ByteDance | Seed

PXDesign: Fast, Modular, and Accurate De Novo Design of Protein Binders

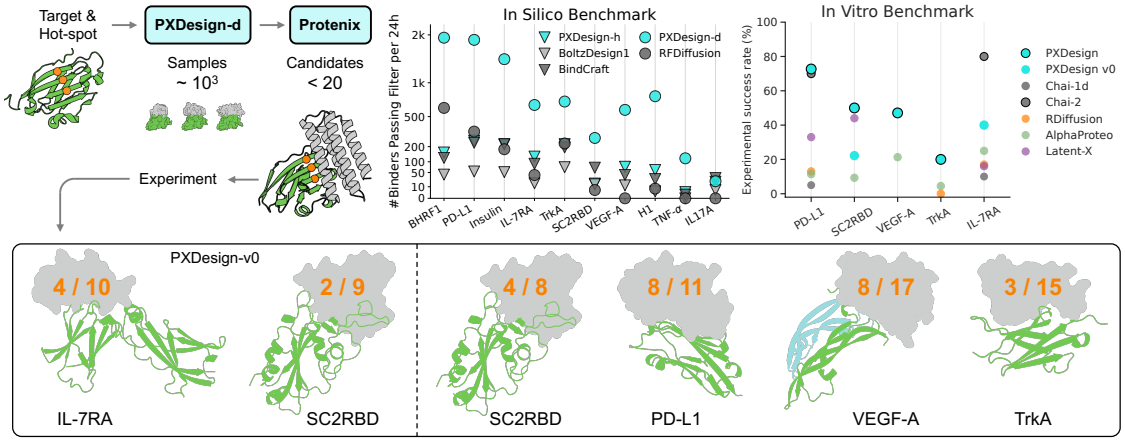
Protenix Team[1]

[1]ByteDance Seed

Abstract

PXDesign achieves nanomolar binder hit rates of 20–73% across five of six diverse protein targets, surpassing prior methods such as AlphaProteo. This experimental success rate is enabled by advances in both binder generation and filtering. We develop both a diffusion-based generative model (PXDesign-d) and a hallucination-based approach (PXDesign-h), each showing strong *in silico* performance that outperforms existing models. Beyond generation, we systematically analyze confidence-based filtering and ranking strategies from multiple structure predictors, comparing their accuracy, efficiency, and complementarity on datasets spanning *de novo* binders and mutagenesis. Finally, we validate the full design process experimentally, achieving high hit rates and multiple nanomolar binders.

To support future work and community use, we release a unified benchmarking framework at https://github.com/bytedance/PXDesignBench, provide public access to PXDesign via a webserver at https://protenix-server.com, and share all designed binder sequences at https://protenix.github.io/pxdesign.



Date: September 2, 2025

Correspondence: Xinshi Chen and Wenzhi Xiao at chenxinshi,xiaowenzhi@bytedance.com

Contents

1 Introduction . . . 3

2 Filtering and Ranking: Accuracy, Efficiency, and Diversity . . . 3
- 2.1 Filter Design and Setup . . . 4
- 2.2 Main Findings . . . 5

3 Generators and In silico Benchmarking . . . 5
- 3.1 Development of Diffusion and Hallucination . . . 7
- 3.2 *In silico* Benchmarks . . . 7
  - 3.2.1 Unconditional Protein Monomer Benchmark . . . 7
  - 3.2.2 Conditional Protein Binder Benchmark . . . 8

4 In vitro Experiments . . . 8

5 Limitations and Challenges . . . 9

6 From Protein Binders to a Unified Model for Molecular Design . . . 11
- 6.1 Demonstration Across Modalities . . . 11
- 6.2 Cyclic Peptide Binder Benchmark . . . 11

7 Discussion . . . 12

8 Contributions and Acknowledgements . . . 18

A Filtering Methodology and Benchmark Evaluation . . . 19

B PXDesign-h (hallucination) Details . . . 21

C PXDesign-d (diffusion) Details . . . 23

D Generator Benchmarking Details . . . 24

E Details of Wet-lab Results . . . 25

F Cyclic Peptide Details . . . 26

1 Introduction

PXDesign is a model suite for de novo protein-binder design that has been validated on six targets, achieving nanomolar binder hit rates of 20–73% on five of them and outperforming strong baselines such as AlphaProteo [58] by 2–6×. These results reflect a systematic approach targeting two essential aspects of binder design success: proposing structurally complementary candidates, and prioritizing those most likely to bind effectively.

While recent methods have advanced both stages, important gaps remain. Diffusion-based generators [24, 28, 53] and hallucination-based optimization methods [3, 13, 20, 35, 38, 39, 54] have shown promise in proposing viable binder candidates, but direct head-to-head comparisons under consistent evaluation are lacking. Likewise, confidence scores such as pLDDT or interface pTM from the AlphaFold series and related variants [1, 2, 14, 19, 27, 30, 32, 32, 40, 46, 55, 56] are widely used for filtering, but important questions remain: How accurate and generalizable are these scores? Can alternative predictors improve enrichment, diversity, and yield? Addressing these questions is essential to building a robust, high-hit-rate method.

Here, we present PXDesign, a unified framework integrating both binder generation and confidence-based filtering. Our main contributions are:

- **Generation via two strategies**: We develop both a diffusion-based (PXDesign-d) and a hallucination-based (PXDesign-h) generation method, with architectural and algorithmic enhancements for scalability and efficiency. Each achieves state-of-the-art *in silico* performance, and our head-to-head comparison reveals their respective strengths. We also show our diffusion model's strong capacity for diverse, designable protein generation in unconditional tasks.
- **Filtering strategies**: We construct and evaluate confidence-based filters from Protenix [12] and AlphaFold-based models [1, 27] using datasets including Cao data [11], RFDiffusion [53], EGFR [16] and SKEMPI [25, 34, 36]. Our results highlight the value of alternative predictors and show how filtering strategies influence diversity and yield in complementary ways.
- **Experimental validation across targets**: We evaluated PXDesign on six diverse protein targets via *in vitro* expression and binding assays. PXDesign achieves high hit rates across five targets (Table 1). Notably, a substantial improvement the initial version (PXDesign v0) is observed on SC2RBD, where the hit rate increased from 2 out of 9 to 4 out of 8. PXDesign outperformed AlphaProteo [58] across all evaluated targets, achieving 2- to 6-fold higher hit rates on IL-7RA, PD-L1, VEGF-A, SC2RBD, and TrkA.

Altogether, PXDesign delivers both high *in vitro* success rates and methodological insights, guiding future applications and enabling reproducible evaluation through open-sourced tools and a public webserver.

Table 1 Experimental hit rates (%) for designed binders across different targets and methods. "v0" indicates results from the initial PXDesign version. See Appendix E for detailed explanations.

Method	IL7RA	PD-L1	VEGF-A	SC2RBD	TrkA	TNF-α
PXDesign	–	**72.7**	**47.1**	<u>50.0</u>	**20.0**	0.0
PXDesign v0	<u>40.0</u>	–	–	22.2	–	–
RFDiffusion [53]	17.0	13.0	–	–	0.0	–
AlphaProteo (by HTRF) [58]	25.0	11.4	21.3	9.3	4.5	0.0
Chai-1d [47]	10.0	5.0	–	–	–	–
Chai-2 [47]	**80.0**	<u>70.0</u>	–	–	–	**21.0**
Latent-X (by display) [9]	26.0	49.0	–	**52.0**	10.0	–
Latent-X (by HT_BLI) [9]	16.0	33.0	–	44.0	–	–

2 Filtering and Ranking: Accuracy, Efficiency, and Diversity

Filtering and ranking are essential for turning large sets of generated designs into a focused shortlist of promising candidates. In this section, we evaluate **AF2-IG** [53] and **Protenix** variants (full, Mini, and Mini-Templ) [23] across multiple datasets to answer three key questions:

1. **Accuracy**: Which filters enrich true binders most effectively?
2. **Efficiency**: Can we reduce computational cost without sacrificing positive predictive value?
3. **Diversity**: Do different filters capture complementary regions of design space?

2.1 Filter Design and Setup

We constructed confidence-based filters using per-sequence and interface-level scores from AF2-IG and Protenix variants. All Protenix models use a 2-step ODE diffusion sampler for efficiency [23]. Thresholds were tuned via

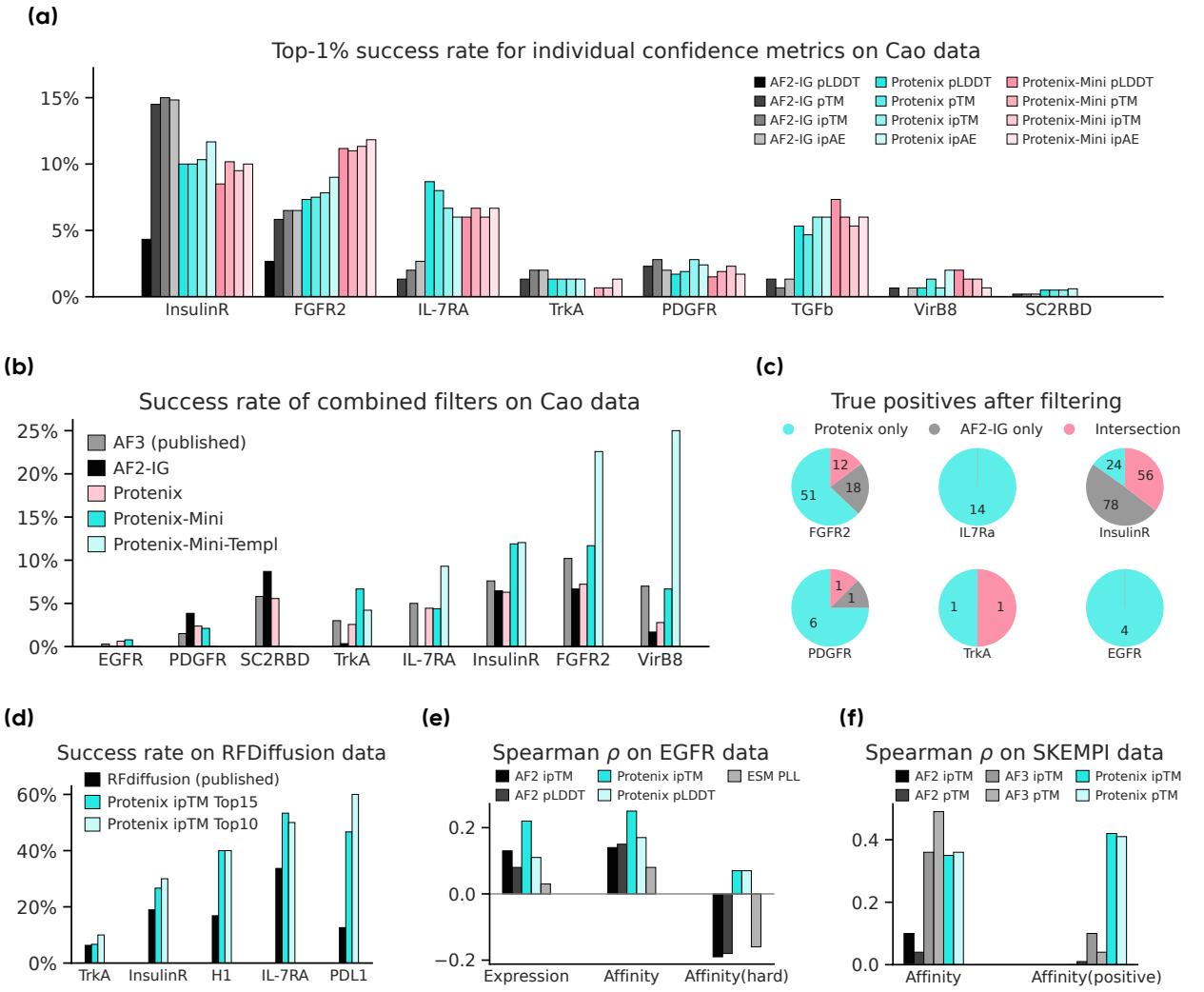


Figure 1 Filtering and ranking power comparison. (a) Top-1% success rates for individual confidence scores derived from AF2-IG, Protenix, and Protenix-Mini(-Templ) on Cao data. **(b)** Combined filter success rates across targets. "AF3 (published)" refers to results reported by Zambaldi et al. [58], while other results are computed via our unified pipeline. Filter thresholds for each model are listed in Table 2. **(c)** Venn diagrams show limited overlap between true positives retained by Protenix and AF2-IG filters, highlighting complementary design space coverage. **(d)** Re-ranking RFDiffusion binders [53] using Protenix ipTM improves success rates across all targets. RFDiffusion [53] generated 95 binder designs per target across five targets. "RFDiffusion (published)" shows the original experimental success rates, based on filtering with AF2-IG. "Protenix ipTM Top10(15)" reports success rates after re-ranking the same designs by Protenix ipTM and selecting the top 10 or 15, consistently improving hit rates across all targets. **(e)** On the EGFR competition dataset [16], Protenix better ranks expression and affinity, especially among binders with measurable affinity ("Affinity(hard)"). **(f)** On the subset of SKEMPI2.0 [25, 34], Protenix outperforms AF2 and matches AF3 [34] in ranking correlation. On positive-binding mutations ("Affinity(positive)"), Protenix shows pronounced advantage.

Table 2 Thresholds for combined filters. Each filter represents a model-specific combination of score thresholds used for binary selection. "AF2-IG-easy" reflects thresholds proposed by BindCraft [38]. "AF2-IG" denotes thresholds selected via our own grid search on Cao data; these match the values independently reported in Zambaldi et al. [58]. "AF3" refers to thresholds grid-searched in Zambaldi et al. [58]. "Protenix" denotes a unified threshold set applied to Protenix, Protenix-Mini, and Protenix-Mini-Templ, while "Protenix-basic" represents a relaxed criterion used on challenging targets (e.g., TNF-α) in wet-lab experiments to preserve diversity.

Filter Name	Confidence Thresholds	Structure Thresholds
AF3 [58]	min ipAE < 1.5, binder pTM > 0.8	complex RMSD < 2.5 Å
AF2-IG [58]	ipAE < 7.0, pLDDT > 0.9	binder RMSD < 1.5 Å
AF2-IG-easy [38]	ipAE < 10.85, ipTM > 0.5, pLDDT > 0.8	binder bound/unbound RMSD < 3.5 Å
Protenix [12]	binder ipTM > 0.85, binder pTM > 0.88	complex RMSD < 2.5 Å
Protenix-basic [12]	binder ipTM > 0.8, binder pTM > 0.8	complex RMSD < 2.5 Å

grid search on Cao data [11], a retrospective benchmark of experimentally tested designs without AlphaFold-based pre-filtering. The search process is described in Appendix A, and Table 2 summarizes the resulted thresholds.

We also tested ranking and enrichment performance on three additional datasets:

- EGFR challenge [15]: expression and binding phenotypes for de novo binders from Adaptyv Bio.
- SKEMPI subset [25]: point mutations with measured ddG, curated to retain high-confidence AF3 structures by Lu et al. [34].
- RFdiffusion wet-lab set [53]: binary binding outcomes for published designs.

2.2 Main Findings

(1) Accuracy — Protenix achieves higher binder enrichment. On Cao data, Protenix-derived confidence metrics outperform AF2-IG across most targets when used individually (Figure 1a) or in combination (Figure 1b). AF2-IG performs better on a few specific targets, but overall Protenix's precision and AUC are higher (see Appendix A). On the EGFR challenge, Protenix scores correlate more strongly with experimental expression and binding affinity than AF2 or ESM [30, 44] (Figure 1e). On the SKEMPI subset, Protenix matches or exceeds AF3's performance, especially in the subset of positive binding ddG mutations (Figure 1f). Importantly, Protenix-based filtering can also improve hit rates on previously published designs. Similarly, in re-ranking the 95 RFdiffusion designs [53], replacing AF2-IG with Protenix ipTM for top 10 or 15 selection substantially increases observed success rates (Figure 1d).

(2) Efficiency — strong performance at lower cost. Protenix-Mini and Protenix-Mini-Templ achieve significant runtime reductions compared to the Protenix full model while maintaining comparable enrichment quality, enabling practical large-scale screening and rapid triaging of candidate pools.

(3) Diversity — complementary coverage of design space. On Cao data, the overlap between sequences retained by Protenix and AF2-IG is surprisingly small (Figure 1c). Each model captures different subsets of true positives, suggesting that their inductive biases are complementary. This motivates the use of multiple predictors to improve coverage and robustness in real-world applications.

3 Generators and In silico Benchmarking

Among various generative strategies explored for protein binder design, two have emerged as the most widely adopted and experimentally validated:

- **Diffusion (training-based):** Generative models, typically denoising diffusion models, trained to sample binder structures conditioned on the target. These approaches require either fine-tuning or training from scratch. Examples include RFDiffusion [28, 53], GeoFlow-v2 [45], Chai-2 [47], Latent-X [9], etc.

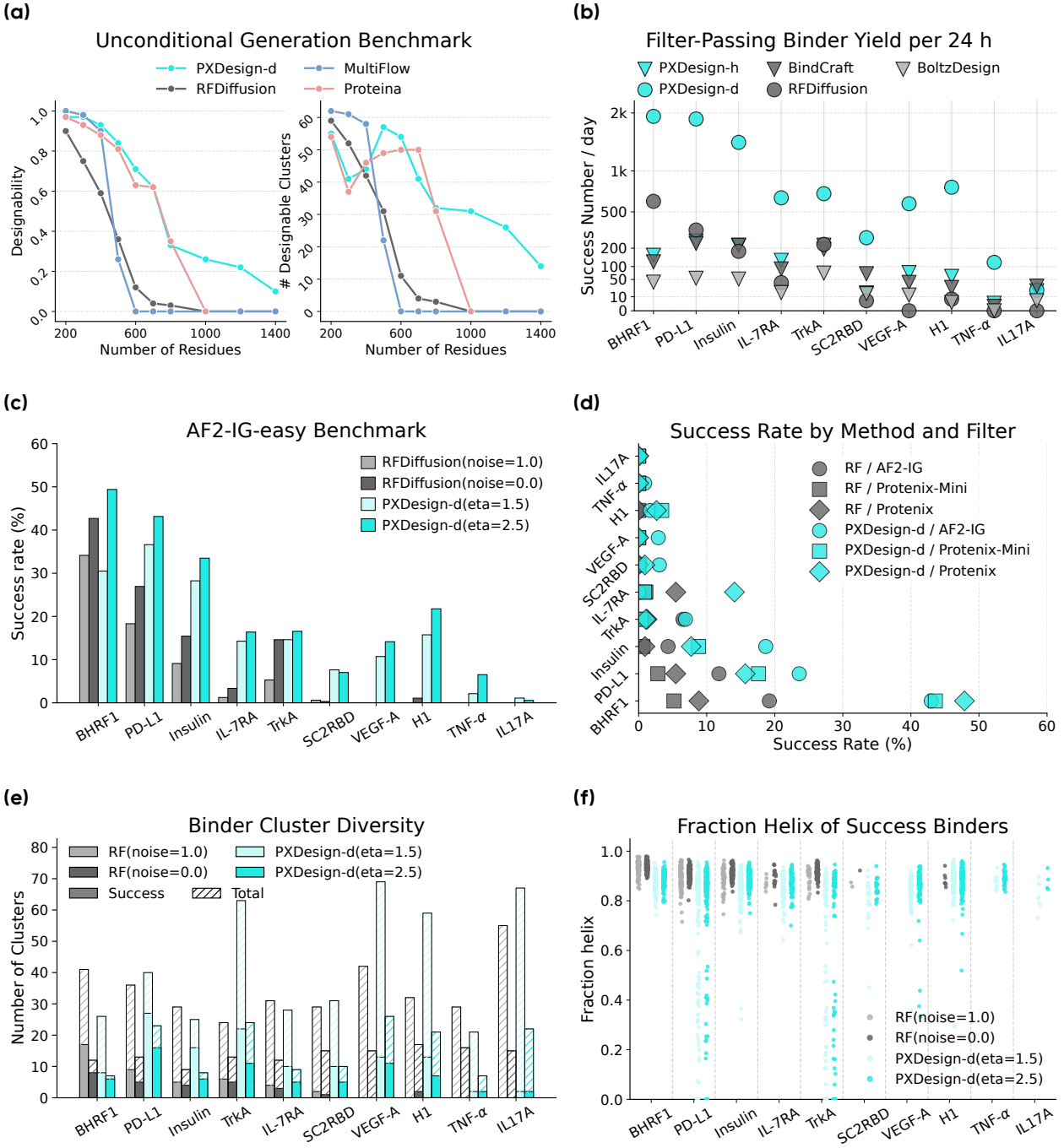


Figure 2 In silico evaluation of PXDesign on unconditional protein monomer design and conditional protein binder design tasks. (a) Unconditional monomer benchmark. Left: The ratio of designable proteins (scRMSD $<$ 2). Right: The number of desigable structure clusters (TM-score $<$ 0.5). **(b)** The time required for PXDesign-h and PXDesign-d to generate one sample passing AF2-IG-easy under default settings respectively. **(c)** Success rate of binders under the AF2-IG-easy filter defined in Table 2 across different methods on 10 representative protein targets. **(d)** Success rate of binders across different methods and filters defined in Table 2. **(e)** The number of structure clusters (TM-score $<$ 0.5) of all binders and successful binders under AF2-IG-easy filter. **(f)** The fraction α-helix of successful binders.

- **Hallucination (training-free):** Direct sequence optimization via backpropagation through a frozen structure predictor, targeting high-confidence scores such as pLDDT or ipTM. Representative tools include BindCraft [38, 39] and BoltzDesign1 [13].

Both strategies have shown strong empirical results, yet few studies have systematically compared them under matched conditions. Here, we present a side-by-side evaluation of diffusion- and hallucination-based generation, with both methods implemented and optimized in-house to enable a fair comparison. In this section, we demonstrate that each method achieves state-of-the-art performance on *in silico* metrics, outperforming existing baselines. Building on this foundation, we compare the two strategies directly, highlighting their respective strengths and trade-offs for general-purpose protein binder design.

3.1 Development of Diffusion and Hallucination

Diffusion. We developed the protein design model, named PXDesign-d, built upon the Protenix structure prediction framework. PXDesign-d supports accurate, rapid, and programmable structure-based protein binder design. Details are described in Appendix C. Several key enhancements distinguish our approach:

- Efficient architecture: Unlike prior diffusion-based methods that rely on $SE(3)$-equivariant models or AlphaFold2-style frame representations [28, 53, 58], PXDesign-d directly generates Cartesian atom coordinates. Inspired by AlphaFold3, it employs a Diffusion Transformer (DiT) backbone without using expensive triangle updates during diffusion. This architectural simplification offers several-fold speedup in long-sequence generation while maintaining high structural fidelity, enabling scalable virtual screening and design of large proteins.
- Unified multi-target training: Although this work focuses on designing protein binders for protein targets, PXDesign-d is trained to support a wide range of molecular target types, including proteins, small molecules, nucleic acids (DNA/RNA), and post-translational modifications.
- Controllable generation: PXDesign-d supports a wide range of conditional inputs, including multiple sequence alignments (MSAs), structural priors from the target, target-specific hotspots, and user-defined preferences such as secondary structure or solvent-accessible surface area (SASA). While all results in this work were generated without applying user-defined preferences, these capabilities make PXDesign-d adaptable for future tasks requiring controllable generation.

Hallucination. We implemented PXDesign-h, a gradient-based sequence optimization pipeline using frozen Protenix predictors (Appendix B). Several enhancements are introduced to improve this approach:

- End-to-end differentiation: AlphaFold3-style models use a 200-step diffusion decoder by default, making backpropagation infeasible. We reduce this to a 2-step ODE-based sampler in Protenix [12], enabling efficient end-to-end differentiation through the entire structure prediction process.
- Protenix-Mini for faster optimization: We develop Protenix-Mini [23], a lightweight variant that offers similar predictive performance to the full model but significantly faster runtime, making it ideal for iterative sequence optimization.
- Ensemble for robustness: To improve generalization and avoid overfitting to any single predictor, we optimize sequences against an ensemble of five Protenix-based models, sampled randomly at each step.

3.2 In silico Benchmarks

We evaluate our model under two distinct settings: **unconditional protein monomer generation** and **conditional protein binder generation**. The unconditional generation task serves to demonstrate the model's general capability to generate diverse and structurally realistic proteins without external constraints. In contrast, the conditional generation task reflects real-world applications in binder design, where the goal is to generate protein sequences and structures that can bind to a given target with high affinity and specificity.

3.2.1 Unconditional Protein Monomer Benchmark

PXDesign-d outperforms or matches prior baselines (RFdiffusion [53], MultiFlow [10], Proteina [22]) in both designability and diversity across various lengths up to 1400 residues, with the performance gap widening as sequence length increases. Despite being trained only with a crop size of 640 residues, PXDesign-d maintains

strong structural fidelity and diversity even for sequences exceeding 1000 residues, where competing methods show substantial degradation (Figure 2a). We note that recent works [17, 21], released after our benchmarking was completed, report improved performance on similar tasks. These results reflect the rapid progress in the field. Detailed datasets, evaluation metrics, and sampling hyperparameters are provided in Appendix D.

3.2.2 Conditional Protein Binder Benchmark

To extensively evaluate our method, following previous works [58], we use 10 protein targets with diverse structural properties as test set. These targets are not only biologically important, but also cover the difficulty of successfully designing binders for the proteins in the Protein Data Bank (PDB). Full benchmarking protocols, including dataset definitions, evaluation metrics, and runtime efficiency comparisons with hallucination-based methods, are described in Appendix D.

Quality. Across this benchmark, PXDesign-d consistently achieves higher success rates than RFdiffusion on the AF2-IG-easy criterion (Figure 2c). When applying alternative filters such as AF2-IG, Protenix-Mini, and Protenix (Figure 2d), PXDesign-d maintains both higher mean success rates and broader target coverage, with notable advantages on challenging cases like VEGF-A, IL17A, and TNF-α.

Structural Diversity. PXDesign-d generates a greater number of distinct structural clusters than RFdiffusion for nearly all targets (Figure 2e). Analysis of secondary structure composition (Figure 2f) reveals that PXDesign-d produces binders spanning a broad range of α-helix content, while RFdiffusion outputs are heavily α-helix-biased. This indicates broader coverage of fold space and greater structural versatility.

Diffusion vs. Hallucination. We directly compare our diffusion-based PXDesign-d with hallucination-based approaches, including our Protenix-powered PXDesign-h, BindCraft, and BoltzDesign1, under identical evaluation settings. Runtime analysis (Figure 2g) shows that PXDesign-d delivers more successful designs within 24h than any hallucination method, owing to its faster generation speed and higher pass rates. While hallucination remains competitive for targeted, small-scale optimization, diffusion is better suited for large-scale, exploratory campaigns.

Given these advantages, all subsequent wet-lab experiments and the PXDesign webserver deployment are based on PXDesign-d (diffusion).

4 In vitro Experiments

We experimentally validated PXDesign on six protein targets: Interleukin-7 Receptor Alpha (IL-7RA), SARS-CoV-2 receptor-binding domain (SC2RBD), Programmed Death-Ligand 1 (PD-L1), Tropomyosin receptor kinase A (TrkA), Vascular Endothelial Growth Factor A (VEGF-A), and Tumor Necrosis Factor-α (TNF-α), chosen for both biological relevance and diverse design challenges (Table 3).

For each target, we used the PXDesign-d to generate a diverse set of *in silico* designs (60–160 aa) and applied

Table 3 Targets and experimental results of PXDesign-d binders. For SC2RBD and IL-7RA, "v0" indicates results from the initial PXDesign version. A binder is counted as successful if $K_D < 1000$ nM.

Target	PDB ID	Crop	Hot-spot	Candidates Tested	Successful Binders
IL-7RA	3DI3	B17-209	B58, B80, B139	10 (v0)	4 (v0)
SC2RBD	6M0J	E333-526	E485, E489, E494, E500, E505	9 (v0) + 8	2 (v0) + 4
PD-L1	5O45	A17-132	A56, A115, A123	11	8
TrkA	1WWW	X282-382	X294, X296, X333	15	3
VEGF-A	1BJ1	V14-107, W14-107	W81, W83, W91	17	8
TNF-α	1TNF	A12-157, B12-157, C12-157	A31, A32, A113, C73, C87	16	0

a filtering pipeline: Candidates were required to pass both AF2-IG and Protenix filters (the Protenix ipTM cutoff was relaxed to 0.80 for VEGF-A, SC2RBD, and TNF-α to maintain diversity). Passing designs were clustered by structural similarity using Foldseek [48], and the highest-ranked representative by Protenix ipTM from each cluster was selected, yielding 8–18 candidates per target.

Selected binders were expressed in an *E. coli* cell-free system with an N-terminal Strep-tag and purified via Strep-Tactin affinity chromatography. Designed proteins with expression yields greater than 0.3 mg/mL (determined by A280) or 0.2 mg/mL (determined by the Bradford assay[8]) were subjected to BLI affinity measurements. An initial screening using biolayer interferometry (BLI) was conducted at a single concentration of 1000 nM to identify potential binders. Candidates exhibiting response signals above 0.06 nm were selected for multi-concentration BLI assays to determine K_{D}. Binders with $K_{\mathrm{D}} < 1000$ nM were considered successful. Both expression and BLI assays were performed by GenScript (Nanjing, China), and target proteins were purchased from Sino Biological (Beijing, China).

PXDesign v0 has already achieved strong performance on IL-7RA, surpassing all baselines except Chai-2. On SC2RBD, the updated PXDesign shows a marked improvement over v0, increasing the hit rate from 22.2% to 50.0%. Moreover, on PD-L1, PXDesign achieves a hit rate of 72.7% (8 out of 11), which is comparable to Chai-2 (about 70.0%) and well above all other baselines.

PXDesign can design structurally diverse protein binders that exhibit potent target binding (Figure 3). For each target, PXDesign produces binders with distinct fold and diverse secondary structure compositions, as evident from the range of α-helical, β-strand, and mixed folds. This diversity reflects the framework's ability to explore broad structural space while still achieving high-affinity binding, enabling robustness across heterogeneous design challenges.

5 Limitations and Challenges

While confidence-based filtering shows strong potential, several challenges remain.

Score variability across targets. Confidence scores exhibit substantial variability across targets (Figure 4), which makes tuning and selecting unified thresholds challenging. We further observe that success rates among different targets have Pareto tradeoffs: improvements on one target often degrade performance on others, and Pareto improvements are difficult to achieve. For example, in Figure 4, raising the pTM threshold from 0.75 to 0.92 boosts the FGFR2 success rate from 5% to 100%, while the SC2RBD success rate falls from 0.1% to 0%. Effective filtering strategies must account for such target-specific shifts, as well as the constraints of limited evaluation datasets. Model-agreement-based approaches may help mitigate some of these weaknesses.

Dataset limitations. The Cao data contains very few strong binders, which limits its utility for evaluating high-precision filtering strategies. In particular, certain filters, such as Protenix-Mini on SC2RBD, exhibit zero success rate (SR) due to the sparsity of validated positives. While Protenix also shows low SR on the EGFR subset of Cao data (Figure 1b), it achieves substantially higher AUC and success rate on the EGFR competition dataset (Figure 9). Additionally, our wet-lab experiments indicate that intersection filters can be effective in practice, even though they appear overly stringent on Cao data. Despite these issues, Cao remains one of the only large-scale, publicly available benchmarks for binder filtering, underscoring the need for richer datasets to support fair and consistent evaluation.

Pipeline limitations. Our current pipeline relies on multiple sequential steps—generation, structural prediction, filtering, and clustering—each with its own computational cost and potential error propagation. While this modular design improves interpretability and flexibility, it can be resource-intensive for very large design spaces and may limit throughput in time-sensitive campaigns. Integrating filtering more tightly with generation or exploring early-stage, low-cost triage methods could improve scalability.

Experimental limitations. Our current experimental assessment of success rate is constrained by the throughput of the BLI assay. Incorporating display-based screening methods could substantially increase throughput, enabling the evaluation of approximately 100 designs per target. Similar to AlphaProteo, our pipeline encountered difficulty with one challenging target, TNF-α, despite achieving remarkable improvements in success rates for the other five targets. For PD-L1, six designs were excluded from testing due to low expression

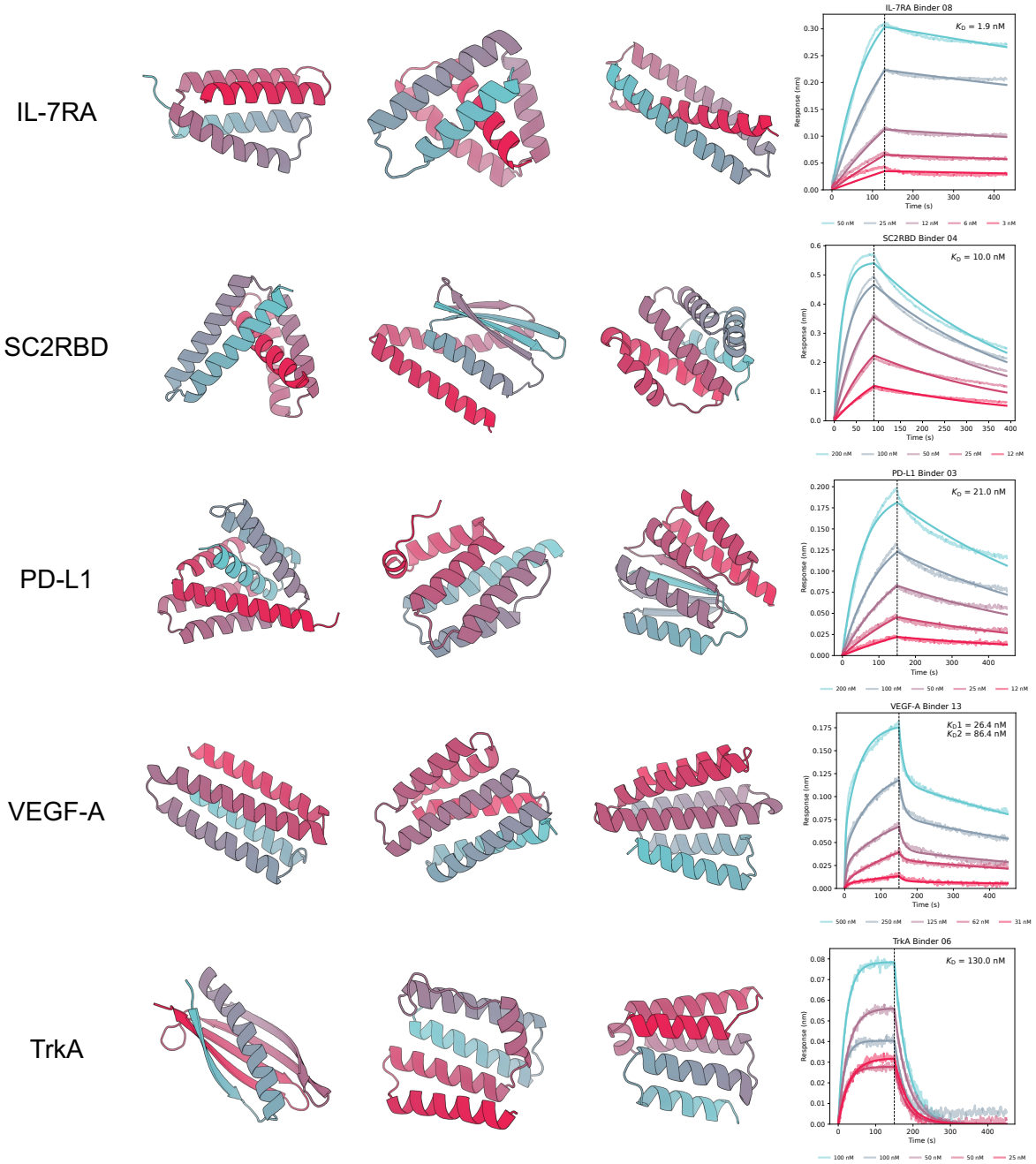


Figure 3 Designed binders and characterization. For each row, three structures of nanomolar binders were showcased. The right columns are BLI sensorgrams at multiple analyte concentrations; the vertical dashed line marks the switch from association to dissociation phases.

yields; however, nearly all of the remaining designs were confirmed to be functional binders. In addition to failures among our own designs, we occasionally encountered positive controls that did not successfully expressed. We are actively developing a robust experimental platform to enable reliable characterization of designed binders at scale.

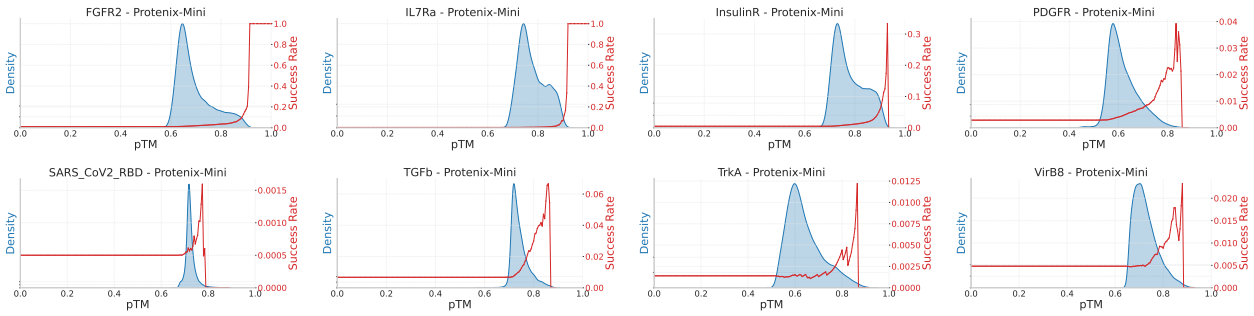


Figure 4 The Protenix-Mini pTM score distributions and success rates on different targets. The left y-axis shows the density of pTM scores, while the right y-axis displays the success rate across different pTM thresholds. Across different targets, the best success rates correspond to distinct pTM thresholds.

6 From Protein Binders to a Unified Model for Molecular Design

While our current *in silico* and wet-lab validations focus on protein binder design, the underlying framework is readily extensible to a broad range of molecular targets. Systematic benchmarking and experimental validation of these additional modalities represent important future directions.

6.1 Demonstration Across Modalities

While the primary focus of this work has been on protein binder design, our design model naturally extends beyond these tasks. Since both our diffusion and hallucination generators are built on (or closely derived from) the Protenix structure prediction framework, they inherit the ability to model diverse biomolecular targets, including nucleic acids, small molecules, and post-translationally modified proteins. In principle, the same generative–filtering pipeline can be applied to these modalities.

At present, these broader applications have not undergone the same level of rigorous evaluation as protein binders. We have conducted case studies (Figure 5 and Figure 6c) that illustrate the feasibility of generating binders for nucleic acids, small molecules, and cyclic peptides, but these remain qualitative demonstrations.

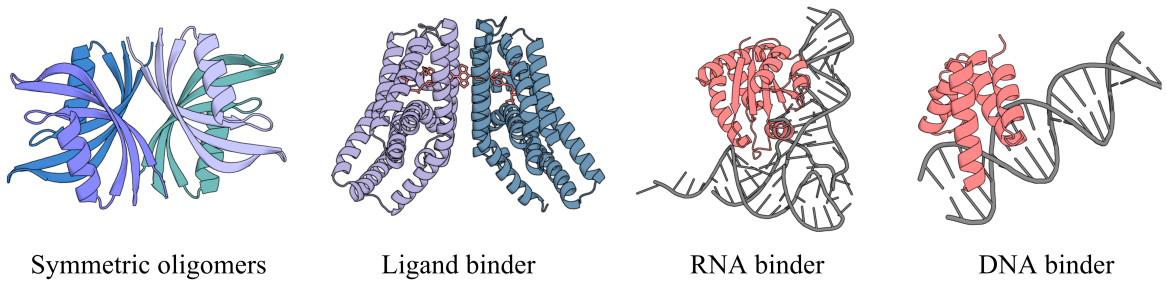


Figure 5 Case studies illustrating the versatility of our unified design model. Example designs for symmetric oligomers, ligand binder, RNA binder, DNA binder.

6.2 Cyclic Peptide Binder Benchmark

Cyclic peptides are emerging as a promising therapeutic modality due to their enhanced membrane permeability, oral bioavailability, synthetic tractability, low immunogenicity, and capacity for specific binding to protein surfaces previously considered "undruggable" [26, 29, 37, 41, 41, 42, 42, 50–52, 59]. Cyclic peptide binder design is structurally and biophysically similar to protein binder design, allowing reuse of *in silico* metrics with

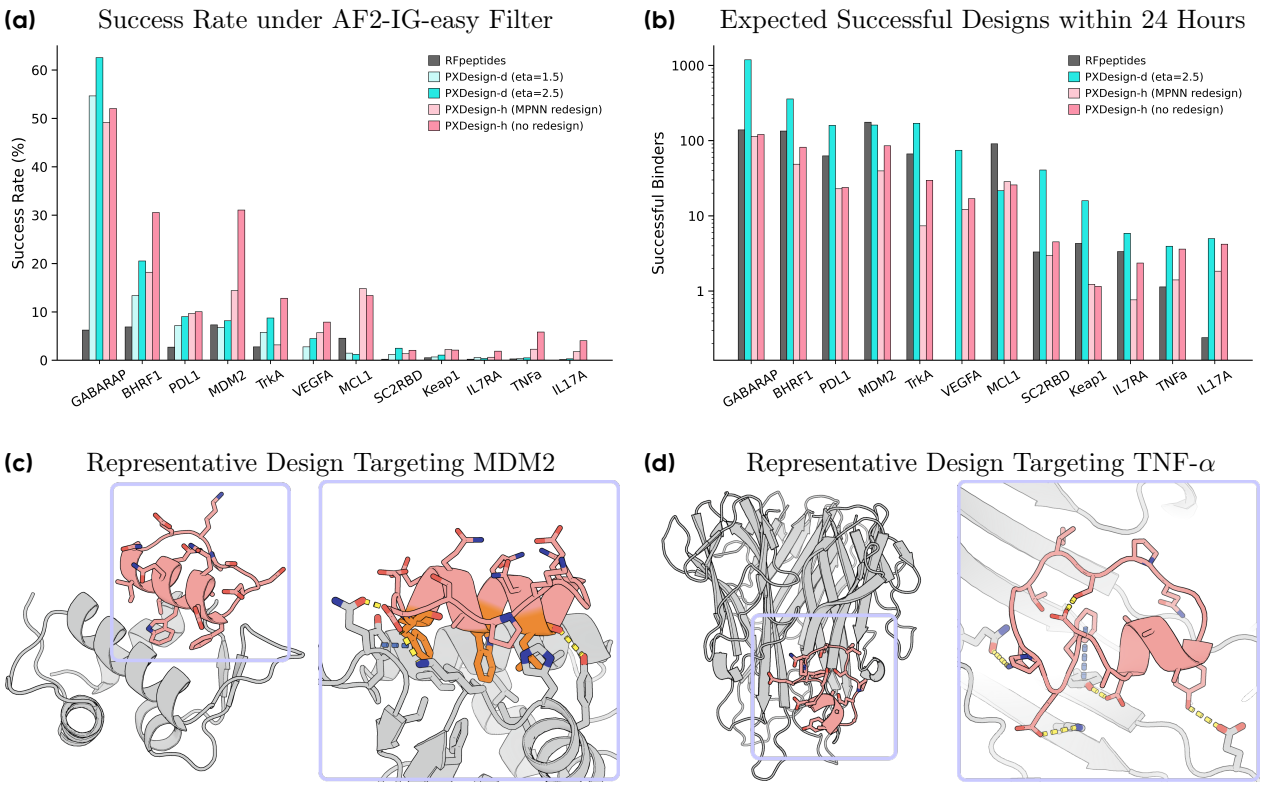


Figure 6 In silico evaluation of PXDesign on zero-shot cyclic peptide binder design. **(a)** Success rates of various methods across targets, evaluated using the AF2-IG-easy filter. **(b)** Expected number of successful cyclic peptide binders generated within a 24-hour period for each method and target. **(c)** Representative design targeting MDM2. Left: Protenix-predicted complex structure. Right: zoomed-in view of the binding interface. The conserved "F—W–L" triad is highlighted in orange. Key interactions are shown. **(d)** Representative design targeting TNF-α. Left: Protenix-predicted complex structure. Right: zoomed-in view of the binding interface. Key interactions are shown.

minimal modification. As a first cross-modality benchmark, we extended both PXDesign-d and PXDesign-h for zero-shot cyclic peptide binder generation targeting 12 diverse proteins, including AlphaProteo targets [58] and proteins with previously reported binders [41, 42], spanning sequence lengths of 8–18 residues.

Our models consistently outperform RFpeptides [41] in success rates under the AF2-IG-easy criterion (Figure 6a) and in efficiency measured by the expected number of successful designs per 24h (Figure 6b), with particularly strong gains on challenging targets such as TNF-α. While PXDesign-h is slower, it achieves superior performance on specific targets (e.g., MDM2, MCL1, IL17A, TNF-α), and its performance further improves when using native sequences without ProteinMPNN redesign—highlighting its ability to generate viable sequences directly (Figure 12).

Detailed benchmarking protocols and per-length analyses are given in Appendix F. Representative designs for MDM2 and TNF-α (Figure 6c–d) illustrate structurally plausible interfaces, with the MDM2 design recovering the conserved "F—W–L" interaction triad critical for hydrophobic binding [7].

7 Discussion

We have presented PXDesign, a unified framework for structure-based molecular design that integrates scalable diffusion-based generation (PXDesign-d) with complementary hallucination-based search, coupled to multi-filter prioritization using orthogonal structure predictors (AF2-IG and Protenix). Across extensive *in silico* benchmarks and wet-lab validation on six biologically diverse protein targets, PXDesign-d demonstrates

higher throughput, higher pass rates, and broader structural diversity than hallucination approaches, making it well-suited for large-scale exploratory campaigns. Guided by these findings, both our experimental pipeline and public web server are currently built on PXDesign-d.

Our filtering analyses reveal that no single confidence metric generalizes across all targets—success rate optima vary, and ensemble filtering provides broader design space coverage. Current public benchmarks, such as the Cao dataset, have sparse positive signal, constraining statistical resolution and sometimes misrepresenting practical performance. Improved, community-shared datasets with richer annotations are essential for accelerating method development and enabling fair comparisons.

While this work has focused on protein–protein binder design, the framework generalizes naturally to other modalities. Our case studies and initial benchmarks on cyclic peptide binders demonstrate that minimal modifications allow zero-shot generalization beyond proteins, and the same generative–filtering pipeline is in principle applicable to nucleic acids, small molecules, and post-translationally modified proteins. Scaling experimental validation to these modalities—particularly where high-quality benchmarks and structure predictors are available—is an important next step.

From a modeling standpoint, our results underscore the advantages of generative architectures that are both high-throughput and conditioning-flexible. Diffusion provides efficient large-scale sampling, while hallucination retains value for targeted, small-scale optimization. As structure prediction continues to advance, deeper integration of prediction and design—where predictors act not only as filters but also as gradient-informed guides—could enable faster, more accurate, and more interpretable workflows. Protenix already illustrates this potential by serving both as a scoring function and as a backbone for generative models.

In summary, PXDesign offers a practical and extensible pipeline for large-scale structure-based design, validated in wet-lab experiments and accessible via a public web server. By combining complementary generative paradigms, orthogonal filtering, and open benchmarks, we aim to support a broader shift toward unified, general-purpose molecular design systems that bridge computational discovery and experimental realization.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, 630(8016):493–500, 2024.

[2] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. Nature methods, 21(8):1514–1524, 2024.

[3] 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.

[4] Alexandru V Asimit, Valeria Bignozzi, Ka Chun Cheung, Junlei Hu, and Eun-Seok Kim. Robust and pareto optimality of insurance contracts. European Journal of Operational Research, 262(2):720–732, 2017.

[5] 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.

[6] 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.

[7] Angelika Böttger, Volker Böttger, Carlos Garcia-Echeverria, Patrick Chène, Heinz-Kurt Hochkeppel, Wayne Sampson, Kiah Ang, Stephanie F Howard, Steven M Picksley, and David P Lane. Molecular characterization of the hdm2-p53 interaction. Journal of molecular biology, 269(5):744–756, 1997.

[8] Marion M Bradford. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Analytical biochemistry, 72(1-2):248–254, 1976.

[9] Alex Bridgland, Jonathan Crabbé, Henry Kenlay, Daniella Pretorius, Sebastian M Schmon, Agrin Hilmkil, Rebecca Bartke-Croughan, Robin Rombach, Michael Flashman, Tomas Matteson, et al. Latent-x: An atom-level frontier model for de novo protein binder design. arXiv e-prints, pages arXiv–2507, 2025.

[10] Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In International Conference on Machine Learning, pages 5453–5512. PMLR, 2024.

[11] 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.

[12] 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, 2025. doi: 10.1101/2025.01.08.631967.

[13] 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, pages 2025–04, 2025.

[14] Nathaniel Corley, Simon Mathis, Rohith Krishna, Magnus S Bauer, Tuscan R Thompson, Woody Ahern, Maxwell W Kazman, Rafael I Brent, Kieran Didi, Andrew Kubaney, et al. Accelerating biomolecular modeling with atomworks and rf3. bioRxiv, pages 2025–08, 2025.

[15] Tudor-Stefan Cotet, Igor Krawczuk, Martin Pacesa, Lennart Nickel, Bruno E Correia, Nikhil Haas, Ahmad Qamar, Chance A Challacombe, Patrick Kidger, Constance Ferragu, et al. Crowdsourced protein design: Lessons from the adaptyv egfr binder competition. bioRxiv, pages 2025–04, 2025.

[16] Tudor-Stefan Cotet, Igor Krawczuk, Martin Pacesa, Lennart Nickel, Bruno E Correia, Nikhil Haas, Ahmad Qamar, Chance A Challacombe, Patrick Kidger, Constance Ferragu, et al. Crowdsourced protein design: Lessons from the adaptyv egfr binder competition. bioRxiv, pages 2025–04, 2025.

[17] Giannis Daras, Jeffrey Ouyang-Zhang, Krithika Ravishankar, William Daspit, Costis Daskalakis, Qiang Liu, Adam Klivans, and Daniel J. Diaz. Ambient proteins: Training diffusion models on low quality structures. bioRxiv, 2025. doi: 10.1101/2025.07.03.663105.

[18] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. Science, 378(6615):49–56, 2022.

[19] 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, pages 2021–10, 2021.

[20] Christopher Frank, Ali Khoshouei, Lara Fuβ, Dominik Schiwietz, Dominik Putz, Lara Weber, Zhixuan Zhao, Motoyuki Hattori, Shihao Feng, Yosta de Stigter, et al. Scalable protein design using optimization in a relaxed sequence space. Science, 386(6720):439–445, 2024.

[21] 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, 2025.

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

[23] Chengyue Gong, Xinshi Chen, Yuxuan Zhang, Yuxuan Song, Hao Zhou, and Wenzhi Xiao. Protenix-mini: Efficient structure predictor via compact architecture, few-step diffusion and switchable plm. arXiv preprint arXiv:2507.11839, 2025.

[24] John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. Nature, 623(7989):1070–1078, 2023.

[25] Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics, 35(3):462–469, 2019.

[26] Xinjian Ji, Alexander L Nielsen, and Christian Heinis. Cyclic peptides for drug development. Angewandte Chemie, 136(3):e202308251, 2024.

[27] 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.

[28] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with rosettafold all-atom. Science, 384(6693):eadl2528, 2024.

[29] Qiuzhen Li, Efstathios Nikolaos Vlachos, and Patrick Bryant. Design of linear and cyclic peptide binders from protein sequence information. Communications Chemistry, 8(1):211, 2025.

[30] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv, 2022.

[31] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

[32] Lihang Liu, Shanzhuo Zhang, Yang Xue, Xianbin Ye, Kunrui Zhu, Yuxin Li, Yang Liu, Jie Gao, Wenlai Zhao, Hongkun Yu, et al. Technical report of helixfold3 for biomolecular structure prediction. arXiv preprint arXiv:2408.16975, 2024.

[33] Shih-Ching Lo, Xuchu Li, Michael T Henzl, Lesa J Beamer, and Mark Hannink. Structure of the keap1: Nrf2 interface provides mechanistic insight into nrf2 signaling. The EMBO journal, 25(15):3605–3617, 2006.

[34] Wei Lu, Jixian Zhang, Jiahua Rao, Zhongyue Zhang, and Shuangjia Zheng. Alphafold3, a secret sauce for predicting mutational effects on protein-protein interactions. bioRxiv, pages 2024–05, 2024.

[35] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. Nature methods, 19(6):679–682, 2022.

[36] Iain H Moal and Juan Fernández-Recio. Skempi: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics, 28(20):2600–2607, 2012.

[37] Markus Muttenthaler, Glenn F King, David J Adams, and Paul F Alewood. Trends in peptide drug discovery. Nature reviews Drug discovery, 20(4):309–325, 2021.

[38] 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, pages 2024–09, 2024.

[39] 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, pages 1–10, 2025.

[40] 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, pages 2025–06, 2025.

[41] 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, et al. Cyclic peptide structure prediction and design using alphafold2. Nature Communications, 16(1):4730, 2025.

[42] 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, et al. Accurate de novo design of high-affinity protein-binding macrocycles using deep learning. Nature Chemical Biology, pages 1–9, 2025.

[43] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, et al. Mgnify: the microbiome sequence data analysis resource in 2023. Nucleic acids research, 51(D1):D753–D759, 2023.

[44] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. PNAS, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

[45] BioGeometry Team. Geoflow-v2: A unified atomic diffusion model for protein structure prediction and de novo design. bioRxiv, pages 2025–05, 2025.

[46] Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhonikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. BioRxiv, pages 2024–10, 2024.

[47] Chai Discovery Team, Jacques Boitreaud, Jack Dent, Danny Geisz, Matthew McPartlon, Joshua Meier, Zhuoran Qiao, Alex Rogozhnikov, Nathan Rollins, Paul Wollenhaupt, et al. Zero-shot antibody design in a 24-well plate. bioRxiv, pages 2025–07, 2025.

[48] 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.

[49] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. 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.

[50] Alexander A Vinogradov, Yizhen Yin, and Hiroaki Suga. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. Journal of the American Chemical Society, 141(10):4167–4181, 2019.

[51] Fanhao Wang, Tiantian Zhang, Jintao Zhu, Xiaoling Zhang, Changsheng Zhang, and Luhua Lai. Target-based de novo design of cyclic peptide binders. bioRxiv, pages 2025–01, 2025.

[52] Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. Signal transduction and targeted therapy, 7(1):48, 2022.

[53] 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, et al. De novo design of protein structure and function with rfdiffusion. Nature, 620(7976):1089–1100, 2023.

[54] Basile IM Wicky, Lukas F Milles, Alexis Courbet, Robert J Ragotte, Justas Dauparas, E Kinfu, S Tipps, Ryan D Kibler, Minkyung Baek, Frank DiMaio, et al. Hallucinating symmetric protein assemblies. Science, 378(6615): 56–61, 2022.

[55] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, et al. Boltz-1 democratizing biomolecular interaction modeling. BioRxiv, 2024.

[56] Jinbo Xu, Matthew Mcpartlon, and Jin Li. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nature machine intelligence, 3(7):601–609, 2021.

[57] Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. In Proceedings of the 40th International Conference on Machine Learning, pages 40001–40039, 2023.

[58] 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.

[59] Cheng Zhu, Sen Cao, Tianfeng Shang, Jingjing Guo, An Su, Chengxi Li, and Hongliang Duan. Predicting the structures of cyclic peptides containing unnatural amino acids by highfold2. Briefings in Bioinformatics, 26(3): bbaf202, 2025.

8 Contributions and Acknowledgements

Core Contributors †

Milong Ren*
Jinyuan Sun*
Jiaqi Guan
Cong Liu*[1]
Chengyue Gong

Project Lead

Xinshi Chen

Team Lead

Wenzhi Xiao

Other Contributors

Yuzhe Wang* (Cyclic peptide analysis)
Lan Wang (Engineering support)
Qixu Cai[2] (Wet-lab advising)

† Equal contribution; order after first is random.
* Work done during an internship at ByteDance.

External Affiliations

[1] University of Amsterdam, Netherlands
[2] Xiamen University, China

Acknowledgements

We thank members of our team for their support, including Wenbo Lin, Jiazheng Zhou, and Fangzhou Guo for their work on webserver development, Wenzhi Ma for guidance on data processing, and Yuxuan Zhang for assistance with internal tooling. This project builds upon the internal Protenix codebase[1], and we thank all contributors to its development.

[1] https://github.com/bytedance/Protenix

Appendix

A Filtering Methodology and Benchmark Evaluation

Filter Development

To identify general-purpose, model-specific filters, we perform a grid search over combinations of confidence scores. We construct and evaluate confidence-based filtering strategies from two structure predictors: **Protenix** [12] and **AF2-IG** [53]. AF2-IG is a variant of AlphaFold2 adapted for binder design where "IG" denotes "initial guess" [5, 53]. We also include **Protenix-Mini**, a compact variant with reduced model size, and **Protenix-Mini-Templ**, which uses the target structure as a fixed template without relying on MSA input. All Protenix models utilize a 2-step ODE sampler for efficiency, replacing the standard 200-step diffusion sampler.

We design a two-stage filtering strategy to identify high-quality protein complex predictions generated by Protenix. In the first stage, we identify informative evaluation metrics and select an optimal triplet combination based on their ranking performance across targets. In the second stage, we perform a grid search over the cutoff values of the selected metrics to determine optimal thresholds for filtering.

Stage 1: Metric Selection and Combination. We first evaluate the ranking performance of individual metrics using a Top-1% threshold classifier and compute the Enrichment Factor (EF) to measure how effectively each metric identifies high-quality predictions. For each target, we select the Top-3 metrics with the highest EF values. Based on frequency analysis, we identify eight most frequently selected metrics (some with the same EF score): binder pTM, complex pLDDT, interface pLDDT, binder ipTM, complex pTM, binder chain pLDDT, complex gpDE, and interface gpDE. From these eight metrics, we evaluate all possible triplets ($C(8,3) = 56$ combinations) as filters. For each triplet, we apply the Top-1% quantile as a threshold and classify a sample as positive only if it satisfies all three thresholds simultaneously. We compute the Success Rate (SR) for each combination on every target and rank them accordingly. The final optimal combination is determined by aggregating rankings across all targets and selecting the one that achieves the most "Top-1" positions. The final optimal combination is determined by aggregating rankings across all targets and selecting the one that achieves the highest ranking among all targets. Ultimately, we find that a simplified filter using only two metrics, binder pTM and binder ipTM, achieves comparable or better performance than the full triplet, while offering improved robustness and interpretability. Therefore, we adopt {binder pTM, binder ipTM} as the final filtering metric set.

Stage 2: Cutoff Grid Search. Finally, to refine the filtering process, we perform a grid search to determine the optimal cutoff values for the two selected metrics, binder pTM and interface ipTM. We formulate the filter selection problem as an optimization problem,

$$\max_{\mathbf{x}\in\mathcal{X}} \left(\mathrm{SR}_1(\mathbf{x}), \mathrm{SR}_2(\mathbf{x}), \ldots, \mathrm{SR}_k(\mathbf{x})\right), \tag{1}$$

in which SR_i denotes the success rate on the target i, and the set $\mathcal{X}$ is the feasible set of confidence score threshold combination. Typically, there is no feasible solution that can maximize all objective functions simultaneously. Consequently, the focus is the solutions where improving any objective cannot be achieved without deteriorating at least one other objective, which is defined as Pareto Frontier.

Definition: A solution $x_1 \in \mathcal{X}$ dominates x_2, if

$$\forall i, \mathrm{SR}_i(x_1) \geq \mathrm{SR}_i(x_2); \exists i, \mathrm{SR}_i(x_1) > \mathrm{SR}_i(x_2). \tag{2}$$

A solution $x^* \in \mathcal{X}$ is Pareto optimal if there does not exist another solution x that dominates it. The set of Pareto optimal is called Pareto Frontier.

For each threshold combination, we compute the success rate (SR) on each target. Following the definition of Pareto Frontier, the search algorithm can come to a set of optimal points (as demonstrated in Figure 7). To distinguish the final solution, we tend to the solution which has minimal shifts across different targets, known as robust selection or risk aversion policy [4]. We calculate the rank of each combination's SR within each

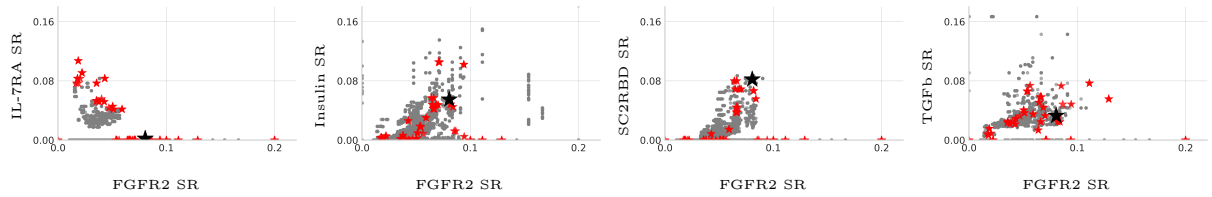


Figure 7 The performance of the AF2 confidence score filters. The SR for each confidence combination is plotted as one gray dot. The pareto frontier filters are highlighted as red stars, and the selected one is marked as a black star.

target and then take the average rank across all targets. This average rank serves as the overall "score" for the threshold combination. As demonstrated in Figure 7c, we come to a balanced SR solution on FGFR2 and SC2RBD. Ultimately, we select the combination with the highest score as the final filtering criteria.

Additional Benchmark Results

Per-Score Filter Evaluation

To complement the Top-1% SR analysis in the main text (Figure 8), we provide additional evaluation of individual confidence scores using two standard ranking metrics: AUC (area under the ROC curve) and AP (average precision, or area under the PR curve). These metrics reflect how well each score discriminates binders from non-binders across a range of thresholds, independent of any fixed selection cutoff.

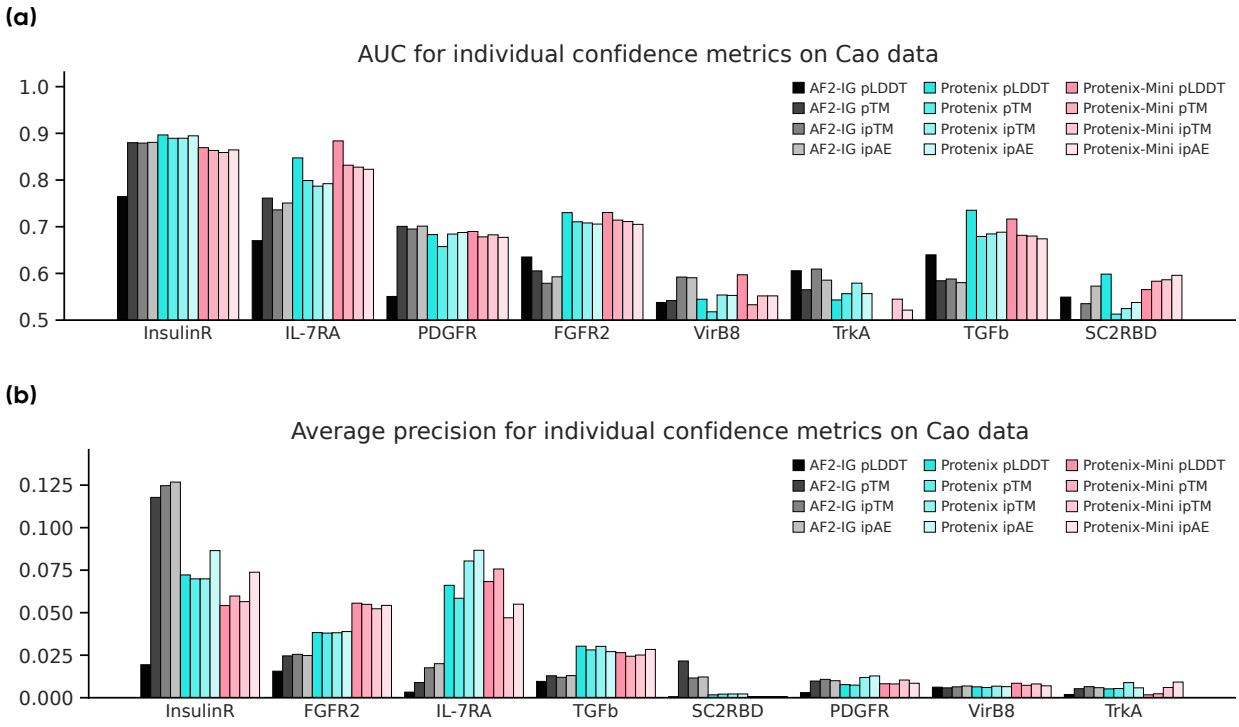


Figure 8 AUC and Average Precision scores for individual confidence metrics on Cao data. (a) Higher values indicate better global discrimination between binders and non-binders. **(b)** Similar to AUC, but more sensitive to top-ranking false positives.

We report AUC and average precision scores for individual confidence metrics across diverse design targets (Figure 8). These results are generally consistent with the Top-1% SR trends, reinforcing that Protenix-derived scores outperform AF2-based scores in most cases. However, no single metric is universally optimal—performance varies by target and model. For instance, AF2 shows stronger precision on Insulin and

SC2RBD, as well as higher AUC on TrkA.

Ranking Accuracy

We further assess whether confidence scores can effectively prioritize designs by binding strength, beyond binary filtering. As shown in Figure 9, Protenix-derived scores consistently achieve higher AUCs across two distinct datasets: **(a)** On the EGFR binder challenge [16], which includes 400 designed binders with expression and binding affinity annotations, Protenix outperforms AF2 and ESM across multiple ranking metrics. **(b)** On the SKEMPIv2 subset filtered by AF3 [34], Protenix matches or exceeds AF3 in AUC overall and shows particularly strong performance on affinity-increasing mutations.

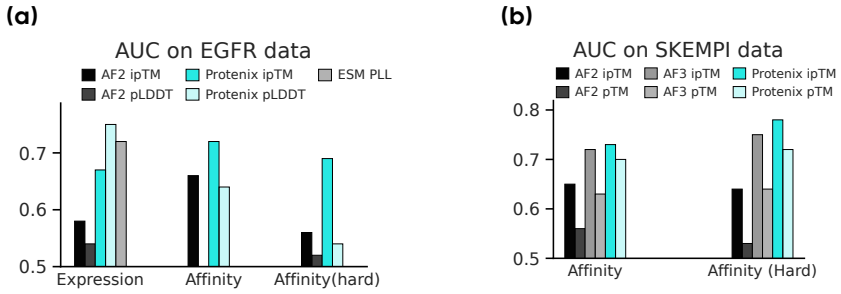


Figure 9 AUC on EFGR and SKEMPI data.

B PXDesign-h (hallucination) Details

Protenix, as a structure prediction model, not only excels at predicting complex protein structures but also provides reliable confidence estimates for those predictions. By treating the input binder sequence logits as trainable parameters, we can leverage these confidence scores to perform backpropagation through Protenix, effectively navigating the sequence space to discover high-quality binder candidates.

The hallucination process consists of two key components: (1) backbone models that supply gradient signals for updating the sequence parameters, and (2) optimization strategies that effectively utilize this gradient information. In this section, we first describe the backbone models employed in our framework, followed by an overview of the loss functions and optimization strategies used to guide the discovery of promising binder sequences.

Backbone structure prediction models

We employ five variants of the Protenix model family to provide gradient signals: Protenix, Protenix-Mini, Protenix-Mini-v2, Protenix-All-Data, and Protenix-Template. Among these, Protenix-Mini-v2 denotes a version of Protenix-Mini trained with different random seeds for its confidence module; Protenix-All-Data refers to a Protenix-Mini variant trained on a larger and more diverse dataset; and Protenix-Template indicates a Protenix-Mini model trained with the additional guidance of provided template structures. To prevent overfitting to any single model during the hallucination process, we randomly select one of these structure prediction models at each optimization step to perform the forward and backward passes, thereby generating the necessary gradient information for sequence updates.

Loss and Optimization

Thanks to its two-step diffusion architecture, the Protenix-based hallucinator enables end-to-end backpropagation of gradients from all confidence metrics, rather than being limited to contact loss derived from Pairformer outputs [13]. In our implementation, we combine multiple loss components with the following weights: pLDDT (0.15), pAE (0.4), ipAE (0.1), contact loss (1.0), interface contact loss (1.0), helix loss (-0.3), and radius of gyration of the designed binder (0.3). The core ingredients in hallucination process can be described as:

$$s_t = s_{t-1} - \gamma \cdot \nabla_s \mathcal{L}\left(f_\theta(s_{t-1}, c)\right), \tag{3}$$

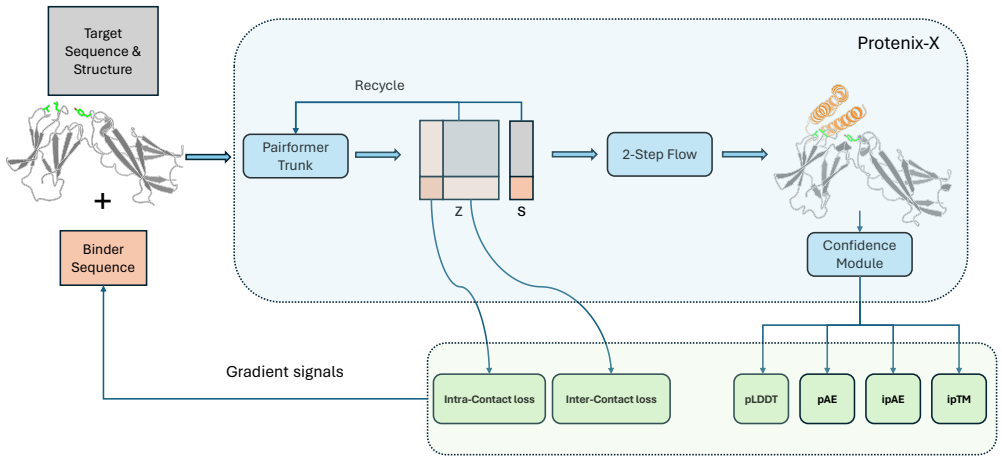


Figure 10 Overview of one-step optimization in Protenix-Hallucination. At each optimization step, a backbone structure prediction model, denoted as Protenix-X, is randomly selected from five candidates: Protenix, Protenix-Mini, Protenix-Mini-v2, Protenix-Mini-All-Data, and Protenix-Mini-Template, to prevent overfitting to the gradient preferences of any single model. During the forward pass, the target structure, target sequence, and binder sequence are input into Protenix-X to generate sequence representations s and pairwise representations z. Subsequently, an ODE-based sampler with 2 sampling steps efficiently predicts complex structures, and a loss defined on the generated structures is computed. Finally, gradients from this end-to-end differentiable process are used to update the protein binder sequence logits.

where s_t represents binder sequence logits at time step t, $\mathcal{L}$ is the loss functions defined on the outputs of structure prediction model f_θ, c denotes the conditions, e.g. target structures and sequences and γ is the step size for optimization. For binder hallucination, we adopt a 4-stage relaxed-to-hard sequence optimization process inspired by Pacesa et al. [38] and Cho et al. [13], which is detailed below.

Stage 0: Softmax Warm-up To avoid unstable exploration in the continuous logits space and to encourage the generation of realistic binder sequences, we begin with randomly initialized binder sequence logits. During this stage, we perform gradient descent updates guided by the backbone model's feedback, followed by a softmax operation on the logits. This procedure ensures that the logits represent a valid amino acid distribution at each step.

Stage 1: Soft logits update We use logits from Stage 0 to re-initialize our binder sequence logits and start soft-logit updates during this stage. As the optimization steps increase, we continuously transfer binder sequence representation from soft logits s to softmax logits $p = \text{softmax}(s)$ by linear interpolating soft logits and softmax logits. For example, we have in total T steps in this stage, at time step t, binder sequence is represented as $(1-\lambda)s + \lambda p$, where $\lambda = \frac{t+1}{T}$.

Stage 2: Softmax logits update In this stage, the goal is to transfer sequence logits representation from continuous representation to the final one-hot representation. To do so, we slowly decrease the temperature τ to get temperature-conditioned softmax logits $p = \text{softmax}(\frac{s}{\tau})$. For a T-step stage 2, temperature at time step t is given by $\tau_t = 0.01 + (1 - 0.01) \times (1 - (t+1)/T)^2$. Furthermore, we also use τ_t to scale our learning rate $\gamma_t = \gamma \cdot \tau_t$ to stabilize the final update of the binder sequence representation that are close to one-hot representation.

Stage 3: One-hot sequence update Finally, we directly get one-hot sequences by taking argmax from softmax sequence logits. To update one-hot sequence, we update softmax sequence using the gradients on it. The hard

sequence s^h at time step t is given by $s_t^h = \text{stop\_grad}(\text{argmax}(\text{softmax}(s_t)) - \text{softmax}(s_t)) + \text{softmax}(s_t)$.

C PXDesign-d (diffusion) Details

Model Architecture

PXDesign-d is a diffusion-based protein design model built as a direct extension of the Protenix all-atom structure prediction framework. To enable generative capabilities, we introduce a special token `[xpb]` to denote residues to be designed. Each `[xpb]` token consists of four backbone atoms (N, CA, C, O). During training, the coordinates of all atoms are perturbed with noise and subsequently denoised through a learned diffusion process. Target residues are soft-conditioned through pairwise features. Specifically, the single-token condition s is initialized by embedding basic residue-level features, such as amino acid identity, hotspot annotations, etc. The pairwise condition z is initialized by embedding binned pairwise distances derived from the target structure. If no distance information is available for a residue pair, we assign it to a special bin.

Because binned pairwise distances offer strong structural constraints, we find it unnecessary to freeze the coordinates of target residues during training. Instead, the model learns to recover structure directly from these embedded pairwise signals. For the condition-based generation task, unlike previous methods [53] that relied on inpainting, PXDesign-d directly generates the coordinates of all atoms from noisy structure. Throughout the process, no additional constraints are imposed on the noise. This approach preserves the topology of the condition region while allowing flexibility in its side chains.

The overall architecture consists of two components: the prior module and the diffusion module. The diffusion network component includes 4 layers of atom-level attention encoders, 16 layers of a token-wise transformer, and 4 layers of atom-level attention decoders.

Training

PXDesign-d is trained on a large and diverse dataset that integrates both experimentally resolved structures and distilled data derived from predictive models. The dataset composition closely follows the training corpus used for Protenix, but with several extensions. We curated a subset of the Protein Data Bank (PDB) [6] up to May 1, 2021, and categorized complexes based on their molecular context. These include protein monomers as well as complexes involving proteins, ligands, DNA, or RNA. For each complex, we define task-specific design objectives by specifying which residues are subject to design, allowing for fine-grained control over sampling across different structural and functional categories. To enhance coverage and diversity, we supplement the experimental data with high-confidence structures predicted by AlphaFold2. Specifically, we incorporate monomer structures from the AlphaFold Protein Structure Database (AFDB) [49] and MGnify [43].

PXDesign-d is trained from scratch in two stages to facilitate both general structural modeling and target-conditioned generation. In the first stage, we upweight monomer-only distillation data, enabling the model to learn the fundamentals of protein backbone geometry in an unconditional setting. In the second stage, we gradually shift the sampling distribution toward experimentally resolved PDB complexes and target-conditioned design tasks.

The model is trained end-to-end by minimizing a weighted combination of loss functions. Specifically, we apply an MSE loss over all heavy-atom coordinates and a smooth LDDT loss, as used in Protenix for structure prediction. Additionally, we find that introducing a distogram loss on projected token embeddings helps improve local consistency and geometric plausibility. For the distogram loss and the smooth LDDT loss function, during training, we masked the samples with a noise scale less than 4Å. The total loss is:

$$\mathcal{L} = (0.03\mathcal{L}_{\text{disto}} + 1.0\mathcal{L}_{\text{LDDT}}) \cdot \mathbb{I}(\hat{t} < 4\text{Å}) + 4.0\mathcal{L}_{\text{MSE}}. \tag{4}$$

We use a crop size of 640 residues, a batch size of 64, and a diffusion batch size of 8. The model is optimized using Adam with a learning rate of 0.0005. The diffusion noise schedule follows the same formulation as Protenix.

Sampling and Evaluation in Benchmark

We perform diffusion-based sampling with 1000 steps for monomer generation and 400 steps for binder generation. We set the sampling parameters $\gamma_0 = 1.0$ and $\gamma_{\min} = 1.5$ via a preliminary hyperparameter grid search.

As shown in Figure 11b, the parameter η controls the trade-off between designability and diversity: higher values of η generally lead to higher-quality structures but reduced structural diversity. An ablation study of different η schedules shows that both linear and piecewise schedules outperform fixed-η variants, particularly for long-sequence generation. Based on these results, we adopt a piecewise schedule in our final model, where $\eta = 1.0$ for $t < 0.65$ and $\eta = 2.0$ for $t \geq 0.65$.

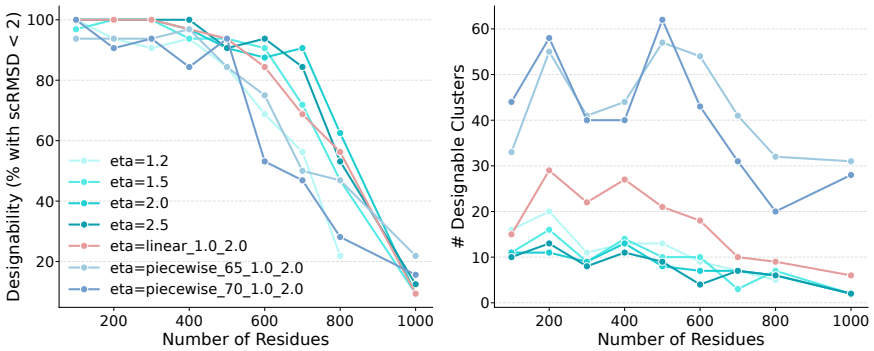


Figure 11 Ablation on η schedule in unconditional design. η increase from 1.0 to 2.0 in the linear and piecewise schedules. Piecewise schedules increase η at $t = 0.65$ or $t = 0.7$.

D Generator Benchmarking Details

Unconditional Monomer Generation Protocol

For each sequence length (200–1400 residues), we generate 100 monomer backbones per method (RFdiffusion, MultiFlow, Proteina, PXDesign-d). Sequence design is performed using ProteinMPNN-CA [18] with default settings, followed by structure prediction using ESMFold [31]. Since PXDesign-h is specialized for binder design, we exclude it from this comparison.

Designability metric: A backbone is considered designable if the best self-consistency RMSD (scRMSD), computed over 8 independently designed sequences, is below 2 Å.

Diversity metric: Following Yim et al. [57], designable backbones are clustered using a TM-score threshold of 0.5, and the number of resulting clusters is reported.

The same evaluation protocol and random seeds are used across all methods to ensure fair comparison.

Conditional Binder Generation Protocol

Following Zambaldi et al. [58], we use a benchmark set of 10 protein targets with diverse structural properties. For each target, we generate binders using RFdiffusion (noise = 0.0 and 1.0), PXDesign-d, or hallucination-based methods, then perform sequence design with ProteinMPNN [18] at temperature 0.0001.

Since structure generation speed differs substantially between diffusion- and hallucination-based methods, we adopted evaluation settings that reflect realistic computational trade-offs. For diffusion-based designs, each generated structure was paired with a single ProteinMPNN sequence. For hallucination-based designs, where structure generation is much slower, we followed the BoltzDesign1 protocol [13] and sampled 8 sequences per structure to increase the chance of success. A design was deemed successful if at least one sequence met the filter criteria defined in Table 2.

Diversity metrics: We cluster generated binders using TM-score < 0.5 and report both the number of clusters and the number of successful clusters.

Secondary structure analysis: Secondary structure content (α-helix percentage) is computed using DSSP on the folded structures, aggregated per target.

Runtime efficiency: To better reflect real-world efficiency, we measured both generation and evaluation times for different methods. Specifically, we measure the number of successful designs generated within 24 hours (including generation + evaluation) for diffusion- and hallucination-based methods (PXDesign-h, BindCraft, BoltzDesign1). we used the default experimental settings for each method. The length of generated binders strictly followed the previous work. For various targets, we generated 1,280 to 2,880 samples using NVIDIA A800 GPUs and recorded the average number of AF2-IG-easy filter-passing candidates produced within 24 hours.

BindCraft adjustments: Since BindCraft integrates hallucination and evaluation in a single pipeline, we removed evaluation time from our measurement to enable a fairer comparison. Specifically, we measured the time consumed by the `binder_hallucination` function (https://github.com/martinpacesa/BindCraft/blob/main/bindcraft.py, Lines 109-111). Within this function, we counted only the time taken to hallucinate the binder, excluding the time spent on trajectory checking (https://github.com/martinpacesa/BindCraft/blob/main/functions/colabdesign_utils.py, Lines 177-233).

A complete list of the GitHub repositories and commit hashes for all compared methods is provided in Table 4.

Table 4 Details on running the compared methods.

Method	GitHub repository and Commit
RFdiffusion	https://github.com/RosettaCommons/RFdiffusion/tree/main fa340147b9006156b251d1ad0391e3ea8e5f73eb
BindCraft	https://github.com/martinpacesa/BindCraft/tree/main 05702c435e2172a99c2b3faf87487badb6e54727
BoltzDesign	https://github.com/yehlincho/BoltzDesign1 627c0cc7bab41e56f544c5d15467b2dbeb490168

E Details of Wet-lab Results

1. **PXDesign**: We report two metrics for PXDesign:
   - *bind over tested*: the number of binders with BLI-measured affinity within 1000 nM divided by the number of candidates successfully measured by BLI,
   - *bind over designed*: the number of binders with BLI-measured affinity within 1000 nM divided by the number of candidates sent for wet-lab expression. Some designs were not successfully expressed, which reduces this metric; however, expression failures are not necessarily indicative of true lack of expressibility, as we observed similar failures even among known positive controls reported by AlphaProteo.

Table 5 Binding performance metrics for PXDesign across different analytes.

	IL-7RA	PD-L1	VEGF-A	SC2RBD	TrkA
PXDesign bind over tested	40.0	72.7	47.1	35.3	20.0
PXDesign bind over designed	40.0	47.1	44.4	33.3	16.7

2. **AlphaProteo**: Table 1 of the main text reports success rates based on yeast display, which is known to have frequent false positives. We therefore corrected these rates using the true positive ratio from HTRF validation results (Table S3).

3. **Chai-1d / Chai-2**: Results were obtained by manual reading of Figure 2a.

4. **Latent-X**: The success rates reported in the main text are based on mammalian display, which does not fundamentally differ from yeast display. We therefore supplemented these with success rates corrected using high-throughput BLI measurements.

F Cyclic Peptide Details

Benchmark Settings

Target information. We benchmarked the performance of PXDesign in cyclic peptide binder design against 12 diverse protein targets: 8 minibinder targets from AlphaProteo [58] (BHRF1, SC2RBD, PD-L1, TrkA, IL-7RA, IL17A, VEGF-A, TNF-α), 3 targets with experimentally determined structures from RFpeptides [42] (MCL1, MDM2, GABARAP), and 1 target from AfCycDesign [41] (Keap1). For the AlphaProteo and RFpeptides targets, we adopted the cropping schemes and hotspot selections described in the respective original studies. For Keap1, as AfCycDesign primarily employed hot-loop grafting rather than *de novo* design, we generated cyclic peptides using the structure with PDB ID 2FLU [33], applying a cropping range of A325–609 and hotspot residues A334, A380, A382, A415, A483, and A530, according to previous structural biology research [33].

Method setup. For PXDesign, we incorporated the Type 2 cyclic offset as described in AfCycDesign [41] into the positional encoding module on the cyclic chain. All other settings followed the minibinder design protocols detailed in Appendix B and Appendix C. For the baseline method, RFpeptides [42], we used the RFdiffusion release specified in Table 4. Each method was used to generate 128 peptide sequences of lengths 8–18, considering synthetic feasibility, developability, structural stability, and prior successful designs.

To assess real-world efficiency, we estimated the number of successful designs within 24 hours on an NVIDIA V100 GPU. The *in silico* success rate for cyclic peptide binders is determined using the AF2-IG-easy filtering criteria, encouraged by prior work leveraging AlphaFold2-based filters [41, 42]. As large-scale experimental binding data for cyclic peptide binders remain unavailable to our knowledge, we leave the validation and refinement of these *in silico* filters for future investigation.

Additional Results

In addition to the overall *in silico* success rates for cyclic peptide binder design tasks shown in Figure 6a, we report detailed per-length success rates across different targets in Figure 12.

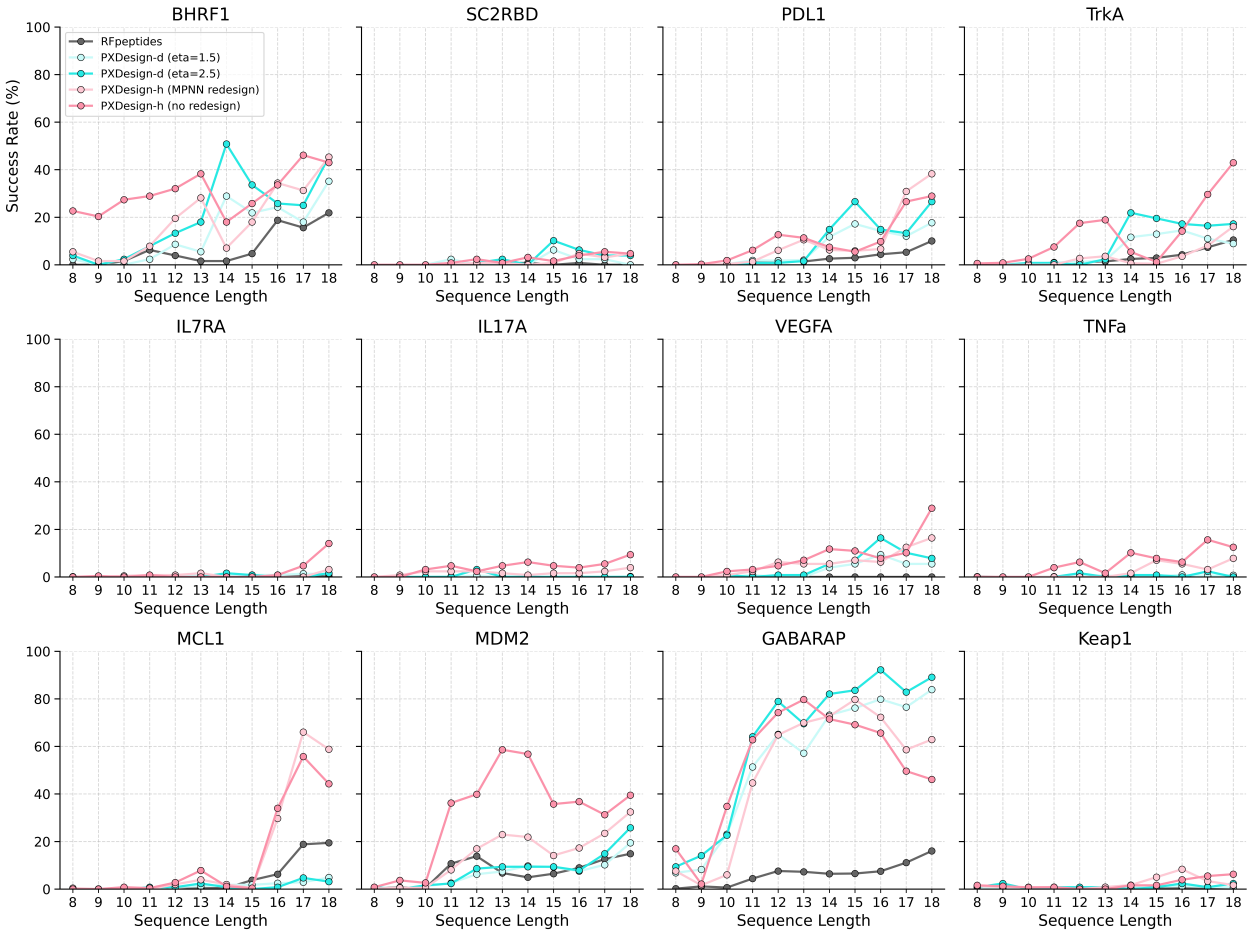


Figure 12 Per-length in silico success rates for zero-shot cyclic peptide binder design. Success rates are assessed under the AF2-IG-easy filtering criterion.