_a)\mathbf{v}_a^\theta(\mathbf{h})$ for $a \in \{x, z\}$ with the scaling factors $\gamma_a(t_a) = \big[(1 - t_a)^2 + t_a^2\sigma_a^2\big] / \big[(1 - t_a)^2 + t_a^2\sigma_a^2/k_a\big]$, and the modality-specific parameters k_a, σ_a chosen to emulate low-temperature density concentration. Starting from $(\mathbf{x}_0^{C_\alpha} + \vec{d}\,\mathbf{1}_N, \mathbf{z}_0)$, we jointly integrate $d\mathbf{y}_a = \widetilde{\mathbf{v}}_a^\theta(\mathbf{h}) dt_a$ along schedules taking (t_x, t_z) from $(0, 0)$ to $(1, 1)$, denoising carbon alpha coordinates faster than latent variables, and then decode the terminal representation.

Related Work. Deep learning-based protein binder design frameworks use either flow- or diffusion-based generative models (Watson et al., 2023; Ingraham et al., 2023; Team et al., 2025c; Butcher et al., 2025; Stark et al., 2025; Zhang et al., 2025; Lin et al., 2026; Didi et al., 2026b) or hallucination-based methods (Goverde et al., 2023; Pacesa et al., 2025; Cho et al., 2025; Candido et al., 2026) that directly optimize sequences on structure predictors' confidence metrics (Jumper et al., 2021; Abramson et al., 2024; Corley et al., 2025). RL post-training is common in natural language modeling (Ouyang et al., 2022; Team et al., 2025b; Khatri et al., 2025; Liu et al., 2025b) and visual synthesis (Black et al., 2024; Liu et al., 2025a; Zheng et al., 2026; Xue et al., 2026) to align outputs with user preferences, enhance reasoning, or improve aesthetic quality. RL and related alignment methods have also been used in protein language modeling (Angermueller et al., 2020; Jain et al., 2022; Stocco et al., 2024; Su et al., 2026; Zhang et al., 2026), for inverse folding (Wang et al., 2025b; Xue et al., 2025; Xu et al., 2025) and antibody design (Zhou et al., 2024), as well as to improve protein fitness and stability (Sun et al., 2025; Wang et al., 2025a; Widatalla et al., 2026). However, these works concern protein sequence generation; to our knowledge, no prior work addresses RL post-training of modern flow- or diffusion-based protein binder generators, which typically model fully atomistic structures.

3 Proteína-Reforza

3.1 Reinforcement Learning for De Novo Protein Binder Design

Objective. We build on Advantage Weighted Matching (AWM) (Xue et al., 2026) for post-training. AWM is a policy gradient RL method for diffusion and flow models that approximates intractable likelihoods with evidence lower bound (ELBO) estimates. For each target condition $\mathbf{c}^{\text{target}}$, we generate

a group of $\mathcal{G}$ samples, or rollouts, under policy $\pi_{\theta_{\text{old}}}$. The rollout parameters θ_{old} are an exponential moving average (EMA) of the optimized parameters θ of the learned policy π_θ. Adapting AWM to Proteína-Reforza's partially latent representation yields the following policy optimization objective

$$\begin{aligned}
\mathcal{J}_{\text{AWM}}^{\text{Reforza}}(\theta) &= \mathbb{E}_{\mathbf{c}^{\text{target}},\{(\mathbf{x}_i^{C_\alpha},\mathbf{z}_i)\}_{i=1}^{\mathcal{G}}}\left[\frac{1}{\mathcal{G}}\sum_{i=1}^{\mathcal{G}}\Big(A_i\exp\big(\ell_{\theta_{\text{old}},i}-\ell_{\theta,i}\big)-\beta\,\mathcal{K}_{\theta,\theta_{\text{ref}},i}\Big)\right],\\
\ell_{\theta,i} &= \mathbb{E}_{\xi_i}\left[\left\|\mathbf{v}_x^\theta(\mathbf{h}_i)-\mathbf{u}_{x,i}\right\|^2+\left\|\mathbf{v}_z^\theta(\mathbf{h}_i)-\mathbf{u}_{z,i}\right\|^2\right],\\
\mathcal{K}_{\theta,\theta_{\text{ref}},i} &= \mathbb{E}_{\xi_i}\left[\left\|\mathbf{v}_x^\theta(\mathbf{h}_i)-\mathbf{v}_x^{\theta_{\text{ref}}}(\mathbf{h}_i)\right\|^2+\left\|\mathbf{v}_z^\theta(\mathbf{h}_i)-\mathbf{v}_z^{\theta_{\text{ref}}}(\mathbf{h}_i)\right\|^2\right],
\end{aligned} \tag{1}$$

where $\mathbf{h}_i$ follows its definition in Sec. 2 and includes the noised alpha carbon coordinates $\mathbf{x}_{t_x,i}^{C_\alpha}$ and latent variables $\mathbf{z}_{t_z,i}$, with targets $\mathbf{u}_{x,i}=\mathbf{x}_i^{C_\alpha}-[\mathbf{x}_{0,i}^{C_\alpha}+\vec{d}_i\,\mathbf{1}_N]$ and $\mathbf{u}_{z,i}=\mathbf{z}_i-\mathbf{z}_{0,i}$, ξ_i collects fresh times and noise draws, and $\mathcal{K}_{\theta,\theta_{\text{ref}},i}$ estimates the Kullback–Leibler divergence between the learned policy π_θ and a reference $\pi_{\theta_{\text{ref}}}$, weighted by β. Advantages A_i are calculated as (Shao et al., 2024)

$$A_i=\text{clip}\left(\frac{R_i-\bar{R}_{\mathbf{c}^{\text{target}}}}{\max(\sigma_{\text{batch}},\varepsilon)},a_{\min},a_{\max}\right),\qquad \bar{R}_{\mathbf{c}^{\text{target}}}=\frac{1}{\mathcal{G}}\sum_{i=1}^{\mathcal{G}}R_i,\qquad R_i=R\Big(\mathcal{D}(\mathbf{x}_i^{C_\alpha},\mathbf{z}_i),\mathbf{c}^{\text{target}}\Big), \tag{2}$$

where σ_{batch} is the reward std. dev. of the batch across groups, $a_{\min}$, $a_{\max}$ and ε are hyperparameters and rewards R_i are calculated on decoded binder C_α coordinates $\mathbf{x}_i^{C_\alpha}$ and latents $\mathbf{z}_i$ together with target $\mathbf{c}^{\text{target}}$ (see Algorithm 1). We do not use an exact ELBO weighting and omit weighting the L_2 terms, consistent with Proteína-Reforza's pretraining objective, as also done by Xue et al. (2026) in practice.

General vs. target-specific RL. We consider two RL settings for protein binder design. General RL aims to improve binder generation across a broad distribution of targets, analogous to general-purpose RL fine-tuning in language and image generation. Target-specific RL specializes the model to individual targets of interest. This setting is relevant to drug discovery applications, where the target is typically known in advance and experimental validation is often the bottleneck, motivating additional computational effort to improve candidate quality for that target. We study both settings.

RL Target Sets. For general RL, we curated training sets of binder targets, $\mathbf{c}^{\text{target}}$ above, analogous to the prompt sets used for RL fine-tuning of large language models. To construct a set of protein targets, we selected complexes from the Protein Data Bank (PDB) (Berman et al., 2000) with high-quality interfaces and well-resolved interacting chains, applied clash and minimum-contact filters, and constructed both single- and multi-chain targets. Filtering for sequence redundancy, followed by hotspot assignment and structural clustering, yielded 52,492 protein targets in 15,841 clusters. For small molecule targets, we used the PLINDER training set (Durairaj et al., 2024), retaining only systems with single ligands. We excluded low-quality crystal structures and applied filters for clashes, unresolved heavy atoms, and minimum ligand size. We then extracted and deduplicated the ligands, yielding 8,795 small molecules for RL training. Training and data curation details in the Appendix.

3.2 Rewards for Protein Binder Generation

All rewards $R_i(\mathbf{x}_i^{C_\alpha},\mathbf{x}_i^{\neg C_\alpha},\mathbf{s}_i,\mathbf{c}^{\text{target}})$ for RL training Proteína-Reforza are evaluated on decoded binder sequences and structures $\mathbf{x}_i^{C_\alpha},\mathbf{x}_i^{\neg C_\alpha},\mathbf{s}_i=\mathcal{D}(\mathbf{x}_i^{C_\alpha},\mathbf{z}_i)$ together with the target $\mathbf{c}^{\text{target}}$.

Folding Success Reward. Structure prediction model confidence scores are commonly used as filters for candidate selection, making them a natural choice as rewards for RL. Following Didi et al. (2026b), we consider a binder sequence $\mathbf{s}$ as successful *in silico* if the complex prediction satisfies $f_{\text{pLDDT}}(\mathbf{s})>0.90$, $f_{\text{ipAE}}(\mathbf{s})\leq 7.0$Å, and $f_{\text{Binder-RMSD}}(\mathbf{s})<1.5$Å using AlphaFold2-Multimer (Evans et al., 2022) (AF2). For small molecule targets, we instead use RosettaFold-3 (RF3) (Corley et al., 2025), and define success by $f_{\text{min-ipAE}}(\mathbf{s})<2$Å, $f_{\text{Binder-RMSD}}(\mathbf{s})<2$Å and $f_{\text{Ligand-RMSD}}(\mathbf{s})<5$Å from the top-ranked of five RF3 predictions. We then define a folding success reward as

$$R_i^{\text{succ}}=1-\frac{1}{3}\begin{cases}\dfrac{[\text{ipAE}_i-7]_+}{7}+\dfrac{[0.90-\text{pLDDT}_i]_+}{0.10}+\dfrac{[d_i-1.5]_+}{1.5}, & \text{protein},\\[2ex] \dfrac{[\text{min-ipAE}_i-2]_+}{1}+\dfrac{[d_i-2]_+}{2}+\dfrac{[l_i-5]_+}{5}, & \text{ligand},\end{cases} \tag{3}$$

where $[a]_+=\max(a,0)$, and using the appropriate folding model for protein and ligand targets, respectively, with best-of-5 selection for the latter. The quantities d_i and l_i denote binder C_α scRMSD

and ligand all-atom scRMSD after binder alignment, respectively; all ipAE and RMSD values are in Å. The pLDDT is averaged over the predicted complex on a $[0, 1]$ scale. The dense success reward reaches one when all thresholds are met and decreases linearly with their normalized violations. This thresholding encourages the model to produce successful candidates, but avoids overoptimization.

Force Fields & Physics. Molecular interactions are governed by physics, yet generative pre-training does not explicitly optimize binding energetics, and structure prediction confidence scores do not quantify them. This motivates alignment with physics-based rewards. In fact, Rosetta force field-based metrics (Alford et al., 2017) already support candidate filtering in binder design pipelines (Pacesa et al., 2025). Therefore, we introduce a *Rosetta binding-energy reward*, evaluated using the GPU-accelerated TMol implementation (Leaver-Fay et al., 2025) and applied to refolded complexes. Specifically, we refold the decoded binder $\mathbf{x}_i = \mathcal{D}(\mathbf{x}_i^{C_\alpha}, \mathbf{z}_i)$ together with its target $\mathbf{c}^{\text{target}}$ and relax the complex using TMol. We denote the resulting binder and target conformations by $\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{x}}_i^{\text{tgt}}$, respectively. We then evaluate the interface energy, adjusted by desolvation/solvation corrections, which can be written as the following energy difference with both partners held at their relaxed complex conformations:

$$\begin{aligned} E_i^{\text{Rosetta-}\Delta\Delta G} &= E_{\text{TMol}}(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_i^{\text{tgt}}) - E_{\text{TMol}}(\widetilde{\mathbf{x}}_i) - E_{\text{TMol}}(\widetilde{\mathbf{x}}_i^{\text{tgt}}) \\ &= \underbrace{\Delta E_{i,\texttt{fa_ljatr}} + 0.55\,\Delta E_{i,\texttt{fa_ljrep}} + \Delta E_{i,\texttt{fa_elec}} + \Delta E_{i,\texttt{hbond}}}_{\text{direct interactions}} \\ &\quad \underbrace{\begin{aligned} &+ \Delta E_{i,\texttt{fa_lk}} - 0.38\,\Delta E_{i,\texttt{lk_ball_iso}} + 0.92\,\Delta E_{i,\texttt{lk_ball}} \\ &- 0.33\,\Delta E_{i,\texttt{lk_bridge}} - 0.33\,\Delta E_{i,\texttt{lk_bridge_uncpl}} \end{aligned}}_{\text{solvation/desolvation}} \end{aligned} \tag{4}$$

Here, $\Delta E_{i,k} = E_k(\widetilde{\mathbf{x}}_i, \widetilde{\mathbf{x}}_i^{\text{tgt}}) - E_k(\widetilde{\mathbf{x}}_i) - E_k(\widetilde{\mathbf{x}}_i^{\text{tgt}})$ is the contribution of energy term k to interface formation; the coefficients are Tmol's default `beta2016` weights. The direct interaction terms describe Lennard–Jones attraction and repulsion, electrostatics, and hydrogen bonding. The solvation terms comprise isotropic Lazaridis–Karplus solvation, directional solvation corrections, and water-bridge contributions. Internal conformational and reference terms cancel. We define our reward as $R_i^{\text{Rosetta-}\Delta\Delta G} = -E_i^{\text{Rosetta-}\Delta\Delta G}/\max(L_i^{\text{int}}, 1)$, where L_i^{int} denotes the binder's number of interface residues. The normalization rewards the creation of strong interactions and not merely large interfaces.

In Silico Affinity Predictors. For protein binder design against small molecule targets, we additionally use Nesso-1 (Shenoy et al., 2026), a cofolding model trained on experimental protein–ligand affinity and potency data. Nesso-1 predicts affinity from protein sequences and ligand representations using learned inter-residue and residue–atom distance distributions, without requiring atomistic structure generation. Let $\mathbf{s}_i$ denote the amino acid sequence of the decoded binder $\mathbf{x}_i = \mathcal{D}(\mathbf{x}_i^{C_\alpha}, \mathbf{z}_i)$, and let $\mathbf{m}$ denote the small molecule target specified by $\mathbf{c}^{\text{target}}$. We define the *Nesso binding-affinity reward* as

$$R_i^{\text{Affinity}} = -a_{\text{Nesso}}(\mathbf{s}_i, \mathbf{m}), \tag{5}$$

where a_{Nesso} is Nesso-1's ensemble-averaged continuous affinity prediction. The negative sign assigns higher rewards to protein sequences predicted to bind more strongly to the specified small molecule.

Diversity Reward. We cluster generated binder structures of successful designs within each target's rollout group $\mathcal{G}$ using FoldSeek (Van Kempen et al., 2024) at a TM-score threshold of 0.5, and assign

$$R_i^{\text{div}} = \begin{cases} 1/|\mathcal{C}_i^{\text{success}}|, & \text{if design } i \text{ is successful and has a valid cluster assignment,} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The set $\mathcal{C}_i^{\text{success}}$ contains the successful designs assigned to the same fold cluster as design i, including i itself; its size is therefore at least one whenever the bonus is nonzero, and the reward allocates equal total credit to each successful fold cluster across the rollout group $\mathcal{G}$. This novel diversity reward incentivizes the generation of diverse yet successful candidates in separate, small clusters. Maintaining high generation diversity during post-training is particularly important in drug discovery applications to ensure diverse candidates can be experimentally tested.

Stability Reward. In rare cases, we observed that RL post-training led to unrealistic, extended binder structures. To address this, we propose a stability reward that encourages burial of hydrophobic residues in the binder itself, which has been shown to correlate with protein stability (Cho et al., 2026):

$$R_i^{\text{stability}} = \text{clip}\left(\frac{\overline{\text{NPSA}}_i - 40}{8}, -1, 0\right), \qquad \overline{\text{NPSA}}_i = \frac{1}{N_i}\sum_{j \in \mathcal{H}_i} [A_{\max}(s_{ij}) - \text{SASA}_{ij}]_+. \tag{7}$$

Model	# Folding successes ↑			TMol energy ↓		Interface	Stability	Time	Novelty ↓
	Total	Foldseek	MMseqs	$\Delta\Delta G$	Int.	H-bonds ↑	# Pass ↑	[s] ↓	
Genie3 (Lin et al., 2026)	13.05	9.00	12.18	-69.51	-97.52	5.52	12.27	19.56	0.88
BoltzGen (Stark et al., 2025)	14.91	7.36	11.68	-79.04	-108.2	7.70	14.27	45.33	0.87
ODesign (Zhang et al., 2025)	10.82	6.77	9.55	-83.49	-117.8	7.44	9.05	18.83	0.84
RFDiffusion3 (Butcher et al., 2025)	26.18	11.18	24.00	-75.29	-106.4	6.10	24.32	29.08	0.92
PXDesign (Team et al., 2025c)	29.36	7.59	26.59	-68.29	-96.71	5.72	27.82	16.66	0.91
ESMFold2 (Candido et al., 2026)	50.41	17.64	48.73	-94.46	-125.6	7.56	36.41	596	0.92
BindCraft (Pacesa et al., 2025)	45.73	27.82	45.68	-79.49	-109.3	7.77	39.64	1274	0.83
Proteína-Complexa (Didi et al., 2026b)	22.05	14.18	20.09	-73.57	-103.7	5.90	21.27	**0.45**	0.85
Proteína-Reforza *(ours)*	**91.14**	**61.82**	**83.91**	**-110.5**	**-159**	**10.07**	**85.77**	**0.45**	**0.81**

Table 1: **Protein binder performance.** Foldseek and MMseqs count structural and sequence clusters of successes. Energies and H-bonds average over successes; stability counts those passing its filter. Int. denotes interaction energy.

Here $\mathrm{clip}(a, -1, 0) = \min(0, \max(-1, a))$, and $\overline{\mathrm{NPSA}}_i$ is the buried nonpolar surface-area proxy per residue, in Å^2/residue. The binder length is N_i, s_{ij} is residue j's decoded amino acid identity, $\mathcal{H}_i$ indexes hydrophobic residues (A, F, I, L, M, V, W, Y), and $A_{\max}$ is the residue-type-specific maximum solvent-accessible surface area on the Tien scale. The residue surface area SASA_{ij} is measured on the AF2-refolded binder after removing the target, using a 1.4 Å solvent probe.

We train Proteína-Reforza with different reward combinations; see Sec. 4 and Appendix Tab. 18.

3.3 Cyclic Peptide Generation and Reinforcement Learning

Cyclic peptides are an important drug discovery modality with the potential to combine strong, selective binding to challenging targets with chemically tunable stability and delivery properties. We extend the Proteína-Complexa base model to head-to-tail and disulfide cyclic peptides by conditioning the flow-matching network on the closure type and bonded residue positions. A typed pair feature encodes the bonded pair and its topology. Head-to-tail closure additionally uses cyclic positional encoding; disulfide closure retains linear positional encoding. We fine-tune using either TED-derived disulfide crops or a mixture of protein dimers and CPSea head-to-tail and disulfide complexes (Sec. B.3).

Starting from each supervised fine-tuned (SFT) model, we apply RL to disulfide-conditioned peptides of 12-18 residues using the PDB-based RL target set. For this version of Proteína-Reforza, we use separate rewards tailored to cyclic peptides: Rewards include AlphaFold2 initial-guess (AF2-IG) structure confidence and self-consistency terms with sequence diversity, TMol disulfide scoring, and a bonus for satisfying all folding criteria. A disulfide validity gate assigns a reward of -2 to designs that fail the required cysteine and bond checks. Training details and evaluation criteria in Secs. D and H.3.

3.4 Reward Models from Experimental Data

While structure prediction confidence scores correlate with experimental binding (Overath et al., 2025), they estimate structural accuracy rather than directly predict experimental success. Inspired by RL with human feedback (Ouyang et al., 2022), we train a reward model directly on experimental data to classify binding success.

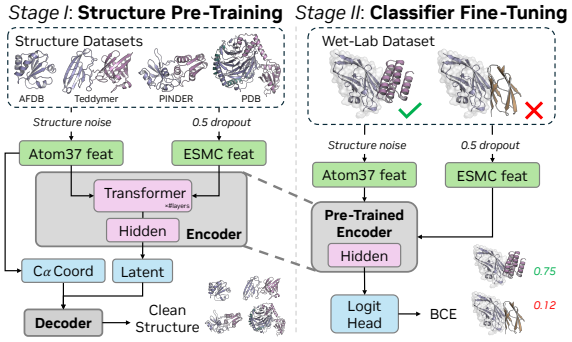


Figure 3: **Two-stage classifier training pipeline.**

We train the classifier in two stages (Fig. 3). First, we pre-train a transformer encoder-decoder on protein structures using a denoising objective, learning general representations of protein geometry informed by ESMC protein language model embeddings (Candido et al., 2026). Next, we attach a classification head to the encoder and fine-tune with cross-entropy on binary binding success outcomes from a phage-display campaign from Didi et al. (2026a), using 646k binder designs across 125 targets.

The classifier serves as a proxy for experimental success, guiding generation towards designs predicted to succeed in the lab. We restrict RL training and evaluation to a filtered subset of the classifier's training targets, with holdout test targets. The *classifier-only reward* $R_i^{\mathrm{Clf}} = \min\{\mathrm{sigmoid}(\mathrm{Clf}(\mathbf{x}_i)), 0.7\}$, where Clf denotes the classifier logit, is combined with AF2 sigals to form the *wet-lab reward* as $R_i^{\mathrm{WL}} = 0.5R_i^{\mathrm{Clf}} + 0.5R_i^{\mathrm{succ}}$. This tests generalization to targets unseen during RL while remaining within the classifier's target coverage. This proof of concept of reinforcement learning with wet-lab feedback showcases the potential for systematic model refinement using experimental data, including emerging lab-in-the-loop frameworks. See Sec. E for more details.

Table 2: **Folding and TMol RL.** Mean success, TMol and interface metrics over 22 targets.

Model / RL setting	Foldseek	TMol energy ↓		Interface
	# Successes ↑	$\Delta\Delta G$	Int.	H-bonds ↑
Proteína-Complexa (base)	14.18	-73.57	-103.7	5.90
Proteína-Reforza $_{\text{Folding only}}$	**61.82**	-110.5	-159.0	10.07
Proteína-Reforza $_{\text{TMol only}}$	32.36	-99.2	-140.1	8.04
Proteína-Reforza $_{\text{Folding+TMol}}$	56.55	**-128.5**	**-184.1**	**11.41**

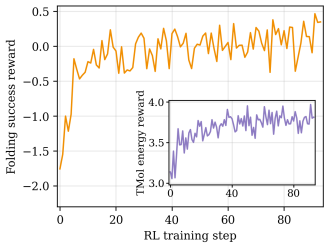


Figure 5: **Training.** AlphaFold2 folding success reward and TMol energy reward (inset) throughout RL training on general RL PDB target set. Higher values indicate better rewards.

4 In silico Experiments

4.1 Protein Targets

General RL. We trained a Proteína-Reforza model with only the folding success reward R_i^{succ} against protein targets and compare performance to the Proteína-Complexa base model and the strongest, publicly available baseline methods Genie3, BoltzGen, ODesign, RFDiffusion3, PXDesign, ESMFold2 and BindCraft; some of these methods have previous versions, which are now outdated and hence omitted. Following Didi et al. (2026b), in silico success is quantified via thresholded AF2-based confidence scores and refolding (Sec. 3.2), considering both total and clustered successes. We also measure TMol-based Rosetta $\Delta\Delta G$ energy score, only the direct interaction term, and the number of interface hydrogen bonds. We further introduce a stability metric that quantifies the geometric and biophysical integrity of the generated binders (see Sec. D), and we measure per-sample generation time and novelty. During RL training, we randomly use 4 sets of AF2 weights, similar to Pacesa et al. (2025); evaluation uses a separate held-out weight set, as detailed in Sec. D. In Sec. G.2, we further show that the RL gains generalize across different folding models. Evaluation is averaged over the 22 benchmark targets previously used by Didi et al. (2026b), considering 200 generated binder candidates for each target. Evaluation and baseline details in Secs. D and I. The results in Tab. 1 demonstrate that Proteína-Reforza outperforms all baselines, including the base model, by a large margin. Post-training on success metrics does not only boost folding success on held-out AF2 weights, but also translates to improve TMol scores and stronger hydrogen bonding. Meanwhile, our model is the fastest by a large margin; its network architecture, inherited from the base model, only uses efficient transformer networks without expensive triangle layers, and it also does not require iterative folding model backpropagation like hallucination methods. Proteína-Reforza also produces the most novel binders.

Target-specific RL. We show target-specific RL for five targets. Fig. 4 compares the number of Foldseek-unique successful designs among 200 candidates generated by the base model, general RL, and target-specific RL. General RL results use the Proteína-Reforza checkpoint evaluated above; target-specific models are obtained through separate post-training runs on top and additionally use the new diversity reward. Target-specific RL increases successes over general RL on all five targets, including from 2 to 23 for H1 and from 0 to 3 for TNFα. This highlights the value of post-training for challenging targets: general and target-specific RL can substantially improve success rates even when the base model yields no successes.

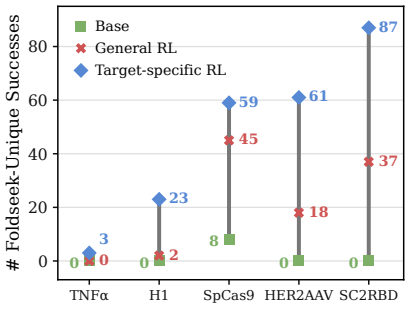


Figure 4: **Target-specific RL increases structurally unique successes.**

Force Field RL. Fig. 5 shows successful reward optimization during RL training of Proteína-Reforza $_{\text{Folding+TMol}}$, which combines folding success and TMol energy rewards. Joint optimization substantially improves folding success metrics, and it also yields more favorable interface energy and hydrogen-bonding scores (Tab. 2) than both base model and Proteína-Reforza $_{\text{Folding only}}$. Training with the TMol reward alone also outperforms the base model, but combining both rewards performs better overall. This proof of concept demonstrates that RL can align molecular generative models with force field energetics, supporting the broader integration of physics-based objectives into molecular design.

Stability and Diversity. Tab. 1 shows that the post-trained Proteína-Reforza model maintains diversity and stability: It leads in unique success rate and its stability pass rate almost matches its success rate, indicating stable reward fine-tuning. In Fig. 24, we also report the change in amino acid distribution; the changes are single-digit percentage points, implying the model does not drastically change its output, despite significantly enhanced performance. We also report a set of physical and geometric interface metrics in Fig. 25; Proteína-Reforza shows enhanced hydrogen bonding, improved

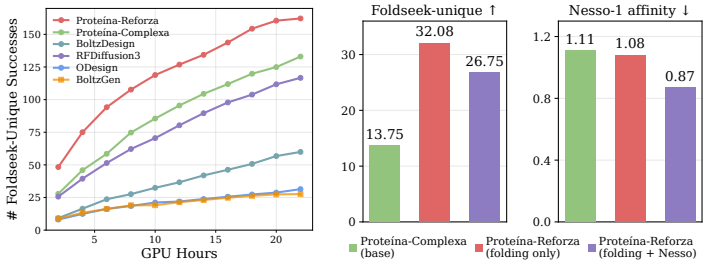


Figure 6: **Ligand scaling and reward comparison.** (*left*) Mean Foldseek-unique successes versus A100 GPU hours, including generation and evaluation. (*right*) Mean unique successes and Nesso-1 affinity. Both panels show means over 12 targets; affinity is averaged over successful designs.

shape complementarity, and considerably larger interfaces. Overall, these findings are consistent with improved binding without indications of reward hacking, in line with analyses in Sec. G.2 showing that in silico success boosts generalize across folding models. Nonetheless, in rare cases we found that RL can lead to unrealistic binders not passing stability criteria. An example is shown in Fig. 22 (a). However, additional training with our stability reward $R_i^{\text{stability}}$ can mitigate this effect; see Fig. 22 (b).

4.2 Small Molecule Targets

RL and Inference-time scaling. Next, we evaluated Proteína-Reforza for ligand targets, trained with both folding success and diversity rewards. Highlighting our model's efficiency, we compare to previous methods under normalized compute, reporting unique successes as resources scale; see Fig. 6 (left). Compute budgets include both generation and scoring with RF3. We find that the post-trained model outperforms both the base model as well as the baseline methods by a large margin (Tab. 14 in Sec. F.2 reports full comparison; setup and per-target analyses in Sec. G.7).

Nesso Affinity Reward. Fig. 6 (*right*) compares the base model with folding success-reward RL and its Nesso variant. Note that lower Nesso affinity scores indicate stronger predicted binding. Adding the Nesso-1 reward significantly improves mean predicted affinity among successful designs on all 12 held-out test targets (Fig. 12) against both base and folding-reward RL model. Foldseek-unique success counts are slightly lower than for the folding-only RL model, but remain much higher than for the base model. Per-target affinity distributions and unique-success counts in Sec. F.2 show per-ligand results. These results showcase the promise of complementary affinity RL, which we believe to be increasingly relevant with emerging affinity predictors (Passaro et al., 2025; Shenoy et al., 2026).

4.3 Terminal-Disulfide Peptide Binders

We evaluate terminal-disulfide peptide design (Sec. 3.3) on 12 protein targets using our fine-tuned Proteína-Complexa base model as well as the Proteína-Reforza RL model (see Tab. 3). Note that evaluation here requires both folding and closure criteria (see Sec. D). RL increases mean successes per target from 12.75 to 19.75 (1.55×), compared with 1.42 for BoltzGen. The unique success rates similarly increase, indicating overall improved computational yield and diversity. For experimental validation, see Sec. 5. Additional in silico evaluations for head-to-tail cyclic peptide design are presented in Sec. G.9. The results confirm that our proposed RL framework not only works for linear minibinder design but also for cyclic peptides.

Table 3: **Disulfide cyclic peptide design.** Average results per target for 200 designs per target across 12 targets.

Model	# Folding successes ↑			TMol energy ↓		Interface
	Total	Foldseek	MMseqs	$\Delta\Delta G$	Int.	H-bonds ↑
BoltzGen (Stark et al., 2025)	1.42	1.08	0.83	-53.09	-56.85	2.58
Proteína-Complexa $_{\text{Disulfide SFT}}$ *(ours)*	12.75	5.42	4.25	**-53.89**	-63.43	2.23
Proteína-Reforza $_{\text{Disulfide}}$ *(ours)*	**19.75**	**6.17**	**4.58**	-52.69	**-65.13**	**2.98**

4.4 Experimental Data Reinforcement Learning

We benchmark wet-lab reward-guided post-training performance on holdout target crops that Proteína-Reforza does not see during RL, but for which the classifier was trained. We train models with classifier reward R_i^{Clf}, folding success reward R_i^{succ}, mixed folding/classifier reward R_i^{WL} and an SFT baseline, see Tab. 4. Compared to the success reward alone, the wet-lab-informed post-training run achieves better generation quality, with increases across all metrics. Furthermore, the wet-lab classifier scores double compared to the base model, demonstrating the effectiveness of optimizing the wet-lab classifier reward. In contrast, the success-only reward model lags in classifier scores, and the SFT baseline fails to optimize classifier scores despite high success counts. Meanwhile, the classifier-only reward achieves a drastic improvement in classifier scores over all other variants but at the cost of fewer successes, highlighting the trade-off in reward choices. This proof-of-concept of RL with experimental feedback highlights the potential for wet-lab-informed generative model training.

Table 4: **Proteína-Reforza's performance for RL on wet-lab classifier rewards.** Metrics averaged across 50 holdout crops, each with 100 samples. Classifier scores use the same checkpoints as RL training.

Model	Classifier Score ↑			# Folding successes ↑			TMol energy ↓		Interface	Stability	Novelty ↓
	Prob	Logit	# $p > 0.5$	Total	Foldseek	MMseqs	$\Delta\Delta G$	Int.	H-bonds ↑	# Pass ↑	
Proteína-Complexa (base)	0.037	-14.63	0.80	1.26	1.08	1.04	-61.85	-76.04	3.62	1.26	0.86
Proteína-Complexa (SFT)	0.016	-13.93	1.26	8.74	3.70	7.98	-63.59	-79.01	4.30	8.70	0.87
Proteína-Reforza w/ R^{WL}	0.085	-11.84	1.54	**13.88**	**9.00**	**13.42**	**-85.29**	**-113.90**	**6.50**	**12.68**	0.85
Proteína-Reforza w/ R^{succ}	0.054	-14.20	1.40	13.70	7.86	13.16	-83.74	-112.13	6.23	12.16	0.86
Proteína-Reforza w/ R^{Clf}	**0.136**	**-8.46**	**6.68**	0.96	0.68	0.94	-71.51	-82.59	3.52	0.94	**0.84**

5 Wet-Lab Validation

Protein Binders. We compared Proteína-Complexa with a preliminary version of Proteína-Reforza on the PD-L1 target in experiments using Surface Plasmon Resonance (SPR). We deliberately generated a small candidate pool of only 2,000 designs per model, which we filtered down to 60 test candidates from each using comprehensive filtering criteria, prioritizing strict-filter candidates and filling any shortfall with relaxed-filter candidates. All 60 selected RL model designs passed the strict criteria, compared with 9 base model designs. In the experiments, the RL-based candidates yielded 7 successful binders, versus zero for the base model. The best binder, visualized in Fig. 2 (a), had an overall apparent equilibrium dissociation constant K_D of 64.7 nM. Its concentration-dependent SPR responses are shown in Fig. 7. These results support improved experimental hit discovery under a limited sampling budget: Proteína-Reforza's improved in silico metrics translate to better experimental performance in the lab. Training, selection, and experimental details in Sec. J.1.

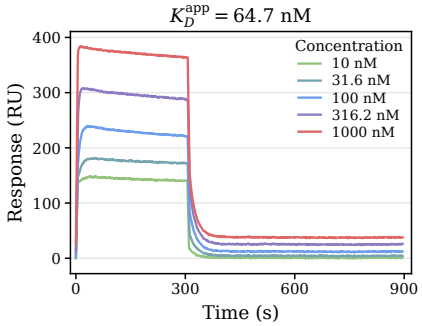


Figure 7: **SPR response of the best PD-L1 binder.** Representative measurements at 5 concentrations (10–1,000 nM).

Cyclic Peptide Binders. We compared terminal-disulfide peptide designs by Proteína-Complexa$_{\text{Disulfide SFT}}$ and Proteína-Reforza$_{\text{Disulfide}}$ for the GABARAP, PD-L1, and TROP2 targets using mRNA display under equal testing budgets. For each target, we selected 10,000 candidates per model using the same tiered filtering, combining structure prediction and interface energy metrics. Among the combined top 50 candidates ranked by target-specific read counts, RL contributed 28 designs versus 22 for SFT on GABARAP, 27 versus 23 on PD-L1, and 29 versus 21 on TROP2. Fig. 8 shows TROP2 results; GABARAP and PD-L1 results in Fig. 37. We again find that RL post-training's in silico boosts translate to the laboratory. The results describe within-target assay enrichment; direct binding-affinity measurements are planned for top designs. Selection and experimental details in Sec. J.2.

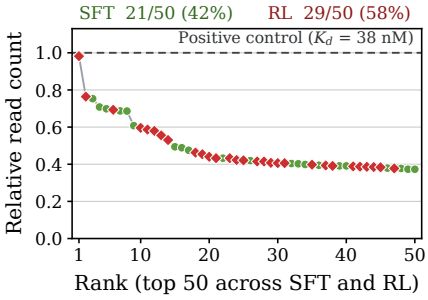


Figure 8: **Cyclic-peptide screening against TROP2.** Top 50 candidates are shown and ranked by read count.

See appendix for many additional analyses, ablations and in silico experiments, including diversity reward ablations, analyses of interfaces learnt by RL, hotspot commitment studies, and more.

6 Conclusions

Proteína-Reforza establishes RL post-training as a practical route to aligning protein generators with design objectives. We demonstrate substantial gains across minibinder and cyclic peptide design, with experimental validation confirming increased binding success. Structure, sequence and biophysical analyses support design plausibility, with no systematic evidence of reward hacking. The framework supports general and target-specific alignment through computational, physical, and experimentally learned rewards. Post-training improves candidate yield and reduces reliance on extensive sampling and filtering. More accurate physical models, broader experimental datasets, and iterative laboratory feedback could improve binding affinity, specificity, and developability. Coupling generation, experimental testing, and reward-model refinement could enable a closed learning loop where each campaign informs the next, allowing molecular design models to improve as evidence accumulates.

AI USE STATEMENT

In this work, we used generative AI tools and large language models (LLMs) to assist us. We utilized LLMs to draft sections of the appendix, to improve writing style and to catch typographical and grammatical errors. Moreover, coding assistants were used to accelerate research execution and code development, and we also used AI tools for retrieval and discovery of related work. All AI-assisted writing and work was reviewed by the authors and we take responsibility for the final content of this manuscript, including text, claims or artifacts produced with the aid of generative AI. No LLMs were used for main scientific research ideation.

ETHICS STATEMENT

Reinforcement learning and post-training offer opportunities to direct protein generation towards beneficial applications in therapeutics, diagnostics, and basic research. By incorporating computational and experimental feedback into generative models, these approaches could improve discovery efficiency, reduce the resources spent on unsuccessful candidates, and unlock new applications beyond current design capabilities. Despite this promise, generative models for molecular design also carry risks of misuse. We advocate careful assessment of intended uses, appropriate oversight, and proportionate safeguards for model deployment.

REPRODUCIBILITY STATEMENT

We support independent reproduction of our results by documenting the complete workflow, from training data construction and model implementation to training, generation, and evaluation. Please see our comprehensive appendix. These descriptions will be complemented by public releases of the code and model weights.

Training Data and Evaluation Targets. Training data and target sets for reinforcement learning are discussed in Sec. B. Evaluation targets are listed in Sec. C.

Model Implementation and Training. Sec. H presents complete training protocols and hyperparameters for reinforcement learning post-training and inference. Moreover, Sec. E discusses the details of the wet-lab binding success classifier and the corresponding reinforcement learning details.

Generation and Benchmarking. Evaluations presented in this paper are supported by metric definitions in Sec. D. Baseline method evaluation is discussed in Sec. I.

Experimental Testing. Details on the experimental testing of candidates in the lab are provided in Sec. J.

Code and Model Release. Upon acceptance, we will release the source code and trained model weights under permissive licenses.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Zemgulyte, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Zidek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. 3, 27

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3

Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Jr. Dunbrack, Roland L., Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 06 2017. doi: 10.1021/acs.jctc.7b00125. 2, 5

Christof Angermueller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell. Model-based reinforcement learning for biological sequence design. In *International conference on learning representations*, 2020. 3

Anonymous. Proteina-pronto: Fast de novo protein design and structure prediction via distillation. *Under review at ICLR 2027*, 2026. 3

Rolf Apweiler, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai-Su L. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 32(suppl_1):D115–D119, 2004. 22

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000. 4, 22

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025. doi: 10.1101/2025.09.18.676967. URL https://www.biorxiv.org/content/early/2025/09/18/2025.09.18.676967. 3, 6, 21, 51

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://www.biorxiv.org/content/early/2026/06/04/2026.06.03.729735. 3, 6, 22, 52

Yehlin Cho, Martin Pacesa, Zhidian Zhang, Bruno E. Correia, and Sergey Ovchinnikov. Boltzdesign1: Inverting all-atom structure prediction model for generalized biomolecular binder design. *bioRxiv*, 2025. doi: 10.1101/2025.04.06.647261. URL https://www.biorxiv.org/content/early/2025/04/06/2025.04.06.647261. 3, 23, 52

Yehlin Cho, Kotaro Tsuboyama, Theodore J. Litberg, Michelle D. Jung, Adunoluwa Obisesan, Qian Wang, Claire M. Phoumyvong, Jane Thibeault, Sergey Ovchinnikov, and Gabriel J. Rocklin. Accurate protein stability prediction for small domains using mega-scale experiments. *bioRxiv*, 2026. doi: 10.64898/2026.05.19.726285. URL https://www.biorxiv.org/content/early/2026/05/20/2026.05.19.726285. 5

Nathaniel Corley, Simon Mathis, Rohith Krishna, Magnus S. Bauer, Tuscan R. Thompson, Woody Ahern, Maxwell W. Kazman, Rafael I. Brent, Kieran Didi, Andrew Kubaney, Lilian McHugh, Arnav Nagle, Andrew Favor, Meghana Kshirsagar, Pascal Sturmfels, Yanjing Li, Jasper Butcher, Bo Qiang, Lars L. Schaaf, Raktim Mitra, Katelyn Campbell, Odin Zhang, Roni Weissman, Ian R. Humphreys, Qian Cong, Jonathan Funk, Shreyash Sonthalia, Pietro Liò, David Baker, and

Frank DiMaio. Accelerating biomolecular modeling with atomworks and rf3. *bioRxiv*, 2025. doi: 10.1101/2025.08.14.670328. URL https://www.biorxiv.org/content/early/2025/08/15/2025.08.14.670328. 3, 4

Justas Dauparas, Ivan V. Anishchenko, Nathaniel R. Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Robbert J. de Haas, Neville P. Bethel, Philip J. Y. Leung, Timothy F. Huddy, Samuel J. Pellock, Doug K Tischer, Frank Chan, Brian Koepnick, H. Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615): 49–56, 2022. 51

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using ligandmpnn. *Nature Methods*, 22(4):717–723, 2025. 51

Kieran Didi, Danny Reidenbach, Matthew Penner, Supriya Ravichandran, Marshall Case, Mike Nichols, Erik Swanson, Alex Reis, Maggie Prescott, Yue Qian, Dongming Qian, Jingjing Yang, Weiji Li, Le Li, Daichi Shonai, Sean Gay, Bhoomika Basu Mallik, Ho Yeung Chim, Liurong Chen, Miguel Atienza Juanatey, Hubert Klein, Dominic Rieger, Phillip Schlegel, Anna U. Macintyre, Maxim Secor, Daniele Granata, Sooyoung Cha, Zhonglin Cao, Guoqing Zhou, Tomas Geffner, Xi Chen, Micha Livne, Zuobai Zhang, Tianjing Zhang, Kyle Gion, Michael M. Bronstein, Martin Steinegger, Kristine Deibler, Scott Soderling, Clara T. Schoeder, Alena Khmelinskaia, Florian Hollfelder, Christian Dallago, Emine Kucukbenli, Arash Vahdat, Pierce Ogden, and Karsten Kreis. Latent generative search unlocks de novo design of untapped biomolecular interactions at scale. *bioRxiv*, 2026a. doi: 10.64898/2026.09.12.751118. URL https://www.biorxiv.org/content/early/2026/09/18/2026.09.12.751118. 1, 6, 21, 52

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b. 1, 2, 3, 4, 6, 7, 20, 21, 22, 23, 26, 40, 46, 50, 51

Roland L Dunbrack Jr. Rēs ipsae loquunt: What's wrong with alphafold's iptm score and how to fix it. *Biorxiv*, 2025. 28

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Danny Kovtun, Emanuele Rossi, Guoqing Zhou, Srimukh Veccham, Clemens Isert, Yuxing Peng, Prabindh Sundareson, Mehmet Akdel, Gabriele Corso, Hannes Stärk, Zachary Carpenter, Michael Bronstein, Emine Kucukbenli, Torsten Schwede, and Luca Naef. Plinder: The protein-ligand interactions dataset and evaluation resource. *BioRxiv*, pp. 2024–07, 2024. 4, 20

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034. URL https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034. 4

Theodore L Fobe, Andrei Kazakov, and Demian Riccardi. Cys. sqlite: a structured-information approach to the comprehensive analysis of cysteine disulfide bonds in the protein databank. *Journal of chemical information and modeling*, 59(2):931, 2019. 20

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. 22

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. In *The Fourteenth International Conference on Learning Representations*, 2026. 1, 2, 46

Casper A. Goverde, Benedict Wolf, Hamed Khakzad, Stéphane Rosset, and Bruno E. Correia. De novo protein design by inversion of the alphafold structure prediction network. *Protein Science*, 32(6):e4653, 2023. doi: https://doi.org/10.1002/pro.4653. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4653. 3

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 623:1070–1078, 2023. 3

Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. Biological sequence design with gflownets. In *International conference on machine learning*, pp. 9786–9801. PMLR, 2022. 3

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 3, 22

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*, 2025. 3

David E. Kim, Davin R. Jensen, David Feldman, Doug Tischer, Ayesha Saleem, Cameron M. Chow, Xinting Li, Lauren Carter, Lukas Milles, Hannah Nguyen, Alex Kang, Asim K. Bera, Francis C. Peterson, Brian F. Volkman, Sergey Ovchinnikov, and David Baker. De novo design of small beta barrel proteins. *Proceedings of the National Academy of Sciences*, 120(11):e2207974120, 2023. doi: 10.1073/pnas.2207974120. 26

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 2

Daniel Kovtun, Mehmet Akdel, Alexander Goncearenco, Guoqing Zhou, Graham Holt, David Baugher, Dejun Lin, Yusuf Adeshina, Thomas Castiglione, Xiaoyun Wang, Céline Marquet, Matt McPartlon, Tomas Geffner, Gabriele Corso, Hannes Stärk, Zachary Carpenter, Emine Kucukbenli, Michael Bronstein, and Luca Naef. Pinder: the protein interaction dataset and evaluation resource. *bioRxiv*, pp. 2024–07, 2024. 22

Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A. Sutherland, Andrew Hitchcock, C. Neil Hunter, Alex Kang, Evans Brackenbrough, Asim K. Bera, Minkyung Baek, Frank DiMaio, and David Baker. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693): eadl2528, 2024. doi: 10.1126/science.adl2528. URL https://www.science.org/doi/abs/10.1126/science.adl2528. 23

Nathan Lambert. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*, 2026. 2

Andy M Lau, Nicola Bordin, Shaun M Kandathil, Ian Sillitoe, Vaishali P Waman, Jude Wells, Christine A Orengo, and David T Jones. Exploring structural diversity across the protein universe with the encyclopedia of domains. *Science*, 386(6721):eadq4946, 2024. 20, 21, 22

Andrew Leaver-Fay, Jeff Flatten, Alex Ford, Joseph Kleinhenz, Henry Solberg, David Baker, Andrew M. Watkins, Brian Kuhlman, and Frank DiMaio. tmol: a gpu-accelarated, pytorch implementation of rosetta's relax protocol. https://github.com/uw-ipd/tmol, 2025. 5

Yeqing Lin, Minji Lee, Aakarsh Vermani, Ellena Jiang, Sebastiaan De Cooman, Matej Špeťko, and Mohammed AlQuraishi. Fast and ultra-capable protein design: Advancing the frontier through atomistic se(3)-equivariance with genie 3. *bioRxiv*, 2026. doi: 10.64898/2026.05.01.722168. URL https://www.biorxiv.org/content/early/2026/05/05/2026.05.01.722168. 3, 6, 51

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a. 3

Mingjie Liu, Shizhe Diao, Jian Hu, Ximing Lu, Xin Dong, Hao Zhang, Alexander Bukharin, Shaokun Zhang, Jiaqi Zeng, Makesh Narsimhan Sreedhar, Gerald Shen, David Mosallanezhad, Di Zhang, Jonas Yang, June Yang, Oleksii Kuchaiev, Guilin Liu, Zhiding Yu, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. Scaling up rl: Unlocking diverse reasoning in llms via prolonged training, 2025b. 3

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 1, 2, 3, 6

Max D Overath, Andreas SH Rygaard, Christian P Jacobsen, Valentas Brasas, Oliver Morell, Pietro Sormanni, and Timothy P Jenkins. Predicting experimental success in de novo binder design: a meta-analysis of 3,766 experimentally characterised binders. *BioRxiv*, pp. 2025–08, 2025. 6

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025. 1, 3, 5, 6, 7, 23, 52

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707. 8

Stephen A. Rettie, Katelyn V. Campbell, Asim K. Bera, Alex Kang, Simon Kozlov, Yensi Flores Bueso, Joshmyn De La Cruz, Maggie Ahlrichs, Suna Cheng, Stacey R. Gerben, Mila Ya-Lan Lamb, Analisa Murray, Victor Adebomi, Guangfeng Zhou, Frank Dimaio, Sergey Ovchinnikov, and Gaurav Kumar Bhardwaj. Cyclic peptide structure prediction and design using AlphaFold2. *Nature Communications*, 16:4730, 2025a. doi: 10.1038/s41467-025-59940-7. 24, 45

Stephen A. Rettie, David Juergens, Victor Adebomi, Yensi Flores Bueso, Qinqin Zhao, Alexandria N Leveille, Andi Liu, Asim K. Bera, Joana A. Wilms, Alina Üffing, Alex Kang, Evans Brackenbrough, Mila Ya-Lan Lamb, Stacey R. Gerben, Analisa Murray, Paul M. Levine, Maika Schneider, Vibha Vasireddy, Sergey Ovchinnikov, Oliver H. Weiergräber, Dieter Willbold, Joshua A. Kritzer, Joseph D. Mougous, David Baker, Frank Dimaio, and Gaurav Kumar Bhardwaj. Accurate de novo design of high-affinity protein-binding macrocycles using deep learning. *Nature Chemical Biology*, 21:1948–1956, 2025b. doi: 10.1038/s41589-025-01929-w. 24

Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017. doi: 10.1126/science.aan0693. 26

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 4

Nikhil Shenoy, David Errington, Emmanuel Bengio, Kacper Kapuśniak, Kerstin Klaeser, Yui Tik Pang, Vladimir Radenkovic, Prudencio Tossou, Therence Bois, Andrew Wedlake, and Francesco Di Giovanni. Nesso-1: Accelerating open-source binding affinity predictions. *bioRxiv*, 2026. doi: 10.64898/2026.08.01.742196. URL https://www.biorxiv.org/content/10.64898/2026.08.01.742196v1. 2, 5, 8

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. Boltzgen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494. URL https://www.biorxiv.org/content/early/2025/11/24/2025.11.20.689494. 3, 6, 8, 21, 51

Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):2542, 2018. 19

Filippo Stocco, Maria Artigues-Lleixa, Andrea Hunklinger, Talal Widatalla, Marc Guell, and Noelia Ferruz. Guiding generative protein language models with reinforcement learning. *arXiv preprint arXiv:2412.12979*, 2024. 3

Xingyu Su, Xiner Li, Masatoshi Uehara, Sunwoo Kim, Yulai Zhao, Gabriele Scalia, Ehsan Hajiramezanali, Tommaso Biancalani, Degui Zhi, and Shuiwang Ji. Iterative distillation for reward-guided fine-tuning of diffusion models in biomolecular design. In *International Conference on Learning Representations*, volume 2026, pp. 139443–139469, 2026. 3

Haoran Sun, Liang He, Pan Deng, Guoqing Liu, Zhiyu Zhao, Yuliang Jiang, Chuan Cao, Fusong Ju, Lijun Wu, Haiguang Liu, Tao Qin, and Tie-Yan Liu. Accelerating protein engineering with fitness landscape modelling and reinforcement learning. *Nature Machine Intelligence*, 7(9):1446–1460, 2025. 3

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *BioRxiv*, pp. 2025–01, 2025a. 51

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu,

Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025b. 3

Protenix Team, Milong Ren, Jinyuan Sun, Jiaqi Guan, Cong Liu, Chengyue Gong, Yuzhe Wang, Lan Wang, Qixu Cai, Xinshi Chen, and Wenzhi Xiao. Pxdesign: Fast, modular, and accurate de novo design of protein binders. *bioRxiv*, 2025c. doi: 10.1101/2025.08.15.670450. URL https://www.biorxiv.org/content/early/2025/08/16/2025.08.15.670450. 3, 6, 24, 51

Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PLOS ONE*, 8(11):e80635, 2013. doi: 10.1371/journal.pone.0080635. 26

Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024. 5, 19, 22

Mihály Váradi, Stephen Anyango, Mandar S. Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yu Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John M. Jumper, Ellen Clancy, Richard Green, Ankur Vora, M. Wasilul Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard J. Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022. 20

Chenyu Wang, Masatoshi Uehara, Yichun He, Amy Wang, Avantika Lal, Tommi Jaakkola, Sergey Levine, Aviv Regev, Hanchen Wang, and Tommaso Biancalani. Fine-tuning discrete diffusion models via reward optimization with applications to dna and protein design. In *International Conference on Learning Representations*, volume 2025, pp. 47871–47899, 2025a. 3

Ziwen Wang, Jiajun Fan, Ruihan Guo, Thao Nguyen, Heng Ji, and Ge Liu. Proteinzero: Self-improving protein generation via online reinforcement learning. *arXiv preprint arXiv:2506.07459*, 2025b. 3

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023. 1, 3, 21

Talal Widatalla, Ashir A Borah, Samuel H King, Claudia L Driscoll, Rafael Rafailov, and Brian L Hie. Aligning protein-generative models to experimental fitness with proteindpo. *Nature Methods*, pp. 1–9, 2026. 3

Junde Xu, Zijun Gao, Xinyi Zhou, Xingyi Cheng, Le Song, Guangyong Chen, Pheng-Ann Heng, and Jiezhong Qiu. Protein inverse folding from structure feedback. *Advances in Neural Information Processing Systems*, 38:24274–24299, 2025. 3

Yanbo Xu, Yu Wu, Sungjae Park, Zhizhuo Zhou, and Shubham Tulsiani. Temporal score rescaling for temperature sampling in diffusion and flow models. *arXiv preprint arXiv:2510.01184*, 2026. 3

Fanglei Xue, Andrew Kubaney, Zhichun Guo, Joseph K Min, Ge Liu, Yi Yang, and David Baker. Improving protein sequence design through designability preference optimization. *arXiv preprint arXiv:2506.00297*, 2025. 3

Shuchen Xue, Chongjian GE, Shilong Zhang, Yichen Li, and Zhi-Ming Ma. Advantage weighted matching: Aligning RL with pretraining in diffusion models. In *Forty-third International Conference on Machine Learning*, 2026. 1, 2, 3, 4, 46

Ziyi Yang, Hanyuan Xie, Yinjun Jia, Xiangzhe Kong, Jiqing Zheng, Ziting Zhang, Yang Liu, Lei Liu, and Yanyan Lan. Cpsea: large-scale cyclic peptide-protein complex dataset for machine learning in cyclic peptide design. *Advances in Neural Information Processing Systems*, 38, 2026. 20

Vinicius Zambaldi, David La, Alexander E. Chu, Harshnira Patani, Amy E. Danson, Tristan O. C. Kwan, Thomas Frerix, Rosalia G. Schneider, David Saxton, Ashok Thillaisundaram, Zachary Wu, Isabel Moraes, Oskar Lange, Eliseo Papa, Gabriella Stanton, Victor Martin, Sukhdeep Singh, Lai H. Wong, Russ Bates, Simon A. Kohl, Josh Abramson, Andrew W. Senior, Yilmaz Alguel, Mary Y. Wu, Irene M. Aspalter, Katie Bentley, David L. V. Bauer, Peter Cherepanov, Demis Hassabis, Pushmeet Kohli, Rob Fergus, and Jue Wang. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024. 23, 24

Odin Zhang, Xujun Zhang, Haitao Lin, Cheng Tan, Qinghan Wang, Yuanle Mo, Qiantai Feng, Gang Du, Yuntao Yu, Zichang Jin, Ziyi You, Peicong Lin, Yijie Zhang, Yuyang Tao, Shicheng Chen, Jack Xiaoyu Chen, Chenqing Hua, Weibo Zhao, Runze Ma, Yunpeng Xia, Kejun Ying, Jun Li, Yundian Zeng, Lijun Lang, Peichen Pan, Hanqun Cao, Zihao Song, Bo Qiang, Jiaqi Wang, Pengfei Ji, Lei Bai, Jian Zhang, Chang yu Hsieh, Pheng Ann Heng, Siqi Sun, Tingjun Hou, and Shuangjia Zheng. Odesign: A world model for biomolecular interaction design. *arXiv preprint arXiv:2510.22304*, 2025. 3, 6, 51

Yulin Zhang, He Cao, Zihao Jiang, Chenyi Zi, Zhipeng Zhou, Zijing Liu, Yu Li, Jia Li, and Ziqi Gao. Proteinopd: Towards effective and efficient preference alignment for protein design. *arXiv preprint arXiv:2605.10189*, 2026. 3

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026. 3

Xiangxin Zhou, Dongyu Xue, Ruizhe Chen, Zaixiang Zheng, Liang Wang, and Quanquan Gu. Antigen-specific antibody design via direct energy-based preference optimization. *Advances in Neural Information Processing Systems*, 37:120861–120891, 2024. 3

Appendix

A Limitations and Future Work **19**

B Training Data and Target Sets for Reinforcement Learning **19**

B.1 Training Data and Targets for Protein Minibinder Design 19

B.2 Training Data and Targets for Ligand Minibinder Design 20

B.3 Training Data and Targets for Disulfide Peptide Design 20

B.4 Training Data and Targets for Wet-Lab Alignment 21

B.5 Training Data for Wet-Lab Binding Classifier 21

C Evaluation Targets **23**

C.1 Protein Test Targets for Minibinder Design 23

C.2 Ligand Test Targets for Minibinder Design 23

C.3 Protein Test Targets for Peptide Design 24

C.4 Protein Test Targets from Wet-Lab Data 24

D Evaluation Metrics **24**

E Wet-Lab Binding Classifier and RL with Wet-Lab Feedback **26**

E.1 Structure Pre-Training Setup 26

E.2 Classifier Fine-Tuning Setup 27

E.3 Wet-Lab Classification Performance 28

F Individual Target Results **29**

F.1 Minibinder RL for Protein Targets 29

F.2 Minibinder RL for Ligand Targets 31

F.3 Terminal Disulfide Peptide RL for Protein Targets 33

G Additional Experiments **34**

G.1 Ablation Study: Reward Choices 34

G.2 Ablation Study: Transfer Across Folding Models 35

G.3 Ablation Study: Ligand SFT and Diversity Reward 37

G.4 Generalization to Structurally Dissimilar Targets 37

G.5 Target-specific RL with Stability Reward 38

G.6 Interface, Sequence and Bioinformatics Metrics Analysis 39

G.7 Inference-time Scaling for Proteína-Reforza 40

G.8 Hotspot Commitment and Hotspot Independence for Proteína-Reforza 43

G.9 Transfer across Peptide and Minibinder Design Tasks 44

H Model Training and Inference Details **46**

H.1 Minibinder RL for Protein Targets 46

H.2 Minibinder RL for Ligand Targets 49

H.3 Cyclic Peptide RL for Protein Targets 49

H.4 Wet Lab Reward Model RL 50

H.5 Inference Hyperparameters 50

I Baseline Information **51**

I.1 Diffusion Baseline 51

I.2 Hallucination Baseline 52

J Wet Lab Validation **52**

J.1 Minibinder Design to PD-L1 52

J.2 Cyclic Peptide Design to PD-L1 and GABARAP and TROP2 56

A Limitations and Future Work

Proteína-Reforza relies on *in silico* rewards, which are imperfect proxies for experimental outcomes. More generally, computational methods to score and rank generated designs can have biases with respect to real-world wet-lab performance. Reinforcement learning can exploit such imperfections, increasing rewards without corresponding gains in design quality. Nevertheless, we observed adversarial reward optimization only in very rare cases and our analyses reveal no evidence of systematic reward hacking in the evaluated settings: Designs generated by our post-trained models retain high structural and biophysical quality, performance gains transfer between different structure prediction models used for training and evaluation, and experimental validation confirms improved binding success in the laboratory.

Future work could broaden the range of rewards and apply Proteína-Reforza at larger scale. Our RL experiments with TMol-based Rosetta binding energy scores provide a proof of concept for physics-based alignment of molecular generative models. Incorporating feedback from learned force fields or molecular dynamics simulations could enable more detailed and accurate physical scoring during post-training. Another key direction is extending RL with experimental feedback to iterative design–test–learn cycles, allowing new measurements to refine reward models and guide subsequent generation. Finally, the framework could be adapted to additional molecular modalities beyond the protein binders and cyclic peptides studied here.

B Training Data and Target Sets for Reinforcement Learning

B.1 Training Data and Targets for Protein Minibinder Design

We construct the RL target training dataset from 246,752 PDB entries with 11,717,083 interface annotations. Each entry contains a target-only protein structure and optional specified interface residues (*hotspots*) identified from a native partner (removed during training).

Target selection pipeline. We exclude protein interfaces that have annotated clashes or covalent modifications. Entries with nucleic acids or small ligands are also removed, leaving only protein-protein interaction pairs. We require an X-ray resolution ≤ 3.5 Å or an EM resolution ≤ 4.0 Å, at least 90% resolved residues on each side, and at least 50 annotated contacts. These criteria select well-resolved interfaces in protein-only contexts. For dimer structures in the PDB entries, we choose either chain as the target with equal probability during training. Multi-chain targets containing 2 or 3 chains are eligible if they contact a common partner. We limit the target to 50-450 residues total, each chain at least 4 residues, and the full length at most 500 including the binder.

We remove sequence redundancy using MMseqs2 (Steinegger & Söding, 2018) at 100% identity and 80% coverage of target and binder sequences, respectively. Multi-chain targets are grouped by component sequence clusters and stoichiometry. We rank entries by resolution, alternate conformations, and interface contacts, with heteromeric interfaces ranking higher than homomeric ones. The highest-ranked passing representative within each target group is retained. These targets are further cleaned such that non-canonical amino acids and unresolved residues are stripped.

We identify target interface residues that have at least one heavy atom within 8 Å of its native binder. The interface residues are then ranked by the number of contacting atom pairs, with ties broken by the shortest distance to the binder. Up to 6 hotspots per chain in the rank order are selected, spaced by at least 4 residues to avoid adjacent selections. If fewer than 3 hotspot residues are selected, spacing is relaxed until 3 are selected or the interface is exhausted. We discard targets with invalid geometry or no hotspots, leading to a final count of 40,890 single-chain targets and 11,602 multi-chain targets.

Target data pipeline. To balance target structure diversity during training, we cluster the single-chain and multi-chain targets separately with FoldSeek (Van Kempen et al., 2024) using a TM-score threshold of 0.7 and 80% coverage of both structures. Multimer clustering additionally requires a per-chain TM-score of 0.5, assignment of all chains, and at least 2 aligned chains. We obtain 8,063 single-chain and 7,778 multi-chain clusters. By default, RL training samples a cluster uniformly, then a target uniformly within it. In this way, the model does not favor larger clusters.

During training, we further augment targets and hotspot annotations to make the model robust to different crops and binding-site specifications. For each rollout, the full target is used with probability

0.5. Otherwise, a random hotspot is chosen as the anchor, and a crop size is uniformly sampled from $[64, 128]$ to spatially crop the nearest residues to the anchor by Cα distance. In this case, the interface geometry is preserved, exposing the model to both full targets and local binding sites. Sequence, coordinates, chain membership, and hotspot annotations are updated accordingly and the visible target is centered. Each hotspot residue, if selected, is then independently dropped with probability 0.5. Binder length is randomly sampled from the pre-specified range.

B.2 Training Data and Targets for Ligand Minibinder Design

We construct the RL target training dataset for ligand binder design using PLINDER (Durairaj et al., 2024), a large-scale, comprehensive database of protein-ligand interactions. From the official train and test sets, we apply the following quality filters:

- Exactly one ligand in the system.
- *holo* structure only.
- Complex crystal structure quality filters: Structure resolution ≤ 3.5 Å, R-factor ≤ 0.4, R-free ≤ 0.45, and the difference between the latter two is less than 0.05.
- Ligand quality filters: No unresolved heavy atoms. No clash outliers. No atom in crystal contacts. ≤ 1 alternate conformer.
- No covalent attachment between ligand and protein in complex.
- ≥ 2 carbon atoms and ≥ 5 heavy atoms in the ligand.
- No oligopeptide ligand.

After deduplicating ligands by both CCD code and SMILES string, and removing 6 samples that fail the featurization pipeline, the final dataset contains 9,111 ligands, which are further split into 8,795/316 for the train/test sets, respectively. The test set will be used to curate the test target set in Sec. C.2. Additionally, we create a self-distillation dataset for supervised fine-tuning (SFT) as a better base model. We generate 20 designs for each ligand target in the filtered PLINDER training set (175,900 designs in total), then filter them using RF3 and the metrics detailed in Sec. D to keep only the successful ones. The final SFT dataset contains 13,251 samples across 5,507 targets.

B.3 Training Data and Targets for Disulfide Peptide Design

We use two SFT datasets to initialize disulfide-peptide RL. The main computational benchmark (Tab. 3) uses TED-derived disulfide crops, whereas the wet-lab experiments (Secs. 5 and J.2) use CPCore complexes from CPSea. The main text uses the short labels Disulfide SFT and Disulfide for the SFT and RL models in each experiment. In appendix comparisons, these unqualified labels denote the TED-derived pair; CPSea Disulfide SFT and CPSea Disulfide explicitly identify the CPSea-derived pair. Each RL model starts from its corresponding SFT initialization.

TED-derived SFT data. We extract disulfide examples from the Teddymer dataset (Lau et al., 2024; Didi et al., 2026b) so that peptide and general protein data share a single structural source. We extract contiguous 12–18-residue segments containing exactly 2 cysteines, 1 at each terminus, and require their Sγ-Sγ distance to lie between 1.8 and 2.3 Å (Fobe et al., 2019). Each peptide must contact a different original chain of the same Teddymer entry, which serves as the target and contains at least 50 residues. We require at least 3 distinct residues on each side to make peptide-target C_α-C_α contacts at distances < 8 Å. Residues outside the selected segment on the source chain are excluded, so only contacts with the partner chain count. Additional quality checks enforce backbone continuity and geometry and exclude internal or peptide-target clashes. Filtering and exact-peptide-sequence deduplication yield 2,638 training crops. During SFT, Teddymer dimers and these disulfide crops are sampled at a ratio of 85:15.

CPSea-derived SFT data. For the CPSea variant, we use CPCore, the 71,867-complex subset of the AlphaFold-derived CPSea collection that passes its binding-affinity, physicochemical and backbone-geometry filters (Yang et al., 2026; Váradi et al., 2022). We exclude 59,275 complexes with unsupported isopeptide linkages, leaving 8,402 head-to-tail and 4,190 disulfide complexes. We further remove cis-like head-to-tail closures and retain disulfide examples that pass cysteine-pair detection and autoencoder cysteine-recovery checks. After these filters and selection of the training partition, the dataset contains 5,891 head-to-tail and 3,913 disulfide complexes. During SFT,

Table 5: **Wet-lab targets.** RL splits have disjoint clusters; all crops belong to the classifier corpus.

Dataset	Crops	Proteins	Clusters
RL Training	203	82	72
RL Test	50	23	20
RL Total	253	105	92
Classifier Training	819	125	110

Table 6: **Task difficulty.** Experimental cutoffs are 1% / 0%; computational cutoffs are 10% / 0%.

Split	Criterion	Easy	Med.	Hard
Train	Experimental	23	180	0
	Computational	66	40	97
Test	Experimental	6	44	0
	Computational	18	13	19

Table 7: **Structure pre-training data.** Counts after filtering and sampling probabilities.

Source	Type	Structures	p_{sample}
PINDER	Exp.	420,496	0.35
Teddymer	Pred.	510,376	0.25
PDB	Exp.	61,446	0.25
AFDB	Pred.	735,014	0.15
Total		1,727,332	

Exp.: experimental; Pred.: predicted.

Table 8: **Wet-lab classifier data.** Binder fold-family split: design counts and experimental hit rates.

	Train	Val.	Test	Total
Designs	612,544	16,455	17,023	646,022
Positives	4,063	594	613	5,270
Negatives	608,481	15,861	16,410	640,752
Hit rate	0.66%	3.61%	3.60%	0.82%
Targets	125	22	22	125

Teddymer dimers (Lau et al., 2024; Didi et al., 2026b), CPCore head-to-tail complexes, and CPCore disulfide complexes are sampled at a ratio of 70:15:15, respectively.

RL targets. Both RL runs use the same 52,492 protein targets, prepared structures, and hotspot annotations as the protein-binder RL training set described in Sec. B.1. For both runs, peptide lengths are sampled from 12–18 residues, and generation is conditioned on disulfide closure.

B.4 Training Data and Targets for Wet-Lab Alignment

To create a wet-lab alignment dataset, we use the multiplexed phage display campaign of Didi et al. (2026a), which screened 1,093,990 de novo binders against 125 targets. The binders were designed using Proteína-Complexa (Didi et al., 2026b), RFDiffusion (Watson et al., 2023), RFDiffusion3 (Butcher et al., 2025), and BoltzGen (Stark et al., 2025). For RL target curation and classifier training, we use the 646,022 Proteína-Complexa designs with self-generated sequences, covering 125 targets and containing 5,270 experimental hits.

The campaign uses 1–12 hotspot crops per target rather than full target structures. The 10 AlphaProteo targets use the hotspots in Tab. 9. The remaining 115 targets plus one repeated PD-L1 entry use domain crops from the corresponding AFDB entries. The selected designs span 819 crops across 125 targets, with crop lengths of 8–321 residues. Each RL task specifies a target crop and its hotspot. We retain the campaign's task definitions and select the RL training and test tasks as follows.

We first filter out the 10 AlphaProteo targets and cluster the remaining 115 targets plus the AFDB PD-L1 target using FoldSeek with a TM threshold of 0.5 on the full target structures from AFDB. The resultant 101 target clusters are split into disjoint training and test sets for an 80:20 split at the crop level while balancing experimental and computational success rates. At the task level, *experimental difficulty* is easy for an on-design hit rate of at least 1%, medium for a nonzero rate below 1%, and hard for zero hits. *Computational difficulty* is easy for a success rate of at least 10%, medium for a nonzero rate below 10%, and hard for zero successes using the same *in silico* filter for protein binder design. We then retain crops with at least one on-design experimental hit, yielding a final split of 203 training crops and 50 test crops, covering 82 and 23 targets, respectively. The statistics are summarized in Tab. 5, and the difficulty stratification is demonstrated in Tab. 6. We follow the original definition of each task to keep their crops and hotspots (exactly one hotspot per crop). Binder length is uniformly sampled from $[50, 88]$.

B.5 Training Data for Wet-Lab Binding Classifier

B.5.1 Structure Pre-Training

Dataset Description. To provide a strong prior for the encoder representations, we pretrain on a mixture of four protein structure sources to cover both real and predicted interfaces while retaining a monomeric fold prior. The detailed statistics for each dataset are available in Tab. 7.

- **PINDER** (Kovtun et al., 2024), a large-scale curated corpus of experimentally determined protein-protein interaction complexes. We filter PINDER to retain complexes with at least 10 interface residues in both chains, yielding 420,496 complexes clustered into 6,955 structure clusters.
- **Teddymer** (Didi et al., 2026b), a large-scale dataset of synthetic binder-target pairs by splitting the TED domains (Lau et al., 2024) in the predicted AFDB structures, with 510,376 dimer structures.
- **PDB** (Berman et al., 2000), the archive of experimentally resolved structural data of biological macromolecules, which also contains oligomers with more than two chains. We follow the filtering procedure in Didi et al. (2026b) to retain 61,446 complex structures.
- **AFDB** (Jumper et al., 2021), a large-scale database of AlphaFold2-predicted monomer structures providing broad coverage of UniProt (Apweiler et al., 2004) sequences. We follow the filtering procedure in Geffner et al. (2025) to retain 735,014 structures.

Data Pipeline. During structure pre-training, structures are cropped to 384 residues. For multimers, we use an interface-seeded crop that preserves *all* partner chains. A random chain is selected as the binder chain and cropped to a random length drawn from $[50, 230]$; the remaining budget is then allocated to partner residues in order of increasing distance to the binder, using an 8 Å Cα-Cα interface criterion. AFDB uses a random contiguous crop over the sequence. For PDB entries, we remove nucleic acids, ligands, waters, and non-canonical residues at the source before converting the structure to the atom37 representation. Training examples are drawn according to probability p_{sample} in Tab. 7 among the four datasets. For PINDER, we use *cluster-uniform* sampling, whereas for the remaining three datasets we apply structure-uniform sampling.

To further utilize the sequence information, we attach frozen per-residue representations from ESMC-6B (Candido et al., 2026). Embeddings are precomputed once, offline, over the $1{,}826{,}890$ unique chain sequences in the structure pre-training corpus. ESMC always takes the full chain during precomputation, and extracts per-residue embeddings on the fly using the cropping indices during training. This ensures ESMC always sees contiguous sequences regardless of gaps in structure cropping. The 2,560-d embedding is projected into a 128-d per-residue feature via a learnable adapter.

As data augmentation for the denoising autoencoder pre-training, we apply noise and dropout to both structure and sequence features. Gaussian coordinate noise is added per example with a standard deviation σ drawn from a log-uniform distribution over $[0.05, 1.0]$ Å. The model does not receive σ, making the encoder a blind multi-scale denoiser rather than one tuned to a single noise level. The entire ESMC feature adapter output is zeroed with probability 0.5 per example, so a single pre-trained encoder can be deployed with or without sequence embeddings.

B.5.2 Wet-Lab Binding Classification

Experimental Binding Data. We train the classifier on the Proteína-Complexa designs from the wet-lab campaign described in Sec. B.4, using both experimental hits and non-hits. The classifier dataset covers all 125 targets, including those held out from RL training. Its training, validation, and test sets are separated by binder structure within each target, as described below.

Binder Holdout Split. For in-distribution post-training with the wet-lab classifier, we let the classifier see all targets to prevent out-of-distribution structures. Therefore, we apply binder structure clustering to construct the wet-lab classifier training data splits to prevent the classifier from collapsing into a retrieval model based on redundant binder structures. Within each target, all binders are clustered structurally with FoldSeek (Van Kempen et al., 2024) at a TM-score threshold of 0.6, and whole clusters are assigned to a single split, including both positives and negatives. A target is eligible for holdout if it has at least 30 positives, for which 22 of the 125 targets qualify, and these contribute 15,174 fold families in total (median 642 per target, range 218-1,773). Family assignment is stratified so that approximately 15% of positives and 15% of negatives go to validation and 15% of each to test, with the remainder to training. Tab. 8 summarizes the resulting splits.

Data Pipeline. The data pipeline for structure pre-training is mostly adopted for classifier fine-tuning, so the prior complex structure knowledge can be transferred from the pre-training stage. We apply a few different setups to adapt to the classifier fine-tuning task. First, the target-binder complex is never cropped, and the structures are noised with a smaller *fixed* $\sigma = 0.20$ Å. As in pre-training, we attach frozen ESMC-6B representations per residue, and during fine-tuning we zero out the entire language-model block with probability 0.5.

Table 9: **Protein targets for minibinder evaluation.** AF2 denotes AlphaFold2-predicted target structures.

Target	Structure	Target residues	Hotspots	Binder length	Source
BBF-14	9HAG	A1–112	–	70–250	BindCraft
Betv1	AF2	A1–159	A24	70–185	BindCraft
CbAgo (PAZ)	AF2	A180–276	A268	70–160	BindCraft
CD45	AF2	A1–88	A1	80–200	BindCraft
Claudin1	AF2	A1–188	A31, A46, A55, A152	80–175	BindCraft
CrSAS6	AF2	A1–145	–	90–200	BindCraft
Derf21	AF2	A1–118	A10	70–185	BindCraft
Derf7	AF2	A1–196	A115	70–185	BindCraft
HER2-AAV	1N8Z	C23–101, C111–268	C165, C170	60–100	BindCraft
IFNAR2	2LAG	A1–103	A45, A73, A75, A77, A89, A91	60–175	BindCraft
PD-1	AF2	A1–115	A33, A95, A97, A102	80–150	BindCraft
SpCas9	AF2	A96–174, A306–446	A360	70–150	BindCraft
BHRF1	2WH6	A2–158	A65, A74, A77, A82, A85, A93	80–120	AlphaProteo
H1	5VLI	A5–325, B501–670	B521, B545, B552	40–120	AlphaProteo
IL7RA	3DI3	B17–209	B58, B80, B139	50–120	AlphaProteo
IL17A	4HSA	A17–29, A41–131, B19–127	A94, A116, B67	50–140	AlphaProteo
Insulin	4ZXB	E6–155	E64, E88, E96	40–120	AlphaProteo
PD-L1	5O45	A17–132	A56, A115, A123	50–120	AlphaProteo
SC2RBD	6M0J	E333–526	E485, E489, E494, E500, E505	80–120	AlphaProteo
TNF-α	1TNF	A12–157, B12–157, C12–157	A113, C73	50–120	AlphaProteo
TrkA	1WWW	X282–382	X294, X296, X333	50–120	AlphaProteo
VEGFA	1BJ1	V14–107, W14–107	W81, W83, W91	50–140	AlphaProteo

C Evaluation Targets

C.1 Protein Test Targets for Minibinder Design

We evaluate protein minibinder design on the 22-target collection assembled for Proteína-Complexa (Didi et al., 2026b). The collection combines 12 targets from BindCraft (Pacesa et al., 2025) with 10 from AlphaProteo (Zambaldi et al., 2024), using the AlphaProteo PD-L1 setup to avoid including the two additional BindCraft PD-L1 variants. Our evaluation retains the full collection, including H1, IL17A, and TNF-α, which were excluded from Proteína-Complexa's main 19-target benchmark after no evaluated methods produced successful designs for them within a 20-hour budget.

Tab. 9 shows the target structure sources, residue selections, hotspot residues, and binder-length ranges. For CD45, the input is cropped from the d2 extracellular domain, with A1 serving as the placeholder hotspot specification. BBF-14 and CrSAS6 have no specified hotspot residues.

C.2 Ligand Test Targets for Minibinder Design

We include the 4 ligand targets evaluated in RFDiffusion-AllAtom (Krishna et al., 2024), BoltzDesign1 (Cho et al., 2025), and Proteína-Complexa (Didi et al., 2026b), plus 8 PLINDER test targets selected using the quality filters in Sec. B.2 and the criteria below.

Chemical diversity. We select distinct Murcko scaffolds and ring systems, covering 14–35 heavy atoms, lipophilicities ($\mathrm{clog}P$), and sp^3-carbon fractions (F_{sp^3}).

Data separation. The 8 PLINDER test ligands have no CCD or SMILES overlap with training data.

Binder utility. We prioritize ligands for which a high-affinity minibinder has a practical use. The panel covers scavenging toxic or overdosed compounds (UA7, UAJ, CLU, D16, TUD), monitoring drugs with narrow therapeutic windows (95E, 8MM), and quantifying a biomarker (NGC). This selection avoids choosing ligands solely because they inhibit their native protein partners.

Tab. 10 lists the ligand structures and heavy-atom counts. We generate 200 designs of 100 residues per target.

Table 10: **Ligand targets.** Names are CCD codes; HA counts heavy atoms.

Target	PDB	HA	Source
UA7	6WVH	35	PLINDER
TUD	6GW2	34	PLINDER
D16	8GKF	32	PLINDER
UAJ	6WV7	27	PLINDER
CLU	2PNC	14	PLINDER
8MM	5US1	34	PLINDER
95E	5NR3	14	PLINDER
NGC	7EE5	22	PLINDER
SAM	7C7M	27	RFDiffusion-AA
OQO	7V11	38	RFDiffusion-AA
FAD	7BKC	53	RFDiffusion-AA
IAI	5SDV	33	RFDiffusion-AA

C.3 Protein Test Targets for Peptide Design

We adopt the 12-target cyclic-peptide panel curated by PXDesign (Team et al., 2025c), comprising eight targets from AlphaProteo (Zambaldi et al., 2024), three from RFpeptides (Rettie et al., 2025b), and Keap1 from AfCycDesign (Rettie et al., 2025a). Tab. 11 lists the original benchmark sources and the prepared structures used for each target.

Table 11: **Protein targets for cyclic peptide design.** AlphaProteo structures use the PDB crops in Tab. 9.

Target	Residues	Chains	Panel source	Structure input	Hotspots
BHRF1	157	1	AlphaProteo	AlphaProteo	A65, A74, A77, A82, A85, A93
GABARAP	117	1	RFpeptides	3WIM	A46, A48, A49, A50, A60, A63
IL-17A	213	2	AlphaProteo	AlphaProteo	A94, A116, B67
IL-7RA	193	1	AlphaProteo	AlphaProteo	B58, B80, B139
Keap1	285	1	AfCycDesign	2FLU	A334, A380, A382, A415, A483, A530
MCL1	145	1	RFpeptides	2PQK	A228, A231, A235, A253, A263, A270
MDM2	83	1	RFpeptides	4HFZ	A54, A58, A61, A62, A72, A73, A96, A100
PD-L1	115	1	AlphaProteo	BindCraft	A39, A98, A106
SC2RBD	194	1	AlphaProteo	AlphaProteo	E485, E489, E494, E500, E505
TNF-α	438	3	AlphaProteo	AlphaProteo	A113, C73
TrkA	101	1	AlphaProteo	AlphaProteo	X294, X296, X333
VEGF-A	188	2	AlphaProteo	AlphaProteo	W81, W83, W91

C.4 Protein Test Targets from Wet-Lab Data

Besides the test target selection pipeline described in Sec. B.4, we provide more details in Tab. 12.

D Evaluation Metrics

We generate 200 designs per target for the protein, ligand, and cyclic-peptide benchmarks in Sec. C. Wet-lab reward RL uses 100 designs per crop across 50 target crops. Both Proteína-Complexa and Proteína-Reforza use their generated sequences without redesign. Inference settings are in Sec. H.5.

Protein folding. AF2-Multimer provides 5 weight sets, indexed 0–4. RL training randomly selects among sets 1–4; set 0 is the held-out evaluator. Evaluation uses ColabDesign with 3 recycles, target templates, and initial-guess conditioning (AF2-IG). Interface predicted aligned error (ipAE) averages binder-target pAE entries, excluding target-target pairs. Complex pLDDT averages residue Cα confidence over the target-binder complex on a $[0, 1]$ scale. Binder scRMSD is the Cα RMSD between generated and refolded binders after optimal binder alignment. Success requires

$$\text{ipAE} \leq 7\,\text{Å}, \qquad \text{pLDDT} \geq 0.90, \qquad \text{scRMSD} < 1.5\,\text{Å}.$$

The cross-evaluator analysis applies these thresholds to each AF2 checkpoint, retaining the native confidence scores averaged over the holdout checkpoints.

Ligand folding. We use the top-ranked of 5 RF3 predictions. Min-ipAE is the minimum reported chain-pair interface pAE between the binder and target/ligand chains. Binder scRMSD uses the Cα alignment defined above. Ligand scRMSD uses all ligand atoms after binder all-atom alignment, with Cα alignment as a fallback; the ligand is not independently aligned. Success requires

$$\text{min-ipAE} < 2\,\text{Å}, \qquad \text{binder scRMSD} < 2\,\text{Å}, \qquad \text{ligand scRMSD} < 5\,\text{Å}.$$

Folding and closure of terminal-disulfide and head-to-tail cyclic peptides. Cyclic success requires AF2-IG ipAE < 7 Å, binder-only pLDDT > 0.80, and binder Cα scRMSD < 1.5 Å, computed after aligning the predicted and generated binders independently of the target. Peptides must also satisfy the corresponding closure criterion. Terminal-disulfide peptides require exactly two cysteines, one at each terminus, with $d_{\text{SG}-\text{SG}} \leq 2.5$ Å. Head-to-tail peptides require $d_{\text{N}-\text{C}} \leq 1.5$ Å between the N-terminal nitrogen and C-terminal carbonyl carbon.

Total and unique successes. Total successes count designs passing all task-specific gates. Unique successes count clusters of successful binders, computed separately per target. FoldSeek clusters

Table 12: **The 50 holdout wet-lab target crops grouped by target protein (23 proteins).** Each crop is a contiguous or multi-range excerpt of the target chain centered on exactly one hotspot residue. Crops are separated by semicolons, segments within a crop are separated by commas, and the hotspot residue is in parentheses.

Target	UniProt	Crops	Crop ranges (hotspot)
CLU	P10909	3	21–32, 386–407 (21); 205–222 (214); 120–133, 262–301, 332–350 (276)
CNTFR	P26992	2	114–128, 143–181, 193–207, 228–243, 281–292 (172); 115–128, 165–179, 194–210, 224–266, 278–295 (286)
CRTAM	O95727	2	36–83, 96–110 (44); 28–114 (67)
DPP4	P27487	1	236–254, 703–717, 732–754 (244)
FCG3B	O75015	1	109–197 (165)
FcRn	P55899	1	201–296 (246)
FGFR1	P11362	2	418–448 (425); 814–822 (822)
FST	P19883	2	126–142, 156–194, 219–232, 320–332 (169); 198–270 (241)
HLA-B	P01889	1	32–70, 88–121 (41)
HLA-E	P13747	2	20–102, 167–196 (86); 204–303 (296)
IL20	Q9NYY1	2	25–39, 69–92, 114–146, 168–176 (130); 59–95, 116–176 (140)
IL4R	P24394	4	25–88, 101–113 (51); 34–44, 59–121 (67); 32–128, 144–158, 202–216 (91); 123–232 (185)
ITGA6	P23229	2	23–36, 625–646 (633); 662–678, 716–735 (725)
MRC2	Q9UBG0	3	42–172, 769–797 (154); 371–381, 509–583, 640–653 (523); 826–891, 927–951 (881)
NECT1	Q15223	3	1–8 (1); 249–337 (291); 473–489 (481)
NTRK2	Q16620	2	199–288 (238); 289–381 (328)
PSG1	P11464	6	34–142 (73); 34–39, 53–107, 119–139 (132); 151–217, 230–243, 309–323 (162); 143–233 (197); 336–418 (387); 337–352, 379–392, 408–418 (416)
RNF43	Q68DV7	1	610–623 (617)
SERPINA3	P01011	2	226–245, 276–289, 415–422 (236); 80–103, 139–173, 199–209, 323–358 (350)
TNFL9	P41273	1	74–147, 199–214, 231–247 (88)
UBC9	P63279	2	16–27, 42–62, 75–158 (49); 1–158 (59)
UBE2B	P63146	1	20–152 (75)
VATF	Q16864	4	1–104 (14); 8–104 (41); 1–91 (59); 7–107 (94)
Total		50	

structures at TM-score threshold 0.5 with alignment type 1 and no sequence-identity requirement; MMseqs clusters sequences at minimum identity 0.1 and coverage 0.8.

PDB novelty. PDB novelty is the maximum FoldSeek alignment TM-score of the extracted generated binder against the configured PDB database; lower values indicate greater novelty.

TMol energies. TMol $\Delta\Delta G$ sums beta2016-weighted contributions over binder–target residue-block pairs. TMol interaction (Tmol Int.) combines attractive and repulsive Lennard–Jones, electrostatic, and hydrogen-bond terms:

$$E_{\text{Int.}} = E_{\text{fa_ljatr}} + 0.55E_{\text{fa_ljrep}} + E_{\text{fa_elec}} + E_{\text{hbond}},$$

where E terms denote unweighted cross-interface contributions. For noncovalent interfaces, $\Delta\Delta G = E_{\text{Int.}} + E_{\text{solv}}$, with Lazaridis–Karplus, directional LK-ball, and water-bridge solvation terms:

$$\begin{aligned} E_{\text{solv}} = E_{\text{fa_lk}} &- 0.38E_{\text{lk_ball_iso}} + 0.92E_{\text{lk_ball}} \\ &- 0.33E_{\text{lk_bridge}} - 0.33E_{\text{lk_bridge_uncpl}}. \end{aligned}$$

One-body and intramolecular terms are excluded; disulfide, bonded-geometry, and torsion terms contribute only if they have cross-interface entries. Both matrix directions are summed using TMol's storage convention, without the training reward's interface-size normalization. Lower scores are more favorable; these are force-field scores, not experimental binding free energies. Evaluation uses refolded complexes after Cartesian FastRelax with 2 schedule repeats and no soft repulsion. Other metrics use the original refolded coordinates.

Hydrogen bonds and electrostatics. Interface H-bonds count donor–acceptor bonds across the interface, using HBPLUS for protein benchmarks and TMol for PD-L1 selection. TMol energies sum weighted cross-interface `hbond` and `fa_elec` terms for H-bonds and electrostatics, respectively.

Surface properties. SASA uses a 1.4 Å probe and 960 surface points. Interface ΔSASA is $\mathrm{SASA}(B) + \mathrm{SASA}(T) - \mathrm{SASA}(B \cup T)$ in Å^2, summing burial on both partners without division by 2. Binder interface residues have an atom within 4 Å of the target. Interface hydrophobicity is the percentage of these residues in $\{\mathrm{A}, \mathrm{C}, \mathrm{F}, \mathrm{I}, \mathrm{L}, \mathrm{M}, \mathrm{P}, \mathrm{V}, \mathrm{W}, \mathrm{Y}\}$. Surface hydrophobicity is their fraction of isolated-binder SASA, shown as a percentage in Fig. 25. Shape complementarity (SC) measures geometric complementarity of the opposing surfaces; larger values are better.

Binder stability. These filters favor compact binders with buried hydrophobic cores, which correlate with miniprotein stability (Rocklin et al., 2017). We calculate them after removing the target/ligand to exclude interface burial. With binder length L, residue SASA a_i, and Tien theoretical maximum accessible area $A_{\max}(s_i)$ (Tien et al., 2013), the filters are:

- **Exposed residue fraction:** the fraction of binder residues with relative SASA $a_i/A_{\max}(s_i) \geq 0.50$ must not exceed 0.40.
- **Buried NPSA per residue:** $L^{-1} \sum_{i:s_i \in \mathcal{H}} [A_{\max}(s_i) - a_i]_+$, with $\mathcal{H} = \{\mathrm{A}, \mathrm{F}, \mathrm{I}, \mathrm{L}, \mathrm{M}, \mathrm{Y}, \mathrm{V}, \mathrm{W}\}$, required to be ≥ 40 Å^2/residue. This is a hydrophobic-residue burial proxy.
- **Radius-of-gyration ratio:** $R_g/(2.38L^{0.365}) \leq 1.25$, where $R_g = \sqrt{L^{-1} \sum_i \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}$ for binder Cα coordinates in Å.

The cutoffs are heuristics informed by literature and a pilot of 1,000 native partner chains. Among 274 well-buried chains of 50–120 residues, the exposed-fraction 90th percentile was 0.393, motivating the 0.40 limit. The gyration cutoff approximates the pilot's proposed $R_g/L^{1/3} \leq 3.5$ Å boundary over this length range. The burial cutoff is more permissive than the > 50 Å^2/residue filter used in small beta-barrel design (Kim et al., 2023), although differing SASA definitions limit direct comparison. These filters screen structural plausibility rather than establish experimental stability.

Wet-lab binding classifier. Probabilities and logits average 5 fixed classifier checkpoints per design. The #$p > 0.5$ column averages threshold-passing counts over 50 target crops. AUROC and AUPRC are areas under the ROC and precision-recall curves. pt-AUROC averages target-wise AUROCs over all test targets, each containing at least 1 positive by construction. P@50 is the positive fraction among the 50 highest-scoring test pairs. Classifier training and selection follow Sec. E.

Binding affinity (ligand targets). Nesso-1 scores the binder sequence and task ligand independently of RF3. Affinity is $\log_{10}(\mathrm{IC}_{50}/\mu\mathrm{M})$; lower is better, with -3 corresponding to 1 nM and 0 to 1 μM.

Aggregation. Protein, ligand, and cyclic-peptide benchmark tables report per-target averages. Energies, interface H-bonds, Nesso affinity, and novelty are averaged over successful designs.

Compute. Time [s] is A100 generation time per sample, summed across shards, including failed designs. Sequence-redesign baselines include the redesign model's loading and runtime. Protein and ligand scaling curves count unique successes against generation and evaluation GPU hours.

E Wet-Lab Binding Classifier and RL with Wet-Lab Feedback

E.1 Structure Pre-Training Setup

The pre-training stage encourages the encoder to capture the crucial geometric interactions in predicted and experimental protein structures, enabling better representation learning with a structure prior via the denoising objective. The encoder adopts a transformer architecture similar to Proteína-Complexa (Didi et al., 2026b): 12 layers with a hidden dimension of 768 per token and 256 per pair and 12 attention heads. The input sequence featurization with a total dimension of 596 is:

- Binary chain break flag;
- 20 one-hotspot embeddings for 20 amino acid types;
- Absolute atom37 coordinates with mask, totaling 148;
- Relative atom37 coordinates with mask, totaling 148;

Model/Score	AUPRC	AUROC	pt-AUROC	P@50
Proteína-Reforza	**0.134**	**0.806**	0.698	0.22
- ESMC	0.099	0.768	0.661	0.08
- LM dropout	0.123	0.799	0.681	**0.26**
- Structure noise	0.123	0.804	**0.735**	0.24
ipTM	0.103	0.779	0.682	0.20
pTM	0.076	0.736	0.648	0.08
ranking score	0.097	0.775	0.680	0.20
ipSAE_min	0.071	0.731	0.661	0.10

Table 13: **Classification metrics and ablation studies on the binder holdout test split.** For each model, we select the checkpoint with the highest validation per-target AUROC (pt-AUROC). All other metrics are pooled over the complete set of 17,023 target–binder test pairs.

- 3 backbone torsion angles, each binned to 21 bins, totaling 63;
- 4 side-chain torsion angles with mask, each binned to 21 bins, totaling 88;
- Frozen 2560-d ESMC-6B embeddings projected to 128-d, with adapter dropout probability 0.5.

The input pair featurization with a total dimension of 328 is:

- Pairwise distance from Cα to other four backbone atoms, each binned to 21 bins, totaling 84;
- The inter-residue orientation described by 5 torsion angles, each binned to 21 bins, totaling 105;
- Same-chain residue index difference, capped to $[-32, 32]$ + different-chain marker, totaling 66;
- Same-chain and same-residue token index difference, capped to $[-32, 32]$ + difference marker, totaling 66;
- Binary same-entity flag;
- Difference between symmetry ids, capped to $[-2, 2]$ + different-entity marker, totaling 6.

The last 4 relative positional encodings follow AlphaFold3 (Abramson et al., 2024), Alg. S.3. We compute all structure features from the noised all-atom structures to prevent reconstruction leakage.

Unlike Proteína-Complexa, which emphasizes structure reconstruction, we instead use a larger latent dimension of 64 with no variational prediction of the variance. The transformer-based decoder takes the encoder latent and the noisy Cα coordinates as input, with 4 layers, a hidden dimension of 384, and 12 heads for reconstruction. The standard MSE loss is used as the denoising objective. This setup yields an encoder with 129M trainable parameters and a decoder with 12M, encouraging the encoder to capture more meaningful structural representations.

We pre-train the full encoder-decoder architecture on a combination of structure datasets (see App. B.5.1) with a total batch size of 256 for 200k steps. The AdamW optimizer is used with a base learning rate of 10^{-4}, a weight decay of 10^{-2}, and an EMA weight of 0.999.

E.2 Classifier Fine-Tuning Setup

With the pre-trained encoder capturing geometric priors, we then fine-tune its backbone with an additional logit head on the wet-lab labels. The pre-trained 200k EMA encoder checkpoint is loaded, with an additional MLP-based logit head that adds 2M trainable parameters. The featurizations are identical to the pre-training stage. The logit head is built upon the following mean-pooled features:

- Global pooling upon the last encoder hidden;
- Binder pooling upon the last encoder hidden;
- Target pooling upon the last encoder hidden;
- Element-wise absolute difference for the above two;
- Global pooling upon the last encoder pair hidden between binder and target.

The model is fine-tuned on wet-lab structures and labels with a total batch size of 256 for 20k steps, with validation every 2k steps. The AdamW optimizer is used with a base learning rate of 10^{-5} for the encoder and 3×10^{-4} for the logit head. No weight decay or EMA is used.

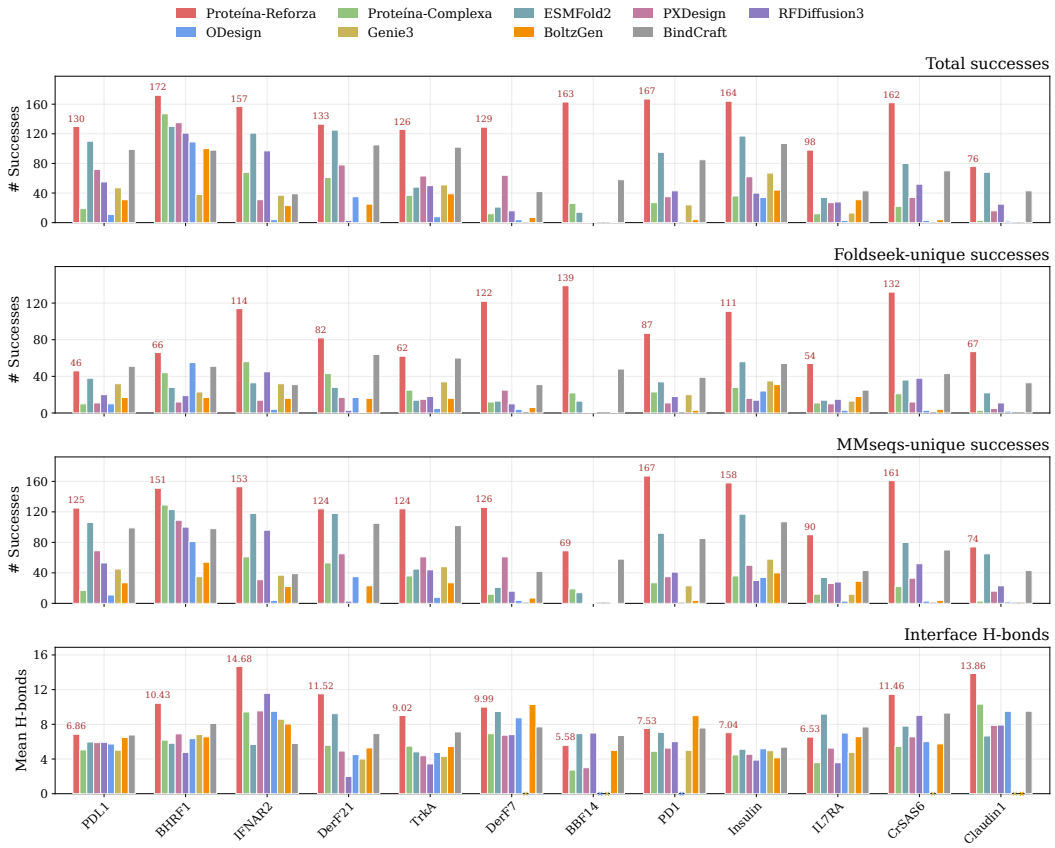


Figure 9: **Per-target protein binder results on easy targets.** Labels show Proteína-Reforza results for 200 designs per model and target. Crosses denote undefined mean hydrogen-bond counts when no designs succeed.

E.3 Wet-Lab Classification Performance

We summarize the classification performance and ablation studies on the holdout test target-binder pairs in Tab. 13, where we also include AlphaFold2-Multimer confidence scores with initial guess as a reference. For each model variant, the checkpoint with the best validation per-target AUROC (pt-AUROC) is selected for test evaluation. AUPRC, AUROC, and P@50 (precision at top 50 candidates) are reported over all 17,023 test pairs. We compare the following ablations and confidence baselines:

- **- ESMC**: No ESMC feature is used (zeroed out).
- **- LM dropout**: ESMC features are always used (no dropout).
- **- Structure noise**: No 0.2 Å structure noise is applied.
- **ipTM**: interface predicted TM-score that measures the interface alignment predicted by AF2.
- **pTM**: predicted TM-score that measures the global alignment predicted by AF2.
- **ranking score**: defined as $0.8 \cdot \text{ipTM} + 0.2 \cdot \text{pTM}$.
- **ipSAE_min** (Dunbrack Jr, 2025): the minimal interface predicted scores from alignment error that better capture the true positive structures.

The best confidence-based classifier (ipTM) achieves only 0.103 AUPRC, leaving many false positives and low precision. With our two-stage pipeline, our classifier captures both the structure and sequence information that contribute to identifying true successes. Our classifier achieves noticeable improvements in both AUROC and AUPRC, potentially providing more accurate signals from the wet lab than confidence-based scores. ESMC embedding plays a crucial role, as we observe that similar sequences with only two residues changed can flip the binding, yet structure prediction models fail

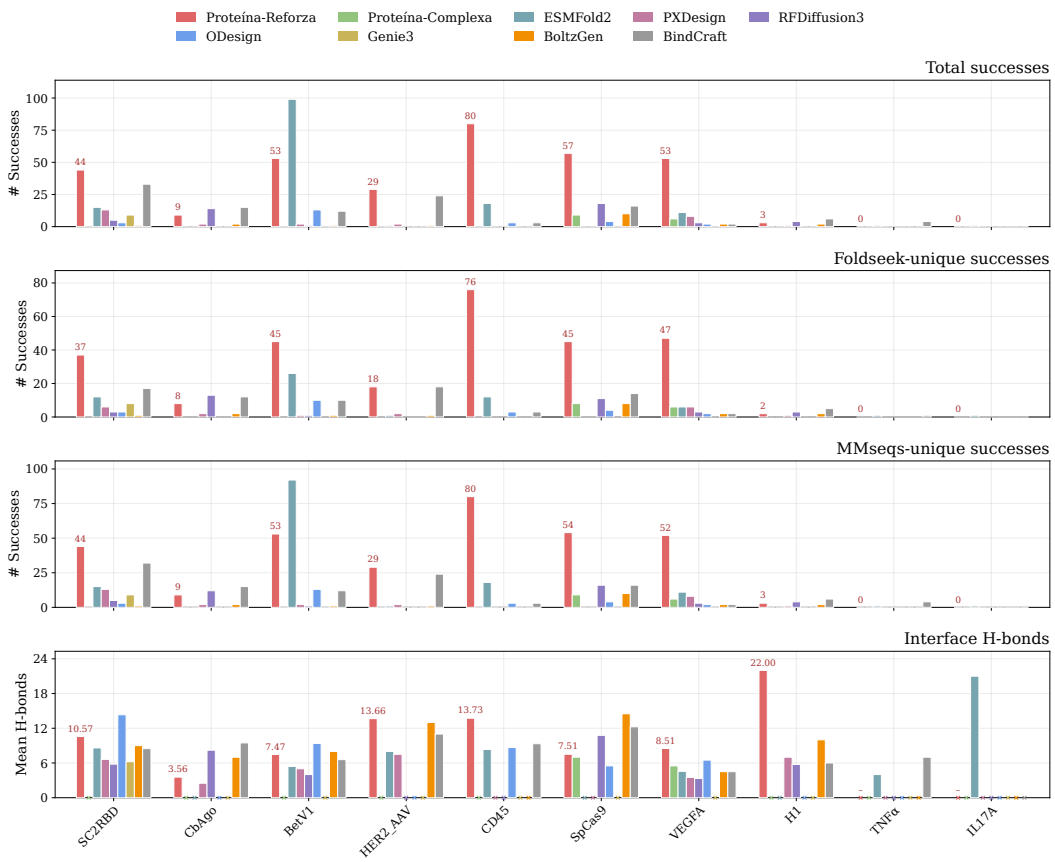


Figure 10: **Per-target protein binder results on hard targets.** Labels show Proteína-Reforza results for 200 designs per model and target. Crosses denote undefined mean hydrogen-bond counts when no designs succeed.

to distinguish between them. Data augmentation on both the sequence and structure sides has less impact, but it generally prevents overfitting to a single modality.

F Individual Target Results

The results in this section compare methods under a fixed budget of 200 samples per target. We break down the aggregate results for Proteína-Reforza and variants with additional energy or affinity rewards; evaluation criteria and training details appear in Secs. D and H. For comparisons at fixed GPU budgets and inference-time scaling across protein and small-molecule targets, see Sec. G.7.

F.1 MiniBinder RL for Protein Targets

General RL. The comparison in Tab. 1 shows that Proteína-Reforza increases mean folding successes from 22.05 to 91.14 and Foldseek-unique successes from 14.18 to 61.82 per target, exceeding all evaluated external baselines on both metrics. Figs. 9 and 10 break down these results across the 22 protein targets, grouped into 12 easy and 10 hard targets for readability. Proteína-Reforza increases both total and structurally unique successes over Proteína-Complexa on 20 of 22 protein targets.

Gains are largest on DerF7, BBF14, and CrSAS6, where Proteína-Reforza yields 122, 139, and 132 Foldseek-unique successes, respectively, compared with 12, 22, and 21 for Proteína-Complexa and at most 31, 48, and 43 for any external baseline. RL thus improves both structural diversity and success counts. The advantage is not uniform: on PDL1, Proteína-Reforza produces more total successes than any baseline, but BindCraft retains more Foldseek-unique successes (51 compared with 46).

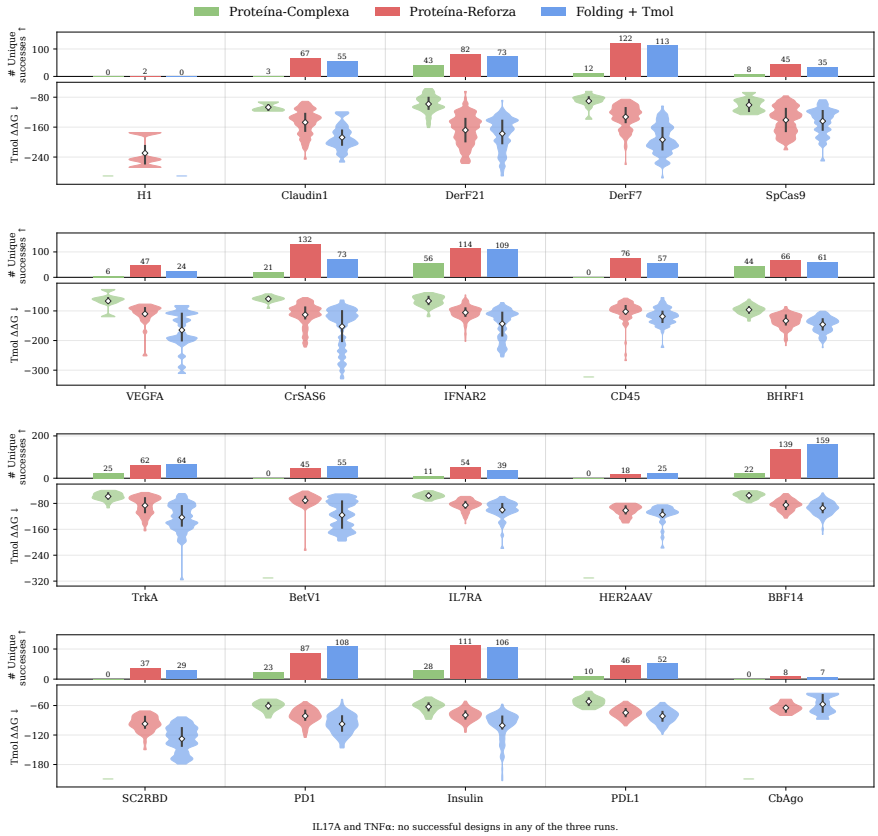


Figure 11: **Per-target Tmol energy–success tradeoffs.** Bars count Foldseek-unique successes; violins show successful-design energies (lower is better). Diamonds mark means and thick lines mark interquartile ranges.

Model	# Folding successes ↑			Stability	Nesso	Time	Novelty ↓
	Total	Foldseek	MMseqs	# Pass ↑	Affinity ↓	[s] ↓	
ODesign	13.00	5.58	11.08	13.00	1.02	13.89	0.84
BoltzGen	32.58	7.42	20.08	32.50	0.99	17.68	0.83
RFDiffusion3	54.75	26.00	53.75	54.75	1.06	25.02	0.82
BoltzDesign	46.08	28.92	46.08	40.25	**0.97**	163.7	0.82
Proteína-Complexa	25.00	13.75	24.42	25.00	1.11	**0.70**	0.84
Proteína-Reforza *(ours)*	**103.2**	**32.08**	**54.75**	**103.2**	1.08	**0.70**	**0.79**

Table 14: **Ligand results.** Averaged over 12 ligand targets. Foldseek and MMseqs count unique structures and sequences among successes. Predicted Nesso affinity averages over successes; stability counts successful designs that pass its filter.

Proteína-Reforza also finds successful designs on targets where Proteína-Complexa has none, including CD45, SC2RBD, and BetV1, with 76, 37, and 45 Foldseek-unique successes, respectively. For BetV1, ESMFold2 has more total successes (99 versus 53), but fewer unique structures (26 versus 45), illustrating why we report both yield and diversity. CbAgo and H1 remain difficult, with only 8 and 2 unique successes under Proteína-Reforza, respectively, and some external baselines perform better. The main-paper target-specific RL experiments address particularly difficult targets: H1 improves from 2 to 23 unique successes and TNFα from 0 to 3 (Fig. 4). IL17A remains the most challenging protein target for our models: we observe no successful designs in any of the evaluated RL runs.

We analyze how RL gains relate to training-target similarity in Sec. G.4 (Fig. 21).

TMol reward: energy and diversity. The aggregate comparison in Tab. 2 shows that combining folding and TMol rewards improves mean relaxed energy while largely retaining the structural diversity of folding-only RL. The two objectives are not inherently in conflict — measured from Proteína-Complexa, folding RL, Tmol RL, and their combination all improve both — but they trade against each other once folding RL has already been optimized. Fig. 11 gives per-target success counts and energy distributions, and Fig. 12(a) the change relative to folding-only RL. Among the 19 targets with successful designs under both runs, TMol $\Delta\Delta G$ improves on 18, with an equally weighted

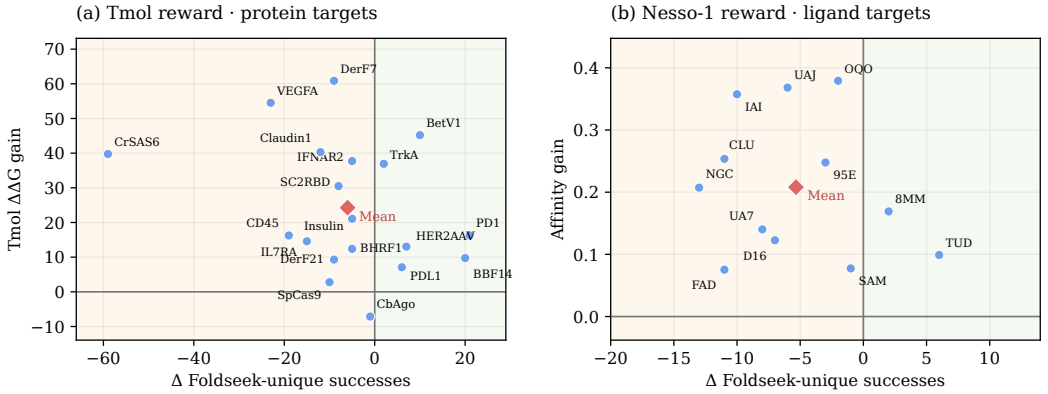


Figure 12: **Reward improvement versus structural diversity.** (a) Tmol reward for proteins; (b) Nesso-1 reward for ligands. Relative to general RL, rightward indicates more Foldseek-unique successes and upward indicates lower mean energy or predicted affinity. Diamonds denote equally weighted target averages.



Figure 13: **Per-target small-molecule binder results.** Labels show Proteína-Reforza values.

mean gain of 24.27. Foldseek-unique successes decrease on 13 targets, with a mean reduction of 6.00 across the 19 targets. The trade-off varies across targets: DerF7 gains 60.87 energy units while losing 9 unique successes, and VEGFA gains 54.52 while losing 23. TrkA improves on both axes, gaining 36.93 energy units and 2 unique successes. These results support assessing energy gains together with structural diversity; see Sec. G.1 for the reward-weight comparison.

F.2 MINIBINDER RL FOR LIGAND TARGETS

General RL. We expand the ligand results discussed in Sec. 4.2, with the full aggregate comparison in Tab. 14 and per-target results in Fig. 13. Proteína-Reforza increases mean total successes from 25.00 to 103.2 and Foldseek-unique successes from 13.75 to 32.08 per target. This exceeds BoltzDesign (28.92), the strongest external baseline, by a smaller margin than for proteins. Across all 12 ligands, Proteína-Reforza improves both total and Foldseek-unique successes over Proteína-Complexa; it leads external baselines in total successes on 10 targets and Foldseek-unique successes on 7. These

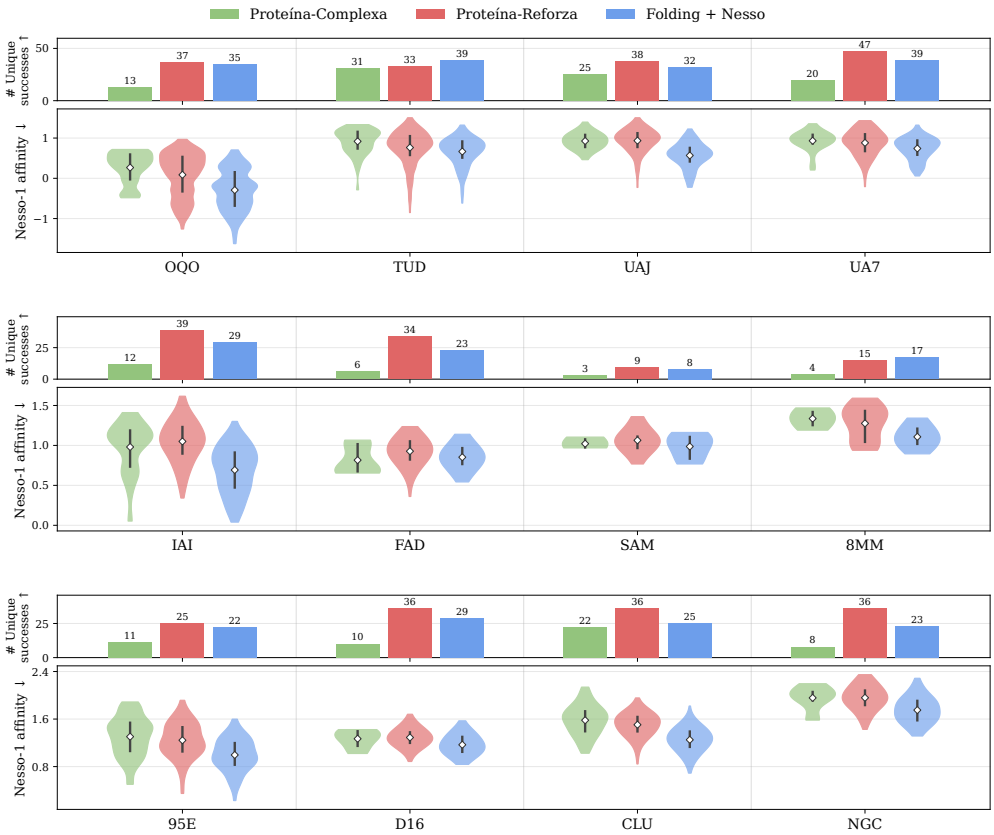


Figure 14: **Per-target Nesso-1 affinity–diversity tradeoffs.** Bars count Foldseek-unique successes; violins show predicted binding affinity (lower is better). Diamonds mark means; thick lines mark interquartile ranges.

gains preserve the base model's fast generation speed: 0.70 seconds per design on an A100, versus 13.89–163.7 seconds for the reported external baselines (approximately 20–234× faster; Tab. 14).

UA7 and IAI reach 47 and 39 unique successes compared with 20 and 12 for Proteína-Complexa and at most 39 and 21 for external baselines. For TUD and CLU, however, 174 total successes each correspond to only 33 and 36 unique structures; RFDiffusion3 obtains 36 and 38 unique successes despite lower total yields. SAM and 8MM remain challenging, with only 9 and 15 unique successes under Proteína-Reforza, below the best external baselines (35 and 20). Aggregate gains coexist with target-specific gaps and substantial redundancy among successful designs on some ligand targets.

Nesso-1 reward RL. We investigate how a binding-affinity predictor can guide RL toward stronger predicted binding while maintaining a high yield of unique successful designs. We add the Nesso-1 affinity term at weight 0.3 to the folding reward and diversity bonus (Sec. H.2). The combined objective improves both predicted affinity and successful generation relative to Proteína-Complexa: mean predicted affinity decreases from 1.11 to 0.87, while Foldseek-unique successes increase from 13.75 to 26.75 per target, a 1.95× gain (Fig. 6 (*right*)). Compared with Proteína-Reforza, the Nesso-1 run improves predicted affinity from 1.08 to 0.87 while retaining 83.4% of unique successes (26.75 versus 32.08), balancing predicted binding strength with the yield of distinct successful designs.

Fig. 14 shows affinity distributions among successful designs, with separate scales for each row; Fig. 12(b) summarizes the changes relative to Proteína-Reforza. Affinity gain is the Proteína-Reforza mean minus the Nesso-RL mean among successful designs, so positive values indicate improvement. Mean predicted affinity improves on all 12 ligands, while unique-success counts decrease on 10. The cost varies across targets: OQO improves by 0.38 affinity units with only 2 fewer unique successes, whereas IAI improves by 0.36 units with 10 fewer. TUD and 8MM improve on both axes, adding 6 and 2 unique successes, respectively; NGC loses 13. Affinity gains can accompany increased or reduced structural diversity, motivating joint evaluation of both metrics for each individual ligand.

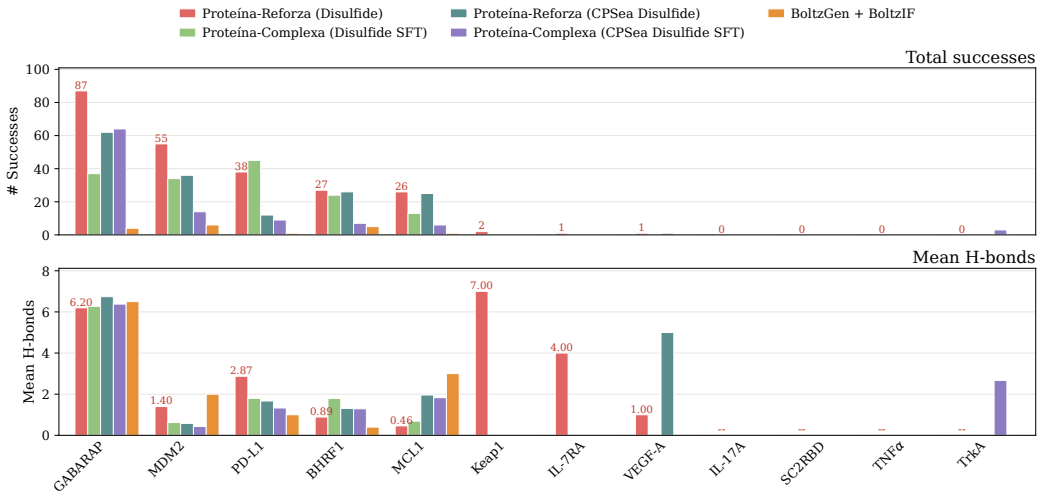


Figure 15: **Per-target terminal disulfide binder results.** Hydrogen-bond counts average over successes, and dashes denote targets where no design succeeded.

The Nesso-1 results illustrate how an affinity predictor can complement folding and diversity rewards, improving predicted binding while preserving much of the successful-design yield of Proteína-Reforza. The practical benefit depends on predictor accuracy: improved predicted affinity must still be validated through independent experimental measurements of protein–ligand binding affinity.

F.3 Terminal Disulfide Peptide RL for Protein Targets

Fig. 15 reports per-target results for the models in Tab. 15, with 200 designs per target under hotspot conditioning. Proteína-Reforza$_{\text{Disulfide}}$ achieves the most folding successes on GABARAP, MDM2, BHRF1, and MCL1. Among the evaluated models, it is also the only model to generate successful designs for Keap1 and IL-7RA. VEGF-A remains challenging: Proteína-Reforza$_{\text{Disulfide}}$ and Proteína-Reforza$_{\text{CPSea Disulfide}}$ each produce only 1 successful design. These results show that the aggregate gains vary across targets, with limited success on several difficult targets.

G Additional Experiments

This section analyzes how RL design choices affect binder success, diversity, and robustness. We first compare reward formulations, test transfer across folding models, and study the roles of SFT initialization and diversity rewards (Secs. G.1 to G.3). We next analyze generalization to structurally dissimilar targets (Sec. G.4) and identify a stability failure mode in target-specific RL and show how the buried NPSA reward mitigates it (Sec. G.5). Further analyses characterize changes in binding interfaces and sequences, and test whether additional inference compute continues to yield distinct successful binders for protein and small-molecule targets (Secs. G.6 and G.7). Finally, we assess performance without hotspot conditioning and retention of linear protein-binder design capability after terminal-disulfide peptide RL.

G.1 Ablation Study: Reward Choices

Folding-success reward. A useful folding reward should encourage candidates to satisfy all success criteria without spending further optimization on metrics that already pass their thresholds. We compare 3 choices: the threshold-capped success reward in Eq. (3), an uncapped reward that continues to reward metric improvements beyond the success thresholds, and an ipAE-only reward that optimizes interface confidence without explicitly enforcing the other criteria. The capped reward supplies a dense learning signal for failed criteria while giving equal reward to all successful designs; the alternatives test the importance of thresholding and of combining complementary folding metrics. We evaluate all 3 RL variants on 22 protein targets, generating 200 designs per target.

Fig. 16 shows that the success reward yields the most structurally unique successes across the benchmark. It produces 2,005 total, 1,360 Foldseek-unique, and 1,846 MMseqs-unique successes, compared with 1,686/1,114/1,518 for the uncapped reward and 1,425/912/1,189 for ipAE-only reward. All 3 improve over the base model, but capping the reward increases structurally unique successes by 22.1% over the uncapped variant and 49.1% over ipAE-only RL. These results support optimizing joint threshold satisfaction: continued improvement of already-passing metrics or interface confidence alone is less effective at expanding the pool of successful designs.

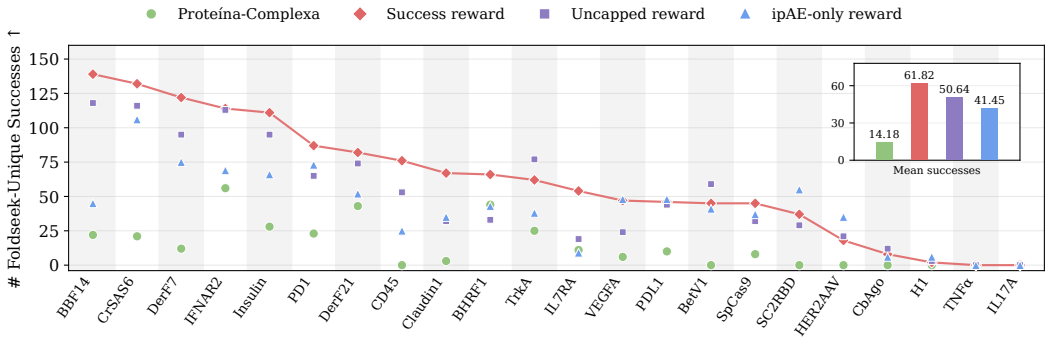


Figure 16: **Folding-reward ablation.** Foldseek-unique successes for 3 reward choices across 22 protein targets.

Diversity reward for target-specific RL. Repeated optimization on a single target can concentrate generation on a few successful folds, limiting the variety of candidates available for experimental testing. We therefore add the successful-fold diversity reward in Eq. (6): each successful design receives the inverse size of its Foldseek cluster within the rollout group, giving each successful cluster equal total credit. Only successful designs receive the bonus, with larger rewards for underrepresented folds. Starting from general RL, we test weights 0/0.5 on HER2AAV and 0/0.2 on SpCas9 (Tab. 19).

On HER2_AAV, adding the bonus preserves a similar number of successes (88 versus 92) while increasing Foldseek-unique successes from 39 to 61 and MMseqs-unique successes from 42 to 68 (Fig. 17). The corresponding unique-to-total ratios rise from 42.4% to 69.3% and from 45.7% to 77.3%, indicating a broader set of successful structures and sequences. On SpCas9, the bonus increases total successes from 60 to 73, Foldseek-unique successes from 46 to 59, and MMseqs-unique successes from 51 to 67. These results support the diversity reward for target-specific adaptation,

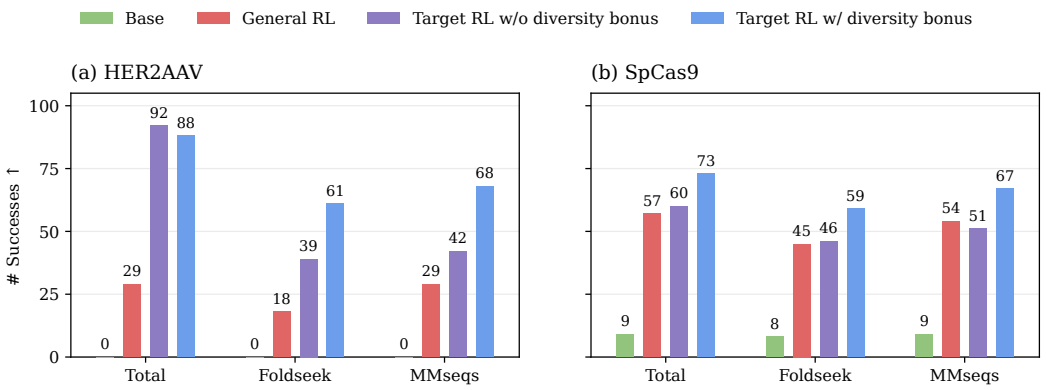


Figure 17: **Target-specific diversity reward.** Total, Foldseek-unique, and MMseqs-unique successes on HER2AAV and SpCas9, comparing general RL and target-specific RL with and without the diversity bonus.

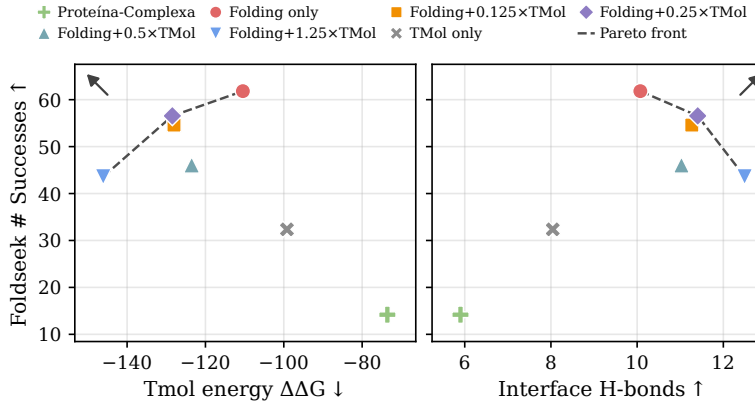


Figure 18: **Trade-off between design diversity and physics metrics.** The pareto front between TMol energy or interface H-bond and the number of unique success. Black arrow indicates optimal direction of the pareto front.

particularly when increased success alone would yield redundant candidates. These comparisons use selected checkpoints with different training durations rather than a matched training budget.

TMol reward. Optimizing the model towards generating minibinder that have unbounded stronger interface (lower TMol energies and more interface hydrogen bonds) can lead to less diverse designs. Empirically, the trade-off between TMol energy term and binder diversity can be explained by sharpen interface residue type distribution towards specific hydrophobic or aromatic residues. We observe that the strength of TMol reward in the RL objective have a significant impact on such energy diversity trade-off. We run several RL experiments with TMol reward, including TMol-only and Folding+TMol rewards with different weighting for the TMol (Fig. 18). When the Proteína-Reforza is trained with TMol-only reward, despite improvement over Proteína-Complexa on both unique success and TMol ddG, it cannot reach the performance of Proteína-Reforza trained using folding reward. Adjusting the weight of the TMol term in the combined folding+TMol reward an effectively move the Proteína-Reforza performance on a pareto front of TMol energy and unique success trade-off, the 0.25 weighting of TMol term strikes a balance. We report this checkpoint in Tab. 2.

G.2 Ablation Study: Transfer Across Folding Models

Motivation and evaluation setup. Optimizing a folding oracle can exploit its scoring preferences: apparent gains may reflect overfitting to that oracle while failing to improve under independent folding models. We test this possibility by evaluating the same generated designs with AlphaFold2-Multimer (denoted AlphaFold2), Boltz2, and ESMFold2. For each generative model, we sample 200 designs for each of 22 protein targets, giving 4,400 designs per evaluator. AlphaFold2 provides 5 weight sets. During AlphaFold2-reward training, we randomly select among the last 4 (configured indices 1–4) and reserve the first (index 0) for evaluation, for cross-validation across weights and folding models.

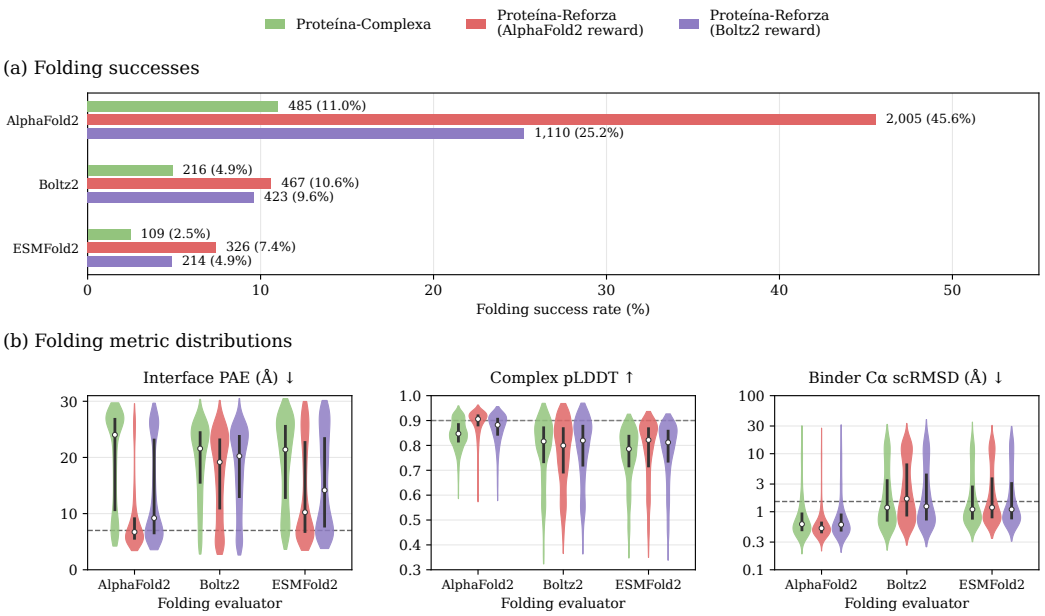


Figure 19: **Transfer across folding models.** (a) Success counts and rates. (b) Folding metrics for all 4,400 designs per model; markers show medians and interquartile ranges, dashed lines show thresholds.

RL training based on AlphaFold2 reward uses initial-guess conditioning and 1 recycle; evaluation uses the held-out weights, target templates, initial guesses, and 3 recycles. Boltz2 evaluation uses 3 recycles and 1 prediction, with target-only templates from the generated target coordinates and no MSA. ESMFold2 uses 10 loops, selecting the highest-ipTM prediction from 5 predictions, without an MSA or target distogram. We apply the same success thresholds to all evaluators: ipAE ≤ 7 Å, complex pLDDT ≥ 0.90, and binder C_α scRMSD < 1.5 Å. These common cutoffs are a reporting convention, not a calibration across models: their metric distributions differ, and other cutoffs could be chosen for different evaluators. We compare counts and metric distributions within each evaluator.

Transfer beyond the training oracle. AlphaFold2-reward RL increases successes under Boltz2 from 216 to 467 and under ESMFold2 from 109 to 326 (Fig. 19). Under the held-out AlphaFold2 weights, successes increase from 485 to 2,005. Thus, the gains extend beyond both the AlphaFold2 weights used for training and the AlphaFold2 model family. Both independent evaluators also show lower median interface PAE after RL, providing evidence of improved interface confidence beyond the thresholded success counts. These results argue against gains arising solely from exploitation of the training oracle, although agreement among predictors does not rule out shared biases.

RL with a Boltz2 reward. To test whether the RL method also works with a different reward model, we replace AlphaFold2 with Boltz2 while reusing the optimization hyperparameters tuned for AlphaFold2. This is an illustration of transferability; Boltz2 may benefit from a different optimal hyperparameter set, which we do not tune here. Boltz2-reward RL increases successes from 216 to 423 under Boltz2, from 485 to 1,110 under AlphaFold2, and from 109 to 214 under ESMFold2. Median interface PAE decreases under all 3 evaluators, while pLDDT and binder scRMSD changes depend on the evaluator. These gains support RL with either folding oracle.

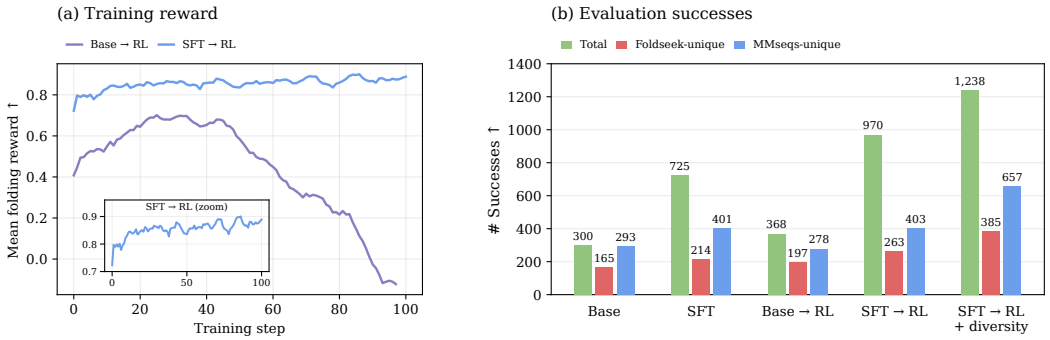


Figure 20: **Ligand SFT and diversity ablation.** (a) Training reward from base and SFT initializations. (b) Total, Foldseek-unique, and MMseqs-unique successes across 12 targets, including RL with the diversity bonus.

G.3 Ablation Study: Ligand SFT and Diversity Reward

Why initialize RL with SFT? RL needs a strong starting policy to improve performance without destabilizing generation. For ligand binders, we first use supervised fine-tuning (SFT) on positive designs, then optimize the RF3 folding reward. SFT trains rank-32 LoRA adapters at learning rate 10^{-5}; RL continues adapter training at 3×10^{-5}, with the remaining backbone and autoencoder frozen (Sec. H.2). We compare base- and SFT-initialized RL without the diversity bonus at a matched training budget, using 200 designs per ligand target and the RF3 criteria in Sec. D.

Fig. 20(a) shows that SFT provides a better starting point, leading to substantially higher training reward throughout RL. Starting from the base model instead yields large initial gains, followed by a decline as training progresses. This behavior reflects the difficulty of improving a weak initial policy: further gains require increasingly large departures from the base model, which can destabilize generation and lead to collapse. KL regularization constrains this departure, but simply removing it to allow further improvement risks collapse as well. A strong SFT initialization reduces this tension by placing RL closer to high-reward solutions, supporting sustained improvement with less policy drift.

Why add a diversity reward after SFT? Fig. 20(b) shows that a better initialization alone does not ensure a diverse set of successful designs. SFT increases total/Foldseek-unique/MMseqs-unique successes from 300/165/293 to 725/214/401. Subsequent RL reaches 970/263/403, compared with 368/197/278 for base-initialized RL. However, its 33.8% increase in total successes over SFT is accompanied by only a 22.9% increase in structurally unique successes and a 0.5% increase in sequence-unique successes. Thus, optimizing folding success alone increasingly produces redundant candidates, motivating an explicit incentive for diversity among successful designs.

We add the successful-fold diversity bonus with weight 0.5 to ligand RL. The final model yields 1,238 total, 385 Foldseek-unique, and 657 MMseqs-unique successes, expanding both structural and sequence coverage in Fig. 20(b). This comparison also uses longer training and does not isolate the bonus; the target-specific ablation in Sec. G.1 provides complementary evidence.

G.4 Generalization to Structurally Dissimilar Targets

RL gains could depend on structural similarity to training targets, limiting their relevance to new targets. To test this possibility, Fig. 21 compares per-target gains with similarity to the eligible RL training pool. For each test chain, we take the highest reported query-normalized TM-score against 52,492 training-target entries, then average these scores with equal weight per chain.

Across 22 targets, similarity has little linear association with absolute Foldseek-unique gains (Pearson $r = -0.066$) and a weak positive association with MMseqs-unique gains ($r = 0.206$). All 3 targets below 0.5 similarity gain Foldseek-unique successes: Claudin1, from 3 to 67 (similarity 0.421); SpCas9, from 8 to 45 (0.259); and CbAgo, from 0 to 8 (0.496). BBF14 improves from 22 to 139 at similarity 0.596. High similarity alone does not ensure success: TNFα and IL17A remain at 0 despite similarities of 0.994 and 0.912. These results support generalization beyond close structural matches to the RL training pool. The comparison uses eligible training targets and does not measure actual RL sampling, pretraining exposure, or interface overlap.

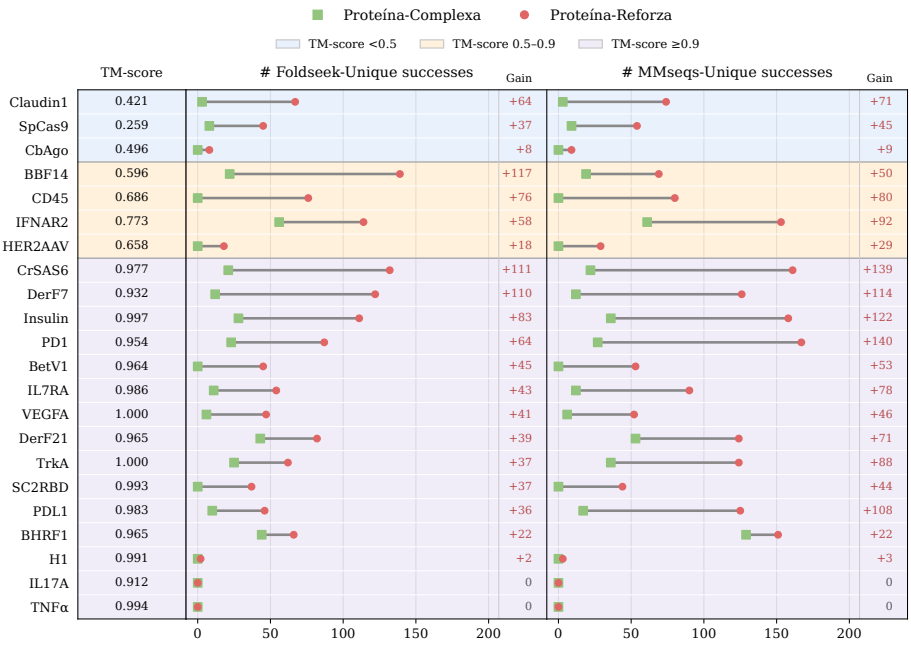
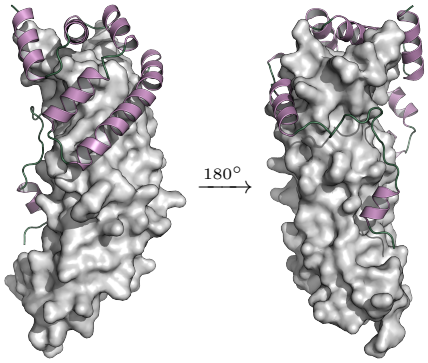


Figure 21: **Per-target RL gains versus training-target similarity.** Connected markers compare base and general RL unique-success counts; gain columns show absolute increases. Similarity is the mean nearest-chain TM-score. Shading marks similarity ranges; rows within each range are ordered by Foldseek-unique gain.

(a) Unrealistic binding mode



(b) NPSA improves filter pass rates

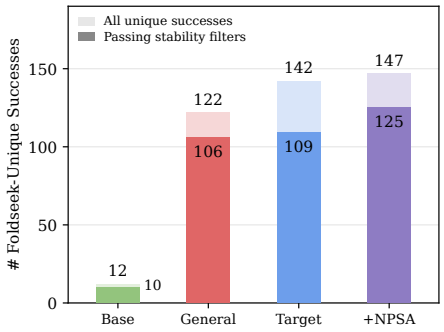


Figure 22: **DerF7 stability reward.** (a) A target-specific RL binder wraps around DerF7 (gray surface) in an unrealistic binding mode despite passing folding criteria; views differ by 180° around the vertical axis. (b) NPSA reward increases Foldseek-unique successes passing our stability screening criteria (solid bar segments).

G.5 Target-specific RL with Stability Reward

Optimizing folding success alone can increase target contacts at the expense of binder stability. For example, the DerF7 binder in Fig. 22(a) is generated by target-specific RL with the same setup as other targets in Fig. 4. This sample passes the folding-success criteria but adopts an unrealistic binding mode, wrapping around the target in an elongated conformation. The 2 views show the same complex rotated by 180° around the vertical axis. The design also fails our stability screening criteria (Sec. D), which serve as structural heuristics rather than a direct measurement of physical stability. This extensive target contact may improve folding scores without ensuring a compact, stable binder, suggesting a potential mode of reward hacking. To mitigate this problem, we introduce the stability reward in Eq. (7), which is based on buried nonpolar surface area (NPSA) and encourages hydrophobic burial within the binder itself.

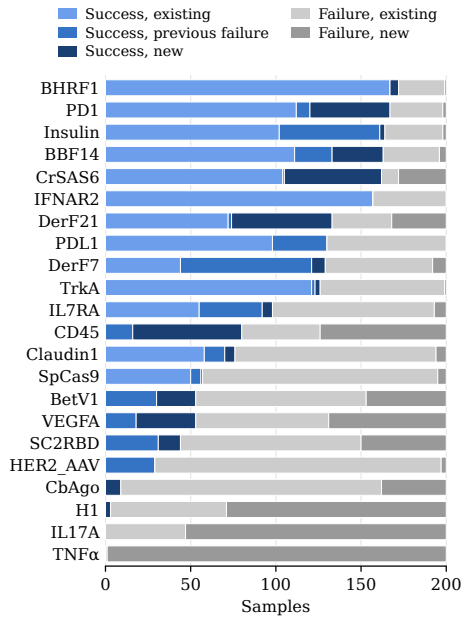


Figure 23: **Interface cluster count breakdown.** Interface is defined having at least one heavy atom within an 8 Å cut-off to the target chains. For each target, the generated interfaces are clustered using FoldSeek multimer with a threshold of 0.3.

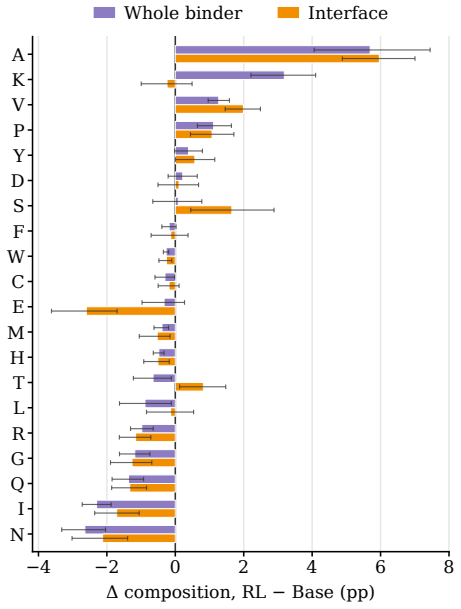


Figure 24: **Amino acid composition shift.** Results are aggregated across all 22 targets, each with 200 samples for Proteína-Reforza, compared to Proteína-Complexa, in percentage points (pp). Intervals are 95% CI.

Both target-specific variants start from the general RL model and use the successful-fold diversity reward; the stability variant additionally applies the NPSA reward with weight 0.2 (Tab. 19). We generate 200 designs per model and count Foldseek-unique folding successes. A successful cluster passes our stability screening criteria if at least 1 of its successful designs meets all criteria in Sec. D, retaining the original cluster assignments. Target-specific checkpoints are selected by Foldseek-unique success. In the figure, Base and General denote Proteína-Complexa and Proteína-Reforza, while Target and +NPSA denote target-specific RL without and with the stability reward.

DerF7 illustrates how the stability reward mitigates this issue (Fig. 22(b)). Target-specific RL increases Foldseek-unique successes from 122 to 142 relative to general RL, but only increases stability-passing clusters from 106 to 109, reducing their fraction from 86.9% to 76.8%. This pattern is consistent with reward hacking through over-optimization of folding objectives at the expense of stability criteria. Adding the buried NPSA reward increases stability-passing clusters to 125 (85.0%) while further increasing Foldseek-unique successes to 147. Explicitly rewarding hydrophobic burial therefore improves the yield of diverse designs that satisfy both folding-success and stability criteria.

G.6 Interface, Sequence and Bioinformatics Metrics Analysis

A higher *in-silico* success rate does not by itself show that the policy designs better binders: it could also mean the policy has collapsed onto a narrow set of outputs, or learned to exploit biases in the evaluation metrics. To better understand the mechanism of RL training, we therefore compare Proteína-Reforza's output distribution with Proteína-Complexa's along three axes: the binding modes, their sequence composition, and the physical properties of the resulting interfaces.

Interface-cluster breakdown. Decomposing each of Proteína-Reforza's 200 hotspot-conditioned designs per target against Proteína-Complexa's 200 for the same target, using FoldSeek interface clusters at TM 0.3, shows that the gain in successes (2,005 versus 485) comes predominantly from binding modes the base model already reaches (Fig. 23 in Sec. G.6). Of Proteína-Reforza's 2,005 successes, 62.4% occupy a mode in which Proteína-Complexa also succeeds, 19.1% occupy a mode the base model samples but never succeeds in, and 18.6% occupy a mode absent from the base model's 200 designs. With hotspot conditioning on, newly-observed interface clusters therefore

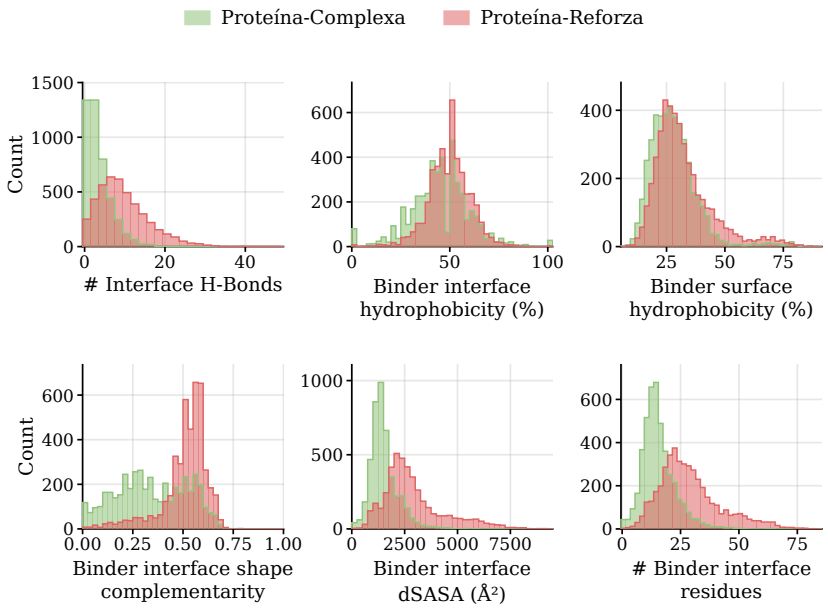


Figure 25: **Interface physical metrics.** Statistics are calculated over all 22 targets, each with 200 samples. AlphaFold2 refolded structures are used for the calculation.

reflect finer-grained structural variation at largely shared epitopes, demonstrating the diversity and exploratory power for discovering new binding modes.

Amino-acid composition. Reinforcement learning shifts binder composition, but the shift is not uniform across the binder and the interface (Fig. 24). Comparing Proteína-Reforza to Proteína-Complexa over all 4,400 hotspot-conditioned designs per model, alanine (A) increases most at both scopes and by almost the same amount. Lysine (K) behaves in the opposite way: it rises sharply over the whole binder yet is unchanged at the contact surface, identifying it as a scaffold effect. The converse also occurs: glutamate (E) is flat over the whole binder but is the most depleted residue at the interface. Serine (S) and threonine (T) are enriched at the interface, with their ability in forming hydrogen bonds to stabilize the complex.

Interface metrics. Across all 4,400 hotspot-conditioned designs per model, Proteína-Reforza produces markedly larger and better-packed interfaces than Proteína-Complexa on the AF2-refolded complexes (Fig. 25). The median interface buries 2,544 Å^2 of surface versus 1,397 Å^2, involves 27 binder residues versus 14, and forms 8 interface hydrogen bonds versus 3. Shape complementarity rises from a median of 0.33 to 0.53, and the interface becomes modestly more hydrophobic (45% $\rightarrow$ 50%), with a smaller shift in overall binder surface hydrophobicity (27% $\rightarrow$ 29%). The distributions are broadly separated rather than merely shifted in the tail, indicating that the improvement extends across the design population. Values are computed over all designs, not only *in-silico* successes.

G.7 Inference-time Scaling for Proteína-Reforza

Motivation and setup. RL could improve success rates by concentrating probability on a small set of high-reward modes, sacrificing the ability to discover additional distinct binders with more inference compute. We test this possibility by measuring how Foldseek-unique successes grow with the compute budget, following the inference-time scaling protocol of Proteína-Complexa (Didi et al., 2026b). At each budget, we cluster successful binders separately for each target at a Foldseek TM-score threshold of 0.5 and report the number of distinct clusters. Continued growth therefore reflects additional successful folds, rather than repeated sampling of the same successful designs.

For proteins, we compare Proteína-Reforza, Proteína-Complexa, and BindCraft on the 22 targets in Tab. 9, with budgets of 2–16 GPU hours per easy target and 4–32 GPU hours per hard target for every method, including both generation and evaluation. The 12 easy targets are PD-L1, BHRF1,

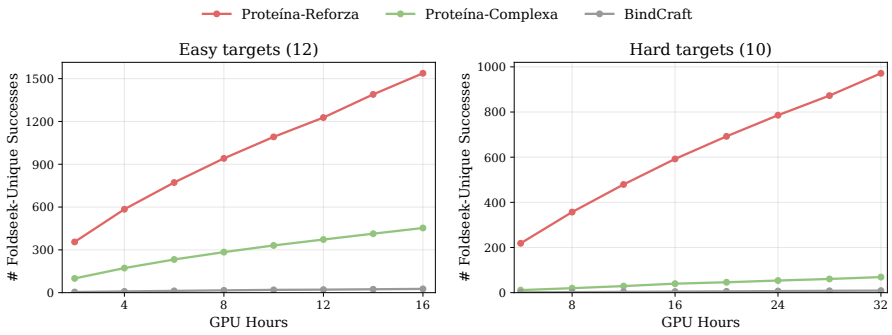


Figure 26: **Inference-time scaling for protein targets.** Mean Foldseek-unique successes for 12 easy targets (left) and 10 hard targets (right), comparing three methods under matched A100 GPU-hour budgets.

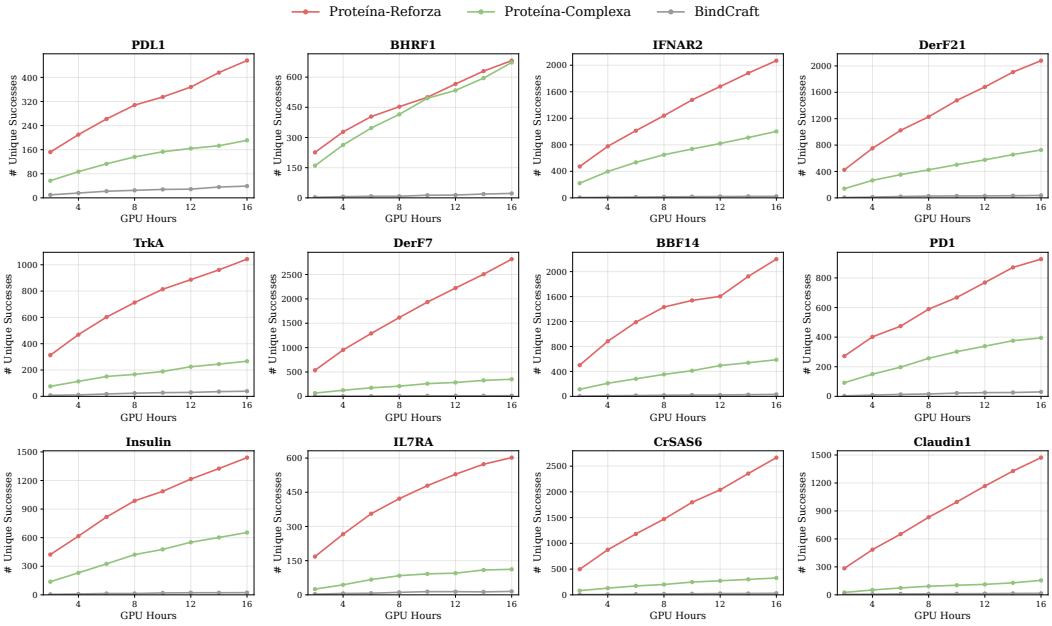


Figure 27: **Results on easy protein targets.** Foldseek-unique successes over 2–16 GPU hours per target.

IFNAR2, DerF21, TrkA, DerF7, BBF-14, PD-1, Insulin, IL7RA, CrSAS6, and Claudin1. The 10 hard targets are SC2RBD, CbAgo, BetV1, HER2AAV, CD45, SpCas9, VEGFA, H1, TNF-α, and IL17A, including the 3 very hard targets studied separately in Complexa. For ligands, we use all 12 targets in Tab. 10 and allocate 2–22 A100 GPU hours per target to each method, including generation and evaluation. We compare against Proteína-Complexa, RFDiffusion3, BoltzGen, ODesign, and BoltzDesign. Both Proteína-Reforza and Proteína-Complexa use self-generated sequences; external baselines use LigandMPNN or their own inverse folding model (Sec. I).

Accelerated sampling. For protein targets, we replace 400-step SDE sampling with 100-step TSR sampling, which gives similar performance under matched compute budgets while substantially reducing generation time. For ligand targets, we find that 100-step TSR sampling degrades performance, so we retain 400-step SDE sampling. We also enable `torch.compile` for generation, applying the accelerated setup to both Proteína-Reforza and Proteína-Complexa. We use only best-of-N sampling for inference-time scaling: additional compute generates and evaluates more candidates, without intermediate search or trajectory steering. Fast generation with TSR and compilation makes this simple strategy practical and tests whether RL retains a broad supply of successful designs.

Protein targets. Fig. 26 shows that RL increases the yield of distinct successful binders throughout the measured compute range. At 16 GPU hours per easy target, Proteína-Reforza averages 1,537.6

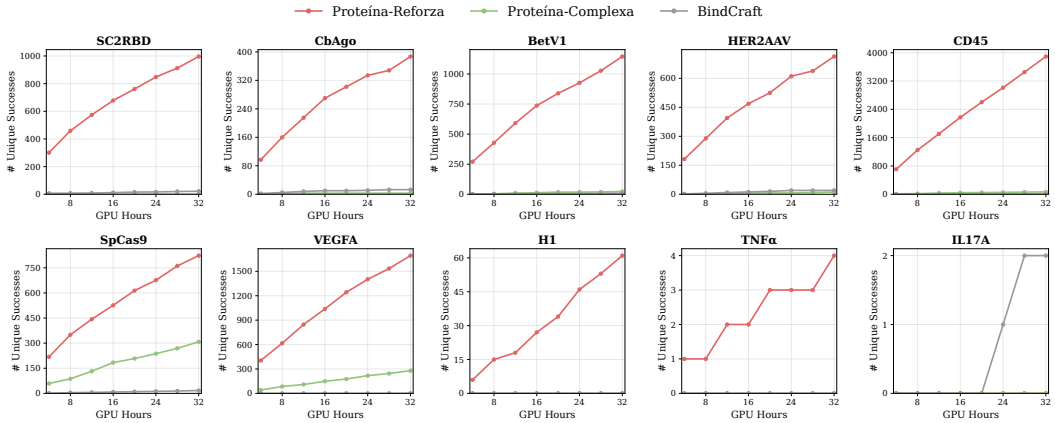


Figure 28: **Results on hard protein targets.** Foldseek-unique successes over 4–32 GPU hours per target.

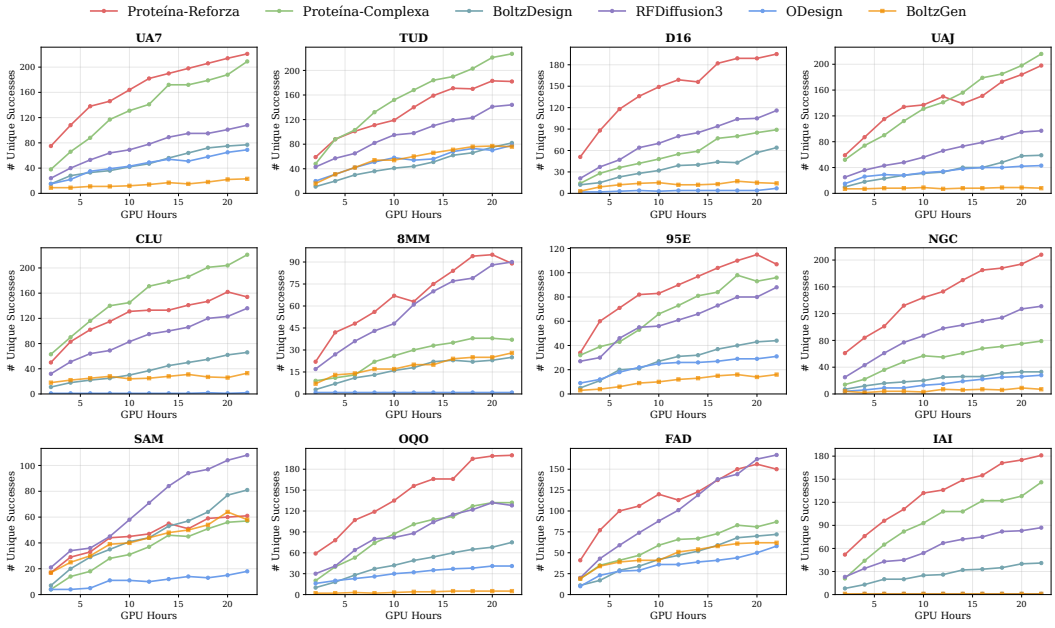


Figure 29: **Inference-time scaling on ligand targets.** Foldseek-unique successes across 12 targets under matched budgets of 2–22 A100 GPU hours per target, including generation and evaluation.

Foldseek-unique successes, compared with 453.3 for Proteína-Complexa and 26.5 for BindCraft. At 32 GPU hours per hard target, the corresponding counts are 971.5, 69.0, and 9.1. These represent $3.4\times$ and $14.1\times$ gains over the base model on the easy and hard groups, respectively. The RL curves continue to rise at the largest budgets, showing that the improved success rate does not come at the cost of exhausting a small set of successful folds.

The per-target curves in Figs. 27 and 28 show substantial gains across both difficulty groups, although the absolute yield remains target-dependent. CD45 and VEGFA illustrate strong scaling on hard targets, while H1, TNF-α, and IL17A remain challenging. RL therefore preserves the benefit of additional sampling, with the largest gains on targets where successful folds remain accessible.

Small-molecule targets. The ligand scaling experiment tests both diversity after RL and practical efficiency against the full set of baselines. The comparison in Sec. 4.2 and Fig. 6 (*left*) shows that Proteína-Reforza achieves the highest mean Foldseek-unique success count at every evaluated budget from 2 to 22 GPU hours per target, ahead of Proteína-Complexa, RFDiffusion3, BoltzGen, ODesign, and BoltzDesign. At 22 GPU hours, it yields 162.2 unique successes per target, versus 133.0 for

Proteína-Complexa and 116.7 for RFDiffusion3, improvements of 22.0% and 39.0%, respectively. RL yields more diverse successful ligand binders at the same compute cost.

Fig. 29 shows that these gains span chemically diverse ligands while retaining target-specific variation. At 22 GPU hours, Proteína-Reforza leads on 6 of the 12 targets, with particularly large gains over Proteína-Complexa on D16 (195 versus 89 unique successes) and NGC (208 versus 79). The base model remains ahead on TUD, UAJ, and CLU, while RFDiffusion3 leads on 8MM, SAM, and FAD. Although some individual curves flatten at larger budgets, the aggregate RL curve continues to grow and remains ahead of all baselines. Together, these results show that RL improves inference-time scaling while retaining the ability to discover distinct successful binders within these budgets.

G.8 Hotspot Commitment and Hotspot Independence for Proteína-Reforza

Hotspot conditioning is one of the most important controls in structure-based binder design: specifying which target residues the binder should contact turns an open-ended search over the whole surface into a localized one, and steers generation toward the functionally relevant epitope rather than merely the most designable patch. But the epitope is not always known: annotations may be incomplete, inferred, or absent on novel targets. We therefore ask whether RL makes the policy more faithful to the given hotspots, and how much performance degrades if removing the condition.

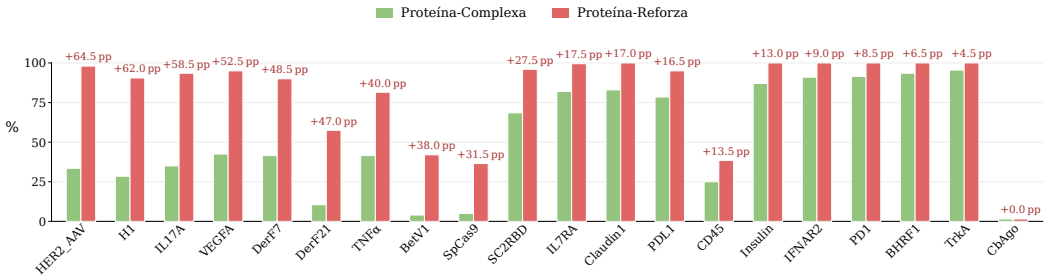


Figure 30: **Hotspot commitment.** Percentage of designs contacting a hotspot within 8 Å across 20 targets.

Hotspot commitment. When comparing Proteína-Reforza with Proteína-Complexa, hotspot commitment improves on essentially every target (Fig. 30) for the hotspot contact with a cut-off distance of 8 Å. We measure the fraction of a model's 200 designs that place a binder Cα within 8 Å of at least one requested hotspot in the refolded complex. The average commitment ratio rises from 51.95% to 80.75%. The gain is positive for 19 of 20 targets and is largest where the base model complies least (HER2_AAV +64.5, H1 +62.0, IL17A +58.5 pp). The two exceptions are informative: CbAgo is unchanged at 1.5%, and CD45 improves only from 25.0% to 38.5%, both targets where neither model reliably reaches the annotated hotspot, indicating a poorly provided condition. Combined with our analysis in Sec. G.6, RL might refine existing binding modes close to the hotspot conditioning, thus improving the hotspot commitment ratio instead of diverging from the provided conditions.

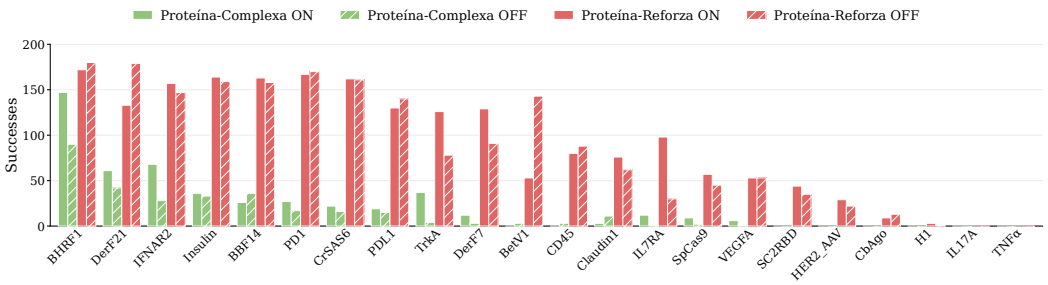


Figure 31: **Impact of hotspot conditioning**. Raw successes are counted among 200 generations per target.

Binder design without hotspot conditioning. Removing hotspot conditioning costs the two models very differently (Fig. 31). Proteína-Complexa falls from 485 to 305 *in silico* successes (-37%),

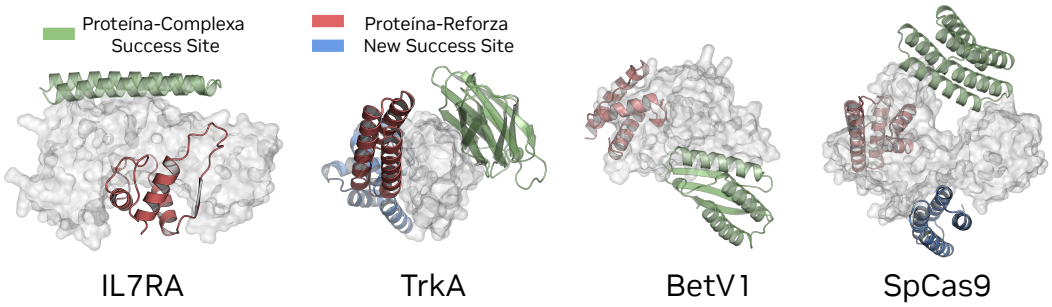


Figure 32: **New binding sites found by Proteína-Reforza where Proteína-Complexa fails to explore.** All binders shown are *in silico* successes, with hotspot off to allow for flexibility in exploring new binding sites.

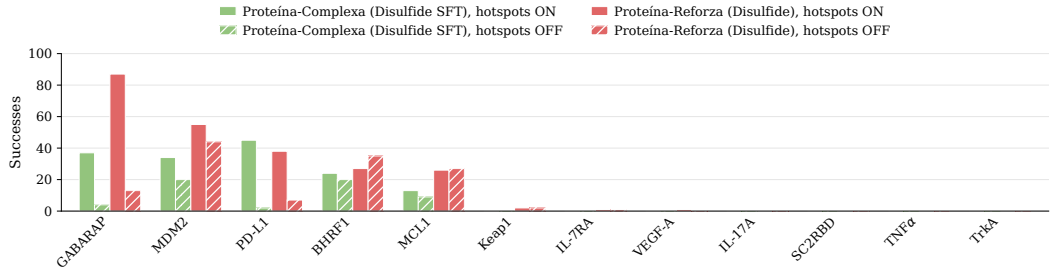


Figure 33: **Impact of hotspot conditioning for Proteína-Reforza (Disulfide) and Proteína-Complexa (Disulfide SFT)**. Raw successes are counted among 200 generations per target.

whereas Proteína-Reforza falls only from 2,005 to 1,955 (-2.5%), suggesting that the RL policy retains most of its design quality without an explicit site condition. Per target, the change has both signs, with BetV1 rising from 53 to 143 successes when the hotspot is removed, while IL7RA falls from 98 to 30. Unconditioned performance thus matches conditioned performance for Proteína-Reforza. This suggests a practical use case on novel targets, where no hotspot can be supplied.

New binding sites. Interestingly, in the no hotspot conditioned setting, we observe that Proteína-Reforza succeeds at interface sites that no Proteína-Complexa success occupies (Fig. 32). Designs are assigned FoldSeek interface clusters at TM 0.3, and clusters are further grouped into binding sites by target contact-footprint overlap. Four representative targets are displayed, where Proteína-Reforza explores new binding sites with *in silico* success. This exploration expands interface diversity when hotspots are poorly defined and can identify binding sites absent from existing annotations.

Terminal-disulfide peptides without hotspot conditioning. We repeat the hotspot ablation for terminal-disulfide peptide design (Fig. 33). Proteína-Complexa (Disulfide SFT) falls from 153 to 55 successes (-64%), and Proteína-Reforza (Disulfide) falls from 237 to 129 (-46%). RL reduces the dependence on an explicit site condition, but hotspot conditioning still helps both models substantially.

G.9 Transfer across Peptide and Minibinder Design Tasks

Specializing a model for disulfide peptides could limit its performance on other binder-design tasks. We test transfer across closure types and retention of minibinder design, using both TED-derived and CPSea-derived SFT initializations (Sec. B.3). All comparisons use 200 designs per target, with 12 targets for cyclic peptides and 22 for minibinders.

Disulfide design across SFT datasets. To test whether RL gains depend on the SFT data source, we compare both initializations on disulfide design (Tab. 15). BoltzGen uses BoltzIF sequence redesign and its own sampling procedure. Mean folding successes per target increase from 8.58 to 13.50 for the CPSea models and from 12.75 to 19.75 for the TED-derived models. RL therefore improves both initializations, with the TED-derived models achieving higher success counts at both the SFT and RL

Model	# Folding successes ↑			TMol energy ↓		Interface
	Total	Foldseek	MMseqs	$\Delta\Delta G$	Int.	H-bonds ↑
BoltzGen	1.42	1.08	0.83	-53.09	-56.85	2.58
Proteína-Complexa$_{\text{CPSea Disulfide SFT}}$	8.58	2.17	2.08	-53.04	-64.30	2.32
Proteína-Reforza$_{\text{CPSea Disulfide RL}}$	13.50	5.25	3.67	**-54.24**	-64.50	2.88
Proteína-Complexa$_{\text{Disulfide SFT}}$	12.75	5.42	4.25	-53.89	-63.43	2.23
Proteína-Reforza$_{\text{Disulfide RL}}$	**19.75**	**6.17**	**4.58**	-52.69	**-65.13**	**2.98**

Table 15: **Disulfide peptide design across SFT datasets.** Means over 12 targets with 200 designs per target. Foldseek and MMseqs count unique structures and sequences; energy and hydrogen bonds are averaged over successful designs.

Model	# Folding successes ↑			TMol energy ↓		Interface
	Total	Foldseek	MMseqs	$\Delta\Delta G$	Int.	H-bonds ↑
Proteína-Complexa	22.05	14.18	20.09	-73.57	-103.68	5.90
Proteína-Reforza$_{\text{Protein RL}}$	91.14	61.82	83.91	-110.45	-158.97	10.08
Proteína-Complexa$_{\text{CPSea Disulfide SFT}}$	23.27	12.77	21.32	-65.74	-95.82	6.19
Proteína-Reforza$_{\text{CPSea Disulfide RL}}$	**33.45**	**19.23**	**31.86**	-66.99	**-96.59**	**6.34**
Proteína-Complexa$_{\text{Disulfide SFT}}$	18.95	11.45	17.91	-64.83	-91.84	5.44
Proteína-Reforza$_{\text{Disulfide RL}}$	28.55	17.95	27.32	**-67.14**	-96.58	6.06

Table 17: **Minibinder design after disulfide-peptide RL.** Means over 22 targets with 200 designs each; bold compares the 4 disulfide models. Base and protein RL rows provide references. Energy and hydrogen bonds are averaged over successful designs.

stages. All 4 models exceed BoltzGen's 1.42 successes per target, supporting RL gains across both SFT datasets under hotspot conditioning.

Transfer to head-to-tail peptides. To test whether improvements extend beyond the closure used during RL, we evaluate head-to-tail peptide generation (Tab. 16). We use cyclic relative positional encoding within the binder, following Rettie et al. (2025a), so that the first and last residues are treated as adjacent. This introduces no parameters and allows minibinder checkpoints to be sampled directly under cyclic conditioning as transfer controls. Without cyclic-peptide training, Proteína-Complexa and Proteína-Reforza$_{\text{Protein RL}}$ achieve 16.08 and 29.17 successes per target, respectively. The CPSea SFT and RL models achieve 11.25 and 13.50, while the TED-derived SFT and RL models achieve 22.92 and 30.50. These results support transfer beyond the disulfide RL task, but head-to-tail generation is not an unseen task for the CPSea models because their SFT mixture includes head-to-tail complexes.

Table 16: **Head-to-tail peptide design.** Mean successes over 12 targets, with 200 designs per target.

Model	# Folding successes ↑		
	Total	Foldseek	MMseqs
BoltzGen	17.42	8.42	11.42
Proteína-Complexa	16.08	8.00	8.33
Proteína-Reforza$_{\text{Protein RL}}$	29.17	**15.17**	**13.17**
Proteína-Complexa$_{\text{CPSea Disulfide SFT}}$	11.25	3.42	3.50
Proteína-Reforza$_{\text{CPSea Disulfide RL}}$	13.50	4.33	3.58
Proteína-Complexa$_{\text{Disulfide SFT}}$	22.92	10.75	10.92
Proteína-Reforza$_{\text{Disulfide RL}}$	**30.50**	13.25	12.17

Retention of minibinder design. Specialization on short cyclic peptides could reduce performance on longer, noncyclic minibinders. We test this possibility in Tab. 17, using the base and protein RL models as references. Bold values compare the 4 disulfide models. Relative to CPSea SFT, RL increases mean successes from 23.27 to 33.45 per target (11.64% to 16.73%), Foldseek-unique successes from 12.77 to 19.23, and MMseqs-unique successes from 21.32 to 31.86. The TED-derived models show the same trend, increasing from 18.95 to 28.55 successes per target (9.48% to 14.28%). Thus, peptide RL improves minibinder success relative to both SFT initializations, although the dedicated protein RL model remains substantially stronger.

Together, these experiments show that RL improvements extend beyond the task used for optimization: protein RL improves head-to-tail peptide generation, while disulfide-peptide RL improves both head-to-tail generation and minibinder design relative to its SFT initializations. Transfer is incomplete, since the dedicated protein RL model remains stronger on minibinders. Task-specific training and the choice of SFT data therefore remain important despite these gains across tasks.

H Model Training and Inference Details

This section describes supervised fine-tuning, reinforcement learning, and inference for our models. All models retain the partially latent flow-matching network and autoencoder architectures of Proteína-Complexa (Didi et al., 2026b). The flow-matching backbone is a pair-biased transformer without triangular layers, and both the encoder and decoder of the autoencoder use standard transformers. This architecture avoids costly triangular operations and supports fast binder generation. The autoencoder remains frozen during minibinder RL. For details, see Didi et al. (2026b) and Geffner et al. (2026).

Model notation. We denote the pretrained Proteína-Complexa initialization by $\mathcal{M}_{\text{base}}$, using the checkpoint for the corresponding target type. We denote the general protein RL model by $\mathcal{M}_{\text{P}}$, its target-specific variants by $\mathcal{M}_{\text{P},t}$, the TMol-reward variant by $\mathcal{M}_{\text{P}}^{\text{TMol}}$, and the wet-lab reward variant by $\mathcal{M}_{\text{P}}^{\text{WL}}$. For ligand targets, $\mathcal{M}_{\text{L}}^{\text{SFT}}$ denotes the supervised initialization, $\mathcal{M}_{\text{L}}$ the folding-reward RL model, and $\mathcal{M}_{\text{L}}^{\text{Nesso}}$ the model with the additional Nesso-1 reward. For cyclic peptides, $\mathcal{M}_{\text{C}}^{\text{SFT}}$ and $\mathcal{M}_{\text{C}}$ denote the TED-derived SFT and RL models; $\mathcal{M}_{\text{C,CPSea}}^{\text{SFT}}$ and $\mathcal{M}_{\text{C,CPSea}}$ denote their CPSea-derived counterparts.

Reward composition. For minibinder RL, we combine the rewards defined in Secs. 3.2 and 3.4 as

$$\begin{aligned} R_i = \lambda_{\text{succ}} R_i^{\text{succ}} + \lambda_{\text{div}} R_i^{\text{div}} + \lambda_{\text{stab}} R_i^{\text{stability}} \\ + \lambda_{\text{TMol}} R_i^{\text{Rosetta-}\Delta\Delta G} + \lambda_{\text{aff}} R_i^{\text{Affinity}} + \lambda_{\text{clf}} R_i^{\text{Clf}}. \end{aligned} \tag{8}$$

The terms measure folding success, successful-fold diversity, stability, normalized TMol interface energy, Nesso-1 affinity, and capped wet-lab classifier probability, respectively. All weights are nonnegative; unused terms have weight 0. The stability reward is already nonpositive, and the TMol and affinity rewards already include the negative sign that favors lower energies and affinity values. Model and target weights are listed in Tabs. 18 and 19; cyclic rewards are defined in Sec. H.3.

Training algorithm. Algorithm 1 summarizes AWM (Xue et al., 2026) in the partially latent representation of Proteína-Complexa (Didi et al., 2026b). Rewards are evaluated on decoded binders, while policy updates operate on their coordinate and latent endpoints using fresh flow-matching noise. The initialization may be a pretrained, SFT, or general RL checkpoint, depending on the task. Gradients flow only through the trainable policy; rewards, advantages, the decoder, and the old and reference policies are held fixed during each update. Here M is the number of accumulated microbatches in Tab. 18.

H.1 Minibinder RL for Protein Targets

General protein-target model. We initialize from the pretrained Proteína-Complexa protein-binder model and optimize with AWM. Tab. 18 summarizes the minibinder and cyclic-peptide RL settings. For the general protein model, we sample PDB targets uniformly over target clusters. We retain the full target with probability 0.5; otherwise, we crop it to 64–128 residues. Each hotspot is independently retained with probability 0.5, with at most 6 hotspots per view. The reward is the dense AF2 folding-success reward in Eq. (3), without additional diversity or stability terms.

Folding and diversity reward configuration. Protein folding rewards use AlphaFold2-Multimer (AF2) with initial-guess conditioning and 1 recycle. For minibinders, we randomly select among 4 AF2 checkpoints to prevent reward hacking towards a specific one. Target-specific minibinder runs additionally use the successful-fold diversity reward in Eq. (6), with Foldseek clustering at a TM-score threshold of 0.5. The DerF7 stability variant also adds the buried-NPSA penalty in Eq. (7). Per-target reward weights are listed in Tab. 19.

Policy optimization. We refresh the rollout buffer at each optimizer update and resample training microbatches from it. Each resampled design receives fresh flow-matching times and noise, so the time samples per design in the table are averages over buffer resampling, distinct from rollout sampling steps. Advantages are centered within each target group and scaled using a rollout-wide reward standard deviation with a fixed floor. We regularize against the fixed reference policy using the settings in Tab. 18. The old policy follows an EMA of the trainable policy: its decay starts at 0 and increases linearly to the tabulated cap. This old-policy EMA is distinct from checkpoint selection, for which we use the trainable policy's non-EMA weights. In practice, importance ratio clipping is set but never triggered, as on-policy training is used with little off-policy from EMA updates.

Table 18: **RL training hyperparameters for minibinders and cyclic peptides.**

Hyperparameter	Protein targets				Ligand targets		Disulfide	
	$\mathcal{M}_{\mathrm{P}}$	$\mathcal{M}_{\mathrm{P},t}$	$\mathcal{M}_{\mathrm{P}}^{\mathrm{TMol}}$	$\mathcal{M}_{\mathrm{P}}^{\mathrm{WL}}$	$\mathcal{M}_{\mathrm{L}}$	$\mathcal{M}_{\mathrm{L}}^{\mathrm{Nesso}}$	$\mathcal{M}_{\mathrm{C,CPSea}}$	$\mathcal{M}_{\mathrm{C}}$
Initialization and optimization								
Initialization	$\mathcal{M}_{\mathrm{base}}$	$\mathcal{M}_{\mathrm{P}}$	$\mathcal{M}_{\mathrm{base}}$	$\mathcal{M}_{\mathrm{base}}$	$\mathcal{M}_{\mathrm{L}}^{\mathrm{SFT}}$	$\mathcal{M}_{\mathrm{L}}^{\mathrm{SFT}}$	$\mathcal{M}_{\mathrm{C,CPS}}^{\mathrm{SFT}}$	$\mathcal{M}_{\mathrm{C}}^{\mathrm{SFT}}$
Learning rate	3×10^{-5}	3×10^{-6}	3×10^{-5}	3×10^{-5}	3×10^{-5}	3×10^{-5}	2×10^{-5}	3×10^{-5}
# GPUs	8	8	8	8	32	32	8	8
Training microbatch / GPU	12	8 or 12	12	12	12	12	16	12
Gradient accumulation	64	64–192	64	64	32	32	64	64
Effective batch size	6,144	6,144 or 12,288	6,144	6,144	12,288	12,288	8,192	6,144
Gradient norm clip	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Rollouts and policy regularization								
# groups / GPU / update	4	1	4	4	1	1	2	4
Group size $\mathcal{G}$	24	96 or 192	24	24	48	48	64	24
New designs / GPU / update	96	96 or 192	96	96	48	48	128	96
Generation batch / GPU	48	48	48	48	24	24	64	24
Sampling scheme	TSR	TSR	TSR	TSR	SDE	SDE	TSR	TSR
Rollout sampling steps	100	100	100	100	400	400	100	100
TSR (k_x, σ_x)	$(16, 3)$	$(16, 3)$	$(16, 3)$	$(16, 3)$	—	—	$(16, 3)$	$(16, 3)$
TSR (k_z, σ_z)	$(4, 8)$	$(4, 8)$	$(4, 8)$	$(4, 8)$	—	—	$(4, 8)$	$(4, 8)$
SDE noise scale	—	—	—	—	0.1	0.1	—	—
Self-conditioning	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time samples / design	8	8	8	8	8	8	8	8
KL weight β	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.10
Advantage clip	$[-5, 5]$	Tab. 19	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$
Reward std. floor	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
EMA increment	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
EMA cap	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Reward and checkpoint								
Folding reward	AF2	AF2	AF2	AF2	RF3	RF3	AF2	AF2
Folding weight λ_{succ}	1.0	1.0	1.0	0.5	1.0	1.0	1.0	1.0
AF2 recycles	1	1	1	1	—	—	1	1
RF3 predictions / design	—	—	—	—	5	5	—	—
Fold diversity weight λ_{div}	0	0.2 or 0.5	0	0	0.5	0.5	0	0.5
Stability weight λ_{stab}	0	0 or 0.2	0	0	0	0	0	0
TMol weight λ_{TMol}	0	0	0.25	0	0	0	0.2	0
Seq. diversity weight λ_{seq}	0	0	0	0	0	0	1.2	0.5
Terminal-Cys bonus weight	—	—	—	—	—	—	0	0.5
Disulfide-validity penalty weight	—	—	—	—	—	—	—	0.2
Success bonus weight λ_{bonus}	0	0	0	0	0	0	0.5	0
Nesso-1 weight λ_{aff}	0	0	0	0	0	0.3	0	0
Classifier weight λ_{clf}	0	0	0	0.5	0	0	0	0
# optimization steps	90	Tab. 19	80	60	114	66	39	24

Effective batch size = GPUs $\times$ microbatch $\times$ accumulation. Dashes denote inapplicable settings.

Algorithm 1 Advantage Weighted Matching for Proteína-Reforza

Require: Initial parameters θ_{init}, frozen decoder $\mathcal{D}$, target conditions $\mathbf{c}^{\text{target}}$, reward R, group size $\mathcal{G}$, reference weight β, and EMA schedule α_k.

Initialize $\theta \leftarrow \theta_{\text{init}}$, $\theta_{\text{old}} \leftarrow \theta_{\text{init}}$, and fixed $\theta_{\text{ref}} \leftarrow \theta_{\text{init}}$.
for each optimizer update k **do**
  Clear the rollout buffer $\mathcal{B}$.
  for each sampled target condition $\mathbf{c}^{\text{target}}$ **do**
    Sample $\{(\mathbf{x}_i^{C_\alpha}, \mathbf{z}_i)\}_{i=1}^{\mathcal{G}} \sim \pi_{\theta_{\text{old}}}(\cdot \mid \mathbf{c}^{\text{target}})$ without gradients.
    Decode binder sequences and structures with $\mathcal{D}(\mathbf{x}_i^{C_\alpha}, \mathbf{z}_i)$ and evaluate rewards R_i.
    Add the endpoints $(\mathbf{x}_i^{C_\alpha}, \mathbf{z}_i)$, condition $\mathbf{c}^{\text{target}}$, and rewards R_i to $\mathcal{B}$.
  end for
  Compute advantages A_i using Eq. (2).
  Initialize accumulated gradient $g \leftarrow 0$.
  for each of M training microbatches **do**
    Resample a microbatch from $\mathcal{B}$ and draw fresh times and noise ξ_i for each design.
    Form corrupted inputs $\mathbf{h}_i$ using the Proteína-Complexa interpolants (Sec. 2).
    Compute the corresponding velocity targets $\mathbf{u}_{x,i}$ and $\mathbf{u}_{z,i}$.
    Estimate matching losses $\widehat{\ell}_{\theta,i}$, $\widehat{\ell}_{\theta_{\text{old}},i}$ and reference penalty $\widehat{\mathcal{K}}_{\theta,\theta_{\text{ref}},i}$ on the same $\mathbf{h}_i$ (Eq. (1)).
    Compute $\rho_i \leftarrow \exp(\widehat{\ell}_{\theta_{\text{old}},i} - \widehat{\ell}_{\theta,i})$.
    Accumulate $g \leftarrow g + \frac{1}{M}\nabla_\theta \operatorname{mean}_i\left[-A_i\rho_i + \beta\widehat{\mathcal{K}}_{\theta,\theta_{\text{ref}},i}\right]$.
  end for
  Update trainable parameters θ with AdamW using g.
  Update the rollout policy: $\theta_{\text{old}} \leftarrow \alpha_k\theta_{\text{old}} + (1-\alpha_k)\theta$.
end for
return Post-trained policy π_θ with frozen decoder $\mathcal{D}$.

Target-specific adaptation. Each target-specific run starts independently from the general protein RL checkpoint at step 90, adapting it to a specific target at a reduced learning rate. We retain the same target cropping augmentation and AF2 reward setup as the general protein model, and add the successful-fold diversity reward in Eq. (6). Tab. 19 summarizes the settings for all 7 runs.

Physics-reward RL. The TMol variant $\mathcal{M}_{\text{P}}^{\text{TMol}}$ initializes from Proteína-Complexa and uses the same optimization, group size, and rollout settings as $\mathcal{M}_{\text{P}}$. Targets are sampled uniformly over clusters from a filtered PDB pool that excludes 200 targets with incomplete backbone atoms. The reward combines AF2 folding success with the normalized TMol interface-energy reward in Eq. (8). TMol scores the refolded complex after 1 repeat of Cartesian FastRelax. Its reward is the negative interface energy normalized by the number of binder interface residues. An interface residue is defined as a residue whose heavy atoms lie within 4.0 Å of any heavy atom on the other chain.

Table 19: **Target-specific protein RL settings.**

Hyperparameter	SpCas9	HER2 AAV	SC2RBD	H1	TNFα	DerF7	DerF7 + stab.
# GPUs	8	8	8	8	8	8	8
# groups / GPU / update	1	1	1	1	1	1	1
Group size $\mathcal{G}$	96	96	96	192	192	96	96
Training microbatch / GPU	12	12	12	8	12	12	12
Gradient accumulation	64	64	64	192	128	64	64
Effective batch size	6,144	6,144	6,144	12,288	12,288	6,144	6,144
Time samples / design	8	8	8	8	8	8	8
Rollout sampling steps	100	100	100	100	100	100	100
Advantage clip	$[-5, 5]$	$[-1, 3]$	$[-1, 3]$	$[-1, 3]$	$[-1, 1]$	$[-1, 3]$	$[-1, 3]$
Diversity weight λ_{div}	0.2	0.5	0.5	0.2	0.2	0.2	0.2
Stability weight λ_{stab}	0	0	0	0	0	0	0.2
# optimization steps	75	100	110	40	85	150	135

H.2 MINIBINDER RL FOR LIGAND TARGETS

Supervised initialization. We first fine-tune the Proteína-Complexa ligand binder model on the positive-design training set derived from PLINDER ligand targets (Sec. B.2). SFT trains only LoRA adapters (rank 32, $\alpha = 64$, dropout 0), with the autoencoder and remaining backbone frozen. We use Adam with learning rate 10^{-5}, 12 examples per GPU on 32 GPUs (global batch size 384), no gradient accumulation or target-conditioning dropout, and a gradient-norm clipping threshold of 1.0. Both ligand RL runs start from SFT step 7,000.

Folding-reward RL. The main ligand RL run continues to optimize only the LoRA adapters using AWM. Targets are sampled uniformly from the PLINDER training split. Optimization, rollout, and reward hyperparameters are summarized in Tab. 18; the advantage normalization and old-policy EMA schedule follow the protein RL setup. RF3 produces 5 predictions per candidate; the top-ranked prediction supplies the folding-success reward in Eq. (3). We add the successful-fold diversity bonus using a Foldseek TM-score threshold of 0.5, with no separate binary success bonus.

Affinity-reward RL. The Nesso-1 run starts from the same SFT checkpoint and uses the same optimization and rollout settings as the folding-reward ligand RL run. This variant adds the Nesso-1 affinity term in Eq. (5) to the RF3 folding reward and diversity bonus. Nesso-1 uses 5 recycling steps, protein refinement, and confidence gating for ligand placement; failed or unconfident predictions receive an affinity value of 4.0 before negation.

H.3 CYCLIC PEPTIDE RL FOR PROTEIN TARGETS

The TED-derived and CPSea-derived variants use the SFT datasets defined in Sec. B.3. The TED-derived pair is used in the main computational benchmark (Tab. 3); the CPSea-derived pair is used in the wet-lab experiments (Sec. J.2). Each RL model starts from its corresponding SFT checkpoint, with both training configurations detailed below.

TED-derived SFT. The main-table models, Proteína-Complexa$_{\text{Disulfide SFT}}$ and Proteína-Reforza$_{\text{Disulfide}}$, share an SFT stage that samples TED dimers and TED-derived disulfide crops at a ratio of 85:15 (Sec. B.3). We initialize from pretrained Proteína-Complexa and fine-tune the flow-matching network while keeping the autoencoder frozen. Training uses Adam with a learning rate of 3×10^{-5}. Each complex is limited to 500 tokens, with the disulfide peptide retained in full. We apply closure-conditioning dropout with probability 0.25 and contact and hotspot dropout with probability 0.50 each. No auxiliary closure losses are used. The checkpoint at step 3,000 serves as both the reported SFT model and the initialization for RL.

CPSea SFT. We initialize Proteína-Complexa and freeze its autoencoder. The mixture contains 587,597 TED dimers, 5,891 curated CPCore head-to-tail complexes, and 3,913 curated CPCore disulfide complexes, sampled with probabilities 0.70/0.15/0.15. We retain complete cyclic peptide chains and train with a 500-token complex budget. Closure-conditioning dropout is 0.25, and contact and hotspot dropout are each 0.50. Training uses a learning rate of 3×10^{-5} on 32 GPUs with 8 examples per GPU (global batch size 256), without auxiliary closure losses. The checkpoint at 6,750 serves as both the SFT model and RL initialization.

TED-derived and CPSea RL. We initialize Proteína-Reforza$_{\text{CPSea Disulfide}}$ from Proteína-Complexa$_{\text{CPSea Disulfide SFT}}$ and the TED-derived Proteína-Reforza$_{\text{Disulfide}}$ from Proteína-Complexa$_{\text{Disulfide SFT}}$. Tab. 18 lists the optimization and rollout settings. Targets are sampled uniformly over 52,492 PDB entries and always cropped to 64–128 residues, with hotspot retention probability 0.5 and at most 6 hotspots. Generated peptides have 12-18 residues and disulfide-closure conditioning.

TED-derived reward configuration. For the TED-derived variant, let $R_i^{\text{fold,C}}$ denote the dense AF2 folding reward, and let c_i indicate whether the peptide contains exactly two cysteines, one at each terminus. Let v_i additionally require an Sγ-Sγ distance in $(0, 2.5]$ Å in the generated peptide. The total reward is

$$R_i^{\text{C}} = \lambda_{\text{succ}} R_i^{\text{fold,C}} + \lambda_{\text{seq}} R_i^{\text{seq}} + \lambda_{\text{div}} R_i^{\text{div}} + \lambda_{\text{Cys}} c_i - \lambda_{\text{invalid}}(1 - v_i). \tag{9}$$

The fold-diversity term, R_i^{div}, is the reciprocal of the number of joint-successful designs in the same Foldseek cluster, obtained by clustering generated binder structures at a TM-score threshold of 0.5. The sequence-diversity term, R_i^{seq}, is the reciprocal of the number of joint-successful designs with the same exact sequence.

CPSea reward configuration. For the CPSea variant, let $R_i^{\text{fold,C}}$ denote the dense AF2 folding reward, R_i^{seq} the sequence-diversity reward, R_i^{SS} the TMol disulfide score, and b_i the indicator that all folding criteria pass. We combine them as

$$R_i^{\text{C}} = \begin{cases} \lambda_{\text{succ}} R_i^{\text{fold,C}} + \lambda_{\text{seq}} R_i^{\text{seq}} + \lambda_{\text{TMol}} R_i^{\text{SS}} + \lambda_{\text{bonus}} b_i, & \text{valid disulfide,} \\ -2, & \text{otherwise.} \end{cases} \tag{10}$$

Here the TMol term scores disulfides. Failed cysteine or bond checks override this sum with a reward of -2. The weights are listed in Tab. 18.

H.4 Wet Lab Reward Model RL

Supervised fine-tuning. We fine-tune the base Proteína-Complexa model on wet-lab successes for the 203 training crops, consisting of 2,469 complex structures that match the crop definition. We exclude the 10 AlphaProteo targets and some other hard targets to match the RL post-training setup (see Sec. B.4 for details). We use a learning rate of 10^{-5} and fine-tune with a total batch size of 64 for up to 5,000 steps. Every 500 steps, we checkpoint the model and report the best model with the most FoldSeek unique successes, corresponding to 1,500 training steps.

Reinforcement learning. We use the general protein RL optimization and sampling settings, with training and test targets restricted to the wet-lab classifier's coverage for reliable scoring. Target crops are sampled uniformly with no cropping. Consistent with the wet-lab target crop definition, we define exactly 1 hotspot per target and always feed it into the generative model as the condition. The binder length is randomly sampled from $[50, 88]$ for each crop.

Wet-lab reward configuration. The wet-lab variant uses Eq. (8) with $\lambda_{\text{succ}} = \lambda_{\text{clf}} = 0.5$ and all other weights set to 0. The classifier reward is $R_i^{\text{Clf}} = \min\{\text{sigmoid}(\text{Clf}(\mathbf{x}_i)), 0.7\}$, as defined in Sec. 3.4. To avoid reward hacking towards a specific classifier checkpoint, we follow a similar approach in Sec. H.1 to use 5 different wet-lab classifier checkpoints as the reward. We select the top 5 checkpoints by validation per-target AUROC. During training, we randomly select 1 of the 5 checkpoints to score the generations during post-training. For evaluation, the average probability and logits over all 5 checkpoints are reported. $\#p > 0.5$ is counted based on average classifier probabilities.

H.5 Inference Hyperparameters

Sampling schemes. We use Temporal Score Rescaling (TSR) sampling for protein-target minibinders and cyclic peptides, and stochastic differential equation (SDE) sampling for ligand-target minibinders. Subscripts x and z denote backbone C_α coordinates and sequence latents, respectively. TSR uses parameters $(k_x, \sigma_x) = (16, 3)$ for the backbone and $(k_z, \sigma_z) = (4, 8)$ for protein minibinders. For ligand SDE sampling, we follow the original Proteína-Complexa setup (Didi et al., 2026b), using a noise scale of 0.1 and a score scale of 1.0 for both components. These settings also apply to the RL rollouts in Tab. 18. All samplers use self-conditioning, a logarithmic backbone time grid and a power-law latent time grid, both with exponent parameter 2.

Minibinder generation. Protein-target models use 100 TSR steps, including the general, target-specific, TMol variants, and wet-lab targets; ligand-target models use 400 SDE steps, including the base, SFT, folding-reward RL, and Nesso-1 RL models. We use the supplied protein hotspots except in the hotspot ablation. We use no reward-based rejection or additional guidance model. Compilation is disabled during minibinder inference. RL checkpoints are listed in Tabs. 18 and 19.

Cyclic-peptide generation. Cyclic-peptide inference uses 100 TSR steps with $(k_x, \sigma_x) = (16, 3)$ and $(k_z, \sigma_z) = (2, 8)$. The latent ratio is $k_z = 2$ at inference and $k_z = 4$ during RL rollouts.

I BASELINE INFORMATION

We run baseline design models using their official codebases and recommended default generation hyperparameters, with the settings and adaptations described below.

I.1 DIFFUSION BASELINE

Diffusion and flow-matching baselines learn distributions over molecular structures and transform noise into protein designs. Diffusion models iteratively denoise random configurations, while flow-matching models learn continuous transformations from noise to data. Depending on the model, generation operates on backbone coordinates, all-atom structures, or joint sequence–structure representations. Target structures, binding sites, and functional motifs provide conditioning for target-specific binder design and other structural constraints.

Unless otherwise specified, we redesign protein binder sequences with ProteinMPNN (Dauparas et al., 2022) and ligand binder sequences with LigandMPNN (Dauparas et al., 2025), following most baselines' default protocols. Timing measurements use a structure-generation batch size of 1 and include sequence-redesign model loading and execution. Both MPNN models are generally fast.

Proteína-Complexa (Didi et al., 2026b). Proteína-Complexa uses partially latent flow matching to jointly generate binder sequences and all-atom structures conditioned on a target. It combines generative pre-training on synthetic and experimental complexes with structure-predictor-guided inference-time optimization. We use the setup in Sec. H.5 with self-generated sequences.

RFDiffusion3 (Butcher et al., 2025). RFDiffusion3 (RFD3) is an all-atom diffusion model that generates protein backbone and side-chain coordinates in the context of proteins, nucleic acids, and small molecules. Although RFD3 can co-design sequences, we follow its recommended sequence-redesign protocol. We use 200 diffusion steps and hotspot conditioning for both protein and ligand binder design, with the recommended lower-temperature setup for protein binders.

BoltzGen (Stark et al., 2025). BoltzGen is an all-atom diffusion model that unifies biomolecular structure prediction and protein design by jointly generating amino-acid identities and atomic coordinates. BoltzGen supports self-generated sequences (co-design) and sequence redesign through its included BoltzIF module. We use BoltzIF with default settings. We use 500 diffusion steps, 3 trunk recycles, and hotspot conditioning for both protein and ligand binders, with the supplied non-uniform noise and sampling schedulers. For terminal-disulfide peptides, we fix the terminal cysteines and specify an SG-SG disulfide bond with hotspot conditioning enabled. BoltzIF preserves these cysteines and excludes cysteine at internal designable positions.

PXDesign (Team et al., 2025c). We use PXDesign's diffusion pipeline, PXDesign-d, which builds on the Protenix (Team et al., 2025a) framework for structure-based protein binder design. Its diffusion transformer generates backbone-only atomic coordinates conditioned on target structures and binding-site information. Although the PXDesign paper reports ligand binder results, its official codebase does not support ligand inputs. We run PXDesign only for protein binders, followed by sequence redesign, using 400 diffusion steps with constant low-temperature noise scaling. For fair timing comparisons, we use PyTorch ProteinMPNN instead of JAX to avoid compilation costs.

ODesign (Zhang et al., 2025). ODesign is an all-atom diffusion framework for biomolecular interaction design that represents proteins, nucleic acids, and small molecules within a shared token representation. ODesign generates backbone-only binder structures. Although its included inverse-folding module supports co-design, we follow the recommended sequence-redesign setup. We use 200 diffusion steps with hotspot conditioning for both protein and ligand binders.

Genie3 (Lin et al., 2026). Genie3 is an SE(3)-equivariant diffusion model that represents proteins as branched polymers, modeling specified motifs and binding interfaces at atomic resolution while generating unspecified regions at backbone resolution. Genie3 generates only binder Cα coordinates; we superimpose the all-atom target structure as the initial guess. It supports only protein binder design, for which we use 100 diffusion steps with hotspot conditioning.

I.2 Hallucination Baseline

Hallucination-based methods optimize sequences using objectives derived from pretrained structure predictors, typically rewarding structural confidence, intra-chain contacts, or inter-chain interactions. These baselines backpropagate through a fixed predictor to update continuous sequence representations, seeking predicted structures that meet the folding and interaction requirements. Costly backpropagation through the full predictor typically makes these methods slower.

BindCraft (Pacesa et al., 2025). BindCraft is a protein binder design pipeline that backpropagates through AlphaFold2-Multimer to optimize binder sequences and their predicted interfaces. We use the official repository's default 4-stage configuration and filter, with hotspots specified in Table 9. For consistency across baselines, we replace BindCraft's original sequence-redesign pipeline with a similar setup in our evaluation pipeline.

ESMFold2 (Candido et al., 2026). ESMFold2 is an all-atom structure predictor built on representations from the ESMC protein language model. Its hallucination-based binder design protocol optimizes sequences using predicted distance distributions to promote intra-chain and inter-chain contacts, together with an ESMC sequence prior. Similar to BindCraft, the ESMFold2-hallucination pipeline optimizes the sequence logits using a combination of ESMFold2 confidence scores. We use 150 optimization steps with a base learning rate of 0.1 and cosine annealing. The folding model uses 1, 50, or 200 diffusion steps, depending on the optimization stage. The final sequence's 200-step ESMFold2 prediction serves as the generated structure. This pipeline supports only protein binder design and does not support hotspot conditioning.

BoltzDesign (Cho et al., 2025). BoltzDesign2 supports both Boltz-1 and Boltz-2 for binder design using distogram-derived contact objectives and optional confidence losses. We use Boltz-2 for ligand binder design, with Gumbel initialization at scale 1, learning rate 0.1, and 75 soft, 45 temperature-annealing, and 5 hard optimization iterations. We use 0 recycles during optimization. The helix-loss weight is uniform on $[-0.3, 0]$. We disable semi-greedy refinement and apo prediction.

J Wet Lab Validation

As *in silico* filters do not necessarily align with true binding behavior, we initiate wet-lab experiments to further validate the binding performance of Proteína-Reforza.

J.1 Minibinder Design to PD-L1

Model training and generation. We aim to compare wet-lab binding performance of Proteína-Complexa with Proteína-Reforza under the same generation and experimental budgets. Therefore, we deliberately limited generation to 2,000 designs per model. The experimental budget permitted testing 60 designs from each pool, requiring a selection procedure that prioritized computational quality while retaining structural diversity. Following the sequence filters used in the Proteína-Complexa wet-lab study (Didi et al., 2026a), we require no cysteines, an isoelectric point (pI) of at most 6, and ESM2-650M pseudo-perplexity (PPPL) of at most 10. Optimizing folding confidence alone does not directly encourage these properties. We therefore added a sequence-quality reward to encourage eligible designs within the limited sampling budget.

We initialized a preliminary version of Proteína-Reforza from the step 70 checkpoint of general-PDB training with the folding reward in Eq. (3), then performed 30 additional RL steps on a 2-target population including PD-L1. Adaptation used a fresh optimizer, the initialized policy as the frozen KL reference, and equal weights for folding and sequence quality:

$$R_i = 0.5R_i^{\text{succ}} + 0.5\left[1 - \frac{1}{3}\left(\frac{[\text{pI}_i - 6]_+}{3} + \frac{[\text{PPPL}_i - 10]_+}{5} + \frac{N_{\text{Cys},i}}{2}\right)\right]. \tag{11}$$

Here N_{Cys} is the binder cysteine count. The sequence term penalizes violations of the experimental sequence gates without rewarding further reductions in pI or PPPL once their thresholds are met. Training used AWM with the AdamW optimizer, a learning rate of 10^{-5}, and a KL coefficient of 0.02. We used the policy after these 30 RL steps for generation.

We used the same PD-L1 target structure, residue range, and hotspots as in the computational benchmark (Tab. 9). Each model generated 2,000 designs with binder lengths sampled uniformly

Table 20: **PD-L1 selection thresholds.** Strict selection uses all criteria; relaxed selection uses all folding and sequence criteria and 5 of 6 interface blocks. H-bond count and energy form 1 block; energies use TMol units.

Family	Metric	Strict	Relaxed
Folding	ipAE (Å)	≤ 7	≤ 10
	Complex pLDDT	≥ 0.90	≥ 0.80
	Binder scRMSD-Cα (Å)	< 1.5	< 2.0
Sequence	Cysteine count	$= 0$	$= 0$
	Isoelectric point	≤ 6	≤ 6
	ESM2 pseudo-perplexity	≤ 10	≤ 10
Interface	Binder surface hydrophobicity	< 0.35	< 0.40
	Shape complementarity	> 0.60	> 0.45
	TMol interface H-bond count	≥ 1	≥ 1
	TMol interface H-bond energy	< -1.0	< -0.5
	TMol interface electrostatic energy	< 0	< 0
	Interface lysine + methionine count	< 3	< 3
	Interface-delta unsatisfied polar atoms	< 4	< 5

from 50–120 residues, using 100 diffusion steps and self-conditioning. Designs were evaluated with AF2-Multimer using the folding metrics in Sec. D. We computed binder sequence metrics, including ESM2-650M PPPL by masking 1 residue at a time, and evaluated interface metrics after refolding.

Filtering and Selection. Tab. 20 lists the selection thresholds. Interface lysine + methionine count is the total number of these residues at the binder interface. Interface-delta unsatisfied polar atoms counts polar heavy atoms that become buried and hydrogen-bond-unsatisfied upon complex formation.

We used strict criteria to prioritize designs with favorable folding, sequence, and interface properties. Since a small pool may contain fewer than 60 strict candidates, relaxed criteria provided a defined fallback for filling the experimental quota. Strict candidates passed all 13 criteria in Tab. 20. Relaxed candidates passed all relaxed folding and sequence criteria and at least 5 of 6 interface criteria, counting H-bond number and energy jointly as 1 criterion. Both models followed the same rule: retain strict candidates first, and use relaxed-only candidates only when needed. The sequence requirements were unchanged under relaxation. Fig. 34 shows that RL shifted the generated pool toward lower ipAE, higher complex pLDDT, and higher shape complementarity. Consequently, Proteína-Complexa yielded 9 strict and 139 relaxed candidates, compared with 93 and 1,325 for preliminary Proteína-Reforza. We also note that the improvement was not uniform across every metric; for example, the RL pool did not have a lower median PPPL.

Before sampling, we deduplicated exact binder sequences and clustered eligible candidates with Foldseek at a TM-score threshold of 0.5. For Proteína-Complexa, we retained all 9 strict candidates and filled 51 slots from relaxed candidates excluding the strict set. We randomly ordered clusters not yet represented, selected 1 random sequence per new cluster, and filled remaining slots round-robin across clusters. For preliminary Proteína-Reforza, the same cluster-stratified procedure selected 60 of the 93 strict candidates, without relaxed-only additions. Clustering used the relaxed pool for Proteína-Complexa and the strict pool for Proteína-Reforza. Each final panel contained 60 distinct sequences, representing 51 and 28 clusters within the respective pools. This procedure prioritized filter quality and cluster coverage rather than ranking designs by a single metric; it used no manual choices or experimental outcomes.

The quality difference persisted after selecting 60 candidates from each small pool (Fig. 35). Relative to Proteína-Complexa, preliminary Proteína-Reforza had lower median ipAE (4.26 versus 6.79 Å), higher complex pLDDT (0.941 versus 0.906), and higher shape complementarity (0.63 versus 0.56). Selected-panel median PPPL was similar (7.24 versus 7.25). Thus, selection from the limited base-model pool did not eliminate the RL advantage in these folding and interface metrics.

Wet-lab protocol and results. Surface plasmon resonance (SPR) measures binding by monitoring changes in the refractive index near a sensor surface bearing one interaction partner. Solutions containing the other partner at a series of concentrations are passed over the surface, producing response curves during association and subsequent dissociation in buffer. These curves characterize the concentration dependence and reversibility of binding. Fitting suitable response curves with

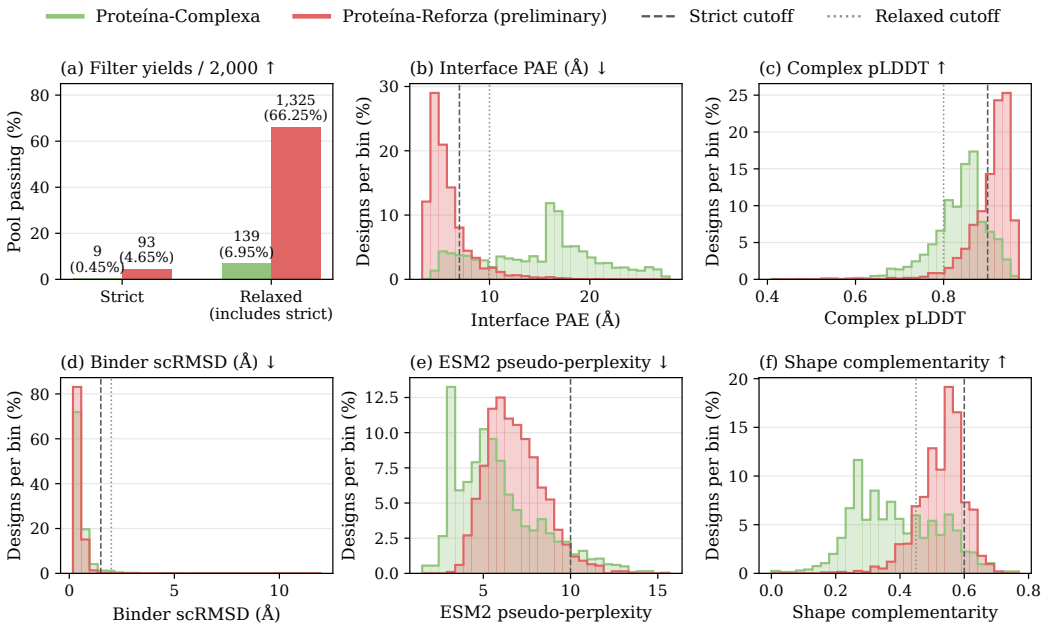


Figure 34: **PD-L1 candidate pools.** Each model generated 2,000 designs. (a) Strict and relaxed filter yields; relaxed includes strict. (b–f) Metric distributions. Arrows indicate the favorable direction (↑: higher; ↓: lower). Dashed and dotted lines mark strict and relaxed cutoffs; coincident cutoffs are shown once. Proteína-Reforza denotes the preliminary version evaluated in this experiment.

a binding model provides estimates of the apparent dissociation constant, K_D, with lower values indicating tighter binding.

Proteína-Reforza produced 7 binding-positive designs among 60 tested (11.7%), including 3 weak hits, whereas Proteína-Complexa produced none among 60. 4 RL designs had reported apparent affinities below 1 μM, including 1 with poor fit quality. The best binder had an overall apparent K_D of approximately 65 nM, with 2 SPR measurements giving 59.6 and 69.5 nM; its measured response curves are shown in Fig. 7. Fig. 36 shows the 6 other hits. These results support improved experimental hit discovery after RL under equal generation and testing budgets. Because the selection rule yielded different strict/relaxed compositions, the comparison measures the full generation-and-selection procedure rather than binding rates at matched computational quality.

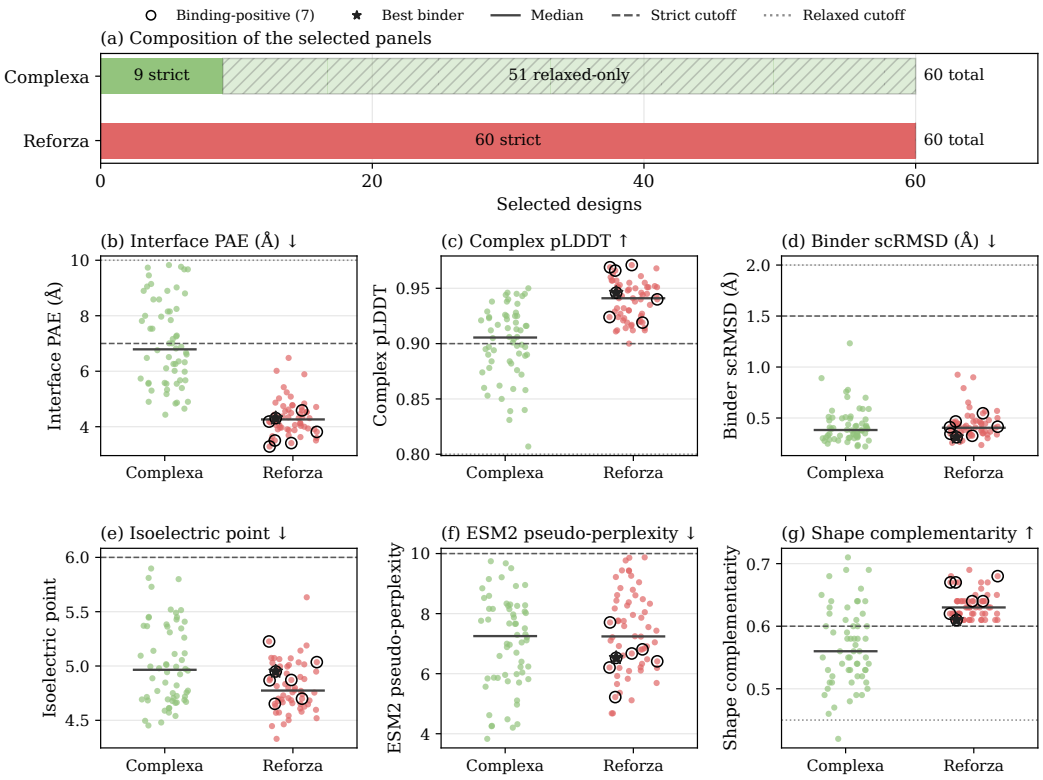


Figure 35: **Selected PD-L1 designs.** (a) Strict (solid) and relaxed-only (hatched) selections. (b–g) All 60 designs per model, with medians shown as short horizontal lines. Rings mark the 7 experimental hits, including 3 weak hits; the star marks the best binder by reported apparent affinity (64.7 nM). Arrows indicate the direction favored by the selection criteria (↑: higher; ↓: lower); for pI, this denotes the experimental upper-bound requirement. Dashed and dotted lines indicate strict and relaxed cutoffs. Experimental highlights were added after selection. Complexa denotes Proteína-Complexa; Reforza denotes preliminary Proteína-Reforza.

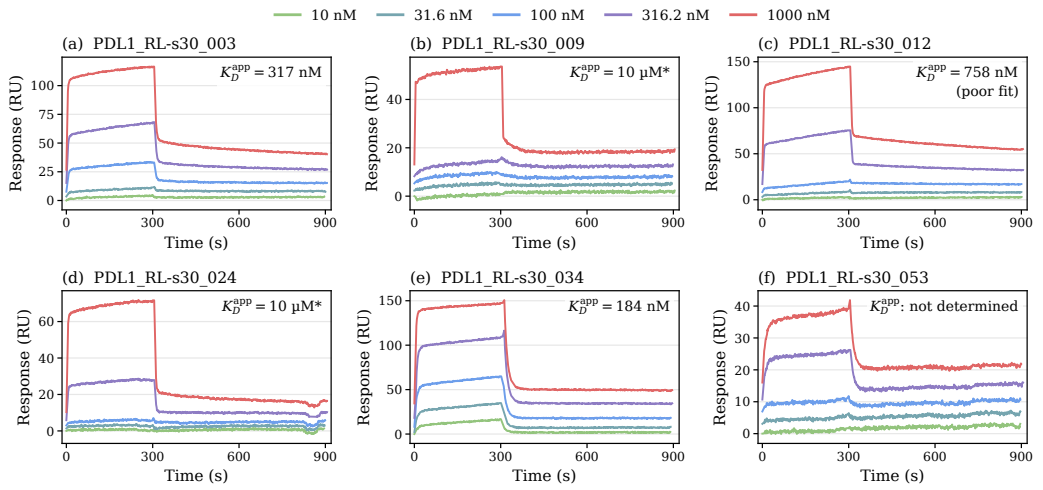


Figure 36: **SPR responses of 6 additional PD-L1 binders.** Panels show representative measurements (SPR replicate 1) at 5 concentrations. Annotations report overall apparent affinities from the experimental summary. Asterisks mark values at the 10 μM fitting limit. Binder 012's fit was rated poor because its RMSE normalized by the maximum response was 5.7%, exceeding the 5% cutoff. Binder 053 showed a weak binding signal, but its K_D could not be determined. Response-axis ranges differ across panels.

J.2 Cyclic Peptide Design to PD-L1 and GABARAP and TROP2

The peptide experiments in Sec. 5 evaluate the CPSea-derived pair: Proteína-Complexa$_{\text{CPSea Disulfide SFT}}$ and Proteína-Reforza$_{\text{CPSea Disulfide}}$ (Secs. B.3 and H.3). These are the models denoted Disulfide SFT and Disulfide in the main-text wet-lab comparison, and SFT and RL in the experimental figures. The computational benchmark (Tab. 3) uses TED-derived models.

Table 21: **Target structures and prepared residues for cyclic peptide design.**

Target	Template PDB or prepared input	Chain and residues
GABARAP	3WIM	A1–117
PD-L1	BindCraft AF2 model	A1–115
TROP2	7PEE	A32–268

Table 22: **Hotspot sets and design counts.** Residues use prepared chain-A numbering.

Target	Set	Hotspot residues	Selected	
			Proteína-Reforza CPSea Disulfide	Proteína-Complexa CPSea Disulfide SFT
GABARAP	H1	46, 48, 49, 50, 60, 63	2,850	3,004
GABARAP	H2	24, 25, 28, 46, 48, 50, 52, 54	3,236	4,474
GABARAP	H3	25, 28, 49, 50, 52	3,914	2,522
PD-L1	H1	39, 98, 106	3,149	3,413
PD-L1	H2	37, 39, 49, 98	3,663	3,669
PD-L1	H3	37, 39, 49, 98, 106	3,188	2,918
TROP2	H1	153, 155, 204, 206, 259, 261	6,236	7,440
TROP2	None	No hotspots	3,764	2,560

Generation and filtering. We selected 10,000 designs per model for each of the three targets, giving 60,000 designs in total. Peptides have 12–18 residues and exactly two terminal cysteines. Target structures, hotspot sets, and design counts are listed in Tabs. 21 and 22.

For design selection, we stratified designs into different tiers using *in silico* filters. Candidates were assigned sequentially to mutually exclusive tiers:

1. **Strict.** Interface $\Delta\Delta G < 0$, a valid disulfide with raw energy < 4, and Protenix minimum inter-chain pAE < 1.5 Å.
2. **Relaxed 1.** Interface $\Delta\Delta G < 0$, a valid disulfide with raw energy < 5, and Protenix minimum inter-chain pAE < 2.0 Å.
3. **Relaxed 2.** Interface $\Delta\Delta G < 0$, a valid disulfide with raw energy < 6, and Protenix minimum inter-chain pAE < 2.5 Å.
4. **Any two.** Pass at least two of the three strict-tier gates after excluding tiers 1-3.
5. **Backfill.** Fill remaining slots by consensus rank from AF2-passed candidates not assigned to tiers 1-4.
6. **AF2-relaxed.** AF2 interface pAE < 10 Å, binder self-consistency RMSD < 2 Å, and binder pLDDT > 0.70.

Tiers 1-5 required AF2 initial-guess interface pAE < 7 Å, binder pLDDT ≥ 0.80, and binder self-consistency RMSD < 1.5 Å. All tiers required exactly two terminal cysteines. Panel counts by tier are shown in Fig. 38.

Experimental readout and analysis. For each target, we jointly ranked measured SFT and RL candidate records by target-specific read count and selected the top 50, excluding positive controls. The positive controls comprised two GABARAP peptides ($K_d = 5$ and 9 nM), one PD-L1 peptide ($K_d \approx 67$ nM), and one TROP2 peptide ($K_d \approx 38$ nM). For visualization, we normalized read counts by the maximum measured count across SFT designs, RL designs, and positive controls for each target (Fig. 37). This scaling preserves within-target rankings and count ratios.

The top 50 included 25, 11, and 26 strict-tier records for GABARAP, PD-L1, and TROP2, respectively (Fig. 39), with the remaining records drawn from other selection tiers.

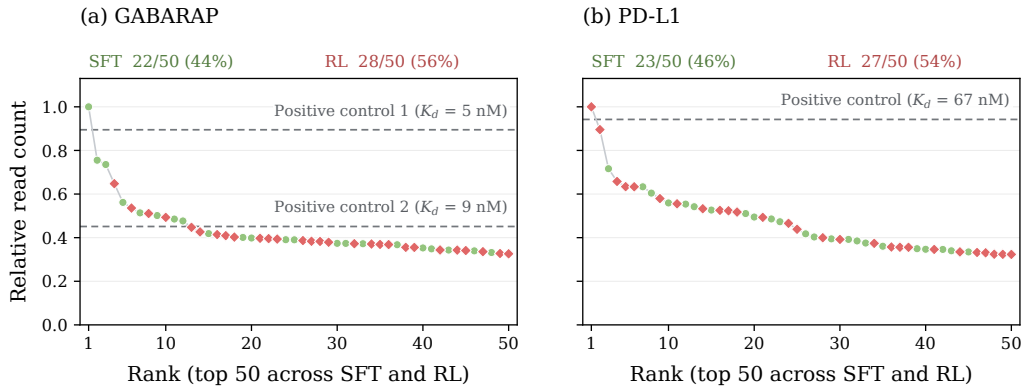


Figure 37: **Cyclic-peptide screening against GABARAP and PD-L1.** Top 50 candidate records per target, pooled across SFT and RL and ranked by read count. Each count is divided by the maximum measured count for that target, including controls. Green circles denote SFT and red diamonds denote RL. Dashed lines show positive controls.



Figure 38: **Selected library composition by target, model, and computational tier.** Each target-model group contains 10,000 candidates.

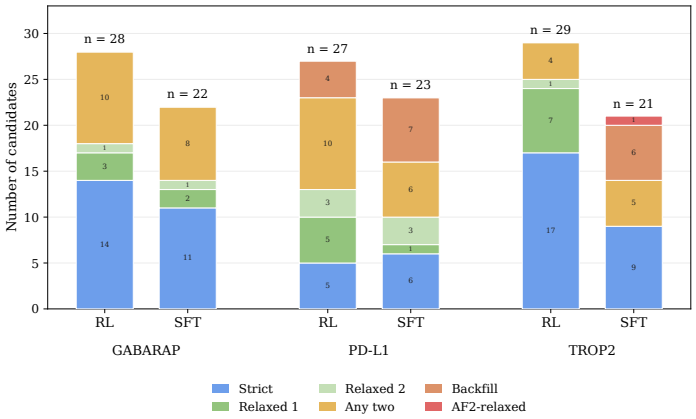


Figure 39: **Computational selection tiers among the combined top 50 candidates.**