PROTEINREPA: STRUCTURE PREDICTORS AS REPRESENTATION TEACHERS FOR GENERATIVE PROTEIN DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern protein design commonly follows a design-and-filter strategy, in which generative models propose candidates and biomolecular structure predictors determine which designs advance. In this paradigm, the predictor's knowledge is primarily used to reject unsuccessful designs rather than to improve the generative process itself. We ask whether internal representations encoding geometric and evolutionary priors can be instead learned by generative protein design models and, if so, which capabilities improve. We introduce ProteinREPA, which aligns the pair representation of RFD3 with that of a frozen, MSA-conditioned Boltz-2 teacher using a lightweight learned projection and cosine-similarity objective. The teacher is used only during training and introduces no additional cost during standard inference. We compare ProteinREPA with matched unaligned controls trained using the same data and optimization schedule, evaluating unconditional sequence–structure generation, protein–protein binder design, protein–ligand binder design, and enzyme motif scaffolding. ProteinREPA improves intrinsic sequence–structure compatibility: the fraction of jointly generated sequence–structure pairs that refold self-consistently rises from 34.5% for the matched control to 43.8% with the affine projection and 51.3% with the bias-free linear projection. Furthermore, training RFD3 with ProteinREPA improves the success rates for the majority of task specific design benchmark cases, such as enzyme–motif scaffolding, protein–protein-, and protein–ligand-binding tasks, demonstrating that alignment also improves functional scaffold generation. Increasing coordinate-mediated self-conditioning from two to ten iterations provides a further and often larger gain across the three benchmark classes; for example, the number of passing BHRF1 binder backbones increases from 63 to 125 out of 200. Together, these results establish biomolecular structure predictors as representation teachers for generative protein models and show that improved learned representations and inference-time geometric search address complementary limitations of current all-atom generators.

1 INTRODUCTION

De novo protein design increasingly combines generative models with biomolecular structure predictors such as AlphaFold 3, Boltz-2, and RoseTTAFold3 (Abramson et al., 2024; Passaro et al., 2025; Corley et al., 2025). A generative model proposes sequence–structure candidates, which are independently refolded and retained only when the predicted structure recapitulates the design, preserves its intended functional geometry, and exceeds predefined confidence thresholds. This design-and-filter strategy has enabled substantial progress, but remains inefficient on demanding problems such as protein–protein interface design, ligand-pocket generation, and enzyme design Yang et al. (2026). Successful generation requires solving two related problems. The proposed sequence must encode the generated structure, and that structure must satisfy the geometric constraints of the design task. Fully atomistic models attempt to address both by generating residue identities and atomic coordinates jointly. Nevertheless, jointly generated sequences need not encode their accompanying structures, and even compatible sequence–structure pairs may fail to realize the required interface, pocket, or catalytic geometry. RFD3, for example, generates sequence and all-atom structure but

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

still obtains substantially higher design success when its generated backbones are redesigned using an inverse-folding model Butcher et al. (2025). Structure predictors detect these failures, yet their knowledge is normally used only after generation. Their pair representations integrate residue compatibility, relative geometry, chemical context, intermolecular interactions, and, when available, evolutionary information from multiple-sequence alignments. Many all-atom generators also use pair representations to condition coordinate denoising, but learn them under a different objective and without access to MSAs during inference. These shared relational representations provide a natural interface for transferring knowledge from a predictor into a generator. We draw inspiration from REPresentation Alignment (REPA) which improves image diffusion models by aligning their intermediate features with those of frozen pretrained encoders (Yu et al., 2024). Related work has aligned residue-wise protein-diffusion features with protein-language and inverse-folding representations (Xu et al., 2026; Wang et al., 2024). It remains unknown whether pair representations from a complete biomolecular structure predictor provide an effective target, whether the mapping between independently learned feature spaces affects transfer, and whether improved representations are sufficient for geometrically constrained protein design. We introduce ProteinREPA, which aligns the pair representation of RFD3 with that of a frozen, MSA-conditioned Boltz-2 teacher. A learned projection maps the student representation into the teacher feature space, where the representations are matched using cosine similarity. The teacher and projection are used only during training and are removed during standard inference. We train matched no-alignment, bias-free linear, and affine-alignment models using the same architecture, data, random seed, and optimization schedule. We extend all arms beyond the published RFD3 base-training horizon, over which the no-alignment control continues to improve. Despite differing by only a bias vector, the affine projection consistently outperforms the bias-free projection. Because the projected features are used only to compute the auxiliary loss and never enter the denoising network, this result indicates that the geometry of the alignment objective materially changes the supervision received by the student representation. ProteinREPA improves both joint sequence–structure generation and constrained design. It produces generated sequences that are more compatible with their accompanying structures, while benchmarks that replace those sequences using ProteinMPNN or LigandMPNN show that the benefit extends to scaffold and functional-geometry generation Dauparas et al. (2022; 2025). Under standard sampling, the affine model outperforms the matched control on most enzyme and ligand targets, but only on two out of ten protein–protein binder cases. We then increase coordinate-mediated self-conditioning from two to ten iterations. Additional refinement improves the aligned model across enzyme, protein–protein, and protein–ligand benchmarks, with gains frequently larger than those obtained from representation alignment under the default sampler. Together, our results show that predictor-derived representation supervision improves the learned generator, while additional self-conditioning can further improve the resolution of constrained geometries at inference. This suggests a path toward models designed jointly for informative internal representations and efficient iterative refinement.

2 Background and Related Work

All-atom protein generation and predictor-guided design. Diffusion and flow-based methods generate protein structures under increasingly detailed functional constraints (Watson et al., 2023; Yim et al., 2023). Many generate backbones and subsequently assign sequences using an inverse-folding model such as ProteinMPNN (Dauparas et al., 2022). More recent models, including ODesign, La-Proteína, Complexa, and RFD3, jointly model sequence and atomistic structure (Zhang et al., 2025; Geffner et al., 2026; Didi et al., 2026; Butcher et al., 2025). RFD3 additionally supports proteins, nucleic acids, small molecules, and arbitrary atomic constraints. ProteinREPA does not modify this generative architecture; it studies how supervision from a biomolecular structure predictor changes its capabilities. Generated candidates are commonly evaluated by independently refolding their sequences using AlphaFold3, Boltz-2, or RoseTTAFold All-Atom Abramson et al. (2024); Passaro et al. (2025); Krishna et al. (2024). Other approaches incorporate predictor-derived confidence or structural objectives directly into sequence optimization or design-time search, including BindCraft, and BoltzDesign (Pacesa et al., 2025; Cho et al., 2025). Such methods demonstrate the value of predictor knowledge but require repeated predictor evaluations; ProteinREPA instead transfers this knowledge into the generator's internal representations during training.

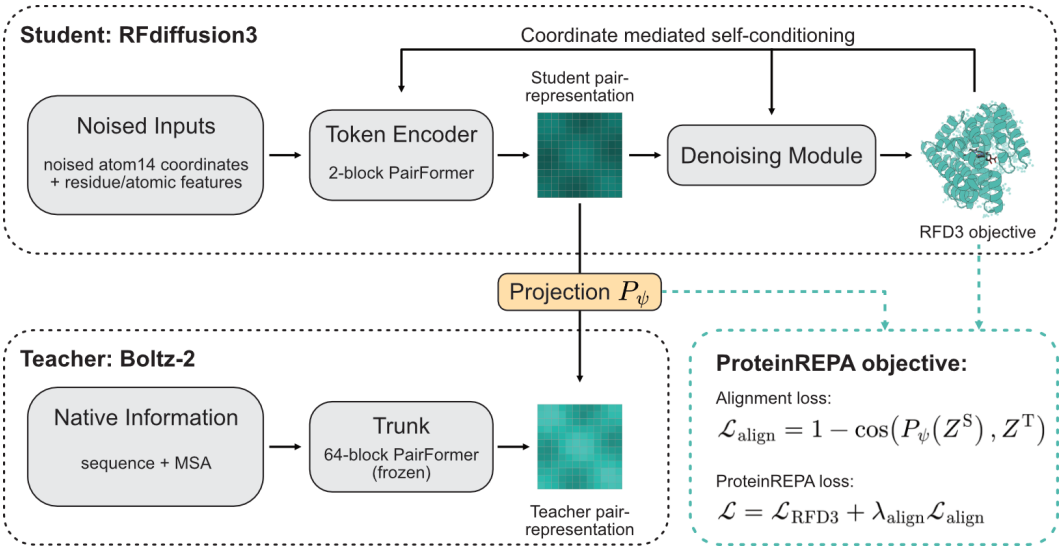

Figure 1: **Overview of ProteinREPA.** RFD3 processes noise-corrupted all-atom coordinates and produces a pair representation Z_ℓ^{S} that conditions coordinate denoising Butcher et al. (2025). A frozen Boltz-2 teacher processes the corresponding sequence and MSA to produce Z^{T}. A learned projection maps Z_ℓ^{S} into the teacher feature space, where the representations are aligned using cosine-similarity. Gradients propagate through the projection and RFD3, while Boltz-2 remains frozen. The teacher and projection are removed after training, leaving standard RFD3 inference unchanged.

Representation alignment and privileged supervision. Knowledge distillation and learning using privileged information transfer knowledge from a teacher that may have access to inputs unavailable to the student at inference (Hinton et al., 2015; Lopez-Paz et al., 2015). REPA applies this principle to image diffusion by aligning projected intermediate features with representations from a frozen self-supervised encoder, improving convergence and sample quality without changing inference (Yu et al., 2024). ProteinREPA aligns relational pair representations with those of a complete, MSA-conditioned biomolecular structure predictor. Pair representations encode relationships between molecular tokens and support reasoning about relative geometry, residue compatibility, and intermolecular interactions. Their presence in both Boltz-2 and RFD3 provides a direct transfer interface, while the teacher's access to native sequences and MSAs makes the objective a form of privileged training-time supervision.

Self-conditioning and inference-time scaling. Protein-design performance also improves with additional inference computation, for example by generating more candidates, evaluating them with external predictors, optimizing task-specific rewards, or repeatedly refining internal predictions (Didi et al., 2026). RFD3 uses coordinate-mediated self-conditioning, whereby geometry from a previous denoised prediction is supplied to the next model iteration (Butcher et al., 2025). We increase the number of these iterations to study iterative geometric refinement without invoking an external prediction model.

3 PROTEINREPA

3.1 OVERVIEW

ProteinREPA aligns pairwise representations from a frozen biomolecular structure prediction model into an all-atom diffusion model, here we used Boltz-2 and RFD3, respectively. Given a clean training example with coordinates $\mathbf{x}_0$ and conditioning information $\mathbf{c}$, we sample a noise level σ from the RFD3 EDM training distribution and corrupt the coordinates to obtain $\mathbf{x}_\sigma$ Karras et al. (2022). For residues designated for generation, $\mathbf{c}$ excludes their amino-acid identities. RFD3 instead

uses a fixed 14-atom layout, with virtual atoms placed at C_β in unused side-chain slots (Butcher et al., 2025). Residue identity is therefore not exposed through atom counts or presence masks and must be recovered from the denoised all-atom geometry. RFD3 processes the noised coordinates and conditioning information with its token encoder, which contains a two-block Pairformer, to produce the student pair representation

$$Z_\ell^{\mathrm{S}}(\mathbf{x}_\sigma, \sigma, \mathbf{c}) \in \mathbb{R}^{I \times I \times d_{\mathrm{S}}}. \tag{1}$$

RFD3 uses coordinate-mediated self-conditioning: the pair representation is recomputed using the detached denoised-coordinate prediction from the preceding iteration, rather than being propagated directly between iterations. All but the final self-conditioning iteration are evaluated without gradient, and we apply the alignment objective to Z_ℓ^{S} from the final, gradient-enabled iteration. The teacher target is precomputed using a fixed Boltz-2 model F_ϕ:

$$Z^{\mathrm{T}} = F_\phi(\mathbf{u}) \in \mathbb{R}^{I \times I \times d_{\mathrm{T}}}, \tag{2}$$

where $\mathbf{u}$ contains the native sequence, MSAs, and molecular composition corresponding to the training example, but no native coordinates or structural templates. We extract Z^{T} from the output of the final Boltz-2 PairFormer block after its final recycling iteration. These teacher inputs are unavailable to RFD3, making Z^{T} a source of privileged training-time supervision. In the experiments reported here, we align only the pair representations.

3.2 Pair-Representation Alignment

We use $d_{\mathrm{S}} = d_{\mathrm{T}} = 128$. Although the representations have equal dimensionality, they occupy independently learned feature spaces. For each token pair, we therefore compute

$$\widehat{Z}_{ij}^{\mathrm{S}} = P_\psi\big(Z_{\ell,ij}^{\mathrm{S}}\big), \tag{3}$$

where P_ψ is a single learned projection from the student to the teacher feature space. The cached Boltz-2 pair representations exhibit a pronounced shared direction across token pairs, consistent with a nonzero feature mean (Appendix B.1). This observation motivated us to test whether allowing the projection to translate, as well as rotate and rescale, the student representation improves alignment. We compare two parameterizations with nearly identical capacity:

$$P_\psi(z) = \begin{cases} Wz, & \text{bias-free linear,} \\ Wz + b, & \text{affine.} \end{cases} \tag{4}$$

The affine variant permits translation between the student and teacher feature spaces. We evaluate below whether the learned bias compensates for their differing feature means or improves alignment of pair-specific information more generally. Teacher representations are computed offline for complete structures and subsequently mapped into the token order of each training crop. Let $\mathcal{M}$ denote the set of token pairs for which both endpoints have a valid cached teacher target. We minimize the mean cosine distance between the projected student and teacher pair representations:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \left[1 - \frac{\widehat{Z}_{ij}^{\mathrm{S}} \cdot Z_{ij}^{\mathrm{T}}}{\left\|\widehat{Z}_{ij}^{\mathrm{S}}\right\|_2 \left\|Z_{ij}^{\mathrm{T}}\right\|_2} \right]. \tag{5}$$

During training, the loss is additionally averaged over the diffusion samples associated with each structure. Because the teacher representations are precomputed and fixed, gradients propagate only through P_ψ and RFD3. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{RFD3}} + \lambda_{\mathrm{align}} \mathcal{L}_{\mathrm{align}}, \tag{6}$$

where $\lambda_{\mathrm{align}} = 0.5$. The alignment loss is active from the first optimization step and is applied across the full RFD3 training noise distribution. The teacher and projection are not required during standard inference.

4 Experiments

We evaluate ProteinREPA by asking three questions. First, does representation alignment improve the intrinsic ability of RFD3 to generate compatible sequences and structures? Second, do these improvements transfer to geometrically constrained protein-design tasks? Third, can the aligned representation be exploited at inference time, either through additional self-conditioning or as an internal signal for evaluating and steering generation trajectories?

4.1 Experimental Setup

We train three RFD3 arms from scratch with matched architecture, data, optimizer, seeds, and schedule: an unaligned control using the original RFD3 objective, ProteinREPA–linear with a bias-free projection from student pair features to frozen Boltz-2 targets, and ProteinREPA–affine with a weight-and-bias projection. Both aligned arms use a constant alignment weight of $\lambda_{\text{align}} = 0.5$, isolating alignment and projection parameterization. Training begins with 256-token, 2,560-atom crops and 32 diffusion samples per structure, then moves to 384 tokens, 3,840 atoms, and 16 samples. Lightweight native-sequence probes track learning as inexpensive diagnostics, not substitutes for the full design benchmarks (Appendix F.1). Full configurations are given in Appendix E.5. For the final benchmarks, we include the released RFD3 checkpoint only as an external reference: its larger, incompletely available training collection precludes a controlled comparison. Our primary comparisons use the matched no-alignment control. To motivate pair-feature alignment, we independently perturb atom-single, token-single, and token-pair features with Gaussian noise scaled to their empirical standard deviations and measure the change in $\hat{\mathbf{x}}_0$ across diffusion noise levels. Single features have the largest effect at low noise, while the pair representation Z^{S} remains influential across a broader range where global structure is resolved. This supports its use as a relational alignment target, without establishing superiority over every combination of internal features (Appendix Fig. 7).

4.2 Representation Alignment Improves Learning and Generation

We monitor training using ProteinMPNN negative log-likelihood (NLL) and native-sequence designability, defined as self-consistent refolding within 2.0 Å scRMSD. These serve as inexpensive probes in place of running the full design benchmarks during training. Each arm and checkpoint is evaluated on 200 length-150 unconditional backbones and 96 designs per target across six enzymes and six binders. Target identities, lengths, protocols, and per-target curves are given in Appendix F.1, we show the results in Figure 2. We assess base-training efficiency by step-normalized area under the learning curve (AULC) and final quality in a separate matched-checkpoint evaluation. During base-context training, ProteinREPA–affine improves designability AULC by 4.5 percentage points over the control (11/15 checkpoints) and lowers NLL AULC by 0.033 nats per residue (13/15 checkpoints). At the last pre-training checkpoint, designability is 54.5% versus 43.5%; the linear model remains near the control on both integrated measures. Across six fine-tuning probes from optimization step 112,500 to 157,500, affine has the highest unconditional designability and lowest NLL at 5/6 checkpoints each. In the larger, separately evaluated final checkpoints we sample 1,000 designs per arm, designability instead reaches 51.3% for linear, 43.8% for affine, and 34.5% for the control, while affine retains the lowest NLL. The probes track learning, but do not determine the final model ranking. For enzymes, affine improves NLL and all-atom lDDT over the control at 24/29 pre-training and 6/6 fine-tuning checkpoints for each metric. For binders, the corresponding counts are 22/29 and 2/6 for NLL, and 19/29 and 3/6 for target-aligned RMSD. The enzyme advantage is more persistent than the binder advantage. We therefore assess final models on the full design benchmarks below using the final fine-tuned checkpoints. All probe results are listed in tabular form in the Appendix F.1.

4.3 Protein-Design Benchmarks

We next evaluate whether the benefits of representation alignment transfer to protein–protein binder design, protein–ligand binder design, and enzyme motif scaffolding tasks. For each target and model arm, we generate 200 backbones and design eight sequences per backbone using ProteinMPNN for protein–protein binder tasks and LigandMPNN for enzyme-, and protein–ligand binder tasks. The resulting sequence–structure pairs are independently evaluated using AlphaFold 3. A backbone is counted as successful if at least one of its eight designed sequences satisfies all task-specific criteria. The success criteria are stated below and described in full, together with the target definitions and sampling parameters, in Appendix F. We evaluate the three matched training arms together with the published RFD3 checkpoint. We additionally sample ProteinREPA–affine using ten coordinate-mediated self-conditioning iterations instead of the default two. This comparison tests inference-time scaling within the affine model. Because the other arms were not sampled using ten iterations, it does not determine whether those models would benefit equally from additional self-conditioning, however, we expect the other arms to behave similarly.

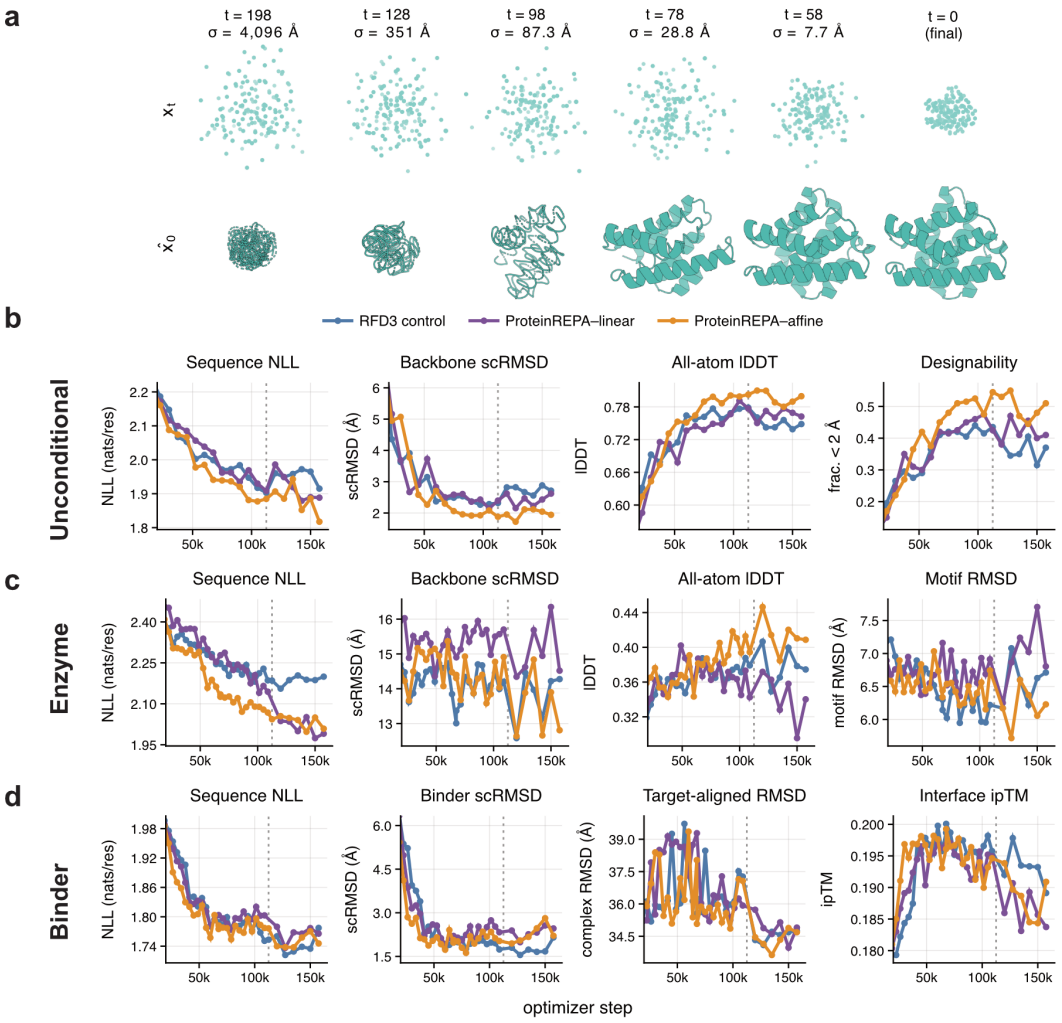

Figure 2: **Learning dynamics and sequence–structure co-generation. a** Representative denoising trajectory showing the noised coordinates $\mathbf{x}_\sigma$ and predicted clean coordinates $\hat{\mathbf{x}}_0$. **b** Unconditional sequence and structure metrics during pretraining, including ProteinMPNN negative log-likelihood, backbone self-consistency, all-atom refold agreement, and native-sequence designability. **c** Corresponding metrics on the enzyme-design monitoring set. **d** Corresponding metrics on the protein–protein binder monitoring set. Lower values are better for negative log-likelihood and RMSD metrics; higher values are better for lDDT, designability, and ipTM. All curves in this figure were obtained during base-context pretraining, before long-context fine-tuning. The grey dotted line indicates the transition from base-training to fine-tuning.

4.3.1 Enzyme Motif Scaffolding

The atomic motif enzyme (AME) benchmark, introduced with RFdiffusion2 (Ahern et al., 2026), tests whether a model can scaffold a prescribed catalytic motif around a bound ligand. For each designed sequence, the predicted structure is aligned to the design using the backbone N, C_α, and C atoms of the catalytic residues. A sequence is considered successful if the resulting RMSD over all catalytic-residue heavy atoms is below 1.5 Å and the design contains no ligand clash, defined as any protein–ligand atom pair separated by less than 1.5 Å. A generated backbone passes if at least one of its eight LigandMPNN sequences satisfies both criteria. Under the default sampler, ProteinREPA–affine achieves a higher pass rate than the matched control on 35 of 41 targets (Fig. 3). Absolute success rates decrease sharply as the number of constrained residue islands increases, reflecting the

increasing difficulty of arranging several disconnected motifs within a single compatible scaffold. Nevertheless, the aligned model retains an advantage across much of the benchmark, demonstrating that the improvements observed during training extend beyond sequence quality to the generation of scaffolds capable of supporting functional geometries. All methods approach the evaluation floor on the most highly constrained cases. Increasing the number of self-conditioning iterations produces a further and often larger improvement. The ten-iteration affine model improves on the default affine sampler on $33/41$ targets and increases the target-macro-averaged pass rate from 25% to 35.6%. Thus, representation alignment and inference-time computation provide complementary gains: alignment improves the learned generator, while repeated self-conditioning gives the model additional opportunities to resolve the constrained geometry.

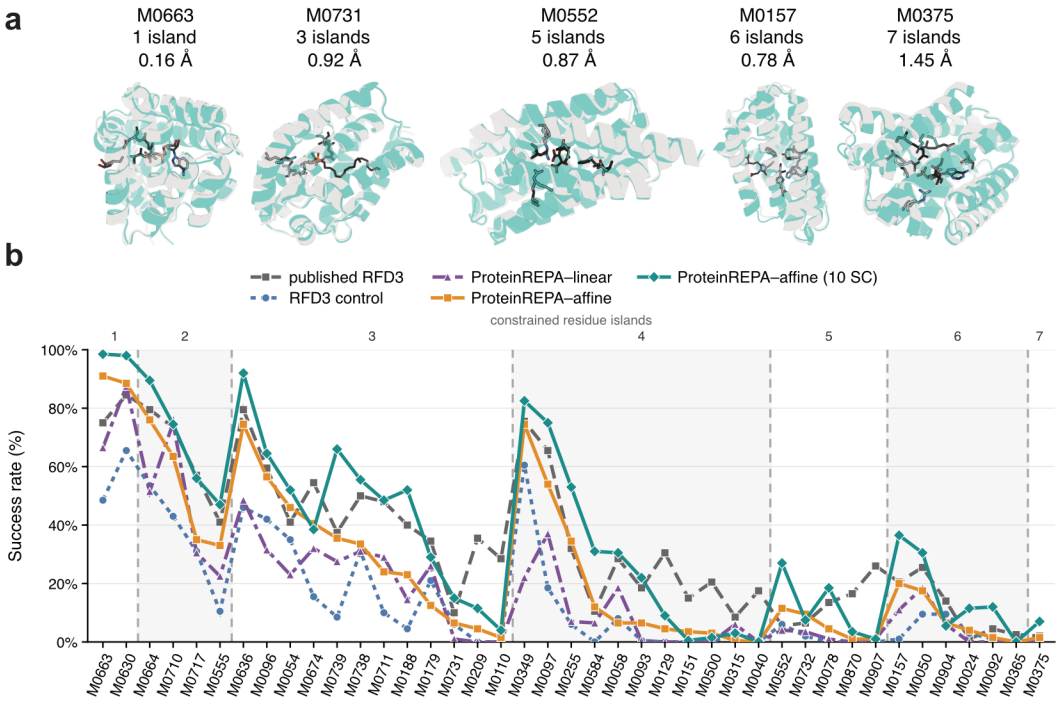

Figure 3: **Enzyme motif-scaffolding performance.** **a** Representative scaffolds generated around prescribed catalytic motifs and bound ligands. **b** Backbone success rate for each of the 41 atomic enzyme motif targets, grouped by the number of discontiguous constrained residue islands.

4.3.2 Protein–Protein Binder Design

To benchmark protein binder design we adopt the ten protein–protein targets introduced by AlphaProteo, including the same deposited target structures, cropped regions, hotspot residues, and target-specific binder-length ranges. The model must generate a binder that folds consistently, forms a confident interface, and contacts the specified epitope. Following the AlphaProteo AF3 evaluation protocol, a refolded sequence is considered successful when its minimum inter-chain pAE is at most 1.5 Å, binder pTM is at least 0.8, and target-aligned binder Cα RMSD is below 2.5 Å. We additionally require at least 75% of the specified hotspot residues to lie within 5 Å of a binder heavy atom. A backbone passes if at least one of its eight ProteinMPNN sequences satisfies every criterion.

We find that ProteinREPA–affine outperforms the matched control on two of the ten targets and ties on four and is worse on four (Fig. 4). No model, including the published RFD3 reference found a solution to IL-17A, TNFα. Notably the linear arm outperforms both the control and the affine arm on three benchmark cases. The asymmetry between the consistent gains observed for enzyme motif scaffolding and protein–binder design benchmark may be attributed to how privileged information is distributed within the teacher representation, explored in Appendix E.4. Additional self-conditioning has a substantially larger effect on several targets. For BHRF1, for example, increasing

from two to ten self-conditioning iterations raises the number of passing backbones from 63 to 125 out of 200. Improvements are smaller on targets already near the evaluation floor, indicating that additional computation can refine promising binding trajectories but does not reliably rescue targets for which a suitable interface geometry is rarely identified.

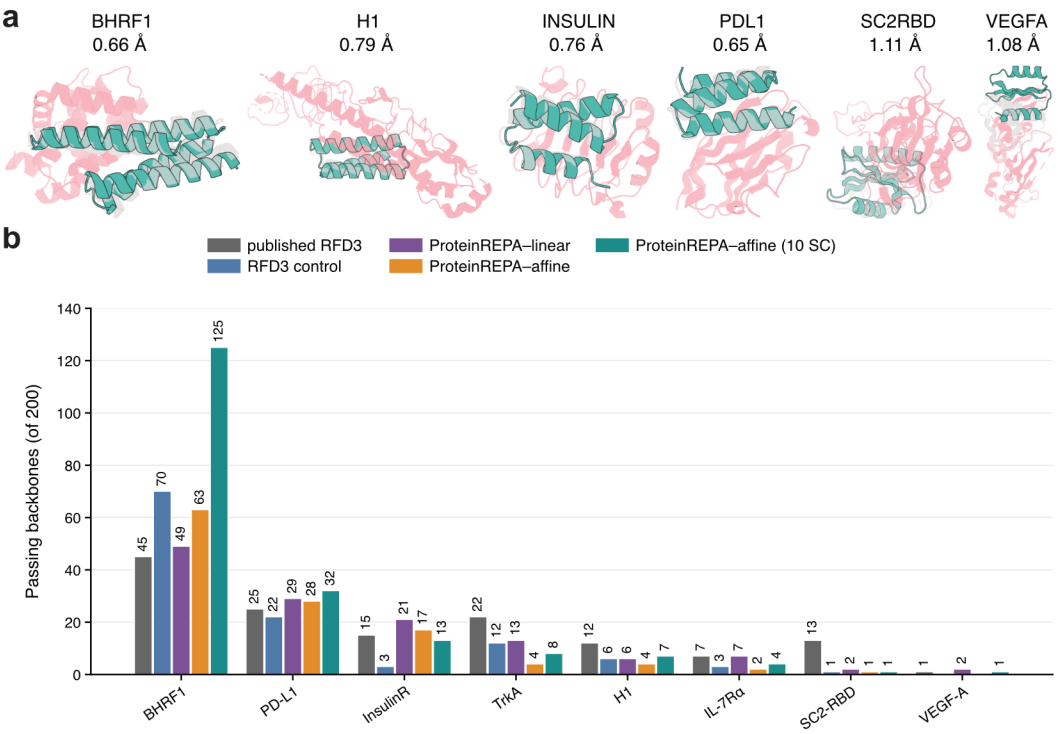

Figure 4: **Protein–protein binder-design performance.** **a** Representative generated binders and their independently predicted complexes. **b** Number of passing backbones among 200 samples per target.

4.3.3 Protein–Ligand Binder Design

We evaluate small-molecule binder design using the four ligand targets shared by Complexa and RFD3: OQO, SAM, IAI, and FAD Krishna et al. (2024). For each target, the model must generate a protein pocket around the specified ligand pose. A refolded sequence is considered successful when the protein structure agrees with the generated design within an scRMSD of 2 Å and the ligand is recovered within 5 Å RMSD of its intended pose after alignment on the protein, and the minimum inter-chain pAE is below 2. A backbone passes if at least one of its eight LigandMPNN sequences satisfies every criterion. ProteinREPA–affine achieves a higher pass rate than the matched control on three of the four ligands (Fig. 5). The published RFD3 checkpoint remains stronger across these targets. However, because published RFD3 and our models differ in both training-data composition and optimization history, this comparison cannot attribute the remaining gap specifically to representation alignment. The controlled comparison between ProteinREPA and the no-alignment arm nevertheless shows that alignment improves ligand-pocket generation under an otherwise matched training protocol. Increasing self-conditioning from two to ten iterations improves the affine model on all four ligand targets in the current evaluation. As in the enzyme and protein–protein benchmarks, these inference-time gains are frequently larger than the difference between the aligned and non-aligned models under the default sampler.

The aligned representation is not a calibrated internal evaluator. Decoding a Boltz-2 distogram from the projected student representation neither ranked trajectories reliably (agreement emerged only late in sampling and was weakly linked to refolding success) nor improved pass rates via

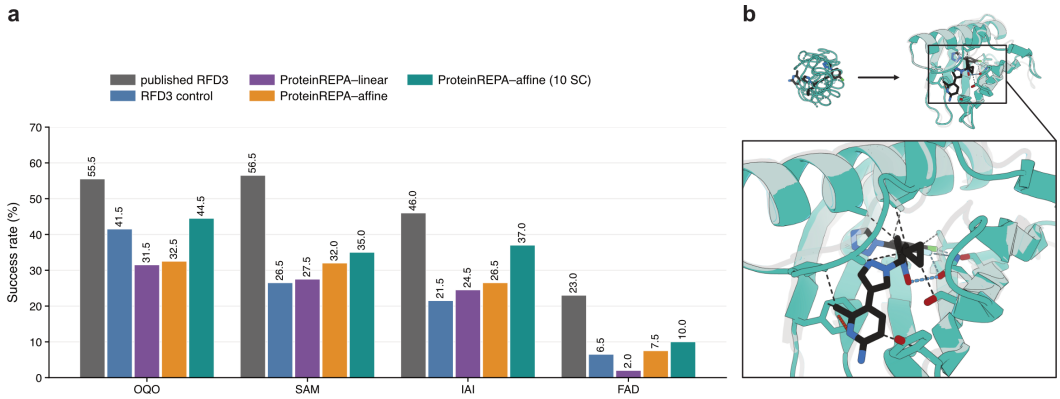

Figure 5: **Protein–ligand binder-design performance. a,** Success rates for OQO, SAM, IAI, and FAD, with 200 generated backbones and eight LigandMPNN sequences evaluated per backbone. **b,** Representative generated ligand-binding protein and close-up of the independently predicted binding pocket.

gradient-based steering, so the representation aids training but is not a calibrated inference-time objective (Appendix B.4).

Synthesis. Under the default sampler, the affine model outperforms the matched control on 35 of 41 enzyme and three of four ligand targets, and on two of ten protein–protein targets. Because benchmark sequences are redesigned with ProteinMPNN or LigandMPNN before independent refolding, these gains cannot stem from RFD3's native sequence generation alone: alignment increases how often the model produces backbones, pockets, and functional geometries that support independently designed sequences. The gains remain target-dependent, and the hardest constrained cases are unresolved—consistent with the teacher representation, whose MSA-dependent component of Z^{T} concentrates in inter-chain and interface pairs that form only a small fraction of the uniform average in $\mathcal{L}_{\mathrm{align}}$ (Appendix E.4). Additional self-conditioning gives a separate, sometimes larger gain, indicating that representation quality and inference-time geometric search address complementary limitations of RFD3-like generators.

5 Conclusion

ProteinREPA demonstrates that pair representations from a biomolecular structure predictor can provide useful training-time supervision for a general-purpose all-atom generator. Under matched training conditions, representation alignment improves the learning trajectory and joint sequence–structure generation, with gains extending most clearly to enzyme motif scaffolding and protein–ligand design. Improvements on protein–protein binder design are more target- and sampler-dependent, indicating that representation transfer improves the learned prior without guaranteeing that sampling finds candidates meeting both the binding geometry and the oracle's criteria. Projection parameterization also matters, although no variant is uniformly superior: the affine projection performs best across base-context training and in sequence likelihood, whereas the bias-free projection attains higher designability after long-context fine-tuning. This interaction shows that representation geometry influences optimization, but its downstream effect depends on the subsequent training regime. Increasing coordinate-mediated self-conditioning from two to ten iterations frequently produces gains larger than those from alignment alone, particularly for enzyme and ligand design. Because this comparison was performed only within the affine model, further controls are needed to determine whether alignment and additional refinement interact. Conversely, distograms decoded from the aligned representation did not provide a reliable ranking or steering objective, showing that representations useful for training are not automatically calibrated for inference-time evaluation. The teacher and projection are used only during training, so ProteinREPA adds no cost to standard generation.

6 AI USE STATEMENT

Generative AI tools, including OpenAI ChatGPT, Codex and Anthropic Claude Code, were used to assist with brainstorming and refining research hypotheses, providing feedback on experimental design, identifying and summarizing relevant literature, inspecting and developing software and analysis code, interpreting and presenting experimental results, and structuring, drafting, and editing the manuscript. The authors independently reviewed the original literature, inspected and tested AI-assisted code, verified reported numerical results against experimental outputs, and revised all AI-assisted text. All experiments and scientific decisions were conducted under author supervision. The authors take full responsibility for the contents, claims, code, and artifacts presented in this work.

7 ETHICS STATEMENT

Protein-generation methods are inherently dual use, although our experiments are limited to established computational benchmarks and do not target pathogens, toxins, or other harmful biological systems. Computational benchmark success does not establish biological activity or safety; experimental use requires appropriate sequence screening, biosafety review, and laboratory validation. Released artifacts will document their intended use and limitations and will be distributed subject to applicable upstream licenses and legal and biosafety requirements.

8 REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce the training, evaluation, and statistical analysis of ProteinREPA in the main text and appendices.

Training details. Appendix E.5 specifies the model architecture, representation taps, projection parameterizations, loss functions and weights, optimizer and learning-rate schedule, crop sizes, effective batch sizes, training stages, random seeds, hardware, checkpoint-selection procedure, and number of optimizer steps. We distinguish optimizer steps from epochs when making training-efficiency comparisons.

Data processing. Appendix D describes the data sources, filtering and preprocessing procedures, training mixture, MSA construction, and computation and storage of the Boltz-2 teacher representations. We use only distillation examples with normalized pLDDT greater than 0.90. We additionally report the teacher input configuration, token-matching procedure used after cropping, and procedures used to assess overlap between training examples and benchmark targets.

Evaluation and statistical analysis. Appendix F provides the complete target lists, conditioning specifications, sampling parameters, sequence-design protocols, structure-prediction evaluators, task-specific success criteria, aggregation conventions, and uncertainty estimates. Per-target results and the outputs underlying aggregate pass rates are included in the supplementary material.

Code, models, and data release. Upon acceptance, we intend to release the training and evaluation code, configuration files, trained model weights, analysis scripts, and benchmark outputs under permissive licenses. Processed datasets and cached teacher representations will be released where permitted by upstream licenses; where redistribution is not possible, we will provide manifests and scripts for reconstructing them from the original sources.

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.

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom-level enzyme active site scaffolding using rfdiffusion2. *Nature Methods*, 23(1):96–105, 2026.

C Bhikadiya, C Bi, L Chen, LD Costanzo, C Christie, S Dutta, Z Feng, S Ghosh, RK Green, V Guranovic, et al. Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic Acids Res.*, 47(D1):D520–D528, 2019.

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

Yehlin Cho, Martin Pacesa, Zhidian Zhang, Bruno E Correia, and Sergey Ovchinnikov. Boltzdesign1: Inverting all-atom structure prediction model for generalized biomolecular binder design. *BioRxiv*, pp. 2025–04, 2025.

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*, 2025.

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.

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

Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael Bronstein, et al. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *International Conference on Learning Representations*, volume 2026, pp. 61056–61109, 2026.

Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life, october 2024. *URL http://biorxiv. org/lookup/doi/10.1101/2024.10*, 10, 2024.

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

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-proteina: Atomistic protein generation via partially latent flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 67350–67395, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

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.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

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.

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

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

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.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Chenglin Wang, Yucheng Zhou, Zijie Zhai, Jianbing Shen, and Kai Zhang. Diffusion model with representation alignment for protein inverse folding. *arXiv preprint arXiv:2412.09380*, 2024.

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.

Junde Xu, Yuansheng Huang, Zijun Gao, Lihang Liu, Xiaoming Fang, Yu Kang, Jiezhong Qiu, and Pheng Ann Heng. Exploring the alignment of generation and understanding in protein structure modeling. *arXiv preprint arXiv:2607.13503*, 2026.

Wei Yang, Shunzhi Wang, Gyu Rie Lee, Jason Z Zhang, Alexis Courbet, David Juergens, Xinru Wang, Thomas Schlichthaerle, Mohamad Abedi, Robert Ragotte, et al. The past, present and future of de novo protein design. *Nature*, 652(8112):1139–1152, 2026.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast protein backbone generation with se (3) flow matching. *arXiv preprint arXiv:2310.05297*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

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.

Odin Zhang, Xujun Zhang, Haitao Lin, Cheng Tan, Qinghan Wang, Yuanle Mo, Qiantai Feng, Gang Du, Yuntao Yu, Zichang Jin, et al. Odesign: A world model for biomolecular interaction design. *arXiv preprint arXiv:2510.22304*, 2025.

A Appendix

B Supplementary analysis

B.1 Dominant Direction in Boltz-2 Pairwise Representations

Before training, we asked what cosine similarity to the Boltz-2 teacher representation would constitute evidence of pair-specific alignment. If teacher vectors share a direction, a predictor that emits the *same vector for every residue pair* could already score well. We therefore normalized each teacher pair vector $u_{ij} = Z_{ij}^{\mathrm{T}}/\|Z_{ij}^{\mathrm{T}}\|$ and fitted the best unit-norm constant predictor: the direction of their mean, $c^\star = \mathbb{E}[u_{ij}]/\|\mathbb{E}[u_{ij}]\|$. For each of 120 multichain complexes, we fitted $c^\star$ on the other 119 complexes, weighting complexes equally, and measured its cosine on the held-out complex. The mean was 0.670 (95% complex-bootstrap interval $[0.660, 0.681]$), essentially the in-fit score of 0.671. As a control that destroys pair identity, cosine between teacher vectors and a within-complex permutation of those vectors was still 0.497 $[0.485, 0.510]$; random directions scored -0.002. The constant score fell to 0.200 for interface pairs, defined as inter-chain token pairs whose representative atoms were within 8 Å in the predicted structure (Fig. 6a). Thus, a raw cosine near 0.67 would not, by itself, establish that a model had learned a pairwise mapping; the shared direction contributes least at interfaces.

The shared direction suggested that an affine projector $Wz + b$ might use its bias to represent the common offset, freeing W to map pair-dependent variation that a bias-free projector Wz must also accommodate. We tested this hypothesis in matched 1,500-epoch ProteinREPA–linear and ProteinREPA–affine runs, comparing both raw cosine and cosine after subtracting the feature means within each training microbatch. The constant-predictor score provides a reference for raw cosine only; after centering, we compare the two arms directly. At the final epoch, raw cosine was 0.8112 (linear) versus 0.8113 (affine), and centered cosine was 0.4964 versus 0.4922 (Fig. 6b). Both raw scores exceeded the constant baseline, but the affine-minus-linear gaps were only $+0.0001$ raw and -0.0042 centered. The bias therefore improved neither the observed alignment nor the centered readout of pair-dependent alignment. These are training-set measurements at random noise draws, not a held-out fixed-noise comparison; centering was a diagnostic, not the objective used to train either arm.

The shared direction suggested that an affine projector $Wz + b$ might use its bias to represent the common offset, freeing W to map pair-dependent variation that a bias-free projector Wz must also accommodate. We tested this hypothesis in matched 1,500-epoch ProteinREPA–linear and ProteinREPA–affine runs, comparing both raw cosine and cosine after subtracting the feature means within each training microbatch. At the final epoch, raw cosine was 0.8112 (linear) versus 0.8113 (affine), and centered cosine was 0.4964 versus 0.4922 (Fig. 6b). Both models exceeded the constant baseline, but the affine-minus-linear gaps were only $+0.0001$ raw and -0.0042 centered. The bias therefore improved neither the observed alignment nor the centered readout of pair-dependent alignment. These are training-set measurements at random noise draws, not a held-out fixed-noise comparison; centering was a diagnostic, not the objective used to train either arm.

B.2 Sensitivity to RFD3 Internal Representations

To assess the contribution of RFD3's internal representations, we perturbed one representation at a time in a trained model with isotropic Gaussian noise. RFD3 maintains six representations: token-single s_i, token-pair Z_{II}, atom-single conditioning c_l, atom-single q_l, atom-pair p_{lm}, and atom-to-token aggregated a_i. At diffusion noise level σ_t, we perturb a representation R as

$$\widetilde{R} = R + \sigma\, \hat{s}(R)\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{7}$$

where $\hat{s}(R)$ is the standard deviation of R over all of its entries, a single scalar per representation, and $\sigma \in \{0.1, 0.5, 1\}$ sets the perturbation magnitude relative to that scale. All other representations remain unchanged. We measure the effect on the denoised prediction using the binder lDDT loss, $1 - \mathrm{lDDT}$, of $\hat{\mathbf{x}}_0$ on 5 protein–protein binder monitoring examples, with 4 diffusion replicas at each of 20 noise levels spanning $\sigma_t \in [6.4 \times 10^{-3}, 2.6 \times 10^{3}]$. Because the baseline loss varies across the noise schedule, we report its change relative to the unperturbed baseline at each noise level.

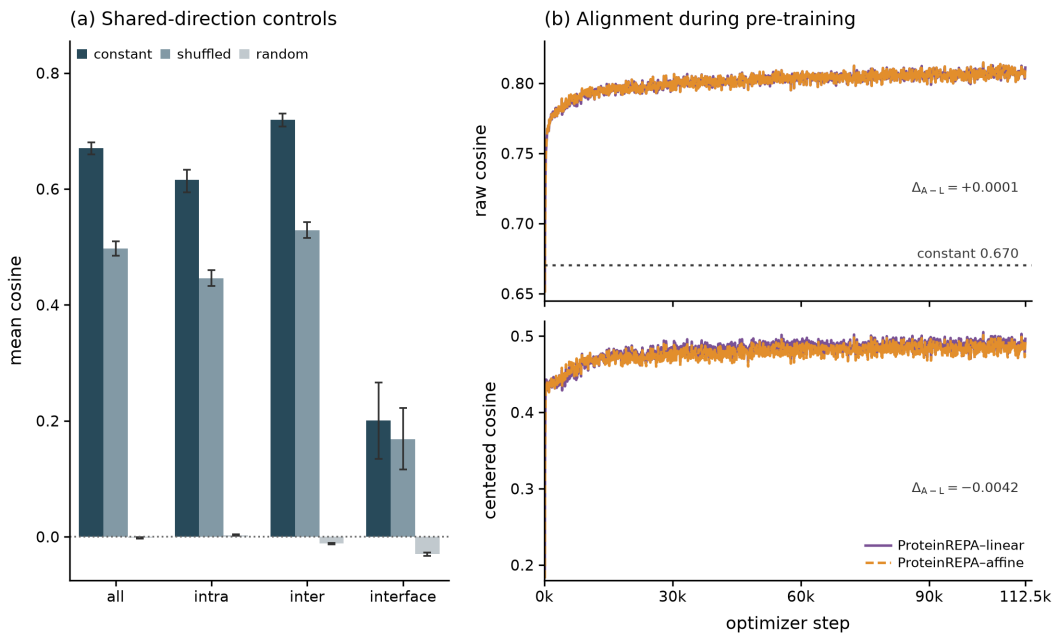

Figure 6: **Shared direction and projector comparison.** (a) Held-out constant-predictor, within-complex shuffled-pair, and random-direction cosine by pair type. Bars are complex-balanced means; error bars are 95% bootstrap intervals over complexes ($n = 120$, except $n = 73$ for interface pairs). (b) Raw (upper) and per-microbatch-centered (lower) alignment during pre-training, shown on separate vertical scales. Linear is solid purple and affine dashed orange; the dotted constant-predictor reference appears only on the raw axis. Annotations give final affine-minus-linear gaps. The alignment curves are training-set measurements, not a held-out fixed-noise evaluation.

Perturbations have small effects at both ends of the noise schedule. Below $\sigma_t \approx 0.15$, the structure is largely determined (mean $|\Delta(1 - \mathrm{lDDT})| \leq 0.01$); above $\sigma_t \approx 300$, the baseline loss approaches the metric ceiling (≥ 0.69) and the perturbation effect falls back to ≈ 0.02. The largest effects occur in the intermediate noise range and differ by representation. The single representations show their largest deviations at the lower-noise end of this range ($\sigma_t \approx 3$–6, maximum $\Delta(1 - \mathrm{lDDT})$ between 0.17 and 0.34), consistent with a role in local atomic refinement after the global arrangement has been resolved. The pair representations peak later and decay more slowly: Z_{II} reaches a maximum $\Delta(1 - \mathrm{lDDT}) = 0.19$ at $\sigma_t \approx 13$ and retains at least half of it out to $\sigma_t \approx 120$, and p_{lm} out to $\sigma_t \approx 190$, compared with $\sigma_t \leq 73$ for every single representation. Z_{II} therefore remains influential at noise levels at which the fold and relative chain placement are still being determined, even though its peak magnitude is not the largest observed (q_l is, at 0.34). Deviations increase with perturbation magnitude, while the $\sigma = 0.1$ curves remain close to baseline.

These results support Z_{II} as a relational alignment target because its perturbation affects denoising across the range in which task-relevant geometry is resolved. They do not establish it as the optimal target: representations differ in dimensionality and in the number of downstream blocks that use them, so equal relative perturbations need not remove equal amounts of information. In particular, Z_{II} is used as an attention pair bias in every subsequent block.

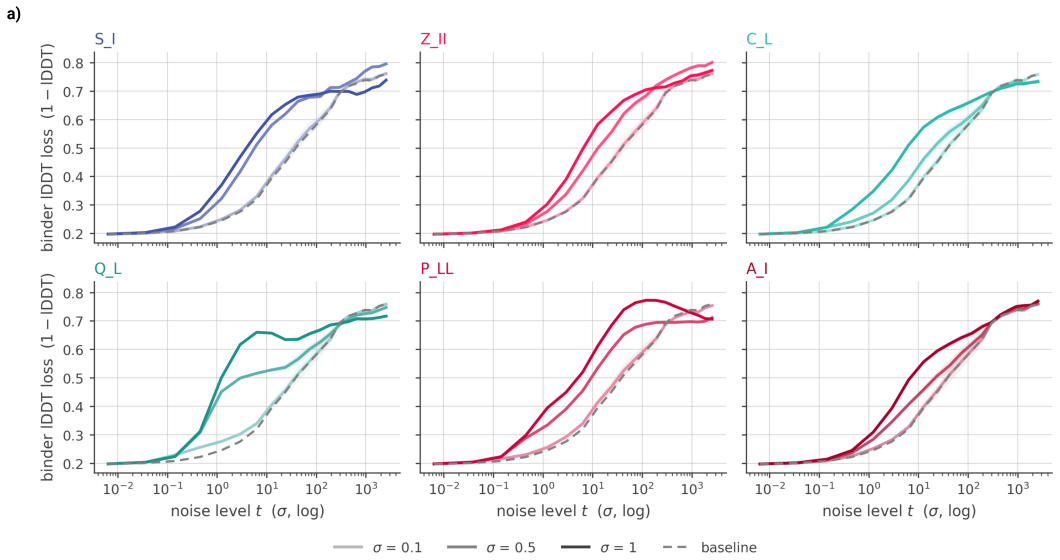

Figure 7: **Sensitivity to perturbations of RFD3 internal representations.** Change in lDDT loss relative to unperturbed baseline after adding channel-scaled Gaussian noise to one representation at a time, across diffusion noise levels and perturbation magnitudes.

B.3 How Much Alignment to the Full-MSA Z^{T} Is Meaningful?

To interpret ProteinREPA's alignment scores in the context of binder design, we examine which components of the teacher's pair representation depend on MSA information and how they relate to interface accuracy.

Because the teacher constructs Z^{T} from sequence and MSA information that RFD3 does not receive at inference, a student–teacher cosine of one is not an appropriate target. We therefore estimate a reference by subsampling the MSA for 20 complexes to depths $d \in \{1, \ldots, 4096\}$ by taking the top-d ranked sequences, recomputing Boltz-2, and measuring the cosine between $Z^{\mathrm{T}}(d)$ and its full-depth representation (mean realised depth 7231). We report this cosine separately for intra-chain, inter-chain, and interface token pairs, where interface pairs are inter-chain pairs in contact under the full-depth prediction. Depth comparisons use the 14 complexes whose MSAs are deep enough to reach every rung, so that a rung is not scored on a different subset of complexes.

The MSA dependence of Z^{T} is non-uniform (Fig. 8a). Intra-chain pairs are already similar to the full-depth representation at $d = 1$ (0.85) and reach 1.00 at $d = 4096$. Interface pairs begin at 0.62, increase more slowly, and plateau at 0.93; they also vary most across complexes (standard deviation 0.29 at $d = 1$, against 0.12 for intra-chain pairs). Thus, MSA-dependent changes are concentrated in inter-chain pairs, particularly at the interface. Consistent with this pattern, Cα lDDT rises from 0.79 to 0.94, while DockQ increases from 0.32 to 0.67 (Fig. 8b). Additional trunk recycling does not recover the missing interface component (Fig. 8f): at $d = 1$, ten recycles raise the inter-chain cosine by only 0.02 ($0.75 \rightarrow 0.78$), and at $d = 4096$ the effect is slightly negative (-0.01). The MSA-sensitive part of the teacher representation is therefore concentrated in the interface, which accounts for a small fraction of pairs and is least directly available to the student.

Against this reference, the aggregate student–teacher cosine plateaus at 0.753 for RFD3 control, 0.784 for ProteinREPA-linear, and 0.785 for ProteinREPA-affine (Fig. 8c), compared with a single-sequence teacher value of 0.835. Because intra-chain pairs dominate the aggregate, these values do not show whether the interface representation is recovered to the same degree.

The relationship between interface cosine and expected DockQ is non-linear (Fig. 8d): below a cosine of approximately 0.6, only 16% of runs reach an acceptable interface (DockQ ≥ 0.23) and the mean DockQ is 0.15, whereas above 0.85 the mean DockQ is 0.73 and 97% are acceptable. The single-sequence interface alignment (0.62) lies near the lower end of this transition, where modest

cosine increases correspond to large changes in expected accuracy. The channel-wise standard deviation of the interface block also separates these complexes by outcome without requiring a full-depth reference (Fig. 8e; Spearman $\rho = 0.82$ with DockQ at fixed depth, against 0.71 for the interface cosine itself). We interpret both trends as associations, since MSA depth covaries with representation similarity, feature scale, and design accuracy.

Together, these measurements suggest why alignment may yield different gains across benchmarks. Fold-level relational features are largely captured by sequence alone, whereas inter-chain contact features are harder to recover and remain a challenge for the student. This pattern is consistent with larger gains in native-sequence designability and enzyme motif scaffolding than in protein–protein binder design. Because interface pairs make up a small fraction of $\mathcal{M}$, a uniform average over valid pairs in $\mathcal{L}_{\text{align}}$ is dominated by intra-chain pairs. Increasing λ_{align} alone may therefore have limited effect on interface alignment. Explicitly weighting inter-chain contacts is a more direct test, which we leave to future work.

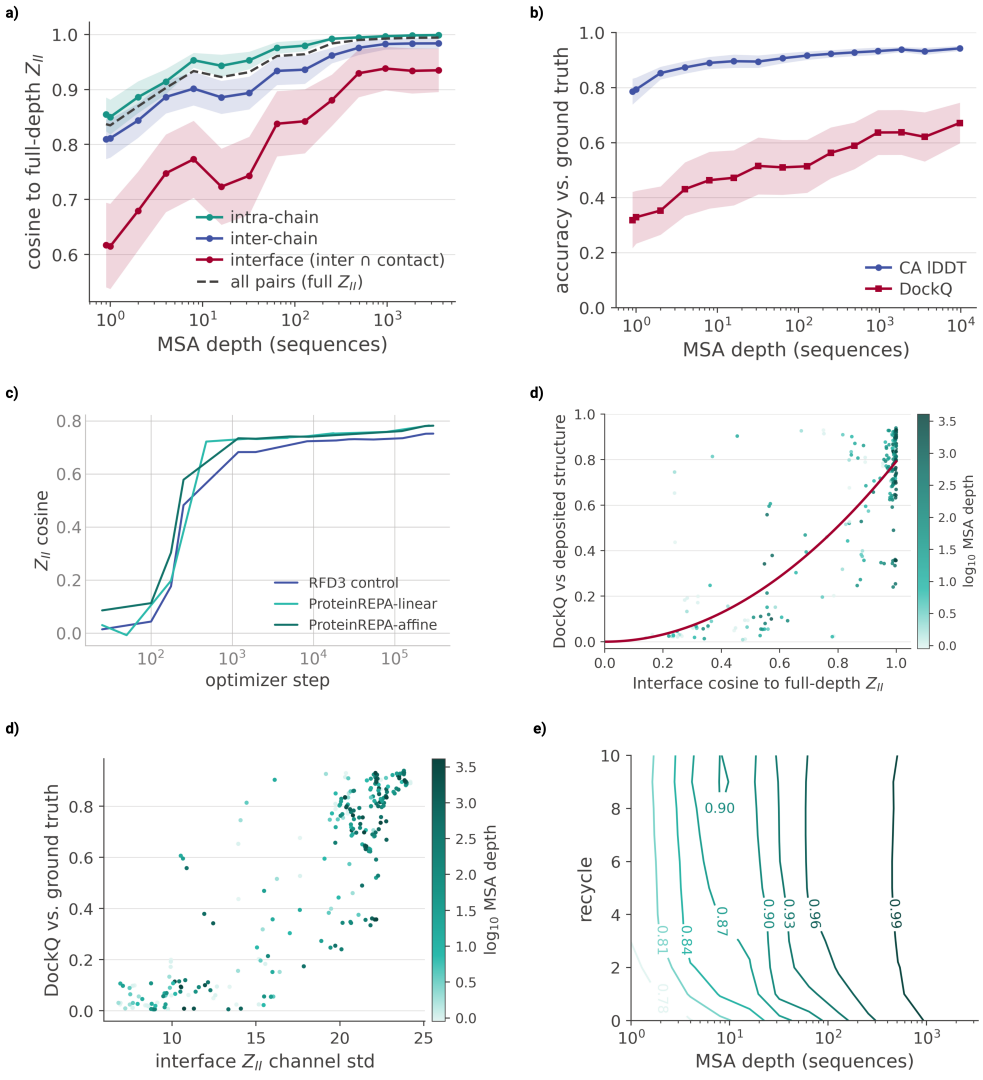

Figure 8: **MSA depth, representation alignment, and design accuracy.** (a) Cosine similarity between subsampled-MSA and full-depth Boltz-2 pair representations, by pair type. (b) Cα lDDT and DockQ across MSA depths. (c) Aggregate student–teacher cosine for n0, n1, and n3, with the single-sequence teacher reference. (d) Expected DockQ as a function of interface cosine to the full-depth teacher. (e) Interface-block channel-wise standard deviation by design outcome. (f) Effect of additional trunk recycling on the interface representation.

B.4 DISTOGRAM-GRADIENT STEERING AT INFERENCE

If the aligned representation captures relational information from the teacher, the trained projection and frozen Boltz-2 distogram head may provide an inference-time score for the student's denoised prediction. At each denoising step, we project the student pair representation with P_ψ, decode distogram probabilities p^b_{ij}, and evaluate the soft negative log-likelihood

$$\mathcal{L}_{\text{dist}}(\hat{\mathbf{x}}_0) = -\sum_{(i,j)} \sum_{b} w_b(d_{ij}(\hat{\mathbf{x}}_0)) \log p^b_{ij}, \tag{8}$$

where w_b softly assigns each predicted distance to bin b. We add $-\lambda \nabla_{\hat{\mathbf{x}}_0} \mathcal{L}_{\text{dist}}$ to [CONFIRM: the denoised prediction, the score, or the sampler drift], using $\lambda \in \{0.05, 0.15, 0.5\}$ on six binder targets. Designs are counted as passing based only on self-consistency (scRMSD < 2 Å), to detect degradation rather than rank designs. These pass rates are not comparable to Fig. 4, which uses additional criteria and a different target set from the AlphaProteo panel.

Steering reduces the self-consistency pass rate on all six targets at every tested weight (Fig. 9a). The decline is already substantial at $\lambda = 0.05$, and generation largely collapses at $\lambda \geq 0.15$. The other panels help characterize this result. The decoded distogram stabilizes no earlier than the coordinates it is intended to assess (Fig. 9b), and its association with the final outcome is weak and late (Fig. 9c). The correlation remains within the null band until $\sigma \approx 1$, then peaks at $|\rho| \approx 0.19$ for self-consistency and [0.25] for interface confidence before declining.

These results do not distinguish between two explanations. Alignment may provide a useful training gradient while producing features that are poorly calibrated as a distance likelihood, making the coordinate gradient an unsuitable descent direction. Alternatively, distograms—including the teacher's—may provide little information about mid-trajectory geometry before that geometry is formed. We did not test the control needed to distinguish these explanations: decoding distograms from Z^{T} on the same trajectories. This negative result limits the inference-time claim: alignment improves the generator during training, but this steering procedure does not provide a usable inference-time objective or uncertainty estimate.

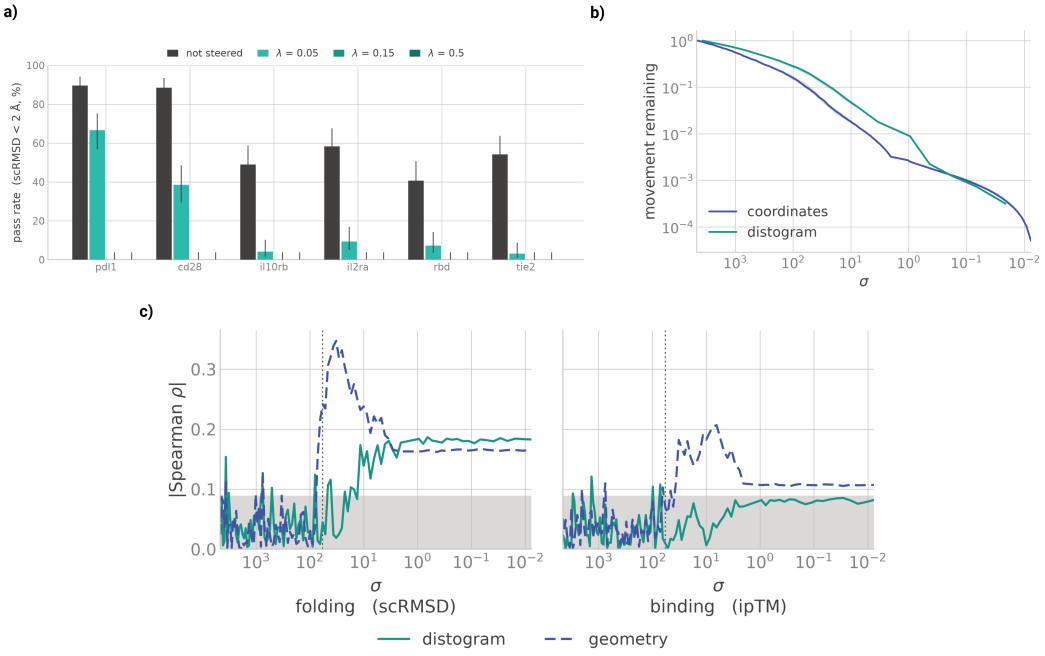

Figure 9: **Distogram-gradient steering during inference.** (a) Self-consistency pass rate across steering weights. (b) Distogram and coordinate stabilization across the denoising trajectory. (c) Association between intermediate distogram scores and final self-consistency or interface confidence.

C Limitations and Future Work

Each training arm corresponds to a single optimization run. Our confidence intervals therefore quantify variability across generated designs or benchmark targets, but not variability across independent training runs. Training multiple seeds was computationally infeasible, and differences that emerge only at individual checkpoints should consequently be interpreted cautiously. The publicly released RFD3 checkpoint is not a controlled baseline because its training data and optimization history differ from ours. In particular, computationally predicted structures account for 23% of our sampling mixture. We therefore use the matched no-alignment arm for causal comparisons and report published RFD3 only as an external reference. ProteinREPA also aligns all valid token pairs uniformly. Because intra-chain pairs greatly outnumber interface pairs, the objective may underweight the MSA-sensitive inter-chain features that are most relevant to binder design. Pair-type-aware weighting or task-adaptive alignment objectives are promising directions for future work. Finally, although the aligned representation improves training, our distogram-based experiments did not produce a calibrated inference-time score. Representation transfer and inference-time evaluation may therefore require different objectives.

D Training Datasets

ProteinREPA is trained on experimentally determined PDB assemblies and computationally predicted AFDB monomers and complexes. All three training arms use identical source pools, sampling probabilities, preprocessing, and teacher-target coverage. In particular, the no-alignment control is restricted to the same sidecar-filtered examples as the aligned arms, preventing target availability from changing the training distribution.

D.1 Experimentally Determined Structures

Experimentally determined structures are obtained from the Protein Data Bank (PDB) (Bhikadiya et al., 2019) and processed as complete biological assemblies with AtomWorks Corley et al. (2025). All deposited assemblies are used rather than only the first, and 27.6% of retained rows come from assembly ≥ 2. Protein chains, nucleic acids, ligands, ions, and other resolved molecular components are retained where supported by the RFD3 preprocessing pipeline. For each protein sequence, we construct a ColabFold MSA (environmental database release 2021-08) for use by the Boltz-2 teacher (Mirdita et al., 2022). The experimental coordinates serve as the clean RFD3 target $\mathbf{x}_0$, but are not supplied to Boltz-2, which receives only sequence, MSA, and molecular composition; no structural templates were supplied in the main training run. Structures are filtered to resolution < 9.0 Å and at most 300 polymer units per assembly. Because the resolution predicate discards entries without a reported resolution, solution and solid-state NMR structures are excluded implicitly; cryo-EM entries are retained and constitute $\sim 5\%$ of the pool. Chains with fewer than four resolved residues are removed. Unresolved atoms are reconstructed from sequence and carry zero occupancy, and residues missing a backbone frame atom are masked in full; no maximum-missing-fraction filter is applied. Where alternate conformations are present, the first is retained. Waters and hydrogens are removed, and selenomethionine is not converted to methionine. Crystallisation additives are excluded using the AlphaFold 3 excluded-ligand list, which removes 11.3% of chain-level rows. We do not deduplicate the training set. Redundancy is instead handled by AlphaFold 3-style inverse-cluster-size sampling weights over MMseqs2 clusters at 40% sequence identity and 80% coverage, with $\beta = 0.5$ and $\alpha_{\text{prot}} = 3$, $\alpha_{\text{nuc}} = 0$, $\alpha_{\text{ligand}} = 1$, $\alpha_{\text{loi}} = 5$. The clustering unit is the chain or non-polymer unit rather than the deposited entry. Note that $\alpha_{\text{nuc}} = 0$ gives nucleic-acid-only examples zero sampling weight, so although nucleic acids are retained within assemblies, pure nucleic-acid examples are effectively never drawn. Two further restrictions follow from requiring a cached teacher representation. Teacher folds were computed only for assemblies of at most $3{,}000$ tokens, and only for assemblies in which every protein chain had a completed MSA. The PDB training pool is therefore the subset of the full metadata whose chain set is fully contained in a successfully folded assembly: $834{,}718$ chain-level rows of $3{,}813{,}611$, and $905{,}717$ interface rows of $6{,}875{,}066$. The second restriction correlates with sequence novelty and is a non-random exclusion.

D.2 Structure-Prediction Distillation Data

The distillation component comprises predicted AFDB monomers and complexes (Jumper et al., 2021; Varadi et al., 2024; 2022; Evans et al., 2021). For the AFDB monomer pool, we retain source structures with mean pLDDT ≥ 90 on the conventional 0–100 scale, as reported per model by the AFDB cluster manifest. Of 2,597,930 cluster representatives, 240,293 (9.3%) pass this threshold; 226,531 of these have a successfully cached Boltz-2 teacher representation, and 226,375 enter the training manifest. This threshold is applied to the source prediction used as $\mathbf{x}_0$, not to a confidence estimate produced by Boltz-2. Within retained monomers, individual atoms with per-atom pLDDT below 70 are additionally masked to zero occupancy. No corresponding pLDDT gate was applied to the AFDB complex pool, and we apply no predicted-aligned-error, clash, compactness, or chain-completeness filters to either pool. Boltz-2 independently processes the sequence and MSA associated with each retained example; it does not receive the source coordinates. The denoising target and teacher representation may therefore describe different predicted conformations. Across 226,276 AFDB monomer examples for which both structures are available, the median Cα RMSD between the source AlphaFold model and the Boltz-2 prediction is 5.12 Å (interquartile range 2.04–10.24 Å), after global Kabsch superposition over all sequence-matched chains, with residues matched by label sequence index (Kabsch, 1976). The distribution is strongly bimodal: 24.6% of examples fall below 2 Å and 26.0% exceed 10 Å. Divergence is substantially larger for the AFDB complex pool, where the median Cα RMSD is approximately 18 Å and only $\sim$1% of heterodimers fall below 1.5 Å, indicating that Boltz-2 frequently does not reproduce the source inter-chain placement. Thus, the teacher target is not a direct encoding of the coordinates used for denoising, although for the complex pool this also means the teacher may supervise inter-chain geometry inconsistent with the denoising target.

D.3 Training-Data Sampling

At each training step, a source is selected according to Table 1, after which a structure and crop are sampled using the RFD3 data loader. The PDB source contains two internally concatenated subsets, `pn_unit` and `interface`; these do not have separately configured source probabilities. These probabilities describe sampling frequency rather than the relative number of cached structures. They also differ from the data mixture used for published RFD3, precluding a controlled comparison of absolute performance with that checkpoint.

Table 1: **Training-data sampling mixture.** Probabilities are taken from the effective configuration shared by all three training arms.

Data source	Structure type	Sampling probability
PDB	Experimental assemblies	0.770
AFDB monomer	Predicted monomers with mean pLDDT ≥ 90	0.209
AFDB complex	Predicted complexes	0.021
Experimental structures		0.770
Distillation structures		0.230

E Architecture and Implementation Details

This appendix provides additional architectural and implementation details for the RFD3 student, the Boltz-2 teacher, and the construction of the representation-alignment targets. The ProteinREPA objective itself is defined in Section 3.

E.1 RFdiffusion3 Architecture and Self-Conditioning

Given clean coordinates $\mathbf{x}_0$, RFD3 samples a noise level from its EDM training distribution Karras et al. (2022),

$$\ln(\sigma/\sigma_{\mathrm{data}}) \sim \mathcal{N}(-1.2, 1.5^2), \qquad \sigma_{\mathrm{data}} = 16 \text{ Å}, \tag{9}$$

and corrupts diffused atoms as $\mathbf{x}_\sigma = \mathbf{x}_0 + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. A single noise level is shared across all atoms in each diffusion sample, while fixed motif atoms remain uncorrupted. ProteinREPA leaves the coordinate augmentation, EDM preconditioning, denoising parameterization, and coordinate loss unchanged from RFD3; the alignment objective is applied across this complete training-noise distribution.

Initial and diffusion-dependent pair representations. RFD3 contains two pair representations with dimensionality 128. The token initializer first constructs a conditioning-dependent pair representation,

$$Z_{II}^{\text{init}} \in \mathbb{R}^{I \times I \times 128}, \tag{10}$$

from token features, relative positional encodings, covalent connectivity, and other supplied conditioning information. This representation is computed once per forward pass and does not depend on the noised coordinates or noise level.

Within the diffusion module, the token encoder combines Z_{II}^{init} with pairwise geometric features derived from the noised coordinates. The encoder contains a two-block Pairformer stack with triangular updates disabled and produces the diffusion-dependent student representation

$$Z_\ell^{\text{S}} \in \mathbb{R}^{I \times I \times 128}. \tag{11}$$

Here, ℓ denotes the fixed output of the diffusion-module token encoder. We do not select among or align multiple RFD3 layers. Subsequent diffusion-transformer and coordinate-decoder blocks consume Z_ℓ^{S} as an attention-pair bias but do not update it.

E.2 Boltz-2 Teacher and Representation Extraction

We use Boltz-2 version 2.2.1 with the `boltz2_conf.ckpt` checkpoint as the frozen structure-prediction teacher (**?**). Boltz-2 constructs single-token and pairwise representations through a 64-block PairFormer trunk. The trunk is evaluated for four iterations, corresponding to three recycling steps.

The teacher target Z^{T} is the pair representation emitted by the final PairFormer block after the final trunk iteration:

$$Z^{\text{T}} \in \mathbb{R}^{I \times I \times 128}. \tag{12}$$

This is the same trunk representation used by the Boltz-2 distogram head and as conditioning for its structure-generation module. Although Boltz-2 also produces single-token and atom-level representations, the reported ProteinREPA models align only Z^{T}.

Teacher inputs. Teacher representations are computed for complete biomolecular assemblies. For protein chains, Boltz-2 receives the native sequence and the corresponding ColabFold MSA. Nucleic-acid chains and the chemical identities of ligands and ions are retained where present. The teacher does not receive the native coordinates $\mathbf{x}_0$, and no structural templates are supplied in the main ProteinREPA training run. Its inputs can therefore be summarized as

$$\mathbf{u}\left(\text{native sequence}, \text{MSAs}, \text{molecular composition}\right). \tag{13}$$

Because the sequence and MSAs of generated residues are unavailable to RFD3 during inference, the teacher representation constitutes privileged training-time supervision.

Relationship to the denoising target. The teacher representation is a Boltz-2 prediction from the native sequence and associated MSA; it is not computed from the coordinates used as the clean RFD3 denoising target. For experimentally determined structures, both are associated with the same native assembly. For examples originating from the AlphaFold Database, $\mathbf{x}_0$ is the deposited AlphaFold model, whereas Z^{T} is obtained from an independent Boltz-2 prediction of its sequence Jumper et al. (2021); Varadi et al. (2024; 2022); Evans et al. (2021). These structures have a median Cα RMSD of approximately 7.15 Å in the current training set. Thus, ProteinREPA transfers the predictor's representation of the sequence rather than directly encoding the target coordinates into the teacher representation.

Table 2: **Cached Boltz-2 teacher representations used for ProteinREPA training.** The template-derived pool was generated for separate experiments and is not included in the main training mixture.

Teacher-representation pool	Number of structures
`repa_pdb_msa_boltz2`	144,985
`repa_asm2plus_boltz2`	74,262
`repa_afdb_msa_boltz2`	226,531
`repa_afdb_complex_boltz2`	23,539
Main training pools, total	469,317

E.3 Teacher-Target Precomputation and Token Mapping

All Boltz-2 teacher representations are computed before RFD3 training. Boltz-2 is never loaded during ProteinREPA optimization. For each successfully processed structure, we store Z^{T} in float32 precision together with stable identifiers that map every teacher token back to its chain and residue or molecular component.

The current teacher-representation pools contain:

RFD3 samples cropped structures during training, whereas the cached teacher representation is computed for the complete assembly. After cropping, the stored token identifiers are used to gather the corresponding rows and columns of Z^{T} into the RFD3 crop order. Let

$$m_i \in 0, 1 \tag{14}$$

indicate whether RFD3 token i has a valid matched teacher token. The valid pair set is

$$\mathcal{M}\,(i, j); :; m_i m_j = 1 \tag{15}$$

Only pairs for which both endpoints are matched contribute to the representation-alignment objective. Cropping therefore changes which elements of the cached representation are supervised but does not require recomputing the Boltz-2 teacher.

E.4 Representation Projection and Alignment

Both the student and teacher pair representations have feature dimension 128. We compare a bias-free linear projection with an affine projection while keeping all other model and training settings fixed. All reported objectives use uncentered cosine similarity: no feature mean is subtracted before normalization.

Table 3: **ProteinREPA training arms.** Internal run identifiers are included for correspondence with released configurations and checkpoints.

Model	Run ID	Projection	λ_{align}
RFD3 control	n0	None	0
ProteinREPA–linear	n1	Wz	0.5
ProteinREPA–affine	n3	$Wz + b$	0.5

For diffusion sample d and valid token pair (i, j), the normalized projected student and teacher representations are

$$\bar{Z}^{\mathrm{S},d}_{ij} \frac{P_\psi\left(Z^{\mathrm{S},d}_{\ell,ij}\right)}{\left|P_\psi\left(Z^{\mathrm{S},d}_{\ell,ij}\right)\right|_2}, \qquad \bar{Z}^{\mathrm{T}}_{ij} \frac{Z^{\mathrm{T}}_{ij}}{\left|Z^{\mathrm{T}}_{ij}\right|_2}. \tag{16}$$

The alignment loss is averaged over both diffusion samples and valid token pairs:

$$\mathcal{L}_{\mathrm{align}} \frac{1}{D} \sum_{d=1}^{D} \frac{1}{|\mathcal{M}|} \sum_{(i,j)\in\mathcal{M}} \left(1 - \bar{Z}^{\mathrm{S},d}_{ij} \cdot \bar{Z}^{\mathrm{T}}_{ij}\right). \tag{17}$$

A single cached teacher target is shared across the D noised diffusion samples drawn for the same structure. Because the teacher targets are precomputed constants, gradients propagate only through the projection and RFD3.

The complete objective is

$$\mathcal{L}_{\mathrm{RFD3}} + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}, \tag{18}$$

with $\lambda_{\mathrm{align}} = 0.5$. The alignment weight is constant throughout training and the loss is applied across the complete RFD3 training noise distribution.

E.5 Model Training

All arms are initialized from random weights and use the same RFD3 architecture, optimizer, data mixture, and seed schedule. ProteinREPA is active from the first optimizer step with constant weight $\lambda_{\mathrm{align}} = 0.5$ and is neither warmed up nor annealed separately. Training proceeds in two stages. Base-context training uses crops of 256 tokens and at most 2,560 atoms, with 32 diffusion samples per structure. Long-context fine-tuning increases these limits to 384 tokens and 3,840 atoms while reducing the number of diffusion samples to 16. The optimizer and learning-rate scheduler are continued rather than restarted (Table 4). The learning curves in Fig. 2 are obtained during base-context training. The pretraining phase is trained for 112,500 optimizer steps, and followed by an additonal 45,000 fine-funing steps. The final unconditional evaluation and all three protein-design benchmarks use the long-context checkpoints after 157,500 steps. These stages are analyzed separately because they use different crop sizes and diffusion batch sizes.

Table 4: **Training configuration shared by the three model arms.**

Setting	Base-context training	Long-context fine-tuning
Crop size	256 tokens	384 tokens
Maximum atoms	2,560	3,840
Diffusion samples per structure	32	16
Effective batch size	32	32
EMA decay	0.999	0.999
Base learning rate	1.8×10^{-3}	Continued
Warmup	1,000 optimizer steps	Not restarted
Decay	Factor 0.95 every 50,000 steps	Continued
GPUs	16 NVIDIA H100	16 NVIDIA H100
Seed	42	42

E.6 Inference and Sampling Parameters

Unless stated otherwise, RFD3 generation uses 200 denoising steps, $\gamma_0 = 0.6$, step scale 1.5, and two recycling iterations. Recycling is implemented through coordinate-mediated self-conditioning: the detached denoised prediction from the preceding iteration is converted into pairwise geometric features, and the diffusion-dependent pair representation is recomputed. Classifier-free guidance and autoguidance are disabled. For each target and model arm, we generate 200 backbones and design eight sequences using ProteinMPNN or LigandMPNN. Each sequence is evaluated using five AlphaFold 3 samples with seed 1. Benchmark success is assigned at the backbone level when at least one designed sequence satisfies every task-specific criterion. The ten-recycle experiments change only the number of RFD3 recycling iterations.

F Benchmarking Details

F.1 Training probe

We used native-sequence probes to monitor the control, linear, and affine models, not to replace the full $k = 8$ benchmarks or the separate 1,000-design endpoint evaluation. Per arm and checkpoint, we generated 200 length-150 unconditional backbones and 96 designs for each of six enzyme targets (AME M0054, M0129, M0157, M0255, M0664, M0907) (Figure 10) and six binder targets (CD28,

Table 5: **Default sampling and evaluation configuration.**

Setting	Value
Denoising steps	200
γ_0	0.6
Step scale	1.5
Recycling iterations	2
Backbones per target	200
Sequences per backbone	8
AlphaFold 3 samples per sequence	5

IL10RB, IL2RA, PD-L1, RBD, Tie2)(Figure 11). Enzyme designs spanned 150–250 residues; designed binders were 100 residues. Conditional pre-training probes were sampled every 50 epochs, unconditional probes every 100, and all three arms every 100 epochs during fine-tuning from 1599 through the matched endpoint at 2099 (75 optimizer steps per epoch). The plots below show pooled curves and individual targets to expose target variability. Conditional refolding used RF3; unconditional refolding used AF3 before fine-tuning and RF3 afterward, with the same scRMSD criterion.

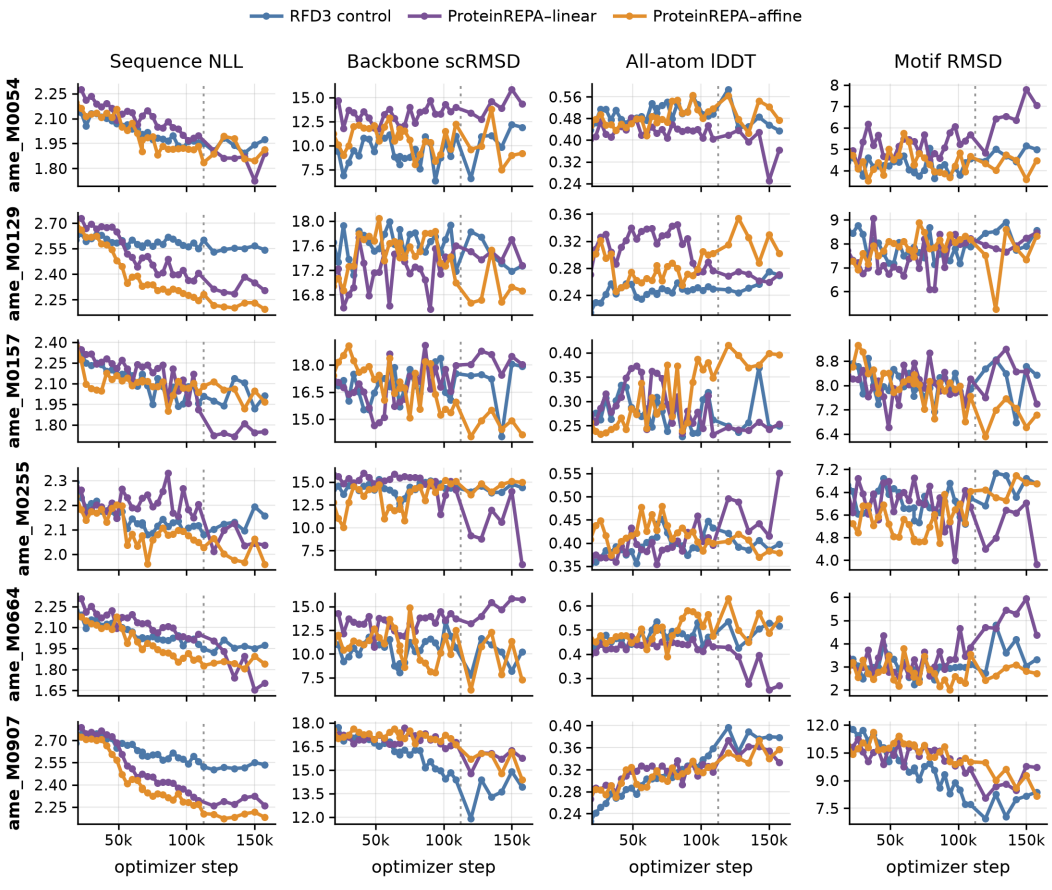

Figure 10: **Enzyme design training probe.** Selected targets for the enzyme design training probe, illustrating target variability. The vertical grey dotted line shows the transition from base-contex training to fine-tuning.

Probe Results The probe results of the experiment sections are listed in the following tables:

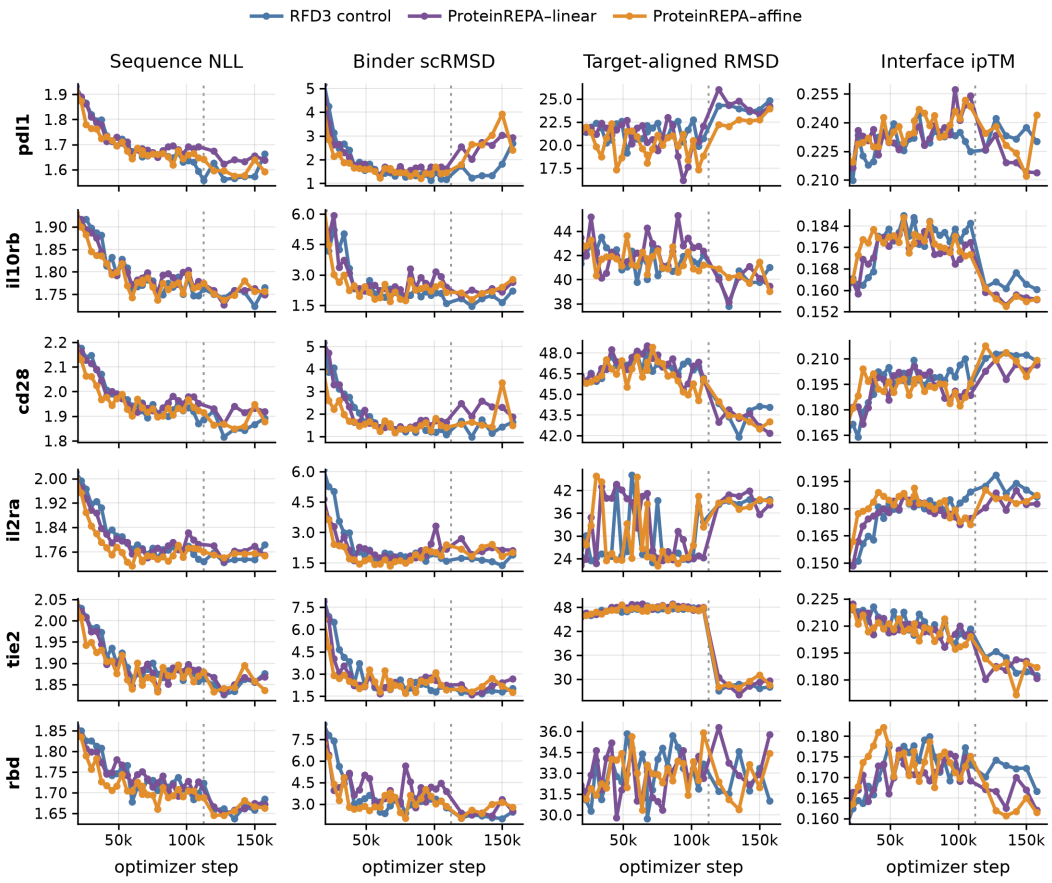

Figure 11: **Protein-protein binder design training probe.** Selected targets for the protein–protein binder design training probe, illustrating target variability. The vertical grey dotted line shows the transition from base-contex training to fine-tuning.

F.2 Enzyme Motif Scaffolding

We evaluate the 41-target atomic motif enzyme benchmark introduced with RFdiffusion2 (Ahern et al., 2026). Each target specifies catalytic residues and a bound ligand around which the model must construct a compatible scaffold. We retain the original ligand coordinates but place each ligand on a separate chain for evaluation. For each backbone, we design eight LigandMPNN sequences and evaluate them with AlphaFold 3 rather than the Chai-1 predictor used in the original benchmark Dauparas et al. (2025); Abramson et al. (2024); Discovery et al. (2024). A prediction is successful when the all-heavy-atom RMSD of the catalytic residues is $\leq$ 1.5 Å after alignment on their backbone N, Cα, and C atoms, and when no non-covalently linked design atom lies within 1.5 Å of a ligand atom. A backbone passes when at least one sequence satisfies both conditions. Scaffold scRMSD, ligand RMSD, and motif-sequence recovery are reported as diagnostics but do not enter the AME success gate.

F.3 Protein–Protein Binder Design

We adopt the ten protein–protein targets introduced by AlphaProteo (Zambaldi et al., 2024), including its deposited structures, cropped target regions, hotspot residues, and target-specific binder-length ranges. Hotspots are supplied as generation constraints and are also used to initialize the binder orientation, but hotspot recovery is not part of the success gate. For each target and model arm, we generate 200 backbones at batch size four and design eight ProteinMPNN sequences per backbone (Table 9). Candidates are independently evaluated using AlphaFold 3. A backbone is suc-

Table 6: Unconditional native-sequence probes. Optimizer step is $75(\text{epoch} + 1)$. Columns follow the corresponding 1×4 probe figure; C, L, and A denote control, linear, and affine, respectively. The full checkpoint grid is tabulated, including early points outside the figures' 20,000-step display window. Each probe uses 200 length-150 backbones per arm; designability is the percentage with self-consistency scRMSD < 2.0 Å. The separate epoch-2099 endpoint uses 1,000 backbones per arm and is not part of the probe trajectories or AULC. Endpoint lDDT was not measured.

Optimizer step	NLL (nats/res)			scRMSD (Å)			All-atom lDDT			Designability (%)		
	C	L	A	C	L	A	C	L	A	C	L	A
Base-context training probes												
7,500	2.454	2.422	2.425	12.70	11.92	11.94	0.410	0.426	0.413	0.5	2.0	1.5
15,000	2.239	2.208	2.217	7.28	7.40	6.99	0.546	0.537	0.558	10.5	11.0	9.5
22,500	2.186	2.172	2.161	4.36	5.17	4.95	0.632	0.586	0.616	19.5	15.0	17.0
30,000	2.147	2.116	2.088	3.63	3.69	5.07	0.693	0.650	0.643	26.5	24.5	22.0
37,500	2.067	2.099	2.074	3.91	2.67	3.77	0.679	0.716	0.674	29.5	35.0	27.0
45,000	2.053	2.085	2.067	2.90	2.91	2.58	0.703	0.712	0.731	27.5	31.0	36.5
52,500	2.003	2.056	1.977	3.15	3.73	2.27	0.715	0.678	0.753	29.0	30.0	42.0
60,000	2.014	2.039	1.985	2.37	2.83	2.70	0.764	0.736	0.751	36.5	34.0	37.5
67,500	1.999	2.021	1.941	2.44	2.50	2.30	0.757	0.745	0.765	41.0	42.0	45.0
75,000	1.971	1.962	1.939	2.49	2.59	2.06	0.762	0.738	0.790	42.0	41.5	48.0
82,500	1.972	1.962	1.937	2.53	2.61	1.96	0.777	0.749	0.799	42.0	42.5	51.0
90,000	1.984	1.935	1.920	2.37	2.36	1.92	0.758	0.749	0.787	40.5	45.0	51.5
97,500	1.947	1.969	1.881	2.27	2.43	1.93	0.768	0.767	0.801	44.0	46.0	52.5
105,000	1.917	1.930	1.877	2.28	2.14	2.08	0.777	0.791	0.799	41.5	47.5	47.5
112,500	1.903	1.909	1.885	2.32	2.40	1.88	0.779	0.777	0.803	43.5	42.5	54.5
Fine-tuning probes												
120,000	1.960	1.986	1.907	2.82	2.62	1.95	0.761	0.750	0.810	38.5	38.0	53.0
127,500	1.949	1.949	1.885	2.82	2.16	1.72	0.742	0.773	0.810	34.5	47.0	55.0
135,000	1.958	1.919	1.943	2.67	2.47	2.12	0.743	0.760	0.788	35.0	41.0	47.0
142,500	1.972	1.879	1.852	2.56	2.23	2.12	0.756	0.777	0.780	40.5	45.5	44.5
150,000	1.965	1.890	1.886	2.88	2.43	2.05	0.739	0.769	0.790	31.5	40.0	48.0
157,500	1.915	1.888	1.818	2.71	2.62	1.94	0.749	0.762	0.800	37.0	41.0	51.0
Independent final evaluation (1,000 designs per arm)												
157,500	1.936	1.929	1.853	5.73	4.20	5.15	–	–	–	34.5	51.3	43.8

cessful when at least one sequence satisfies all three AlphaProteo AF3 criteria: minimum inter-chain pAE ≤ 1.5 Å, binder pTM ≥ 0.8, and target-aligned binder Cα RMSD ≤ 2.5 Å. Hotspot-contact statistics are recorded diagnostically but do not affect the reported pass rates.

Alternative Ways to Benchmark Protein–Protein Binding Benchmarks: Since protein–protein binder design benchmarks vary so widely in literature we list multiple approaches. We chose our benchmarking approach based on the latest published methods we analyzed and aimed to go with the most similar approach to ease comparison (Table 10).

F.4 Protein–Ligand Binder Design

We evaluate the four ligand targets shared by Complexa and RFD3: SAM (PDB 7C7M), FAD (7BKC), IAI (5SDV), and OQO (7V11). The ligand pose is fixed and each generated protein contains 100 residues. We generate 200 backbones per target and design eight LigandMPNN sequences for each backbone. A backbone is successful when at least one independently predicted sequence satisfies all of the following: scaffold scRMSD ≤ 1.5 Å, backbone-aligned ligand RMSD ≤ 5.0 Å, minimum inter-chain pAE ≤ 1.5 Å, and ipTM ≥ 0.8.

Table 7: Enzyme native-sequence probes. Optimizer step is $75(\text{epoch} + 1)$. Columns follow the corresponding 1×4 probe figure; C, L, and A denote control, linear, and affine, respectively. The full checkpoint grid is tabulated, including early points outside the figures' 20,000-step display window. Values are means across six targets of within-target summaries from 96 designs per arm, target, and checkpoint; NLL is scored at temperature 1. The 29 pre-training checkpoints here are the regular 50-epoch grid used for the checkpoint-win counts; two additional irregular, partial archive points shown in the curves are not included.

Optimizer step	NLL (nats/res)			scRMSD (Å)			All-atom lDDT			Motif RMSD (Å)		
	C	L	A	C	L	A	C	L	A	C	L	A
Base-context training probes												
3,825	2.745	2.702	2.691	19.27	18.84	19.53	0.216	0.226	0.238	11.19	10.35	10.08
7,575	2.602	2.565	2.607	16.49	16.21	16.45	0.286	0.302	0.297	8.12	8.07	7.89
11,325	2.469	2.528	2.491	15.12	15.26	15.02	0.297	0.327	0.330	7.40	7.05	6.82
15,075	2.431	2.498	2.453	15.32	15.70	14.48	0.295	0.342	0.349	7.56	6.92	6.74
18,825	2.389	2.456	2.409	14.70	15.94	14.50	0.319	0.329	0.359	7.19	6.89	6.60
22,575	2.377	2.452	2.363	14.48	16.03	14.22	0.334	0.341	0.365	7.21	6.68	6.58
26,325	2.318	2.385	2.303	13.62	14.89	13.66	0.346	0.362	0.376	6.96	6.76	6.42
30,075	2.346	2.406	2.304	13.87	15.51	14.50	0.364	0.358	0.354	6.62	6.75	6.89
33,825	2.356	2.372	2.299	14.39	15.31	15.17	0.350	0.346	0.354	6.78	6.70	6.41
37,575	2.334	2.373	2.287	14.11	15.45	15.05	0.363	0.345	0.343	6.85	6.95	6.64
41,325	2.319	2.376	2.298	14.49	15.58	14.83	0.363	0.368	0.359	6.68	6.69	6.43
45,075	2.316	2.374	2.278	14.56	15.31	15.07	0.355	0.369	0.359	6.89	6.72	6.57
48,825	2.286	2.313	2.290	14.20	14.36	15.15	0.356	0.402	0.368	6.52	6.37	6.84
52,575	2.297	2.339	2.231	14.72	14.97	14.94	0.347	0.388	0.365	6.77	6.39	6.43
56,325	2.279	2.300	2.158	14.59	15.25	14.08	0.361	0.377	0.391	6.69	6.55	6.48
60,075	2.282	2.271	2.188	14.73	15.65	15.37	0.359	0.360	0.343	6.52	6.67	7.03
63,825	2.246	2.253	2.163	13.85	15.17	14.48	0.374	0.387	0.381	6.33	6.72	6.74
67,575	2.249	2.278	2.134	13.37	14.90	14.40	0.387	0.367	0.381	6.14	7.16	6.44
71,325	2.231	2.287	2.123	13.55	15.13	13.68	0.386	0.379	0.387	6.02	6.52	6.54
75,075	2.202	2.246	2.127	14.34	15.77	14.93	0.372	0.363	0.375	6.41	7.05	6.55
78,825	2.242	2.224	2.105	14.18	15.35	14.22	0.365	0.369	0.385	6.72	6.57	6.42
82,575	2.226	2.244	2.121	13.87	15.65	13.87	0.380	0.380	0.405	5.95	6.35	6.21
86,325	2.207	2.233	2.071	14.70	15.94	14.29	0.363	0.350	0.382	6.22	6.88	6.25
90,075	2.242	2.221	2.088	14.67	15.57	14.40	0.365	0.372	0.391	6.40	6.42	6.50
93,825	2.204	2.175	2.105	13.93	15.69	13.58	0.374	0.366	0.418	6.13	6.73	6.25
97,575	2.181	2.205	2.090	14.10	14.97	13.84	0.377	0.363	0.410	6.38	6.56	6.40
101,325	2.212	2.140	2.089	14.46	15.50	14.77	0.371	0.345	0.394	5.97	6.93	6.56
105,075	2.229	2.165	2.079	13.99	15.57	14.04	0.388	0.373	0.413	5.95	6.69	6.15
108,825	2.186	2.144	2.068	14.20	15.70	14.91	0.379	0.342	0.394	6.22	6.91	6.74
Fine-tuning probes												
120,000	2.156	2.037	2.055	12.58	14.34	12.64	0.407	0.372	0.446	6.17	6.25	6.17
127,500	2.187	2.035	2.048	14.46	14.62	13.88	0.349	0.363	0.414	7.08	6.86	5.72
135,000	2.204	2.000	2.041	13.97	15.69	14.84	0.365	0.328	0.385	6.60	7.24	6.60
142,500	2.188	2.051	1.998	13.15	14.94	12.66	0.399	0.358	0.420	6.22	7.02	6.37
150,000	2.187	1.974	2.050	14.22	16.35	13.90	0.378	0.295	0.410	6.63	7.71	6.06
157,500	2.199	1.991	2.009	14.28	14.52	12.81	0.374	0.340	0.409	6.71	6.80	6.23

Table 8: Binder native-sequence probes. Optimizer step is $75(\text{epoch} + 1)$. Columns follow the corresponding 1×4 probe figure; C, L, and A denote control, linear, and affine, respectively. The full checkpoint grid is tabulated, including early points outside the figures' 20,000-step display window. Values are means across six targets of within-target summaries from 96 designs per arm, target, and checkpoint; NLL is scored at temperature 1. The 29 pre-training checkpoints here are the regular 50-epoch grid used for the checkpoint-win counts; two additional irregular, partial archive points shown in the curves are not included.

Optimizer step	NLL (nats/res)			scRMSD (Å)			Target-aligned RMSD (Å)			Interface ipTM		
	C	L	A	C	L	A	C	L	A	C	L	A
Base-context training probes												
3,825	2.260	2.246	2.247	15.09	14.60	15.59	37.41	35.03	34.53	0.144	0.151	0.142
7,575	2.110	2.105	2.109	10.69	10.62	10.17	35.52	35.36	34.83	0.172	0.174	0.173
11,325	2.054	2.049	2.059	9.10	8.51	8.33	35.17	35.10	36.22	0.174	0.176	0.174
15,075	2.015	2.021	2.017	7.36	7.05	6.22	35.85	37.22	35.53	0.178	0.177	0.176
18,825	1.996	1.989	1.985	6.37	6.35	5.43	35.29	35.19	36.06	0.179	0.180	0.181
22,575	1.975	1.967	1.950	5.37	4.98	4.11	36.12	35.24	35.90	0.179	0.183	0.185
26,325	1.954	1.939	1.891	5.22	3.86	2.85	35.19	37.92	36.94	0.182	0.187	0.191
30,075	1.931	1.914	1.871	3.95	3.15	2.62	35.94	35.51	38.39	0.184	0.189	0.197
33,825	1.916	1.902	1.858	3.76	3.25	2.82	35.86	38.50	38.29	0.185	0.190	0.195
37,575	1.906	1.870	1.830	3.39	2.96	2.13	35.77	38.40	35.45	0.187	0.192	0.197
41,325	1.840	1.829	1.820	2.29	2.21	2.09	36.16	39.14	36.30	0.197	0.194	0.197
45,075	1.841	1.826	1.803	2.52	2.38	1.87	39.25	38.75	35.09	0.195	0.197	0.198
48,825	1.821	1.833	1.806	2.17	2.60	2.04	35.76	38.88	35.64	0.198	0.190	0.197
52,575	1.838	1.829	1.823	2.53	2.56	2.15	36.03	38.64	37.19	0.197	0.195	0.194
56,325	1.820	1.811	1.778	2.06	2.12	1.92	39.71	38.65	36.19	0.195	0.195	0.198
60,075	1.781	1.781	1.757	1.78	1.92	1.73	37.41	37.31	39.36	0.200	0.197	0.198
63,825	1.794	1.794	1.804	1.90	2.13	2.41	36.13	38.74	35.38	0.196	0.196	0.193
67,575	1.786	1.804	1.782	1.98	2.20	2.06	35.55	39.28	35.08	0.200	0.197	0.199
71,325	1.783	1.797	1.778	1.99	2.14	2.02	35.62	36.03	35.87	0.199	0.197	0.196
75,075	1.799	1.786	1.777	2.10	2.05	1.80	38.47	35.52	35.63	0.195	0.195	0.197
78,825	1.774	1.774	1.754	1.76	2.24	1.62	35.82	35.20	35.82	0.198	0.195	0.198
82,575	1.781	1.797	1.781	1.93	2.53	2.20	36.01	36.40	35.39	0.197	0.194	0.194
86,325	1.781	1.796	1.772	2.11	1.96	1.94	36.35	36.29	35.96	0.195	0.193	0.195
90,075	1.788	1.811	1.767	2.03	2.52	2.12	36.34	37.28	35.14	0.196	0.191	0.196
93,825	1.801	1.793	1.795	2.05	2.26	2.35	35.70	36.06	35.57	0.196	0.192	0.191
97,575	1.773	1.795	1.770	1.92	2.51	2.18	35.97	35.25	34.86	0.195	0.196	0.194
101,325	1.788	1.820	1.797	1.98	2.74	2.34	36.12	35.81	36.02	0.197	0.192	0.190
105,075	1.768	1.802	1.783	1.96	2.53	2.13	37.52	36.25	37.15	0.198	0.195	0.193
108,825	1.751	1.803	1.777	1.74	2.33	2.03	37.27	35.93	37.07	0.196	0.195	0.195
Fine-tuning probes												
120,000	1.760	1.780	1.740	1.80	2.39	1.94	34.33	35.73	34.46	0.192	0.186	0.194
127,500	1.722	1.737	1.738	1.55	1.99	2.05	34.09	34.88	34.14	0.198	0.191	0.190
135,000	1.729	1.768	1.738	1.73	2.26	2.17	34.71	34.60	33.63	0.193	0.183	0.189
142,500	1.739	1.766	1.758	1.66	2.32	2.47	34.56	35.15	34.34	0.193	0.189	0.184
150,000	1.735	1.777	1.770	1.67	2.56	2.81	34.69	33.97	34.88	0.193	0.186	0.184
157,500	1.777	1.766	1.746	2.15	2.46	2.21	34.75	34.90	34.69	0.189	0.184	0.191

Table 9: **Binder design benchmark targets.** Targets, cropped regions, epitope hotspots and binder length ranges follow AlphaProteo (Zambaldi et al., 2024). All ten targets are scored. Numbering is that of the deposited entry and residue identities are read from the deposited coordinates; hotspots are named by chain only where the target is multi-chain, and are restrained at Cβ, or at Cα where marked * (glycine). *Target res.* counts residues *resolved* in the cropped span, which sets the total complex length and falls below the span width wherever the deposited model has gaps.

Target	PDB	Chains and spans	Target res.	Hotspots	Binder length	Total length
bhrf1	2WH6	A 2–158	157	F65, T74, E77, D82, S85, R93	80–120	237–277
il7ra	3DI3	B 17–209	193	V58, L80, Y139	50–120	243–313
insulin	4ZXB	E 6–155	150	F64, F88, F96	40–120	190–270
pdl1	5O45	A 17–132	116	Y56, M115, Y123	50–120	166–236
sc2rbd	6M0J	E 333–526	194	G485*, Y489, S494, T500, Y505	80–120	274–314
trka	1WWW	X 282–382	101	V294, M296, L333	50–120	151–221
vegfa	1BJ1	V 14–107; W 14–107	188	M81 (W), I83 (W), I91 (W)	50–140	238–328
h1	5VLI	A 1–50, 76–80, 107–111, 258–322; B 501–568, 580–670	281	W521 (B), I545 (B), V552 (B)	40–120	321–401
il17a	4HSA	A 17–29, 41–131; B 19–127	213	Q94 (A), L116 (A), W67 (B)	50–140	263–353
tnfa	1TNF	A 12–157; B 12–157; C 12–157	438	P113 (A), H73 (C)	50–120	488–558

Table 10: **Published in-silico success criteria for de novo binder design.** The thresholds are not comparable across rows: the same nominal quantity appears at 1.5, 7, 10 and 10.85 Å because each protocol is calibrated to a different verifier. Two further pitfalls are worth noting. BindCraft's confidence metrics are on a 0–1 scale, so its `i_pAE` < 0.35 is $0.35 \times 31 = 10.85$ Å and its pLDDT > 0.8 is Bennett's > 80; the AF2-IG-easy row is therefore the same filter as the BindCraft row, expressed in Å. And "RMSD" denotes three distinct measurements: target-aligned binder RMSD (a docking accuracy), binder self-consistency RMSD (a foldability check), and the RMSD between the binder predicted alone and in complex. AlphaProteo defines both the AF3 and the AF2-IG protocol, which is why RFD3 and Complexa can each correctly describe their criteria as AlphaProteo's while using different numbers.

Protocol	Verifier	Thresholds	Source / adopted by
AF3	AF3	min inter-chain pAE ≤ 1.5 Å; binder pTM ≥ 0.8; target-aligned binder Cα RMSD < 2.5 Å	AlphaProteo (**?**); used by RFD3 (**?**)
AF2-IG	AF2 + initial guess	interface pAE < 7.0 Å; complex pLDDT > 0.9; binder scRMSD < 1.5 Å	AlphaProteo (**?**); used by Complexa (**?**)
AF2-IG-easy	AF2 + initial guess	interface pAE < 10.85 Å; ipTM > 0.5; pLDDT > 0.8; binder bound/unbound RMSD < 3.5 Å	BindCraft (**?**), as tabulated by ProtDBench (**?**)
Protenix	Protenix / Protenix-Mini	binder ipTM > 0.85; binder pTM > 0.88; complex RMSD < 2.5 Å	ProtDBench (**?**)
Bennett-style	AF2 + initial guess	$\text{pAE}_{\text{interaction}} < 10$; binder pLDDT > 80	**?**; widely adopted as a convention
BindCraft (full)	AF2 (design-time)	pLDDT > 0.8; pTM > 0.55; i_pTM > 0.5; i_pAE < 0.35; binder RMSD < 3.5 Å; shape complementarity > 0.6; surface hydrophobicity < 0.35; ≥ 7 interface residues; ≥ 3 interface H-bonds; ≤ 4 unsatisfied H-bonds	BindCraft (**?**), `default_filters.json`
This work	AF2 + initial guess	the AF2-IG protocol above, plus a separate epitope term: ≥ 0.75 of the specified hotspot residues within 5 Å of any binder heavy atom	—