Towards an Agentic AI Framework for Generating, Optimizing and Filtering Protein Binders

Mohammad Amaan Sayeed [1] **Nancy A. ElNaker** [1] **Maryam Nadeem** [1] **Boulbaba Ben Amor** [1]

Abstract

De novo protein binder design is increasingly limited by experimental false positives: many candidates appear plausible *in silico* but fail during expression, assay, folding or binding affinity characterization. We present an Agentic AI framework for protein binder generation, optimization, and filtering that explicitly treats candidate selection as a wet-lab risk-minimization problem. The framework combines (i) epitope-conditioned generation with PBIND42, (ii) tri-model optimization using Boltz-2, ProteinMPNN, and a PBIND42 prior, and (iii) deterministic multi-agent filtering system, DYRA, that evaluates structural confidence, assay compatibility, interface quality, and expression plausibility before synthesis. Across two oncology benchmark targets, BHRF1 and PD-L1, the workflow reduces 500 generated candidates per target to compact 24-design plates and recovers two confirmed binders for each target. The best BHRF1 binder reaches a mean K_D of 176 nM, while the best PD-L1 binder reaches a mean K_D of 217 nM, with a second PD-L1 binder at 704 nM. These results show that protein-language-model generation can be made experimentally actionable when coupled to manifold-preserving optimization and agentic pre-synthesis filtering, substantially reducing the number of designs that must be synthesized and assayed.

1. Introduction

De novo design of target-specific protein binders is a foundational problem in therapeutics, diagnostics, and synthetic biology (Cao et al., 2022). The dominant paradigm today is structure-based: diffusion models such as RFdiffusion (Watson et al., 2023) generate backbones around a target's hotspot residues, after which a sequence is recovered by inverse-folding networks (Dauparas et al., 2022). These pipelines have produced experimentally validated binders (Bennett et al., 2023), but they require a high-resolution target structure, are computationally heavy, and-most critically-exhibit low experimental success rates, a manifestation of the same *hallucination* pathology that limits generative AI more broadly. In this paper, we take a deliberately different route. Building on the success of Protein Language Models as foundation models for the protein universe (Lin et al., 2023), we propose a *purely sequence-based, target-aware* binder generator that mirrors the modern LLM recipe: (i) pretraining on a large corpus of natural sequences, (ii) instruction fine-tuning on curated pairs of interacting proteins, and (iii) *in-context epitope conditioning*, in which the target's hotspot residues are injected directly into the prompt to steer generation toward a specific pocket-without ever invoking a 3D structure. To address the hallucination-driven false-positive burden, we wrap the generator in an *agentic AI workflow* that orchestrates structure prediction (Abramson et al., 2024), complex modeling, interface plausibility, and developability checks to autonomously triage thousands of candidates down to a wet-lab-tractable shortlist.

Our contributions are both methodological and practical, organized around three pillars:

1. **Hotspot-conditioned generation.** We introduce PBIND42, an epitope-conditioned instruction-tuned formulation that steers generation toward a user-specified binding surface using only residue identities—no 3D structure required. This reduces off-epitope generation by roughly half compared to unconditioned sampling.
2. **Agentic pre-synthesis filtering (DYRA).** We introduce a four-agent deterministic triage system that evaluates structural confidence, BLI assay compatibility, interface quality, and expression plausibility before any candidate enters the laboratory. DYRA treats candidate selection as a wet-lab risk-minimization problem rather than a single-oracle ranking task.
3. **Experimental validation.** We validate the full pipeline on BHRF1 and PD-L1, reducing 500 generated candidates per target to compact 24-design plates and recov-

[1]Inception Artificial Intelligence, G42, Abu Dhabi, UAE. Correspondence to: Boulbaba Ben Amor <boulbaba.amor@inceptionai.ai>.

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

ering confirmed binders for both targets. As a further demonstration, Figure 1 shows a high-confidence binder generated by PBIND42 to RBX-1 (RING Box Protein 1, PDB: 2LGV), a cancer-relevant E3 ubiquitin ligase component with a highly dynamic RING domain. The predicted complex achieves pLDDT $= 0.933$, ipTM $= 0.903$, iPSAE $= 0.88$, interface $\Delta G = -299.6$ kcal/mol, and ESM-2 masked PLL $= -1.87$, passing all DYRA risk filters at Low Risk. We include this as an *in-silico* illustration on an additional target class, not as experimental evidence; broader experimental generalization remains future work.

The key contribution is not another unfiltered generator. It is an operational framework in which specialist agents identify and remove false-positive modes before wet-lab spend: the structural agent rejects weak or off-epitope predicted complexes; the BLI agent prioritizes immobilization geometry, analyte response, and kinetic observability; the interface-energy agent penalizes brittle contacts; and the expression agent rejects sequences with low protein-model plausibility or low-complexity artifacts. This makes binder design a resource-allocation problem: only candidates that survive multiple independent risk views enter synthesis.

What "Agentic AI" means in this framework. We use "agentic" in the sense of directed multi-model orchestration (Boiko et al., 2023): specialist components operate autonomously within defined roles, distinct from fully closed-loop autonomous agents; we do not claim self-directed planning or closed-loop autonomy. DYRA is a deterministic scoring system; A language model generates explanatory text only, with zero gating influence. This framing highlights a broader design principle: generative pLMs and agentic orchestration layers are not competing paradigms, but composable components of a single protein-design workflow.

Scope of claims. This work is intended as a systems-level proof of concept rather than a claim of state-of-the-art affinity across established structure-conditioned methods. Our central question is whether a protein language model, structure-prediction oracles, and deterministic agentic triage can be composed into a practical binder-design workflow that turns hundreds of generated sequences into a small, 24-design synthesis plate. The BHRF1 and PD-L1 campaigns answer this question affirmatively while exposing remaining limitations in affinity, structural validation, and target diversity. Fig. 1 illustrates a binder to the challenging RBX1[1] (RING Box Protein 1), a highly dynamic component of the Cullin–RING ligase complex whose RING domain samples multiple conformations during catalysis (Harper and Schulman, 2021); such flexible targets are difficult for structure-conditioned methods, motivating a sequence-based approach.

[1] https://www.rcsb.org/structure/2LGV

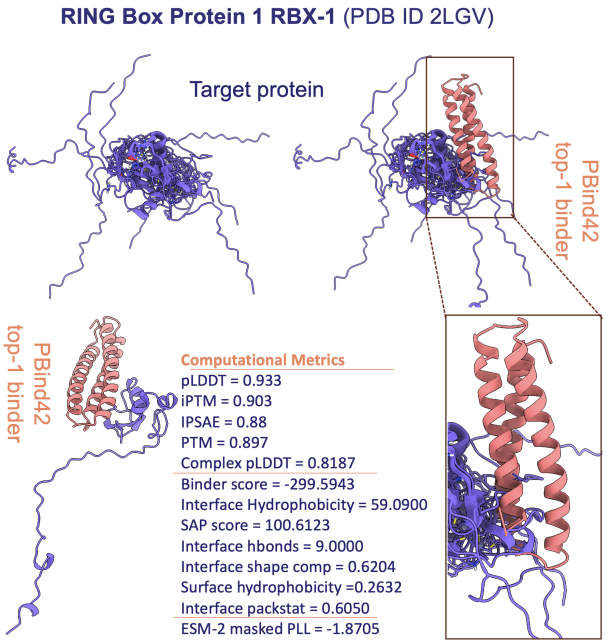

Figure 1. **PBind42 generated qualified binder to RBX-1** (a cancer-relevant E3 ubiquitin ligase component) using our comprehensive pipeline for generating, filtering and optimizing binders.

The rest of this paper is organized as follows. Section 2 compares structure-based and sequence-based approaches to binder design and positions our work relative to the state of the art.Section 3 lays out our complete pipeline. Section 4 describes PBIND42, including the epitope-conditioned instruction-tuning formulation and the Boltz-2 epitope gate. Section 5 presents the tri-model gradient-surgery optimization. Section 6 details the DYRA four-agent filtering system. Section 7 reports experimental results on BHRF1 and PD-L1. Section 8 concludes.

2. Related Work

Structure-based binder design. Most experimentally validated de novo binder-design pipelines operate in structural space. Classical approaches relied on Rosetta-style docking, interface design, and inverse folding, but typically required extensive sampling and experimental selection to recover binders (Cao et al., 2022). More recent deep-learning methods substantially improved this regime by generating binder backbones directly around a target structure. RFDIFFUSION (Watson et al., 2023) uses denoising diffusion over protein backbones and can condition generation on a target structure and specified binding region. ALPHAPROTEO (Zambaldi et al., 2024) couples a structure-conditioned generative model with learned filtering and reports high experimental success rates across multiple benchmark targets. BINDCRAFT (Pacesa et al., 2025) instead performs hallucination-style optimization through frozen AlphaFold2

weights and reports one-shot experimental success across diverse targets. More recently, LATENT-X (Latent Labs Team et al., 2025) introduced an all-atom generative model that jointly models binder sequence, binder structure, and target interactions, achieving strong hit rates and low-nanomolar to picomolar affinities on several mini-binder benchmarks.

These methods define the current state of the art in binder affinity and hit rate, but they share an important operational property: the primary design object is a predicted or generated structure conditioned on target geometry. This creates a powerful route to high-quality interfaces, but it also makes candidate selection highly dependent on structural oracles, post-generation filtering, and assumptions about which predicted complexes are experimentally actionable. In practice, structure-conditioned pipelines still face a wet-lab false-positive problem: high-scoring designs can fail because of poor expression, aggregation, assay geometry, or kinetic observability. Our work is complementary to these systems. Rather than claiming to outperform structure-conditioned methods on affinity, we focus on a risk-controlled sequence-to-lab workflow that turns hundreds of language-model-generated candidates into a compact synthesis plate.

Sequence-based and language-model approaches. Protein language models provide an alternative starting point by learning directly from large corpora of amino-acid sequences. Models such as ESM-2 show that self-supervised sequence learning captures structural and functional information: ESMFold can predict protein structures directly from sequence using representations learned by ESM-2 (Lin et al., 2023). Autoregressive models such as ProtGPT2 (Ferruz et al., 2022) and ProGen/ProGen2 (Madani et al., 2023; Nijkamp et al., 2023) demonstrated that protein-scale language models can generate de novo sequences with natural-like statistical and functional properties. These results suggest that pLMs learn a protein manifold containing information about foldability, solubility, and biochemical plausibility. However, most sequence-only protein language models are not directly target-aware binder generators. They can generate protein-like sequences, score mutations, or support downstream design, but they do not by themselves specify which target surface should be bound or how candidates should be triaged for experimental testing. Recent target-conditioned sequence approaches begin to address this gap: for example, PepMLM fine-tunes ESM-2 for target sequence-conditioned peptide-binder design and validates generated peptides experimentally (Chen et al., 2025). PBIND42/Prot42 (Sayeed et al., 2025) similarly introduced target-conditioned protein-binder generation through a separator-token format, but prior validation remained primarily in silico. Table 1 summarizes representative structure-based and sequence-based binder-design methods.

Table 1. **Comparison of SOTA binder design methods.** Str. = requires target structure or target atomic coordinates as design input. Hits = experimental success rate range. Best K_D = best reported affinity on benchmark targets.

Method	Year	Str.	Hits (%)	Best K_D
RFDIFFUSION (Watson et al., 2023)	2023	✓	0–33	1.6 nM
ALPHAPROTEO (Zambaldi et al., 2024)	2024	✓	9–88	<1 nM
BINDCRAFT (Pacesa et al., 2025)	2025	✓	10–100	<1 nM
LATENT-X (Latent Labs Team et al., 2025)	2025	✓	10–64	pM
BINDENERGYCRAFT (Nori et al., 2025)	2025	✓	n/a	<1 nM
PROT42 (seq-only) (Sayeed et al., 2025)	2025	×	*in silico only*	—
This work (PBIND42 + Agentic AI workflow)	**2026**	×	**8.3**	**176 nM**

Why hotspot residues? A biochemical prior for focused generation. Hotspot residue identities are a compact biochemical prior that concentrates the generator on a desired functional surface. They may come from alanine scanning, NMR epitope mapping, mutational studies, or structures of natural binders. In our prompt (Equation 3), hotspots are supplied as residue identities and positions rather than as full atomic geometry. The Boltz-2 epitope gate (Stage 2) then asks whether a generated binder is predicted to contact the intended surface. This reduces off-epitope false positives early, before expensive optimization and synthesis. When no hotspot information is available, the same generator can be run unconditioned, but the observed gate rejection rate is substantially higher, increasing compute and design-cycle cost.

3. Method Overview

Our pipeline (Fig. 2) has five stages, each described briefly here and in full in subsequent sections.

Stage 1 — Epitope-conditioned seed generation. PBIND42 generates candidate binder sequences using a novel `[CLS]X[SEP]hotspots[SEP]→Y` prompt that conditions the autoregressive decoder on both the full target sequence and the identity of the hotspot residues. This is our first contribution (Section 4.2).

Stage 2 — Boltz-2 epitope gate. Each seed is evaluated by Boltz-2 for all-hotspot contact. Seeds failing the gate are rejected and resampled. This ensures every candidate entering optimisation already contacts the intended epitope, avoiding wasted compute on off-epitope sequences.

Stage 3 — Tri-model gradient-surgery optimization. Accepted seeds enter a 165-step continuous optimization jointly minimising a Boltz-2 structural/interface loss, a ProteinMPNN inverse-folding recovery term, and a PBIND42 NLL manifold prior. The key role of PBIND42 is not merely to regularise language-model likelihood; it keeps the candidate inside the protein model's learned sequence distribution while Boltz-2 and ProteinMPNN pull the same candidate toward a better interface. Per-step gradient surgery (Yu et al.,

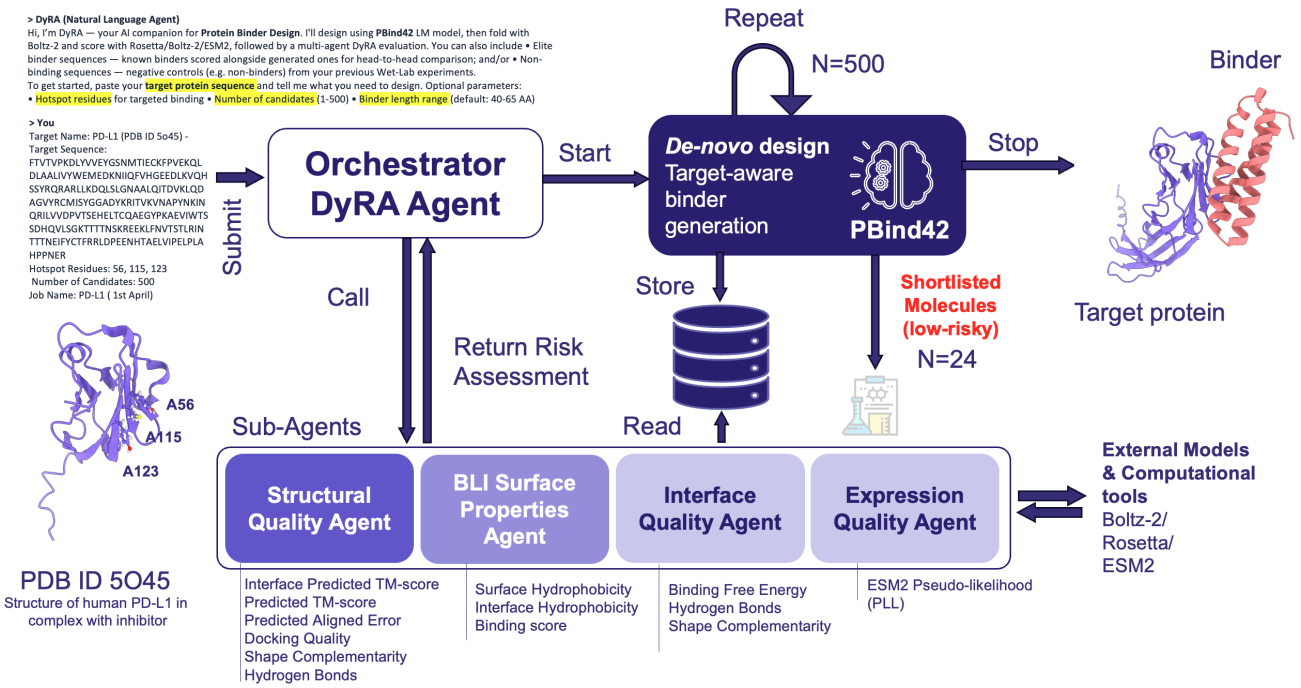

Figure 2. Overview of the proposed Agentic AI workflow for generating, filtering and optimizing protein binders.

2020) projects out the component of the manifold-prior gradient that would oppose structural progress, so the update can improve binding geometry without drifting into artifact-like low-complexity or poorly expressible sequences. This is our second contribution (Section 5).

Stage 4 — Boltz-2 ranking. Discrete candidate sequences are ranked by a Boltz-2 composite of iPTM and interface iPSAE scores (Eq. 1), providing a final in-silico quality ordering before synthesis.

Stage 5 — Dynamic Risk Assessment (DyRA). Top-ranked candidates enter DyRA, a four-agent deterministic scoring system evaluating structural confidence, BLI assay compatibility, interface energetics, and sequence expressibility (ESM2 masked PLL). Only Low/Medium Risk candidates advance to wet-lab synthesis.

A key design choice throughout is that structure prediction and inverse folding are used as *agents inside the funnel*, not as a license to synthesize every high-scoring hallucination. Their outputs become risk signals that are combined with assay and expressibility signals before a design can enter the laboratory.

The ranking composite is:

$$\mathcal{S} = \ell_{iPTM} + 0.5\,\ell_{\text{TB-iPSAE}} + 0.5\,\ell_{\text{BT-iPSAE}}. \qquad (1)$$

The same Boltz-2 model is used for the gate (Stage 2), optimisation (Stage 3), and ranking (Stage 4), ensuring consistency between objective and evaluation.

4. Target-aware binder generation

Protein language models provide a protein-native starting point for binder design. Structure-based pipelines search directly in coordinate space and can produce powerful designs, but they often require extensive filtering because high-scoring structures may correspond to sequences that are difficult to express, unstable, or outside the distribution of natural proteins. A pLM instead begins from sequence space shaped by evolutionary constraints, where foldability, solubility, and biochemical plausibility are already partially encoded. In our framework, PBind42 uses target and hotspot context to generate candidates from this distribution, and structural models are then used as specialist agents for epitope gating, interface optimization, and risk assessment. This lets the workflow combine the diversity and protein-likeness of language-model generation with the geometric precision of structure-based evaluation.

4.1. Prot42 as the Protein Language Model

Prot42 (Sayeed et al., 2025) is a decoder-only protein language model following the LLaMA architecture (Touvron et al., 2023). It has two model variants: **Prot42-B** (500 M parameters, MSL 1,024) and **Prot42-L** (1.1 B parameters, MSL 8,192 after context extension), both pre-trained on UniRef-50 (57.1 M sequences after VSL packing).

Architecture. Prot42 uses the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$), maximal update parametrisation (μP) for hyperparameter transfer from an 81 M proxy model, and

is trained on Cerebras CS-2 accelerators with 850,000 AI-optimised cores. Pre-training uses a 1 M token/batch budget with peak learning rate 4.8×10^{-4} followed by $10\times$ cosine decay.

Context extension. The 1.1 B model undergoes continuous pre-training to extend MSL to 8,192 by progressively increasing context from $1{,}024 \rightarrow 2{,}048 \rightarrow 4{,}096 \rightarrow 8{,}192$ tokens, reducing RoPE rotation angle to preserve long-range attention quality. Validation perplexity (PPL) drops from ~ 9.5 at 1,024 tokens to 5.1 at 8,192, demonstrating improved modeling of long protein sequences.

Sequence generation. At inference, given a prefix $\mathbf{X} = (x_1, \ldots, x_n)$, Prot42 generates $\mathbf{Y} = (y_1, \ldots, y_m)$ autoregressively:

$$p_\theta(\mathbf{Y} \mid \mathbf{X}) = \prod_{t=1}^{m} p_\theta(y_t \mid \mathbf{X}, \text{SEP}, y_{<t}). \tag{2}$$

This base model captures broad evolutionary protein-protein interaction statistics but makes no distinction between different binding epitopes on the same target.

4.2. Epitope-Conditioned Instruction Fine-Tuning

The base Prot42 fine-tuning for binder generation (Equation 2) conditions only on the full target sequence, providing no mechanism to steer generation toward a specific hotspot epitope on the target surface. Different regions of the same target can have very different druggability, and for therapeutics it is essential to bind a pre-specified functional epitope (e.g. the PD-1-binding face of PD-L1, or the Bim-binding groove of BHRF1). Our first contribution is PBIND42: a novel epitope-conditioned instruction-tuned formulation that resolves this.

Epitope-conditioned prompt format We augment the input to include the hotspot subsequence as a second context:

$$\text{input} = [\texttt{CLS}]\ \mathbf{X}\ [\texttt{SEP}]\ \mathbf{X}_{\mathcal{H}}\ [\texttt{SEP}], \tag{3}$$

where $\mathbf{X}_{\mathcal{H}} = (x_h)_{h\in\mathcal{H}}$ is the subsequence of target residues at hotspot indices $\mathcal{H}$. PBIND42 then decodes a binder of length L conditioned on both contexts:

$$p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{X}_{\mathcal{H}}) = \prod_{t=1}^{L} p_\theta(y_t \mid \mathbf{X}, \mathbf{X}_{\mathcal{H}}, y_{<t})\ \mathbf{1}[y_t \neq \text{Cys}], \tag{4}$$

where the Cys indicator suppresses cysteine to avoid disulfide-prone sequences.

This two-context format is a direct extension of the base PROT42 SEP-token format. The hotspot subsequence acts as a *soft binding-site specification*: the model attends to both the full-length target context (for global fold compatibility) and the hotspot residues (for local epitope specificity). Critically, this format imposes no structural constraints - only residue identities at specified positions are used. The inductive bias is positional rather than architectural: the separate SEP-delimited hotspot context gives the decoder an explicit, attendable copy of the epitope identities and their order—which a single concatenated target sequence leaves implicit—so decoding can condition on the epitope signal without inferring which target residues are functionally relevant.

Fine-tuning data and procedure. We fine-tune on 42,112 protein–protein complex structures from DIPS-Plus (Morehead et al., 2023), an enhanced dataset of interacting protein structures derived from the PDB and redundancy-reduced at 30% sequence identity. From each complex we extract the interacting chain pair and retain pairs with length ≤ 250 aa per chain, yielding a final set split 80/20 train/validation. For epitope-conditioned fine-tuning, hotspot residues are identified from the interface contact annotations provided in DIPS-Plus and, for benchmark targets, from the same hotspot lists used by Zambaldi et al. (2024). Fine-tuning uses AdamW (lr 3×10^{-5}, cosine decay, 5 epochs, batch 64, gradient clipping 1.0, fp16 precision).

Boltz-2 epitope gate. Seeds generated by Equation 4 are accepted only if *all* hotspot residues are contacted under Boltz-2:

$$\text{accept}(\mathbf{Y}) \iff \forall\, h \in \mathcal{H}:\ \ell_{\text{BTC}}(\mathbf{Y}, h) \leq \tau, \tag{5}$$

where ℓ_{BTC} is the BinderTargetContact loss for residue h (lower = stronger contact) and τ is the contact threshold. Failed seeds are discarded and resampled. In practice 52–56% of seeds pass the gate, meaning the epitope-conditioned format roughly halves the rejection rate compared to unconditioned generation, which would require far more resampling.

5. Candidates optimization

Accepted seeds enter continuous optimization over a softmax sequence representation. The optimization is deliberately framed as a three-agent negotiation rather than as single-oracle maximization. Boltz-2 supplies a differentiable structural/interface agent, ProteinMPNN supplies an inverse-folding agent that rewards sequences compatible with the predicted binder backbone, and PBIND42 supplies a manifold agent that keeps the design close to the protein-language-model distribution from which the seed was sampled. The design hypothesis is that the PBIND42 distribution already contains binding-capable regions; the two structural agents expose these regions, while the manifold agent prevents the optimizer from taking shortcuts

that score well in silico but fail expression, synthesis, or sequence-quality checks.

(i) Boltz-2 structural/interface loss.

$$\mathcal{L}_{\text{Boltz}} = \ell_{\text{BTC}}(\mathcal{H}) + \ell_{\text{WBC}} + 0.05\,\ell_{\text{TB-PAE}} + 0.05\,\ell_{\text{BT-PAE}} + 0.025\,\ell_{iPTM} + 0.4\,\ell_{\text{WB-PAE}} + 0.025\,\ell_{pTM} + 0.1\,\ell_{pLDDT} - w_{\text{avoid}}\,\ell_{\text{BTC}}(\bar{\mathcal{H}}), \quad (6)$$

where $\bar{\mathcal{H}}$ is the off-target penalty on non-hotspot residues. This agent rewards hotspot contact, high predicted complex confidence, and low inter-chain uncertainty while discouraging off-epitope contacts.

(ii) ProteinMPNN inverse-folding loss ($w_{\text{MPNN}} = 10$). $\mathcal{L}_{\text{MPNN}} = -\log p_{\text{MPNN}}(\mathbf{Y} \mid \text{struct}) + w_{\text{Cys}} \cdot \mathbf{1}[\text{Cys}]$. This agent asks whether the predicted binder backbone can plausibly encode the current sequence, penalizing designs whose interface score is achieved by a sequence/backbone mismatch or undesirable cysteine usage.

(iii) PBind42 manifold prior ($w_{\text{PB}} = 10$). $\mathcal{L}_{\text{PB}} = -\log p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{X}_{\mathcal{H}})$, implemented as a differentiable JAX loss via `pure_callback`/`custom_vjp`. This term is the distributional anchor: it encourages optimized sequences to stay inside the conditional protein-model distribution that generated the seed. In practice this matters because expression and solubility failures are often not caused by missing hotspot contacts alone; they arise when an optimizer leaves the natural-protein manifold and creates repeated motifs, low-complexity stretches, or sequence patterns that are unlikely to express.

Gradient surgery. A naive weighted sum of the three objectives can create a false-positive mode: the structural agent discovers lower predicted interface losses by moving through sequence directions that the protein language model assigns low probability. Such candidates may look strong to a structure oracle yet be poor wet-lab bets. We therefore use per-step gradient surgery (Yu et al., 2020) to couple optimization to risk control:

$$\mathbf{g}_p^* = \begin{cases} \mathbf{g}_p & \text{if } \mathbf{g}_p \cdot \mathbf{g}_s \geq 0, \\ \mathbf{g}_p - \dfrac{\mathbf{g}_p \cdot \mathbf{g}_s}{\|\mathbf{g}_s\|^2}\mathbf{g}_s & \text{if } \mathbf{g}_p \cdot \mathbf{g}_s < 0, \end{cases} \quad (7)$$

where $\mathbf{g}_s = \nabla(\mathcal{L}_{\text{Boltz}} + w_{\text{MPNN}}\mathcal{L}_{\text{MPNN}})$ and $\mathbf{g}_p = \nabla\mathcal{L}_{\text{PB}}$. The total update $\mathbf{g}_{\text{total}} = \mathbf{g}_s + w_{\text{PB}}\mathbf{g}_p^*$ keeps only the component of the PBIND42 prior that is neutral or aligned with structural progress. Equivalently, Boltz-2 and ProteinMPNN are allowed to search for better binding geometry, but PBIND42 prevents that search from leaving the protein-model distribution. This is the optimization analogue of DYRA's wet-lab risk minimization: remove false positives before they consume synthesis and assay budget. Optimization runs for 165 steps across three phases (Phase 1: 100 steps, $\alpha = 0.2\sqrt{L}$; Phase 2: 50 steps, $\alpha = 0.5\sqrt{L}$, log-space; Phase 3: 15 steps, $\alpha = 0.5\sqrt{L}$, log-space), with gradient-norm clipping throughout.

6. Filtering Agentic AI Workflow

DyRA Orchestrator DYRA is the third contribution of this work: a four-agent deterministic triage system that scores each candidate for wet-lab risk before synthesis. The coordinator computes:

$$S = 0.1S_{\text{struct}} + 0.55S_{\text{BLI}} + 0.1S_{\text{iface}} + 0.25S_{\text{expr}}, \quad (8)$$

with risk band Low ($S \geq 55$), Medium ($40 \leq S < 55$), High ($S < 40$). A candidate is Low Risk only if zero sub-agents are High Risk and ≥ 1 is Low Risk. A language model generates explanatory narrative; all gating is deterministic. Equation 8 is a weighted multi-criteria triage in which the weights encode the *relative wet-lab cost* of each failure mode rather than a calibrated probability; we therefore treat DYRA as a risk-prioritised filter, not a formal decision-theoretic utility, with the band thresholds acting as conservative vetoes.

Table 2 summarises each sub-agent; Figure 2 shows how the orchestrator routes candidates through them. The 55% BLI weight reflects our core empirical finding: assay developability - not predicted interface energy - is the dominant wet-lab failure mode. Both confirmed binders on each target had the highest BLI sub-agent scores in their respective panels.

Table 2. **DYRA sub-agent summary.** All scoring deterministic. PLL = ESM2 masked pseudo-log-likelihood (Eq. 9).

Sub-agent	w	Key inputs	Low Risk threshold
Structural	10%	pLDDT, pTM, ipTM (pDockQ excluded)	$pLDDT \geq 0.80$
BLI compat.	55%	Surface hydrophobicity, SAP/CDR-SAP, loop/α composition, ΔG_{int} context, steric clashes	No High-Risk flag
Interface	10%	Rosetta ΔG_{int}, ΔSASA, H-bonds, shape comp., packstat (pDockQ excluded)	$\Delta G_{\text{int}} < -5$ kcal/mol
Expression	25%	ESM2 masked PLL (fully deterministic, no LLM)	PLL ≥ -1.5

The expression sub-agent computes:

$$\text{PLL}(\mathbf{Y}) = \frac{1}{L}\sum_{t=1}^{L} \log p_{\text{ESM2}}(y_t \mid \mathbf{Y}_{\setminus t}), \quad (9)$$

with bands: Low (≥ -1.5), Medium ($-2.0 - -1.5$), High (< -2.0). In our PD-L1 screen, all 8 excluded measurements (yield < 37 μg/mL) had Medium or High expression risk; both confirmed PD-L1 binders had PLL ≥ -1.5 and yield > 450 μg/mL.

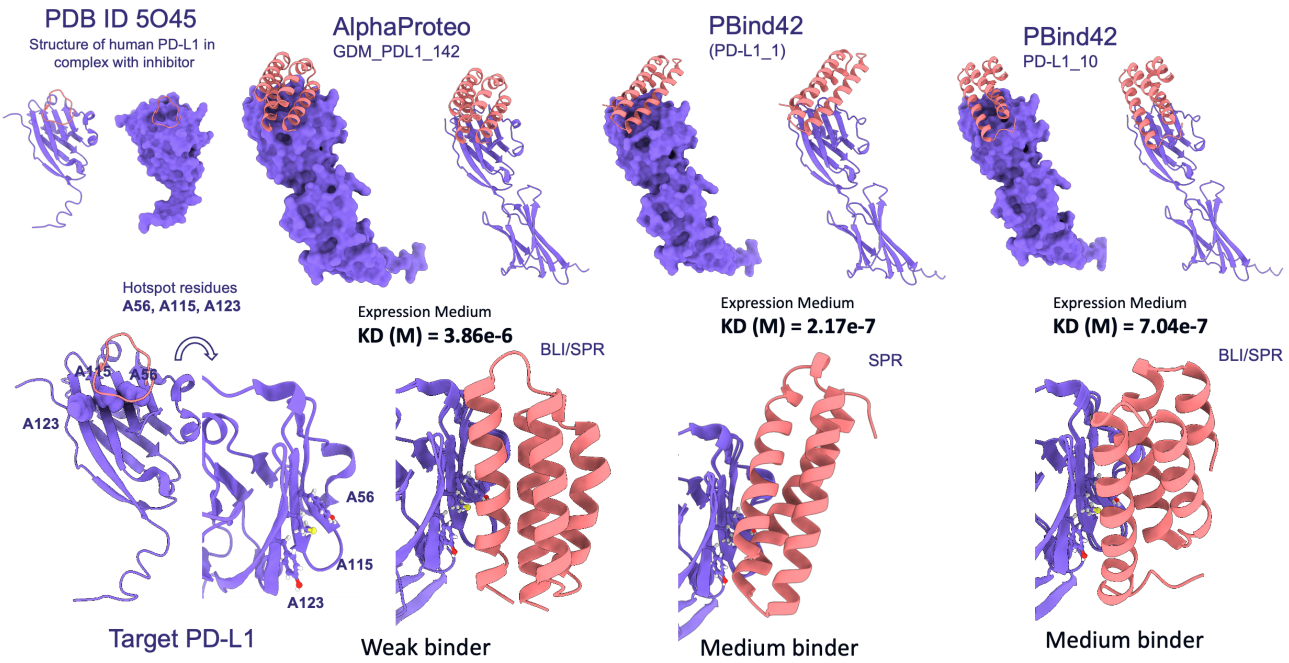

Figure 3. **PD-L1 binders generated by the agentic workflow.** PD-L1 candidates are generated with PBIND42, optimized with tri-model gradient surgery, and filtered by DYRA before synthesis. The figure compares confirmed designs from our workflow with the replicated ALPHAPROTEO reference sequence (Zambaldi et al., 2024).

Structural Quality Agent. The structural agent evaluates whether the predicted complex is globally plausible before more specific assay or interface criteria are considered. It uses Boltz-2-derived confidence metrics including binder pLDDT, complex pTM, ipTM, inter-chain PAE, and interface iPSAE. Candidates with low binder confidence or unstable predicted target–binder geometry are penalized because such designs often produce attractive local contacts without a coherent complex. We exclude pDockQ from the final score because, in our campaigns, it was less informative once pLDDT and interface-confidence metrics were already gated.

BLI Surface Properties Agent. The BLI compatibility agent estimates whether a candidate is likely to produce an interpretable biophysical assay signal. It considers surface hydrophobicity, SAP-like aggregation risk, loop exposure, helix/loop composition near the predicted interface, steric accessibility, and whether the predicted binding orientation is compatible with immobilization and analyte presentation. This agent receives the largest weight in DYRA because our empirical failure modes were dominated by assay developability rather than by predicted interface score alone. In practice, both confirmed binders for each target were among the highest-scoring candidates under this agent.

Interface Quality Agent. The interface agent evaluates the local physical plausibility of the predicted target–binder contact. It uses Rosetta REF15-derived interface energy (Alford et al., 2017), buried surface area, hydrogen-bond count, shape complementarity, and packstat computed on Boltz-2-predicted complexes. The agent penalizes brittle interfaces that rely on a small number of contacts, excessive unsatisfied exposure, or poor packing. Its weight is intentionally lower than the BLI and expression agents because interface metrics computed on predicted structures are noisy; they are useful as veto and ranking signals, but insufficient as the sole basis for synthesis.

Expression Quality Agent. The expression agent estimates whether a binder is likely to be expressible and protein-like. It uses ESM2 masked pseudo-log-likelihood (Lin et al., 2023) as a deterministic sequence plausibility score and penalizes low-complexity motifs, repeated short peptides, homopolymer-like segments, cysteine usage, and sequence patterns that suggest poor solubility or aggregation. This agent is conceptually linked to the PBIND42 manifold prior in optimization: both discourage candidates that obtain strong structural scores by leaving the learned distribution of natural, expressible proteins. In the PD-L1 campaign, candidates excluded for low yield were enriched for Medium or High expression risk, while both confirmed PD-L1 binders had favorable PLL values and high measured yield.

7. Experimental Validation

In Silico performance and DyRA structural metrics PBIND42 achieves sub-μM in-silico K_D (Prodigy ΔG) on all seven ALPHAPROTEO benchmark targets; on BHRF1 ,

PD-L1 , and IL-7Rα it matches or outperforms AlphaProteo in-silico predictions (Sayeed et al., 2025), demonstrating that PBind42 captures binding information latent in the pLM without relying on a precomputed target–binder complex. Before synthesis each candidate is assessed by DyRA using Boltz-2-predicted structural metrics (pLDDT, pTM, ipTM, ipSAE, inter-chain PAE) and Rosetta REF15 interface metrics (ΔG_{int}, ΔSASA, H-bonds, shape complementarity, packstat). The DyRA structural sub-agent ($w = 10\%$) gates on $pLDDT \geq 0.80$; pDockQ is excluded as it correlates poorly with BLI outcomes once pLDDT is gated. The interface sub-agent ($w = 10\%$) uses Rosetta REF15 on Boltz-2-predicted structures; the relatively low weight reflects noise from predicted vs. crystal structures. Metric definitions and thresholds are summarized in Table 2.

Experimental targets and design We validate experimentally on two targets from the AlphaProteo benchmark: **BHRF1** (PDB: 2WH6), an EBV Bcl-2 homolog considered an "easy" target (hydrophobic groove epitope) (Zambaldi et al., 2024); and **PD-L1** (PDB: 5O45), an immune checkpoint with a flat, polar PD-1-binding face classified as a "difficult" target (Zambaldi et al., 2024). Starting from $n = 500$ epitope-conditioned seeds per target, the selection funnel (Figure 4) delivers a 24-design synthesis plate for each target.

Candidate selection funnel Figure 4 shows the full funnel from epitope-conditioned generation to a compact synthesis plate. The computational reduction is approximately $20\times$: from 500 generated seeds to 24 synthesized candidates per target. The funnel reports expressed candidates rather than implying an additional purification-selection stage: $18/24$ BHRF1 designs and $20/24$ PD-L1 designs expressed above the campaign threshold. Both confirmed PD-L1 binders, PD-L1_1 and PD-L1_10, were in the high-yield regime, supporting the use of expression plausibility as a pre-synthesis risk signal.

Main results: binding affinities and hit rates Table 3 presents our central experimental results alongside competing methods, in the style of Latent Labs Team et al. (2025) Table 1. For each target we show the best published K_D from each method, the *replicated* K_D (the same published sequence re-measured in our assay under identical conditions), and our own best K_D. The replicated column is the most controlled comparison available across labs. Full kinetic parameters (k_{on}, k_{off}) for all confirmed binders and replicated references are provided in Table 5, complementing the K_D summary above and the sensorgram geometries in Figure 3.
The BHRF1 screen confirmed two binders: BHRF1_9, the lead (mean $K_D = 176$ nM, clean BLI kinetics in both replicates), and BHRF1_16, a second weak-affinity hit

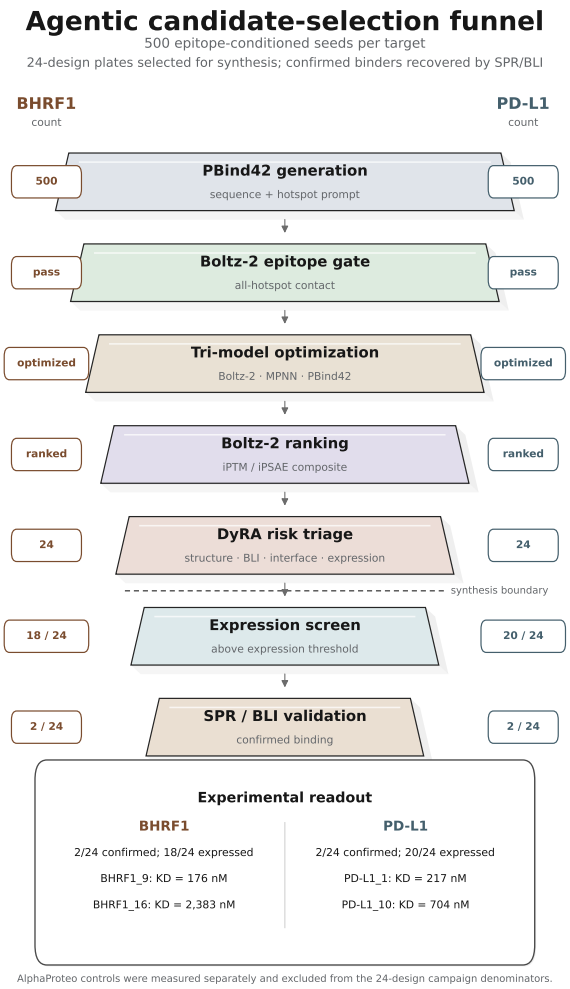

Figure 4. **Candidate selection funnel.** $n = 500$ epitope-conditioned PBind42 seeds are reduced to a 24-design synthesis plate for each target via sequential computational filters. Both confirmed binders on each target passed all stages.

($K_D = 2{,}383$ nM BLI rep 2; concentration-dependent signal confirmed in SPR rep 1 with no K_D fit).

Figure 3 shows the predicted complex geometries for PD-L1_1, PD-L1_10, and the replicated AlphaProteo reference alongside the target. The replicated AlphaProteo references show large inter-lab discrepancies vs. published values: BHRF1 mean 1,175 nM vs. published 8.5 nM ($\sim$140$\times$); PD-L1 mean 3,855 nM vs. published 0.18 nM ($\sim$21,000$\times$). This is consistent with Latent-X (Latent Labs Team et al., 2025) and reflects genuine inter-laboratory variability in target loading, buffer composition, and surface chemistry. The replicated column is therefore the most controlled comparison available in this study: BHRF1_9 (176 nM) is $\sim$7$\times$ tighter than the replicated AlphaProteo reference (1,175 nM), and PD-L1_1 (217 nM) is $\sim$18$\times$ tighter than the replicated AlphaProteo reference (3,855 nM) under matched within-assay conditions.

Table 3. **Experimental comparison to prior binder-design methods.** Published K_D values are from original studies; replicated K_D values are the same published sequences re-measured in our assay. × = no binding; — = not replicated.

Target	Method	Mean K_D (nM) ↓			Hit rate (%) ↑		Designs	Target
		Published	Replicated	Ours	Published	Ours	screened	structure req.
BHRF1	RFDIFFUSION (Watson et al., 2023)	3.4	—	—	21	—	>1,000	✓
	ALPHAPROTEO (Zambaldi et al., 2024)	**8.5**	1,175†	—	**88**	—	55	✓
	LATENT-X (Latent Labs Team et al., 2025)	22.5	—	—	10	—	100	✓
	Ours (PBIND42)	—	—	**176**‡	—	**8.3**	**24**	×
PD-L1	RFDIFFUSION (Watson et al., 2023)	1.6	—	—	13	—	>1,000	✓
	ALPHAPROTEO (Zambaldi et al., 2024)	**0.18**	**3,855**†	—	15	—	159	✓
	BINDCRAFT (Pacesa et al., 2025)	615	—	—	**78**	—	9	✓
	LATENT-X (Latent Labs Team et al., 2025)	5.27	—	—	49	—	100	✓
	Ours (PBIND42)	—	—	**217**‡	—	**8.3**	**24**	×

†ALPHAPROTEO references replicated in our assay: BHRF1 mean 1,175 nM (786, 817, 1,922 nM); PD-L1 mean 3,855 nM (5,111 nM BLI, 2,599 nM SPR).
‡Within-assay comparison: BHRF1_9 and PD-L1_1 are ∼7× and ∼18× tighter than replicated ALPHAPROTEO references, respectively.

Table 4. **Post-hoc ALPHAFOLD 3 complex confidence for confirmed and replicated binders.** All values are from the top-ranked ALPHAFOLD 3 model for each target–binder complex. Predictions were run after wet-lab validation and were not used for generation, ranking, DYRA, or synthesis selection. $\mathrm{PAE}_{\min}$ is the minimum inter-chain predicted aligned error in Å. Contacts are inter-chain residue-token contact-probability pairs with probability ≥ 0.5, counting the symmetric contact matrix once.

Target	Design ID	K_D (nM)	ipTM	$\mathrm{PAE}_{\min}$ (Å)	pLDDT_B	Cts.
BHRF1	GDM_BHRF1_72	1,175	0.91	0.82	91.5	73
BHRF1	BHRF1_9	**176**	0.91	0.85	90.7	**98**
BHRF1	BHRF1_16	2,383	**0.94**	0.90	**93.9**	58
PD-L1	GDM_PDL1_142	3,855	**0.91**	0.97	**91.1**	**64**
PD-L1	PD-L1_1	**217**	0.90	1.04	89.9	53
PD-L1	PD-L1_10	704	0.89	1.12	90.8	48

Table 5. **Full kinetic parameters for all confirmed binders and both ALPHAPROTEO reference sequences replicated in our assay.** k_{on} is reported in $\mathrm{M}^{-1}\mathrm{s}^{-1}$; k_{off} is reported in s^{-1}. SPR values are marked †.

Binder	Rep./Method	K_D (nM)	k_{on}	k_{off}
BHRF1_9	1 / BLI	203.3	1,852	3.76×10^{-4}
	2 / BLI	**148.6**	2,166	3.22×10^{-4}
ALPHAPROTEO ref. (BHRF1)	1 / SPR†	786.2	12,286	9.66×10^{-3}
	2 / BLI	817.0	4,298	3.51×10^{-3}
	3 / BLI	1,921.7	2,343	4.50×10^{-3}
BHRF1_16	1 / SPR†	*binding: yes; K_D not fitted*		
	2 / BLI	2,383.0	2,029	4.83×10^{-3}
PD-L1_1	1 / SPR†	**205.7**	739,808	1.52×10^{-1}
	2 / BLI	227.5	127,497	2.90×10^{-2}
PD-L1_10	1 / SPR†	827.7	210,717	1.74×10^{-1}
	2 / SPR†	580.3	302,388	1.75×10^{-1}
ALPHAPROTEO ref. (PD-L1)	1 / BLI	5,111	62,499	3.19×10^{-1}
	2 / SPR†	2,599	64,238	1.67×10^{-1}

8. Discussion & Conclusion

This work reframes de novo binder design as a risk-controlled handoff from broad in-silico generation to compact wet-lab plates. From 500 epitope-conditioned candidates per target, the framework selects 24 designs and recovers two confirmed binders for both BHRF1 and PD-L1 (leads at 176 nM and 217 nM). We do not claim to outperform structure-conditioned campaigns on affinity or hit rate; rather, protein-language-model generation becomes experimentally actionable when coupled to manifold-preserving optimization and deterministic pre-synthesis filtering. Filtering is thus a design objective, not an afterthought: DYRA combines structural, BLI, interface, and expression risk to catch failure modes a single structure score misses. The post-hoc ALPHAFOLD 3 predictions (Table 4)—from an independent oracle absent from every pipeline stage—support structural plausibility and partially check Boltz-2-specific bias, yet show that high confidence does not monotonically rank measured affinity: predicted structures are useful agents, but wet-lab risk control remains essential.

Limitations remain. Validation covers two oncology targets without experimental co-structures; the DYRA weights (Eq. 8) and eight-term Boltz-2 loss (Eq. 6) are hand-set from these campaigns and not yet cross-validated; and we do not yet report controlled ablations (manifold prior, gradient surgery vs. a naive weighted sum, alternative DYRA weightings), so the manifold-preservation behaviour is a qualitative check rather than a powered comparison. Because Boltz-2 is reused for gating, optimization, and ranking, residual oracle bias is possible and only partially mitigated by the ALPHAFOLD 3 check; and we have not quantified similarity between the DIPS-Plus complexes and the benchmarks, so proximity bias cannot be excluded. We are extending the workflow to further targets (SC2RBD, IL-7Rα, TrkA, VEGF-A, TNF-α, RBX1) with these ablations to test generalization across target classes.

References

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. doi: 10.1038/s41586-024-07487-w.

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

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*, 14:2625, 2023. doi: 10.1038/s41467-023-38328-5.

Daniil A. Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023. doi: 10.48550/arXiv.2304.05332.

Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M. Jude, Iva Marković, Rameshwar U. Kadam, Koen H. G. Verschueren, et al. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, 2022. doi: 10.1038/s41586-022-04654-9.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02761-2.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. 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. doi: 10.1126/science.add2187.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, 2022. doi: 10.1038/s41467-022-32007-7.

J. Wade Harper and Brenda A. Schulman. Cullin–RING ubiquitin ligase regulatory circuits: A quarter century beyond the F-box hypothesis. *Annual Review of Biochemistry*, 90:403–429, 2021. doi: 10.1146/annurev-biochem-090120-013613.

Latent Labs Team, Alex Bridgland, Jonathan Crabbé, Henry Kenlay, Daniella Pretorius, Sebastian M. Schmon, Agrin Hilmkil, Rebecca Bartke-Croughan, Robin Rombach, Michael Flashman, et al. Latent-X: An atom-level frontier model for de novo protein binder design. *arXiv preprint arXiv:2507.19375*, 2025. doi: 10.48550/arXiv.2507.19375.

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. doi: 10.1126/science.ade2574.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Alex Morehead, Chen Chen, Ada Sedova, and Jianlin Cheng. DIPS-Plus: The enhanced database of interacting protein structures for interface prediction. *Scientific Data*, 10:509, 2023. doi: 10.1038/s41597-023-02409-3.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

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. doi: 10.48550/arXiv.2505.21241.

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*, 646:483–492, 2025. doi: 10.1038/s41586-025-09429-6.

Mohammad Amaan Sayeed, Engin Tekin, Maryam Nadeem, Nancy A. ElNaker, Aahan Singh, Natalia Vassilieva, and Boulbaba Ben Amor. Prot42: A novel family of protein language models for target-aware protein binder generation. *arXiv preprint arXiv:2504.04453*, 2025. doi: 10.48550/arXiv.2504.04453.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. doi: 10.48550/arXiv.2302.13971.

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. doi: 10.1038/s41586-023-06415-8.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836. Curran Associates, Inc., 2020.

Vinicius Zambaldi, David La, Alexander E. Chu, Harshnira Patani, Amy E. Danson, Tristan O. C. Kwan, Thomas Frerix, Rosalia G. Schneider, David Saxton, Ashok Thillaisundaram, et al. De novo design of high-affinity protein binders with AlphaProteo. *arXiv preprint arXiv:2409.08022*, 2024. doi: 10.48550/arXiv.2409.08022.