# PROTEOCRAFT: PROTEIN DESIGN AS PROPOSAL–VERIFICATION SEARCH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Protein–protein interactions underpin nearly all biological processes. Yet, designing proteins with desired interactions remains challenging due to the structural complexity governing molecular recognition. This study views protein design as a search optimization problem over a combinatorial sequence space under expensive structural evaluation. Toward this goal, we introduce a novel tree-based search framework, ProteoCraft, that employs approximate structural gradients from diffusion-based structure predictors as effective proposal policies to specify where the search should explore. Complete no-gradient inference of folding models separately assigns values to the resulting sequences, determining what is good. ProteoCraft allocates this verification budget through a coarse-to-fine paradigm: multi-site edits first explore distinct sequence basins, then verified promising states undergo strict single-site refinement. Across fourteen de novo binder-design settings spanning eight target proteins, five immune-relevant TCR-facing targets, and four localized CDR redesign tasks, ProteoCraft improves design quality and high-confidence yield. These improvements persist under independent AF2-IG and Rosetta-based interface evaluation.

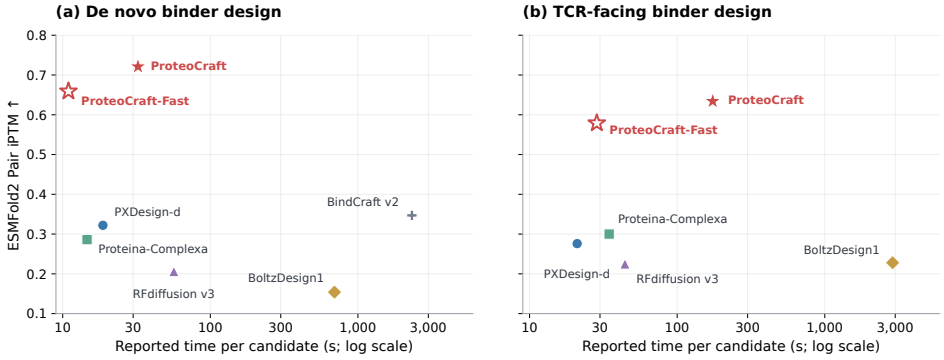



Figure 1: Design quality versus computational cost.

## 1 INTRODUCTION

Protein–protein interactions (PPIs) underlie many cellular processes, and the ability to design proteins that selectively engage a target has broad applications in therapeutics, diagnostics, and synthetic biology (Zambaldi et al., 2024; Cho et al., 2025b). Computational protein design offers a powerful alternative by enabling binders to be tailored to specific targets and epitopes. Early physics-based approaches, such as Rosetta (Cao et al., 2022; Gainza et al., 2023), designed binders through scaffold docking and side-chain optimization. However, these methods typically have low experimental success rates and therefore require screening a large number of candidates (Bennett et al., 2023). Their reliance on docking predefined scaffolds onto rigid target structures can also produce geometric or chemical incompatibilities at the interface, resulting in suboptimal binding or making certain epitopes difficult to target.

Recent advances in deep learning (DL) have transformed biomolecular modeling and design. AlphaFold2/3 (AF2/3) (Jumper et al., 2021; Evans et al., 2021; Abramson et al., 2024), for example, demonstrated remarkable accuracy in predicting monomer and complex structures. AF-based filtering substantially improves binder-design success rates by identifying candidates predicted to form



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.

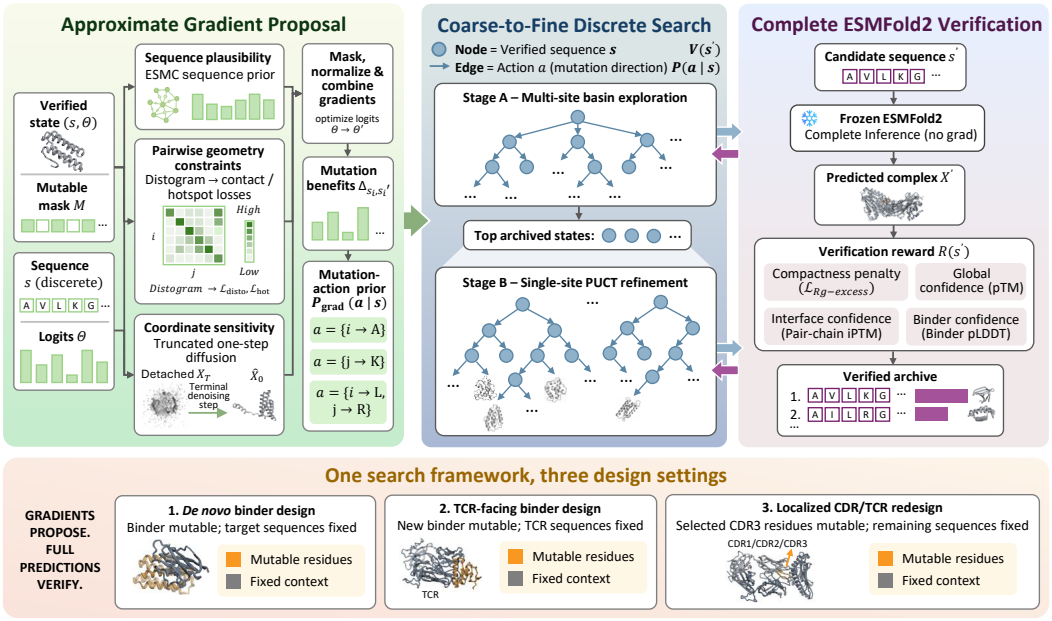



Figure 2: **Overview.** ProteoCraft formulates protein design as proposal–verification search. Given a mutable design region and molecular context, a coarse-to-fine tree search uses truncated diffusion signals to propose candidate sequences, while complete ESMFold2 inference assigns node values.

plausible complexes with their targets (Liu et al., 2026; Team et al., 2025). DL has also enabled the de novo generation of protein binders. A widely used pipeline combines RFdiffusion for backbone generation (Watson et al., 2023) with ProteinMPNN (Dauparas et al., 2022) for sequence design, achieving markedly higher experimental success rates. Nevertheless, it separates backbone generation from sequence design without explicitly modeling full atomic interactions, leaving AF filtering to bridge the gap between structural generation and functional interface design. Building on the success of AF-based filtering, another hallucination-based line, such as BindCraft (Pacesa et al., 2024), demonstrated that structure predictors can also serve as differentiable design engines. By backpropagating through AF2, they directly optimize binder sequences, producing high-affinity designs with substantially reduced experimental screening. This marked a shift from generate-then-filter pipelines toward prediction-guided optimization. However, direct gradient optimization can be myopic and exploit imperfections in the predictive model, motivating search-based approaches that combine gradient-guided proposals with a balanced exploration-exploitation trade-off.

In this paper, we formulate protein design as a combinatorial search problem, where nodes represent discrete sequences, edits define transitions between them, and structure predictors serve as verification oracles. Our central idea is to use gradient-based hallucination as a source of search directions, while using structural verification to decide which sequences to explore further. To this end, we introduce ProteoCraft, a tree-based search framework based on ESMFold2 (Candido et al., 2026). Specifically, gradients through the terminal diffusion step induce priors over residue substitutions, which guide a coarse-to-fine search from coordinated multi-site exploration to single-site refinement. Verification therefore becomes part of the search process rather than only a final filter. Each proposed sequence is independently evaluated by a complete ESMFold2 forward pass, allowing the search to combine gradient-informed mutation preferences with evidence accumulated from previously verified designs.

We evaluate ProteoCraft on de novo binder design, TCR-facing binder design, and localized CDR redesign. Across all settings, ProteoCraft achieves better or competitive average design quality and the yield of high-confidence candidates over baselines and compute-matched variants of our pipeline. These improvements persist under independent AF2-IG evaluation and Rosetta interface-scoring evaluation, suggesting that the gains extend beyond the search-time verifier. Ablations demonstrate complementary contributions from truncated diffusion guidance and tree search. Finally, wet-lab validation confirms binding of selected designs to IFNAR2 and PD-L1, with BLI-measured dissociation constants of 153 nM and 17.1 nM, respectively, without experimental affinity maturation.

# 2 METHOD

## 2.1 PRELIMINARY

**Problem Statement.** Given a target receptor $\mathcal{T}$, we aim to design a binder sequence $\mathbf{s} = (s_1, \ldots, s_L)$. A binary mask $\mathbf{M} \in \{0,1\}^L$ specifies the mutable positions, where $M_i = 0$ indicates a fixed location. The binder identity is parameterized by amino-acid logits $\mathbf{\Theta} \in \mathbb{R}^{L\times 20}$, with continuous probabilities $P_{\mathbf{\Theta}}(i,s) = \mathrm{Softmax}(\mathbf{\Theta}_i/T)_s$ and discrete type $s_i = \arg\max_s \mathbf{\Theta}_{i,s}$, where $T > 0$ is an annealed temperature and $P_{\mathbf{\Theta}}$ exposes differentiable sequence conditioning.

**Tree-based Search for Protein Discovery.** ProteoCraft adopts a modified MCTS framework (Wu et al., 2026) to build a search tree for *de novo* binder design. It regards the protein design process as a finite-horizon Markov Decision Process (MDP), represented by the tuple $\langle S, A, P, R, \gamma\rangle$. Specifically, $S$ is the set of states that contains the entire binder space. $A$ represents possible actions, where each action $a$ corresponds to a different mutation direction. $P(\mathbf{s}' \mid \mathbf{s}, a)$ denotes the transition probability of reaching state $\mathbf{s}'$ from $\mathbf{s}$ after action $a$, $R(\mathbf{s})$ is the reward function to evaluate the quality of the new mutant sequences $\mathbf{s}$, and we use undiscounted reward backup, corresponding to $\gamma = 1$. The root node $\mathbf{s}^0$ represents the wild-type or random binder sequence, and each child node $\mathbf{s}^i$ corresponds to a mutant variant, for which we assign a q-value $q(\mathbf{s}^i)$. Then, a root-to-leaf path ending at a terminal node $\mathbf{s}^{\text{end}}$ forms a trajectory $\mathbf{t} = \mathbf{s}^0 \oplus \mathbf{s}^1 \oplus \mathbf{s}^2 \oplus \ldots \oplus \mathbf{s}^{\text{end}}$, which displays an artificial evolutionary pathway from $\mathbf{s}^0$.

## 2.2 EXPANSION: FROM STRUCTURAL GRADIENTS TO MUTATION PROPOSALS

ProteoCraft iteration consists of four phases: *selection*, *expansion*, *simulation*, and *backpropagation*. A search node stores a verified discrete sequence together with its inherited logits, predicted structure, verification metrics, and search statistics. For a selected verified node $\mathbf{s}^i$, proposal optimization starts from its inherited logits. We pass the corresponding probability matrix $P_{\mathbf{\Theta}}$ through the ESMFold2 `res_type_soft` interface and combine gradients for pairwise geometry, sequence plausibility, and coordinate-level structural sensitivity. Only mutable positions receive logit updates.

We consider three categories of mutation optimization objectives. The foremost type is the folding model distogram prior, formed as $\mathcal{L}_{\text{disto}}(\mathbf{s}^i) = 0.5\mathcal{L}_{\text{intra}} + 0.5\mathcal{L}_{\text{inter}} + 0.2\mathcal{L}_{\text{glob}}$, where $\mathcal{L}_{\text{intra}}$ promotes coherent binder-internal contacts, $\mathcal{L}_{\text{inter}}$ promotes target–binder contacts, and $\mathcal{L}_{\text{glob}}$ regularizes global compactness. Secondly, we use an ESMC masked language-model loss $\mathcal{L}_{\text{ESMC}}$ (Candido et al., 2026). We mask sampled mutable positions and compute cross-entropy against the current soft amino-acid probabilities at those positions. This loss regularizes sequence plausibility during proposal optimization.

Though distogram gradients from $\mathcal{L}_{\text{disto}}(\mathbf{s}^i)$ capture pairwise geometric preferences, they fail to directly expose sensitivity through the coordinate-generating process. To bridge this gap, we differentiate through a single terminal diffusion step to acquire the extra coordinate-sensitive mutation signal. Let $\mathbf{X}_p$ denote the atom coordinates of the modeled complex at noise level $p$, and let $f_\phi^{p\to q}$ denote the reverse-diffusion trajectory from noise level $p$ to $q$. The model parameters $\phi$ remain frozen. Starting from $\mathbf{X}_{\sigma_{\max}} = \sigma_{\max}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we run the trajectory without gradient tracking until noise level $\tau$, then apply one differentiable denoising operation:

$$\begin{aligned} \mathbf{X}_\tau &= \mathrm{sg}\Big[f_\phi^{\sigma_{\max}\to\tau}(\mathbf{X}_{\sigma_{\max}}; \mathbf{c}_{\mathbf{\Theta}})\Big], \\ \hat{\mathbf{X}}_0 &= D_\phi(\mathbf{X}_\tau, \tau; \mathbf{c}_{\mathbf{\Theta}}). \end{aligned} \tag{1}$$

Here, $\mathbf{c}_{\mathbf{\Theta}}$ is the sequence-conditioned representation, sg denotes stop-gradient, and $D_\phi$ predicts clean coordinates at noise level $\tau$. Gradients are retained through the terminal denoising operation and the current sequence conditioning, while the preceding reverse-diffusion trajectory is detached. Here, we define $\mathcal{L}_{1\,\text{step}}\left(\hat{\mathbf{X}}_0\right) = 0.1\mathcal{L}_{\text{pLDDT}} + 0.4\mathcal{L}_{\text{PAE}} + 0.2\mathcal{L}_{\text{iPAE}} + 0.3\mathcal{L}_{\text{pair-chain iPTM}} + 0.1\mathcal{L}_{\text{Rg}} + 0.1\mathcal{L}_{\text{clash}} + 0.1\mathcal{L}_{\text{i-clash}}$, where lower values are preferred for all terms. For $\mu \in \{\text{disto}, \text{ESMC}, 1\,\text{step}\}$, let $\mathbf{g}_\mu$ denote $\nabla_{\mathbf{\Theta}}\mathcal{L}_\mu$ after zeroing the rows corresponding to fixed positions. We normalize each masked gradient separately and obtain the combined proposal gradient $\mathbf{g} = \mathrm{Norm}(\mathbf{g}_{\text{disto}}) + \lambda_{\text{ESMC}}\,\mathrm{Norm}(\mathbf{g}_{\text{ESMC}}) + \lambda_{1\,\text{step}}\,\mathrm{Norm}(\mathbf{g}_{1\,\text{step}})$. The normalization and update schedule are specified in App. B.

Starting from the inherited logits, we perform multiple updates of the form $\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \eta \mathbf{g}$. Let $\mathbf{\Theta}$ and $\mathbf{\Theta}'$ denote the logits before and after this proposal-optimization phase, respectively. In Stage A, the raw argmax sequence $\mathbf{s}' = \arg\max(\mathbf{\Theta}')$ is retained as a candidate and must undergo complete verification before being considered as a final design. We also use the changes in logits to construct discrete mutation proposals. For a substitution $s_i \to s_i'$ at position $i$, we define

$$\mathbf{\Delta}_{s_i,s_i'} = \left(\mathbf{\Theta}'_{i,s_i'} - \mathbf{\Theta}_{i,s_i'}\right) - \left(\mathbf{\Theta}'_{i,s_i} - \mathbf{\Theta}_{i,s_i}\right). \tag{2}$$

A positive $\Delta_{s_i,s_i'}$ indicates that the combined hallucination update increasingly favors residue $s_i'$ over retaining $s_i$. For an admissible single-site action $a = (i, s_i')$, where $M_i = 1$ and $s_i' \neq s_i$, we write $\Delta_a = \Delta_{s_i,s_i'}$. The gradient-derived proposal prior is

$$P_{\text{grad}}(a \mid \mathbf{s}) = \frac{\exp(\Delta_a)}{\sum_{a' \in \mathcal{A}(\mathbf{s})} \exp(\Delta_{a'})},$$

where $\mathcal{A}(\mathbf{s})$ contains the admissible single-site mutation actions from $\mathbf{s}$. Multi-site candidates are constructed from mutation benefits as described in Section 2.4. Applying a proposed mutation action $a$ to $\mathbf{s}$ produces a discrete candidate $\mathbf{s}'$. Candidates undergo complete ESMFold2 verification before selected candidates are inserted as children of $\mathbf{s}$. Each inserted child inherits the optimized logits $\mathbf{\Theta}'$ from the proposal that generated it, rather than the logits before optimization. When necessary, these inherited logits are minimally aligned so that their argmax matches the verified child sequence.

### 2.3 Evaluation and Backpropagation with Full-model Verifier

Each discrete candidate is evaluated by complete frozen ESMFold2 inference without gradient tracking. The baseline verifier loss is $\mathcal{L}_{\text{verify}}(\mathbf{s}) = (1 - \text{pair-chain iPTM}) + 0.2(1 - \text{binder pLDDT}) + 0.1(1 - \text{pTM}) + 0.1\mathcal{L}_{\text{Rg-excess}}$, where lower values are preferred. Pair-chain iPTM is computed between the designated target chains and the binder, while $\mathcal{L}_{\text{Rg-excess}}$ penalizes excessive binder radius of gyration relative to a length-dependent threshold. Additional structural and hotspot statistics are recorded separately. In Stage A, the verification reward is $R(\mathbf{s}) = -\mathcal{L}_{\text{verify}}(\mathbf{s})$.

All successful full-model evaluations are retained in an archive, while only selected candidates are inserted into the tree. Rewards from these selected candidates are backed up along the selected search path without depth discounting. The corresponding visit counts and empirical mean values are updated to guide which sequence branches receive further exploration. These updates do not propagate verifier gradients into proposal optimization. Implementation details are in App. B.

### 2.4 Coarse-to-Fine Multi-stage Selection Strategy

Starting from the root $\mathbf{s}^0$, we progressively adapt the granularity of sequence editing through two regimes: *multi-site exploration* and *single-site refinement*. Early in search, coordinated multi-site mutations provide a larger effective search radius, enabling efficient transitions across distinct sequence basins. After the allocated Stage-A budget is exhausted, the search switches to single-site substitutions for local refinement around selected verified sequences.

**Stage A: multi-site basin exploration** begins from multiple independently initialized sequence–logit states. At each iteration, we first allocate search budget across the initialized roots and then select a verified node within the chosen tree using the Upper Confidence Bound (UCB) criterion (Kocsis & Szepesvári, 2006):

$$U(\mathbf{s}^i) = q(\mathbf{s}^i) + c_{\text{ucb}} \sqrt{\frac{\log N_{\text{total}}}{N(\mathbf{s}^i) + \epsilon}}, \tag{3}$$

where $q(\mathbf{s}^i)$ is the mean verifier reward backed up through node $\mathbf{s}^i$, $N(\mathbf{s}^i)$ is its visit count, and $N_{\text{total}}$ is the total number of visits. Since rewards are defined as negative verifier losses in Stage A, larger $q(\mathbf{s}^i)$ indicates better verified outcomes. The UCB score balances exploitation of promising states with exploration of less-visited regions. After selecting a node, we optimize its inherited logits and construct multi-site candidates from the mutation benefits in Eq. 2. For these benefit-based candidates, the maximum number of edited residues is $K_{\max} = \lceil \rho \sum_i M_i \rceil$. The raw argmax proposal is retained separately and is not constrained by $K_{\max}$. Each candidate undergoes complete verification before selected candidates are inserted into the tree.

**Stage B: single-site refinement.** After Stage A, selected verified sequence–logit states are re-evaluated with a common verifier seed to initialize Stage-B roots. Each newly proposed candidate differs from its parent at a single residue. Stage B uses a balanced reward derived from full-verifier metrics, with additional penalties for unmet hotspot constraints, when specified, and interfaces below a quality floor. This reward guides which sequence branches are explored further. Let $P(a \mid \mathbf{s}^i)$ denote the action prior used by Stage-B search, $q(\mathbf{s}^i, a)$ the mean Stage-B reward backed up through action $a$, and $N(\mathbf{s}^i, a)$ the corresponding backup count. PUCT guides traversal of existing mutation branches (Rosin, 2011):

$$U_{\mathrm{PUCT}}(\mathbf{s}^i, a) = q(\mathbf{s}^i, a) + c_{\mathrm{puct}} P(a \mid \mathbf{s}^i) \frac{\sqrt{N(\mathbf{s}^i)}}{1 + N(\mathbf{s}^i, a)}. \tag{4}$$

At the selected expansion state, we optimize the inherited logits and construct single-site mutation candidates. These candidates undergo complete verification before selected candidates are added to the tree. Their rewards guide subsequent selection between mutation branches. Thus, single-site refinement restricts the number of changed residues per candidate, not the number of candidates evaluated per iteration.

### 2.5 Constrained and Localized Design

ProteoCraft supports different design settings by modifying $\mathbf{M}$ and the interface terms in the optimization objective $\mathcal{L}_{\mathrm{disto}}(\mathbf{s})$. For *de novo* and TCR-facing binder design, all binder residues are mutable while target sequences remain fixed and provide structural context. For localized CDR redesign, the remaining framework sequence participates in structure prediction but is not updated. The framework sequence remains fixed, but its coordinates are not explicitly clamped.

When target hotspots $\mathcal{H} \subseteq \mathcal{T}$ are specified, we define the hotspot constraint loss as $\mathcal{L}_{\mathrm{hot}} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \min_{j \in \mathcal{B}} \ell_{ij}^{\mathrm{contact}}(d_{\mathrm{hot}})$, where $\mathcal{B}$ denotes the set of binder residue indices, $\ell_{ij}^{\mathrm{contact}}$ is the pairwise contact loss, and $d_{\mathrm{hot}}$ is the predefined contact cutoff distance. Then, we replace the standard inter-chain loss term with $\mathcal{L}'_{\mathrm{inter}} = (1 - \lambda_{\mathrm{hot}})\mathcal{L}_{\mathrm{inter}} + \lambda_{\mathrm{hot}}\mathcal{L}_{\mathrm{hot}}$ to encourage the interaction between hotspot residues and the candidate binder. Via this mechanism, hotspot information therefore biases the mutation direction toward a desired target region $\mathcal{H}$.

## 3 Experiments

### 3.1 Experimental Setup

**Design tasks.** We consider three design settings. (1) *De novo* binder design, where the complete binder sequence is generated against a fixed target. Our main benchmark comprises 14 target–hotspot settings across eight target proteins, spanning cellular receptors, allergens, and other structurally diverse proteins. For each method, we generate 15 candidates per setting using three binder lengths and five random seeds, yielding $N = 210$ candidates per method. (2) Immune-relevant TCR-facing binder design, where a new binder is designed against a fixed TCR target. (3) Localized CDR/TCR redesign, where only a designated CDR region is mutable while the remaining receptor sequence provides fixed structural context. Detailed targets, sequence lengths, hotspot annotations, and generation settings are provided in Appx. C.

**Baselines.** We compare against multiple state-of-the-art algorithms, containing PXDesign-d (Team et al., 2025), RFdiffusion v3 (Butcher et al., 2025), Proteina-Complexa (Didi et al., 2026), BoltzGen (Stark et al., 2025), BindCraft (Pacesa et al., 2025), BindCraft v2 (Pacesa Lab, 2026), and BoltzDesign1 (Cho et al., 2025a). We further formulate three matched variants of our own pipeline: *No One-step Diffusion* removes the truncated One-step diffusion gradient while retaining the remaining proposal–verification procedure; *ProteoCraft-Best-of-$N$* uses the same differentiable proposal mechanism and matched candidate budget but selects among independent proposals without tree-search control; and *ProteoCraft-MCTS* uses the full coarse-to-fine tree search. These comparisons separate the contribution of diffusion-based proposal guidance from the contribution of verifier-guided search. All comparisons use matched target inputs and evaluation protocols whenever supported by the corresponding baseline. Notably, for BindCraft v2 on TCR-related binder design, no candidates passed its built-in filters within our sampling budget.

Table 1: **ESMFold2 evaluation of *de novo* binder design on 14 target–hotspot settings across eight target proteins.** Each method contributes 15 candidates per hotspot ($N = 210$). The left block reports mean confidence metrics (higher is better); the right block reports high-confidence design yields as count (percentage). The strict criterion requires both Pair iPTM $\geq 0.75$ and Binder pLDDT $\geq 0.9$. Dashes indicate results not yet reported.

| Method | **Mean confidence ↑** | | | | | **High-confidence yield ↑** | | |
|---|---|---|---|---|---|---|---|---|
| | Pair iPTM | iPTM | pTM | pLDDT | Binder pLDDT | Pair iPTM $\geq 0.5$ | Pair iPTM $\geq 0.7$ | Pair iPTM $\geq 0.75$ Binder pLDDT $\geq 0.9$ |
| RFdiffusion v3 | 0.206 | 0.282 | 0.593 | 0.739 | 0.729 | 20 (9.5%) | 1 (0.5%) | 0 (0.0%) |
| BoltzGen | 0.438 | 0.577 | 0.748 | 0.826 | 0.857 | 105 (50.0%) | 45 (21.4%) | 9 (4.3%) |
| Proteina-Complexa | 0.286 | 0.386 | 0.653 | 0.774 | 0.779 | 48 (22.9%) | 4 (1.9%) | 0 (0.0%) |
| PXDesign-d | 0.322 | 0.433 | 0.683 | 0.801 | 0.832 | 58 (27.6%) | 16 (7.6%) | 1 (0.5%) |
| BoltzDesign1 | 0.154 | 0.220 | 0.562 | 0.638 | 0.490 | 6 (2.9%) | 0 (0.0%) | 0 (0.0%) |
| BindCraft | 0.573 | 0.779 | 0.822 | 0.833 | 0.858 | 162 (77.1%) | 53 (25.2%) | 11 (5.2%) |
| BindCraft v2 | 0.347 | 0.470 | 0.691 | 0.766 | 0.740 | 70 (33.3%) | 7 (3.3%) | 1 (0.5%) |
| No One-step Diffusion | 0.637 | 0.861 | 0.869 | 0.859 | 0.879 | 176 (83.8%) | 85 (40.5%) | 38 (18.1%) |
| ProteoCraft-Best-of-$N$ | 0.660 | 0.885 | 0.891 | 0.866 | 0.880 | 186 (88.6%) | 118 (56.2%) | 40 (19.0%) |
| ProteoCraft-Stage-A-only | 0.691 | 0.906 | 0.895 | 0.876 | 0.906 | 195 (92.9%) | 138 (65.7%) | 85 (40.5%) |
| **ProteoCraft-MCTS** | **0.721** | **0.930** | **0.913** | **0.889** | **0.917** | **195 (92.9%)** | **165 (78.6%)** | **114 (54.3%)** |

**Evaluation.** We propose *model-aligned* evaluation and *independent cross-evaluation*. First, complete frozen ESMFold2 inference reports Pair-chain iPTM, complex iPTM, pTM, pLDDT, and binder-specific confidence. Second, we re-evaluate the same discrete sequences using AF2-IG (Watson et al., 2023) as an independent structure verifier. Third, we use a common Rosetta (Alford et al., 2017) interface-scoring pipeline to measure interface energetics, shape complementarity, packing, hydrogen bonding, and filter pass rates. This separation is important because the search-time verifier is used to allocate search, whereas AF2-IG and Rosetta test whether improvements persist outside the optimization model. Full evaluation definitions and thresholds are in Appx. C.2.

### 3.2 *De Novo* BINDER DESIGN

**Model-aligned Evaluation.** Tab. 1 reports ESMFold2 evaluation over all 210 candidates, where ProteoCraft-MCTS obtains the strongest result across each metric. Relative to the compute-matched Best-of-$N$ baseline, tree-search control improves mean Pair-chain iPTM from $0.660$ to $0.721$ and Binder pLDDT from $0.880$ to $0.917$. Removing the One-step diffusion proposal signal further degrades Pair-chain iPTM to $0.637$, suggesting that the coordinate-level diffusion gradient contributes information beyond the remaining geometric and sequence proposal terms.

Notice that mean scores can hide whether a method consistently produces usable designs. We therefore additionally measure the fraction of candidates satisfying increasingly selective interface-confidence criteria in Tab. 1. The gain becomes larger as the criterion becomes more selective. Under the strictest threshold, ProteoCraft-MCTS yields $114/210$ high-quality candidates ($54.3\%$), compared with $40/210$ ($19.0\%$) for the compute-matched Best-of-$N$ baseline. Thus, tree search does more than improve the average verifier score: it substantially increases the probability that a finite design budget produces a candidate satisfying both interface and folding-confidence requirements.

Per-target results are provided in Appx. 11. ProteoCraft-MCTS improves Pair-chain iPTM over external baselines on essentially all targets and achieves the strongest mean result among the matched ProteoCraft variants on most targets. The remaining difficult cases also illustrate why complete verification is retained throughout search: *proposal guidance improves where the algorithm searches, but individual targets can still exhibit predictor-specific uncertainty or objective mismatch*.

**Independent AF2-IG cross-evaluation.** Tab. 2 evaluates the same 210 sequences from each method using AF2-IG (Watson et al., 2023). ProteoCraft-MCTS again obtains the strongest interface metrics and the highest strict pass rate. The improvement therefore persists under a different structure-prediction model: the strict AF2-IG pass rate increases from $78.6\%$ for Best-of-$N$ to $86.2\%$ for ProteoCraft-MCTS. This result argues against interpreting the ESMFold2 gains solely as optimization of the internal verifier.

Table 2: **Independent AF2-IG and Rosetta evaluation of *de novo* designs.** Each was evaluated on $N = 210$ designs. For AF2-IG, lower iPAE and higher iPTM, pTM, and strict pass rate are better. For Rosetta, lower $\Delta G$ and $\Delta G/\Delta\text{SASA}$ and higher values for the remaining metrics are better. Bold and underlined values indicate the best and second-best results, respectively.

| | AF2-IG | | | | Rosetta interface evaluation | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | iPAE ↓ | iPTM ↑ | pTM ↑ | Strict Pass ↑ | $\Delta G$ ↓ | $\Delta$SASA ↑ | SC ↑ | $\Delta G/\Delta$SASA ↓ | H-bonds ↑ | Packstat ↑ |
| RFdiffusion v3 | 25.426 | 0.142 | 0.550 | 0 (0.0%) | -49.398 | 1834.2 | 0.633 | -2.696 | 4.662 | 0.612 |
| BoltzGen | 26.342 | 0.200 | 0.689 | 0 (0.0%) | -31.138 | 1902.4 | 0.558 | -1.640 | 6.463 | **0.654** |
| Proteina-Complexa | 18.927 | 0.359 | 0.651 | 6 (2.9%) | -28.741 | 1626.0 | 0.557 | -1.671 | 5.314 | 0.597 |
| PXDesign-d | 18.949 | 0.349 | 0.650 | 20 (9.5%) | -50.477 | 2029.6 | 0.603 | -2.469 | <u>8.595</u> | 0.597 |
| BoltzDesign1 | 19.967 | 0.369 | 0.571 | 1 (0.5%) | -52.903 | **3251.0** | 0.607 | -1.611 | 7.743 | 0.542 |
| BindCraft | 7.093 | 0.807 | <u>0.841</u> | 104 (49.5%) | <u>-57.879</u> | 2085.9 | <u>0.665</u> | <u>-2.778</u> | 8.500 | 0.613 |
| BindCraft v2 | <u>7.062</u> | <u>0.804</u> | 0.835 | <u>122</u> (58.1%) | -49.981 | 2029.1 | 0.644 | -2.469 | 6.933 | <u>0.621</u> |
| No One-step Diffusion | 6.162 | 0.848 | 0.864 | 126 (60.0%) | -70.665 | 2585.6 | **0.681** | -2.760 | 11.095 | 0.583 |
| ProteoCraft-Best-of-$N$ | 5.795 | 0.860 | 0.881 | 165 (78.6%) | -68.078 | 2666.7 | 0.641 | -2.588 | 11.395 | 0.576 |
| **ProteoCraft-MCTS** | **4.792** | **0.883** | **0.892** | **181 (86.2%)** | **-81.195** | <u>2933.7</u> | **0.681** | **-2.783** | **12.510** | 0.584 |

Table 3: **TCR-facing binder design.** The left block reports complete ESMFold2 verification and the right block reports independent AF2-IG evaluation.

| | ESMFold2 | | | | AF2-IG | | |
|---|---|---|---|---|---|---|---|
| Method | Pair iPTM ↑ | iPTM ↑ | pTM ↑ | pLDDT ↑ | iPAE ↓ | iPTM ↑ | pTM ↑ |
| RFdiffusion v3 | 0.225 | 0.695 | 0.780 | 0.852 | 27.540 | 0.161 | 0.693 |
| BoltzGen | 0.384 | 0.763 | 0.824 | 0.843 | 26.342 | 0.200 | 0.689 |
| Proteina-Complexa | 0.300 | 0.756 | 0.822 | 0.863 | 26.439 | 0.191 | 0.702 |
| PXDesign-d | 0.276 | 0.717 | 0.795 | 0.873 | <u>26.115</u> | <u>0.207</u> | 0.708 |
| BoltzDesign1 | 0.228 | 0.734 | 0.810 | 0.817 | 26.165 | 0.214 | 0.713 |
| BindCraft | <u>0.513</u> | <u>0.828</u> | <u>0.869</u> | <u>0.884</u> | 27.043 | 0.172 | <u>0.718</u> |
| ProteoCraft-Best-of-$N$ | 0.514 | 0.841 | 0.877 | 0.866 | 25.266 | 0.249 | 0.725 |
| **ProteoCraft-MCTS** | **0.634** | **0.887** | **0.907** | **0.889** | **24.595** | **0.282** | **0.730** |

**Rosetta interface evaluation.** We next test with a common Rosetta pipeline, whose scores assess independently refined interface energetics and geometry. Tab. 2 shows that ProteoCraft-MCTS produces the most favorable mean interface energy, energy normalized by buried surface area, and hydrogen-bond count, while matching the highest shape complementarity. Although BoltzGen obtains the highest mean Packstat, the overall pattern is consistent with the AF2-IG cross-evaluation. Improvements obtained through proposal–verification search persist under evaluators that are not used to guide the search.

### 3.3 Immune-Relevant Protein Discovery

**TCR-Facing Binder Design.** We evaluate ProteoCraft under TCR-facing surfaces, which differ geometrically from the standard *de novo* benchmark. Tab. 3 reports matched ESMFold2 and AF2-IG evaluation. ProteoCraft-MCTS improves over the compute-matched Best-of-$N$ baseline on all reported ESMFold2 metrics, including Pair-chain iPTM from $0.514$ to $0.634$ and pLDDT from $0.866$ to $0.889$. The direction of improvement is also preserved under AF2-IG, where MCTS obtains lower iPAE and higher iPTM and pTM. These results suggest that proposal–verification search is not restricted to conventional globular binder targets.

**Localized CDR/TCR Redesign.** ProteoCraft can also operate on a restricted mutable region while retaining the rest of the complex as structural context. We evaluate four chain-specific redesign settings from two TCR complexes, with ten candidates per setting. Tab. 4 demonstrates that ProteoCraft-MCTS obtains the strongest mean peptide-specific and context-averaged interface confidence. Relative to RFdiffusion v3, mean Peptide Pair iPTM increases from $0.535$ to $0.765$. The improvement is particularly pronounced on the harder 1QRN $\beta$ setting, where ProteoCraft raises Peptide Pair iPTM from $0.421$ to $0.789$.

We again cross-evaluate these designs using AF2-IG and Rosetta in Tab. 5. ProteoCraft-MCTS achieves the lowest AF2-IG iPAE and highest iPTM, while Rosetta reports substantially more favorable mean interface energy, shape complementarity, hydrogen bonding, and packing. Together, these results indicate that restricting the mutable sequence region does not prevent proposal–verification search from exploiting global structural context.

Table 4: **ESMFold2 evaluation for localized CDR redesign.** Peptide Pair iPTM measures confidence between the redesigned chain and the peptide; Mean Pair iPTM averages pairwise confidence between the redesigned chain and all fixed context chains. All metrics are higher-is-better.

| Method | Case | $N$ | Peptide Pair ↑ | Mean Pair ↑ | iPTM ↑ | pTM ↑ | Complex pLDDT ↑ | Chain pLDDT ↑ |
|---|---|---|---|---|---|---|---|---|
| RFdiffusion v3 | 1QRN $\alpha$ | 10 | 0.591 | 0.686 | 0.872 | 0.896 | 0.919 | 0.901 |
| | 1QRN $\beta$ | 10 | 0.421 | 0.560 | 0.824 | 0.857 | 0.905 | 0.912 |
| | 9NMW $\alpha$ | 10 | 0.549 | 0.669 | 0.884 | 0.900 | 0.906 | 0.896 |
| | 9NMW $\beta$ | 10 | 0.579 | 0.687 | 0.909 | 0.920 | 0.914 | 0.918 |
| | **Mean** | – | 0.535 | 0.650 | 0.872 | 0.893 | 0.911 | 0.907 |
| BoltzGen | 1QRN $\alpha$ | 10 | 0.715 | 0.761 | 0.926 | 0.934 | 0.929 | 0.912 |
| | 1QRN $\beta$ | 10 | 0.491 | 0.609 | 0.851 | 0.875 | 0.909 | 0.917 |
| | 9NMW $\alpha$ | 10 | 0.567 | 0.675 | 0.880 | 0.898 | 0.910 | 0.897 |
| | 9NMW $\beta$ | 10 | 0.650 | 0.704 | 0.894 | 0.908 | 0.914 | 0.912 |
| | **Mean** | – | 0.606 | 0.687 | 0.888 | 0.904 | 0.915 | 0.910 |
| BindCraft | 1QRN $\alpha$ | 10 | 0.752 | 0.774 | 0.935 | 0.939 | 0.922 | 0.885 |
| | 1QRN $\beta$ | 10 | 0.464 | 0.594 | 0.842 | 0.871 | 0.898 | 0.887 |
| | 9NMW $\alpha$ | 10 | 0.603 | 0.691 | 0.883 | 0.896 | 0.901 | 0.879 |
| | 9NMW $\beta$ | 10 | 0.720 | 0.754 | 0.935 | 0.939 | 0.925 | 0.928 |
| | **Mean** | – | 0.635 | 0.704 | 0.899 | 0.911 | 0.911 | 0.895 |
| **ProteoCraft-MCTS** | 1QRN $\alpha$ | 10 | 0.816 | 0.818 | 0.951 | 0.953 | 0.939 | 0.919 |
| | 1QRN $\beta$ | 10 | 0.789 | 0.779 | 0.938 | 0.943 | 0.932 | 0.942 |
| | 9NMW $\alpha$ | 10 | 0.682 | 0.732 | 0.922 | 0.923 | 0.913 | 0.880 |
| | 9NMW $\beta$ | 10 | 0.772 | 0.777 | 0.939 | 0.943 | 0.928 | 0.930 |
| | **Mean** | – | **0.765** | **0.776** | **0.937** | **0.940** | **0.928** | **0.918** |

Table 5: **Independent cross-evaluation of localized CDR redesign.**

| | AF2-IG | | | | Rosetta | | | |
|---|---|---|---|---|---|---|---|---|
| Method | iPAE ↓ | iPTM ↑ | pTM ↑ | pLDDT ↑ | $\Delta G$ ↓ | SC ↑ | H-bonds ↑ | Packstat ↑ |
| RFdiffusion v3 | 17.483 | 0.505 | 0.665 | **0.871** | -103.9 | 0.681 | 14.200 | 0.488 |
| BindCraft | 23.294 | 0.360 | 0.602 | 0.642 | -78.3 | 0.571 | 14.750 | 0.618 |
| **ProteoCraft-MCTS** | **14.558** | **0.753** | **0.716** | 0.851 | **-138.3** | **0.702** | **24.575** | **0.673** |

### 3.4 Prospective Experimental Validation of De Novo Binders

To assess whether the computational improvements of ProteoCraft translate into experimentally measurable binding, we prospectively evaluated de novo binders designed against two therapeutically relevant cell-surface targets: interferon alpha/beta receptor 2 (IFNAR2) and programmed death-ligand 1 (PD-L1). Candidate sequences were selected based on the complete ProteoCraft pipeline, integrating ESMFold2 complex confidence, independent structural cross-evaluation, and interface-quality filtering. The selected proteins were synthesized and experimentally tested for direct target binding using biolayer interferometry (BLI).

The lead designs exhibited clear concentration-dependent association and dissociation across analyte concentrations ranging from 31.6 to 1,000 nM (Fig. 3). The IFNAR2 binder reached a dissociation constant of $K_D = 153$ nM, while the PD-L1 binder showed substantially stronger binding with $K_D = 17.1$ nM. Notably, these binders were obtained directly from the computational design and filtering pipeline without experimental affinity maturation. These results provide prospective evidence that the structural confidence and interface-quality improvements achieved by ProteoCraft can translate into physical target binding, including low-nanomolar affinity against PD-L1. Although broader experimental characterization will be required to establish specificity, biological activity, and general success rates across targets, these initial results support ProteoCraft as a practical framework for discovering experimentally functional protein binders.

### 3.5 Ablation Studies and Discussion

**Contribution of One-step diffusion.** Tab. 1 removes only the One-step diffusion gradient. Adding this signal improves mean Pair-chain iPTM from $0.637$ to $0.660$ before tree search and increases the number of candidates above Pair iPTM $0.7$ from $85$ to $118$. This supports the role of truncated coordinate-level sensitivity as a *proposal signal*: it enriches the candidate set before full verification, even though the truncated score itself is never used as the final candidate value.

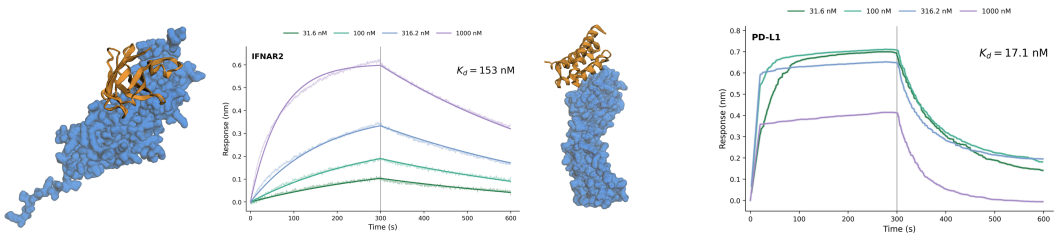



Figure 3: **Wet-lab validation of ProteoCraft-designed binders.** Complex structures and biolayer interferometry sensorgrams for the selected de novo binders against IFNAR2 (left) and PD-L1 (right).

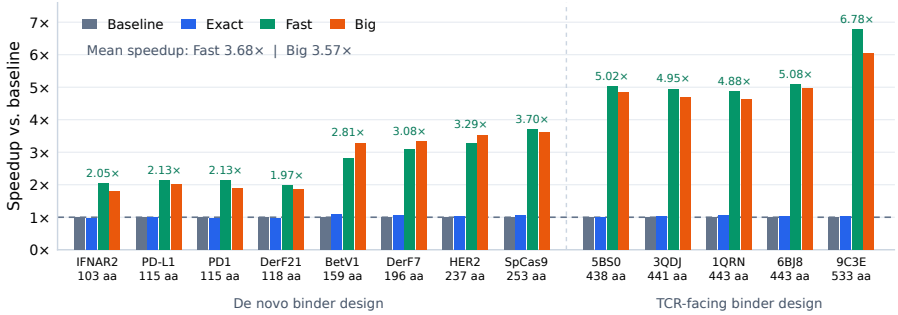



Figure 4: **Accelerating ProteoCraft through inference optimization.**

**Contribution of Stage B refinement.** We assess the benefit of single-site refinement on the 14 de novo binders in Tab. 1. For each setting, both branches resume from the same checkpoint after 70 Stage A steps and receive 40 additional search steps, either continuing Stage A multi-site exploration or switching to Stage B single-site refinement. Stage B increases mean Pair-chain iPTM from 0.691 to 0.721. It also improves all other reported ESMFold2 confidence metrics on average. This supports the benefit of Stage B refinement over extending Stage A under this evaluation setting.

**Runtime and speedup.** Table 6 summarizes per-binder computational costs under method-specific output definitions. ProteoCraft requires an average of 32.34 seconds per qualifying candidate on de novo targets and 174.06 seconds on TCR targets, counting unique search candidates with overall iPTM $> 0.70$. We further evaluate whether inference optimization can accelerate the full ProteoCraft search pipeline using the ESMFold2 optimization kit released by Anthropic.[1] Across 13 targets, Fast and Big achieve mean end-to-end speedups of $3.68\times$ and $3.57\times$ over unoptimized ProteoCraft, respectively, with Fast reaching $6.78\times$ on 9C3E (Fig. 4). These results demonstrate that inference-level improvements can substantially reduce the computational cost of verifier-guided search without performance sacrifice (Fig. 5).

Table 6: **Runtime comparison.** We report mean wall-clock latency (seconds) per binder under each method's native design and selection procedure. Details are in Appendix C.3.

| Method | De novo | TCR |
|---|---|---|
| PXDesign-d | 18.72 | 21.01 |
| Proteina-Complexa | 14.62 | 34.61 |
| RFdiffusion v3 | 56.64 | 44.26 |
| BoltzDesign1 | 693.38 | 2882.20 |
| BindCraft v2 | 2321.50 | — |
| ProteoCraft | 32.34 | 174.06 |

## 4 Conclusion

ProteoCraft shows that approximate gradients from diffusion predictors need not serve as faithful value estimators to be useful for protein design. By using differentiable signals only to propose discrete sequence edits, and reserving complete predictor inference for candidate evaluation, ProteoCraft turns diffusion-based protein design into a proposal–verification search problem. A coarse-to-fine search then allocates expensive verifier calls between broad multi-site exploration and conservative single-site refinement.

[1] https://github.com/anthropics/uplifting-biomolecular-modeling

## AI USE STATEMENT

Generative AI tools were used to assist with language editing, literature discovery, and limited code development and debugging. All suggested references were verified by the authors against the original publications, and all AI-assisted code was manually reviewed and tested for correctness. Generative AI tools were not used to generate experimental data, fabricate results, or autonomously determine the scientific claims or conclusions of this work. The authors designed the methodology, conducted the experiments, analyzed the results, and verified all reported findings. We take full responsibility for the final content of this work, including any text, code, or other artifacts produced with the assistance of generative AI.

## REFERENCES

J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. pp. 1–3, 2024.

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, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14(1):2625, 2023.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I Brent, Yanjing Li, Nathaniel Corley, Paul T Kim, Jonathan Funk, Simon Mathis, Saman Salike, et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, 2025.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, pp. 2026–06, 2026.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M Jude, Iva Marković, Rameshwar U Kadam, Koen HG Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022.

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*, pp. 2025–04, 2025a.

Yehlin Cho, Griffin Rangel, Gaurav Bhardwaj, and Sergey Ovchinnikov. Protein hunter: exploiting structure hallucination within diffusion for protein design. *bioRxiv*, pp. 2025–10, 2025b.

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

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M Bronstein, et al. Scaling atomistic protein binder design with generative pretraining and test-time compute. *arXiv preprint arXiv:2603.27950*, 2026.

Richard Evans, Michael O'neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Pablo Gainza, Sarah Wehrle, Alexandra Van Hall-Beauvais, Anthony Marchand, Andreas Scheck, Zander Harteveld, Stephen Buckley, Dongchun Ni, Shuguang Tan, Freyr Sverrisson, et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, 617(7959):176–184, 2023.

Odessa J Goudy, Amrita Nallathambi, Tomoaki Kinjo, Nicholas Randolph, and Brian Kuhlman. In silico evolution of protein binders with deep learning models for structure prediction and sequence design. *BioRxiv*, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pp. 282–293. Springer, 2006.

Cong Liu, Milong Ren, Jiaqi Guan, Chengyue Gong, Jinyuan Sun, Xinshi Chen, and Wenzhi Xiao. Protdbench: A unified benchmark of protein binder design and evaluation. *arXiv preprint arXiv:2605.04118*, 2026.

Divya Nori, Anisha Parsan, Caroline Uhler, and Wengong Jin. Bindenergycraft: Casting protein structure predictors as energy-based models for binder design. *arXiv preprint arXiv:2505.21241*, 2025.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. Bindcraft: one-shot design of functional protein binders. *bioRxiv*, pp. 2024–09, 2024.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

Pacesa Lab. BindCraft2. `https://github.com/PacesaLab/BindCraft2`, 2026. GitHub repository. Accessed September 25, 2026.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Christopher D Rosin. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence*, 61(3):203–230, 2011.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. Boltzgen: Toward universal binder design. *bioRxiv*, pp. 2025–11, 2025.

Erik Swanson, Michael Nichols, Supriya Ravichandran, and Pierce Ogden. mber: Controllable de novo antibody design with million-scale experimental screening. *bioRxiv*, pp. 2025–09, 2025.

Protenix Team, Milong Ren, Jinyuan Sun, Jiaqi Guan, Cong Liu, Chengyue Gong, Yuzhe Wang, Lan Wang, Qixu Cai, Wenzhi Ma, et al. Pxdesign: Fast, modular, and accurate de novo design of protein binders. *bioRxiv*, pp. 2025–08, 2025.

Protenix Team, Yuxuan Zhang, Chengyue Gong, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, Jiaqi Guan, Lan Wang, Yanping Yang, et al. Protenix-v1: Toward high-accuracy open-source biomolecular structure prediction. *bioRxiv*, pp. 2026–02, 2026.

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

Fang Wu, Weihao Xuan, Heli Qi, Aaron Tu, Ximing Lu, Li Li, and Yejin Choi. Deepsearch: Overcome the bottleneck of reinforcement learning with verifiable rewards via tree-based search. In *International Conference on Learning Representations*, volume 2026, pp. 95559–95584, 2026.

Kiwoong Yoo, Soorin Yim, Kyungwook Lee, Sungjoon Park, Jaehyeong Kong, Doyeong Hwang, Jongseong Jang, and Kiyoung Kim. Search-based inference-time scaling for all-atom protein binder design. In *ICLR 2026 Workshop on Generative and Experimental Perspectives for Biomolecular Design*.

Qilin Yu, Liangyue Guo, Xiayan Qin, Xikun Huang, Baihui Tian, Hongzhun Wang, Yu Liu, Yunzhi Lang, Di Wang, Zhouhanyu Shen, et al. High-affinity protein binder design via flow matching and in silico maturation. *bioRxiv*, pp. 2026–01, 2026.

Vinicius Zambaldi, David La, Alexander E Chu, Harshnira Patani, Amy E Danson, Tristan OC Kwan, Thomas Frerix, Rosalia G Schneider, David Saxton, Ashok Thillaisundaram, et al. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024.

# A Related Work

**Hallucination-based Protein Design.** Neural structure predictors such as AF2 have been repurposed as differentiable design oracles, which optimize protein sequences directly against structural objectives (Anishchenko et al., 2021; Pacesa et al., 2025). BindEnergyCraft introduces a denser energy-based objective for this optimization (Nori et al., 2025), whereas mBER incorporates sequence and structural conditioning to support format-specific antibody design (Swanson et al., 2025). Extending this paradigm to AF3-style predictors (Team et al., 2026; Passaro et al., 2025) is challenging because their iterative diffusion-based inference makes end-to-end differentiation through the complete sampling trajectory computationally expensive. BoltzDesign1 (Cho et al., 2025a) addresses this difficulty by optimizing through the intermediate distogram without gradients from the diffusion sampling. ProteoCraft, instead, differentiates through only the terminal diffusion step and uses truncated gradients as mutation proposals.

**Inference-Time Search and Verification.** Several methods refine protein designs through repeated generation and evaluation. EvoPro (Goudy et al., 2023) applies a genetic algorithm in which discrete sequence variants are evaluated through complete structure-prediction runs. PPIFlow (Yu et al., 2026) couples flow-based generation with an in silico maturation stage, while Proteína-Complexa (Didi et al., 2026) scales inference-time compute through best-of-$N$ sampling, beam search, Feynman–Kac steering, and MCTS over generative trajectories. Search-based inference-time scaling has also been applied to all-atom diffusion models, where inexpensive intermediate rewards guide the selection of denoising trajectories (Yoo et al.). However, existing methods either discard local gradient information or rely on approximate intermediate rewards. ProteoCraft instead uses truncated gradients to propose discrete mutations and leverages coarse-to-fine search that progresses from coordinated multi-site exploration to single-site refinement, focusing expensive evaluations on promising sequence regions.

# B Method Details

Algorithm 1 summarizes the complete procedure.

**Algorithm 1** ProteoCraft proposal–verification search.

**Require:** Fixed context $\mathcal{T}$, mutable mask $\mathbf{M}$, initial logits $\{\mathbf{\Theta}_r\}_{r=1}^{N_{\text{root}}}$, frozen ESMFold2 model
**Ensure:** Best fully verified sequence and structure

initialize an empty global verified archive
**for** $r = 1, \ldots, N_{\text{root}}$ **do**
  $\mathbf{s}_r \leftarrow \arg\max \mathbf{\Theta}_r$; fully verify $\mathbf{s}_r$ and initialize root $\mathbf{s}_r^0$
  add $\mathbf{s}_r^0$ to the verified archive
**end for**
**Stage A: multi-site exploration**
**for** $t = 1, \ldots, B_A$ **do**
  select verified sequence $\mathbf{s}^i$ by Eq. 21
  optimize its logits $\mathbf{\Theta}^i$ using Eq. 13
  compute $\mathbf{\Delta}_{s_i, s_i'}$ and construct multi-site candidates
  fully verify at most eight unique candidates; archive all results
  insert at most two selected children and back up their rewards
**end for**
choose the two best archived states and re-verify them with a common seed
**Stage B: single-site refinement**
**for** $t = 1, \ldots, B_B$ **do**
  select $(\mathbf{s}^i, a_{\mathbf{s}^i})$ by Eq. 23
  apply $a_{\mathbf{s}^i}$ to obtain a one-residue mutant $\mathbf{s}'$
  fully verify $\mathbf{s}'$; archive it and add it as a child
  back up state and action values using Eq. 22
**end for**
**return** the archived $\mathbf{s}$ with minimum $\mathcal{L}_{\text{verify}}(\mathbf{s})$

### B.1 State Representation and Search Statistics

Let $\mathcal{T}$ denote the fixed molecular context and $\mathbf{s} = (s_1, \ldots, s_L)$ the designed sequence. A mask $\mathbf{M} \in \{0,1\}^L$ identifies mutable positions. Each sequence is associated with logits $\mathbf{\Theta} \in \mathbb{R}^{L\times 20}$ and soft residue probabilities

$$P_{\mathbf{\Theta}}(i,a) = \mathrm{Softmax}(\mathbf{\Theta}_i/T)_a, \qquad s_i = \arg\max_a \mathbf{\Theta}_{i,a}, \tag{5}$$

where $T$ is annealed during proposal optimization. Fixed positions and target residues are represented by one-hot identities and never receive logit updates.

A verified node is represented by the same discrete sequence notation used in the main text. For node $\mathbf{s}^i$, we store

$$(\mathbf{s}^i, \mathbf{\Theta}^i, \mathbf{X}^i, \mathcal{V}^i, N(\mathbf{s}^i), q(\mathbf{s}^i)), \tag{6}$$

where $\mathbf{X}^i$ and $\mathcal{V}^i$ are the structure and metrics returned by complete ESMFold2 inference, $N(\mathbf{s}^i)$ is the visit count, and $q(\mathbf{s}^i)$ is the mean backed-up reward. In Stage B, a single-site action $a_{\mathbf{s}^i}$ additionally stores $N(\mathbf{s}^i, a_{\mathbf{s}^i})$ and $q(\mathbf{s}^i, a_{\mathbf{s}^i})$. All state and action values are therefore empirical quantities derived only from fully verified sequences.

### B.2 Differentiable Mutation Proposals

**Soft sequence conditioning.** The probability matrix in Eq. 5 is passed through the ESMFold2 `res_type_soft` interface. The current hard sequence defines a valid residue and atom topology, while the soft representation conditions the trunk and exposes gradients with respect to $\mathbf{\Theta}$. No straight-through estimator is required.

**Distogram objective.** Let $p_{ijr}$ be the predicted probability that residues $i$ and $j$ occupy distance bin $r$. For cutoff $d$, we compute a differentiable contact penalty from the probability mass below $d$,

$$\ell_{ij}^{\mathrm{contact}}(d) = -\log\left(\sum_{r:\, d_r \le d} p_{ijr} + \varepsilon\right). \tag{7}$$

Aggregating Eq. 7 over binder–binder, target–binder, and global binder residue pairs gives $\mathcal{L}_{\mathrm{intra}}$, $\mathcal{L}_{\mathrm{inter}}$, and $\mathcal{L}_{\mathrm{glob}}$, respectively. The geometric proposal loss is

$$\mathcal{L}_{\mathrm{disto}} = 0.5\mathcal{L}_{\mathrm{intra}} + 0.5\mathcal{L}_{\mathrm{inter}} + 0.2\mathcal{L}_{\mathrm{glob}}. \tag{8}$$

**Sequence prior.** Sequence plausibility is regularized with an ESMC masked language-model loss over mutable positions $\mathcal{I}_M = \{i : M_i = 1\}$. Let $\Omega_1, \ldots, \Omega_K \subseteq \mathcal{I}_M$ be sampled nonempty mask sets, and let $p_{\mathrm{ESMC}}^{(k)}(i,a)$ denote the predicted probability of amino acid $a$ at position $i$ under mask set $\Omega_k$. Using the current soft amino-acid probabilities as targets, we compute

$$\mathcal{L}_{\mathrm{ESMC}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\Omega_k|}\sum_{i\in\Omega_k}\sum_{a=1}^{20} P_{\mathbf{\Theta}}(i,a)\log p_{\mathrm{ESMC}}^{(k)}(i,a). \tag{9}$$

The loss is averaged over sampled mask sets and affects proposal optimization only; it is not used to verify or rank a completed candidate.

**Truncated one-step diffusion.** The distogram supplies pairwise geometric information but does not directly differentiate through coordinate generation. We therefore retain gradients through one terminal denoising operation. Starting from $\mathbf{X}_{\sigma_{\max}} = \sigma_{\max}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the reverse trajectory is run without gradient tracking to noise level $\tau$:

$$\mathbf{X}_\tau = \mathrm{sg}\left[f_\phi^{\sigma_{\max}\to\tau}(\mathbf{X}_{\sigma_{\max}}; \mathbf{c}_{\mathbf{\Theta}})\right], \tag{10}$$

$$\widehat{\mathbf{X}}_0 = D_\phi(\mathbf{X}_\tau, \tau; \mathbf{c}_{\mathbf{\Theta}}), \tag{11}$$

where sg denotes stop-gradient and $D_\phi$ is the frozen denoiser that predicts clean coordinates at noise level $\tau$. The coordinate-sensitive loss is

$$\begin{aligned}\mathcal{L}_{\text{1step}} = 0.1\mathcal{L}_{\text{pLDDT}} + 0.4\mathcal{L}_{\text{PAE}} + 0.2\mathcal{L}_{\text{iPAE}} + 0.3\mathcal{L}_{\text{pair-chain iPTM}} \\ + 0.1\mathcal{L}_{R_g} + 0.1\mathcal{L}_{\text{clash}} + 0.1\mathcal{L}_{\text{i-clash}},\end{aligned} \tag{12}$$

with all components written as losses. In particular, $\mathcal{L}_{\text{pair-chain iPTM}} = 1 - \text{pair-chain iPTM}$; the PAE terms measure global and interface alignment error, and the remaining terms penalize low confidence, excessive radius of gyration, and internal or interface clashes. Gradients are retained through the terminal denoising operation and the current sequence conditioning, while the preceding reverse-diffusion trajectory is detached. The structure produced by this truncated calculation is not treated as a verified candidate.

**Gradient combination and soft optimization.** For $\mu \in \{\text{disto}, \text{ESMC}, \text{1step}\}$, let $\mathbf{g}_\mu = \mathbf{M} \odot \nabla_{\mathbf{\Theta}} \mathcal{L}_\mu$. We normalize each signal over active entries, $\text{Norm}(\mathbf{g}) = \mathbf{g}/(\|\mathbf{g}\|_2 + \varepsilon)$, and combine them as

$$\mathbf{g} = \text{Norm}(\mathbf{g}_{\text{disto}}) + \lambda_{\text{ESMC}} \text{Norm}(\mathbf{g}_{\text{ESMC}}) + \lambda_{\text{1step}} \text{Norm}(\mathbf{g}_{\text{1step}}). \tag{13}$$

At each soft step, logits are updated by $\mathbf{\Theta} \leftarrow \mathbf{\Theta} - \eta(T)\mathbf{g}$, with the learning rate scaled by the current temperature. Stage A and Stage B perform 30 and 20 soft steps per expansion, respectively. Importantly, the optimized argmax sequence is only one possible proposal; acceptance always requires complete verification.

### B.3 Hotspot and Region Constraints

For a target hotspot set $\mathcal{H}$ and binder residue set $\mathcal{B}$, we define

$$\mathcal{L}_{\text{hot}} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \min_{j \in \mathcal{B}} \ell_{ij}^{\text{contact}}(d_{\text{hot}}), \tag{14}$$

and replace the interface term in Eq. 8 by

$$\mathcal{L}'_{\text{inter}} = (1 - \lambda_{\text{hot}})\mathcal{L}_{\text{inter}} + \lambda_{\text{hot}}\mathcal{L}_{\text{hot}}. \tag{15}$$

Hotspot distances and coverage do not enter the ranking loss in Eq. 20, but may contribute to the Stage-B search reward. Localized redesign uses the same losses on the complete complex while masking all logit gradients outside the selected region. Thus, fixed chains and framework residues remain structural context without being edited.

### B.4 Converting Logit Updates into Discrete Edits

Let $\mathbf{\Theta}$ and $\mathbf{\Theta}'$ denote the logits before and after one proposal-optimization phase. For the substitution $s_i \to s'_i$ at position $i$, the mutation benefit is

$$\mathbf{\Delta}_{s_i, s'_i} = \left(\mathbf{\Theta}'_{i,s'_i} - \mathbf{\Theta}_{i,s'_i}\right) - \left(\mathbf{\Theta}'_{i,s_i} - \mathbf{\Theta}_{i,s_i}\right). \tag{16}$$

A positive value means that continuous optimization increasingly favors $s'_i$ over retaining $s_i$. Edits at fixed positions and self-substitutions are excluded.

**Stage-A candidates.** Stage A constructs coordinated multi-site edits with

$$K_{\max} = \left\lceil \rho \sum_i M_i \right\rceil. \tag{17}$$

One candidate applies the highest-benefit substitutions at up to $K_{\max}$ distinct positions. Additional candidates sample positions and replacement residues according to their positive benefits. We also retain the raw argmax sequence of $\mathbf{\Theta}'$, which is not constrained by $K_{\max}$. Duplicate and previously evaluated sequences are removed before verification.

**Stage-B actions.** Stage B admits only single-residue substitutions. For an action $a_{\mathbf{s}^i}$ from state $\mathbf{s}^i$, its proposal prior is

$$P(a_{\mathbf{s}^i} \mid \mathbf{s}^i) = \frac{\exp(\mathbf{\Delta}_{a_{\mathbf{s}^i}})}{\sum_{a'_{\mathbf{s}^i} \in \mathcal{A}(\mathbf{s}^i)} \exp(\mathbf{\Delta}_{a'_{\mathbf{s}^i}})}, \tag{18}$$

where $\mathcal{A}(\mathbf{s}^i)$ contains admissible, unevaluated edits from $\mathbf{s}^i$. The prior expresses local gradient preference; it is not an estimate of final candidate quality.

**Logit inheritance.** A child inherits $\mathbf{\Theta}'$ from the proposal that generated it. If the selected discrete child differs from its argmax at an edited position, we minimally align that row:

$$\mathbf{\Theta}^{\text{child}}_{i,s_i^{\text{child}}} = \max_a \mathbf{\Theta}'_{i,a} + \delta, \qquad \delta > 0, \tag{19}$$

leaving all other entries unchanged. This preserves the optimized continuous state while ensuring that the inherited logits encode the verified child sequence.

### B.5 Complete-Model Verification

Every proposed discrete sequence is evaluated by complete frozen ESMFold2 inference. Its default ranking loss is

$$\mathcal{L}_{\text{verify}}(\mathbf{s}^j) = (1 - \text{pair-chain iPTM}) + 0.2(1 - \text{binder pLDDT}) + 0.1(1 - \text{pTM}) + 0.1\mathcal{L}_{R_g\text{-excess}}, \tag{20}$$

In Stage A, the reward is defined as $R(\mathbf{s}^j) = -\mathcal{L}_{\text{verify}}(\mathbf{s}^j)$. Pair-chain iPTM is computed only between the designated target chains and the binder. The final term is a normalized penalty applied only when binder radius of gyration exceeds its length-dependent threshold. Additional confidence, PAE, contact, coordinate, clash, and hotspot statistics are recorded separately and do not add further terms to the ranking loss in Eq. 20.

Stage B uses a separate tree-search reward that balances interface confidence, binder confidence, global confidence, and binder compactness. Additional penalties account for unmet hotspot constraints, when specified, and interfaces below a quality floor. This reward guides which sequence branches are explored further, while the recorded verification loss is used to select candidates for the frozen benchmark cohort.

### B.6 Coarse-to-Fine Tree Search

**Stage A: multi-site basin exploration.** We initialize $N_{\text{root}} = 8$ independent sequence–logit roots and fully verify their argmax sequences. Each root is expanded at least twice before allocating the remaining budget adaptively. Selection uses

$$U(\mathbf{s}^i) = q(\mathbf{s}^i) + c_{\text{ucb}} \sqrt{\frac{\log N_{\text{total}}}{N(\mathbf{s}^i) + \varepsilon}}, \tag{21}$$

where $q(\mathbf{s}^i)$ is the mean backed-up reward. The same criterion is applied at the root level using root-specific visits, yielding hierarchical allocation across initializations and then within each tree. After selection, the inherited logits are optimized, multi-site candidates are constructed and fully verified, and the resulting reward is backed up.

**Reward backup.** For a newly verified child $\mathbf{s}^j$ selected for tree insertion, the trajectory $\mathbf{s}^0 \oplus \cdots \oplus \mathbf{s}^j$ is traversed toward the root. At ancestor $\mathbf{s}^i$, we update

$$N(\mathbf{s}^i) \leftarrow N(\mathbf{s}^i) + 1, \qquad q(\mathbf{s}^i) \leftarrow q(\mathbf{s}^i) + \frac{R(\mathbf{s}^j) - q(\mathbf{s}^i)}{N(\mathbf{s}^i)}, \tag{22}$$

We use undiscounted backup, corresponding to $\gamma = 1$. Stage-B action statistics are updated analogously for every traversed $(\mathbf{s}^i, a_{\mathbf{s}^i})$. Larger $q$ indicates a higher mean backed-up reward under the reward definition used by the current search stage.

Table 7: Default ProteoCraft–ESMFold2 configuration.

| Parameter | Value | Description |
|---|---|---|
| Stage-A outer steps | 70 | Multi-site search budget |
| Stage-A soft steps | 30 | Proposal updates per expansion |
| Stage-B outer steps | 40 | Single-site search budget |
| Stage-B soft steps | 20 | Proposal updates per expansion |
| Sequence temperature | $1.0 \rightarrow 0.01$ | Cosine annealing |
| Base learning rate | 0.05 | Scaled by current temperature |
| One-step noise level $\tau$ | 2.0 | Diffusion truncation level |
| $\lambda_{\text{1step}}$ | 0.25 | Normalized gradient weight |
| Discount factor $\gamma$ | 1 | Tree-backup discount |
| Stage-A mutation ratio $\rho$ | 0.5 | Maximum edited fraction |
| Independent Stage-A roots | 8 | Initial sequence–logit states |
| Stage-A handoff states | 2 | Roots transferred to Stage B |
| Candidates per expansion | $\leq 8$ | Complete verifier calls |
| Inserted children per expansion | $\leq 2$ | Tree branching limit |

**Verified handoff.** After Stage A, the two best verified sequence–logit states in the archive are selected. Since ESMFold2 inference is stochastic, both are re-evaluated with a common verifier seed before becoming Stage-B roots. This prevents differences in verifier sampling from being mistaken for improvement during refinement.

**Stage B: single-site refinement.** At state $\mathbf{s}^i$, action selection maximizes

$$U_{\text{PUCT}}(\mathbf{s}^i, a_{\mathbf{s}^i}) = q(\mathbf{s}^i, a_{\mathbf{s}^i}) + c_{\text{puct}} P(a_{\mathbf{s}^i} \mid \mathbf{s}^i) \frac{\sqrt{N(\mathbf{s}^i)}}{1 + N(\mathbf{s}^i, a_{\mathbf{s}^i})}. \tag{23}$$

The first term exploits mutations whose verified descendants have high reward; the second favors gradient-preferred actions that have received little search budget. After selecting an action, ProteoCraft creates the corresponding single-site mutant, runs complete verification, adds the verified child, and updates node and edge statistics using Eq. 22.

### B.7 Task Instantiations and Configuration

Unless otherwise specified, all task instantiations use the default search configuration in Table 7. Task-specific hotspot definitions and coefficients are fixed within each matched comparison.

**De novo binder design.** All designated binder positions are mutable; target chains remain fixed. Interface losses and pair-chain iPTM are evaluated between the binder and the designated target chains.

**TCR-facing binder design.** The receptor complex is structural context, but only explicitly designated chains contribute to target–binder objectives, pair-chain iPTM, and hotspot masks.

**Localized CDR redesign.** Only residues selected by $\mathbf{M}$ are updated. The remainder of the antibody or TCR framework participates in complete prediction but remains sequence-fixed. Because ESMFold2 does not consume an external template, this setting preserves the framework sequence rather than clamping input backbone coordinates.

## C Experimental Details

This section provides the experimental settings required to interpret and reproduce the supplementary results. We provide benchmark composition, common evaluation protocols, per-target results, independent cross-evaluation, and robustness analyses.

Table 8: Hotspot definitions for eight de novo target proteins and five TCR-facing target systems.

| Case | Group | Hotspot residues |
|---|---|---|
| BetV1 | Normal binder | B24 |
| DerF21 | Normal binder | B28 |
| DerF7 | Normal binder | B132 |
| HER2 | Normal binder | B134, B139 |
| IFNAR2 | Normal binder | B45, B73, B75, B77, B89, B91 |
| PD-L1 | Normal binder | B37, B39, B49, B98 |
| PD1 | Normal binder | B33, B95, B98, B102 |
| SpCas9 | Normal binder | B134 |
| 1QRN | TCR-target binder | A19–A23; A53–A56; A63–A66; A72–A76; B21–B24; B5–B7; B66–B69; B74–B77 |
| 3QDJ | TCR-target binder | A53–A56; A61–A64; A70–A74; B23–B27; B68–B69; B78–B80; B7–B10 |
| 5BS0 | TCR-target binder | A20–A25; A57–A59; A63–A67; A73–A77; B20–B24; B66–B69; B75–B78; B7–B8 |
| 6BJ8 | TCR-target binder | A20–A24; A64–A68; A73–A77; B20–B24; B4–B7; B65–B68; B73–B77 |
| 9C3E | TCR-target binder | A75–A77; A82–A85; A91–A94; B24–B26; B39–B43; B84–B85; B92–B96 |

Table 9: TCR-facing binder-design systems.

| PDB / case | TCR target | Antigen / restriction |
|---|---|---|
| 1QRN | A6 TCR | TAX-A2 |
| 3QDJ | DMF5 TCR | MART1-A2 |
| 5BS0 | A3A TCR | MAGEA3-A1 / Titin-A1 |
| 6BJ8 | TCR55 | HIV-B35 |
| 9C3E | 1G4 TCR | NYESO-A2 |

### C.1 Experimental Setup

**Matched *de novo* binder benchmark.** We evaluate 14 target settings: BetV1, DerF21-B10, DerF21-B28, DerF21-B63, DerF21-B81, DerF21-B92, DerF7-B132, DerF7-B178, HER2, HER2-B139, IFNAR2, PD-L1, PD1, and SpCas9. For each method, we generate five candidates under each of three matched binder-length settings, yielding 15 candidates per target and $14 \times 15 = 210$ candidates per method. All aggregate statistics assign equal weight to each candidate. Because different model families expose incomparable native objectives, method-specific internal losses are never compared directly. Hotspot annotations are listed in Table 8.

**ProteoCraft variants.** **No One-step Diffusion** removes only the truncated one-step diffusion gradient from Eq. 13, retaining the distogram signal, ESMC prior, candidate construction, and full verification. **ProteoCraft-Base / Best-of-$N$** uses the same differentiable proposal mechanism and verifier budget as the full method but generates independent proposal trajectories without tree-search control. **ProteoCraft-MCTS** denotes the complete coarse-to-fine search with Stage-A multi-site exploration and Stage-B single-site PUCT refinement. These variants use identical targets, candidate counts, and downstream evaluators.

**External baselines.** PXDesign-d, RFdiffusion v3, BoltzDesign1, and BindCraft are run from the same target definitions whenever supported by the corresponding method. Candidate counts are matched within each reported comparison. We use each method's native generation procedure, but discard native ranking scores during cross-method evaluation; all retained sequences are re-evaluated by the common protocols below.

**TCR-facing binder design.** We additionally design binders against five immune-relevant TCR surfaces (Table 9). The target TCR remains fixed and the new binder is generated from scratch. The matched comparison contains $N = 42$ designs per method.

**Localized CDR redesign.** We evaluate four chain-specific redesign settings from two scaffolded TCR complexes (Table 10). Only the indicated CDR3 residues are mutable; the peptide, MHC, partner TCR chain, and fixed residues of the redesigned chain remain in the modeled context. We evaluate ten candidates per case for RFdiffusion v3, BindCraft, and ProteoCraft-MCTS, yielding $N = 40$ candidates per method.

Table 10: Localized CDR redesign settings.

| Case | CDR3 sequence | Residue range |
|---|---|---|
| 9NMW $\alpha$ | CAVNNARLMFGDGTQL | 90–105 |
| 9NMW $\beta$ | CASSVAGSPEAFF | 89–101 |
| 1QRN $\alpha$ | CAVTTDSWGKLQF | 88–100 |
| 1QRN $\beta$ | CASRPGLAGGRPEQYF | 91–106 |

Table 11: Mean ESMFold2 Pair-chain iPTM across the 15 candidates for each matched *de novo* target. Higher is better.

| Target | PXDesign-d | RFdiffusion v3 | BoltzDesign1 | No One-step | Base / Best-of-$N$ | ProteoCraft-MCTS |
|---|---|---|---|---|---|---|
| BetV1 | 0.257 | 0.185 | 0.087 | 0.793 | 0.725 | **0.832** |
| DerF21-B10 | 0.462 | 0.241 | 0.214 | 0.698 | 0.746 | **0.771** |
| DerF21-B28 | 0.449 | 0.170 | 0.099 | 0.630 | 0.729 | **0.770** |
| DerF21-B63 | 0.378 | 0.206 | 0.140 | 0.699 | 0.738 | **0.787** |
| DerF21-B81 | 0.392 | 0.140 | 0.178 | 0.600 | 0.725 | **0.756** |
| DerF21-B92 | 0.340 | 0.194 | 0.219 | 0.716 | 0.695 | **0.746** |
| DerF7-B132 | 0.212 | 0.202 | 0.104 | 0.618 | 0.603 | **0.679** |
| DerF7-B178 | 0.312 | 0.215 | 0.121 | 0.587 | 0.632 | **0.714** |
| HER2 | 0.415 | 0.262 | 0.137 | 0.635 | 0.648 | **0.776** |
| HER2-B139 | 0.278 | 0.307 | 0.194 | 0.682 | 0.614 | **0.787** |
| IFNAR2 | 0.182 | 0.233 | 0.109 | 0.644 | 0.675 | **0.704** |
| PD-L1 | 0.435 | 0.316 | 0.246 | 0.773 | 0.795 | **0.798** |
| PD1 | 0.328 | 0.151 | 0.224 | 0.686 | 0.655 | **0.721** |
| SpCas9 | 0.064 | 0.068 | 0.088 | 0.156 | **0.263** | 0.255 |

### C.2 Common Evaluation Protocols

**Complete ESMFold2 evaluation.** Every final discrete sequence is independently re-evaluated by complete frozen ESMFold2 inference. We report Pair-chain iPTM for the designated target–binder pair, complex iPTM, pTM, complex pLDDT, and pLDDT restricted to the designed chain. For localized redesign, Peptide Pair iPTM measures the redesigned chain against the peptide, while Mean Pair iPTM averages its pairwise confidence against all fixed context chains. The verifier loss is defined in Eq. 20. Threshold yields are used only as reporting statistics and do not replace continuous metrics.

**Independent AF2-IG evaluation.** AF2-IG independently predicts each final sequence and is never used for ProteoCraft proposal generation, tree selection, or reward assignment. We report interface predicted alignment error (iPAE; lower is better), iPTM, pTM, and pLDDT where available. Because AF2-IG is outside the search loop, agreement between ESMFold2 and AF2-IG provides evidence that improvements are not restricted to the search-time verifier.

**Rosetta interface evaluation.** All methods use the same post-generation Rosetta pipeline. Each candidate is converted to a common manifest containing its method, case, binder sequence, target structure, binder length, and segment metadata, followed by standardized side-chain refinement and interface analysis. We report interface energy $\Delta G$ (more negative is better), buried solvent-accessible surface area $\Delta\mathrm{SASA}$, Lawrence–Colman shape complementarity (SC), $\Delta G/\Delta\mathrm{SASA}$, interface hydrogen bonds, and Packstat. Rosetta does not participate in sequence generation or tree search.

### C.3 Supplementary Results

**Per-target ESMFold2 interface confidence.** Table 11 reports mean Pair-chain iPTM over the 15 candidates for every target. The ProteoCraft-MCTS column uses the final frozen 210-candidate cohort, independently evaluated with ESMFold2 using 20 recycling loops and 200 diffusion sampling steps. Among the methods listed in the table, ProteoCraft-MCTS obtains the highest mean on 13 of 14 settings; Best-of-$N$ achieves the highest mean on SpCas9.

**Runtime Analysis** Runtime measurements cover 13 shared cases, comprising eight binder design cases and five TCR cases. We measure the wall-clock time required by each method's native

Table 12: Runtime comparison across different targets, reported in seconds per binder. Averages are computed separately over eight de novo and five TCR cases. Besides, — indicates no filter-passing designs within the TCR sampling budget for BindCraft v2.

| Case | Total Length | PXDesign-d | RFdiffusion3 | Complexa BoN+AF2 | BoltzDesign1 | BindCraft v2 | ProteoCraft |
|---|---|---|---|---|---|---|---|
| *De novo binder design* | | | | | | | |
| IFNAR2 | 199 | 13.53 | 49.21 | 14.28 | 458 | 512 | 11.62 |
| PD-L1 | 208 | 14.13 | 49.40 | 14.10 | 504 | 303 | 13.54 |
| PD1 | 220 | 16.40 | 63.64 | 12.05 | 505 | 1128 | 14.96 |
| DerF21 | 257 | 11.49 | 56.46 | 15.05 | 522 | 253 | 17.17 |
| BetV1 | 261 | 14.40 | 48.17 | 16.05 | 686 | 11523 | 19.76 |
| DerF7 | 311 | 18.77 | 55.11 | 17.07 | 782 | 1381 | 36.97 |
| HER2 | 322 | 28.93 | 70.43 | 13.91 | 1021 | 681 | 45.23 |
| SpCas9 | 369 | 32.13 | 60.69 | 14.45 | 1069 | 2791 | 99.50 |
| **Average** | — | **18.72** | **56.64** | **14.62** | **693.38** | **2321.50** | **32.34** |
| *TCR-facing binder design* | | | | | | | |
| 5BS0 | 526 | 19.74 | 42.79 | 33.55 | 2648 | — | 52.60 |
| 3QDJ | 529 | 19.92 | 43.29 | 31.98 | 2616 | — | 52.32 |
| 1QRN | 531 | 20.13 | 46.76 | 33.58 | 2698 | — | 54.30 |
| 6BJ8 | 531 | 20.10 | 42.33 | 33.29 | 2680 | — | 115.02 |
| 9C3E | 621 | 25.17 | 46.15 | 40.67 | 3769 | — | 596.08 |
| **Average** | — | **21.01** | **44.26** | **34.61** | **2882.20** | — | **174.06** |

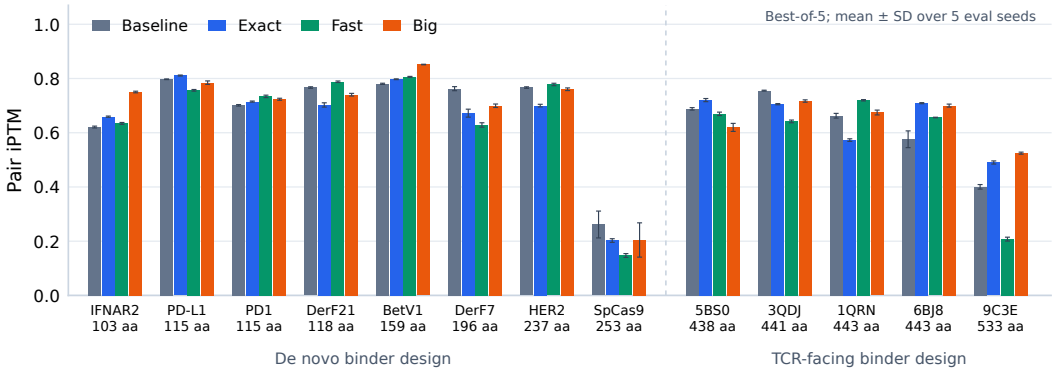



Figure 5: **Design quality with accelerated ESMFold2 inference.**

generation and selection procedure to produce a final binder in Table 12. For trajectory-based hallucination methods, the runtime includes executing the design trajectory and selecting the lowest-loss binder produced by that trajectory. For diffusion-based methods, it is the runtime to generate a binder directly. Thus, the comparison reflects the end-to-end time to obtain one final design under each method's native procedure. Specifically, BoltzDesign1 timings are per generated binder without native redesign or AF3 filtering. For BindCraft v2, timings are summed single-GPU generation durations divided by designs passing its internal filters. ProteoCraft timings are cumulative seconds divided by unique sequence candidates with search-time overall iPTM $> 0.70$, including intermediate candidates, rather than only final outputs.

**Design quality after inference optimization.** Figure 5 compares candidates obtained with different ESMFold2 inference configurations, all re-evaluated using unoptimized ESMFold2. Mean pair-chain iPTM across 13 targets is 0.657 for Baseline, 0.651 for Exact, 0.628 for Fast, and 0.673 for Big. Although individual targets exhibit both improvements and declines, aggregate scores remain broadly comparable. In particular, Big combines a $3.57\times$ mean search speedup with a slightly higher mean pair-chain iPTM than Baseline, while Fast achieves $3.68\times$ speedup with a modest reduction. These results suggest that inference optimization can improve search efficiency while retaining comparable average design quality, although multiple search seeds are needed to establish the robustness of this trade-off.

Table 13: **Sensitivity to the one-step diffusion noise level.** All settings use the same three targets and matched search configuration.

| $\tau$ | Pair iPTM ↑ | iPTM ↑ | pTM ↑ | pLDDT ↑ |
|---|---|---|---|---|
| 1 | 0.682 | 0.871 | 0.896 | 0.886 |
| 2 | **0.777** | **0.930** | **0.939** | **0.921** |
| 4 | 0.730 | 0.914 | 0.931 | 0.909 |

**One-step noise level.** The default One-step diffusion noise level is $\tau = 2$. We vary only $\tau \in \{1, 2, 4\}$ on three matched *de novo* targets. The default $\tau = 2$ gives the strongest mean result across all four metrics, while performance remains competitive at both smaller and larger noise levels.