RECYCLE ONLY WHEN YOU HAVE TO!

Anonymous authors
Paper under double-blind review

ABSTRACT

AlphaFold 3-like models (e.g., Boltz, Protenix, etc.) have pushed the frontier of protein binder design. This architecture is used both for the generation of binder candidates (e.g., by hallucination or converting these models into backbone or full atom generative models) as well as for the screening of these candidates for valid binders. In this paper, we start from the observation that AlphaFold 3's confidence metrics (ipSAE, ipTM) as well as structural consistency metrics (scRMSD) can be predicted directly from the trunk *even at early* recycling. Because our proposed predictor is cheap, differentiable, and available at every cycle, it can be used in two distinct ways: as a decision signal to wisely spend the recycling budget for screening and as an objective for hallucination-based design. We show that both applications lead to increased performance at significantly reduced computational cost.

1 INTRODUCTION

Several methods have recently succeeded in designing *de novo* minibinders (Watson et al., 2023; Mirdita et al., 2022; Pacesa et al., 2025; Stark et al., 2025) and antibodies for specific epitopes (Bennett et al., 2026; Mille-Fragoso et al., 2026; Swanson et al., 2025; Stark et al., 2025). *De novo* binder design typically involves a generative as well as a scoring step. The former creates plausible candidate protein binders, while the latter selects promising candidates for experimental validation. Our work improves both steps.

The screening step typically involves scoring the design by refolding its sequence. Designs are rejected if their refolds have low confidence or are structurally incompatible with the generative step. For state-of-the-art design pipelines, such as RFDiffusion and BindCraft, the computational cost is dominated by this refold. This is because the generative step of these pipelines involves an inverse folding that typically samples multiple sequences per design. As a concrete example, BindCraft typically uses 20 sequences per design. All of these sequences need to be refolded. The screening step has been shown to provide a significant boost to the binding rate (Swanson et al., 2025; Pacesa et al., 2025; Bennett et al., 2026). Furthermore, recent work suggests that sampling more candidates followed by more stringent screening also improves binding (Candido et al., 2026). A significant reduction of the numerical cost at the same refold fidelity is thus highly desirable.

The generative step can be implemented in different ways. A prominent approach involves hallucination Mirdita et al. (2022); Pacesa et al. (2025) for which the designed sequence is obtained by backpropagation through a folding model. This optimization is often quite unstable, which is unsurprising given that it involves large computational graphs. The size of the computational graph arises because the forward pass of the folding model involves recycles as well as a diffusion rollout. Finding ways to reduce the size of the computational graph at the same level of fidelity can thus be expected to improve the design quality as well as speed.

State-of-the-art folding models are based on the AlphaFold 3 architecture Abramson et al. (2024), retaining the fundamental recipe of MSA embedding, the recycled trunk, diffusion and confidence head. In recent hallucination pipelines, the AlphaFold 3 architecture is used in both the generative and scoring steps Candido et al. (2026); Cho et al. (2025).

Our work starts from the observation that confidence as well as structural consistency can be predicted directly from the trunk even in an early recycle. Specifically, we demonstrate that one can train a lightweight *CycleCritic* that can predict these quantities at end of each recycle from the pair

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

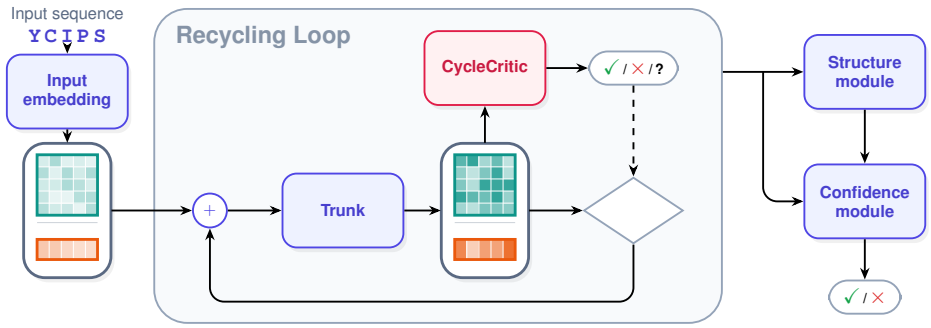


Figure 1: **CycleCritic for efficient screening.** After each recycle, CycleCritic predicts final-depth quality from the current trunk representations. It cancels unpromising refolds before the structure and confidence modules, avoiding further recycling and downstream evaluation. Retained candidates complete the full refold and are accepted only if they satisfy the final screening filters.

representation. We then harness this predictor for both the generative and screening steps, achieving material improvements in compute allocation and binder hallucination in our experiments.

Our experiments demonstrate the benefits of the CycleCritic for screening. We generate 126,000 VHH designs against seven antigens. We then show that the CycleCritic allows us to identify designs that pass both the confidence and self-consistency thresholds at a drastically reduced compute budget. Impressively, this is achieved at very low false positive rate. We also show that the CycleCritic outperforms all state-of-the-art baselines. For example, the CycleCritic is able to identify roughly 2.3x more high quality designs than Protenix-mini per GPU hour. Impressively, it does so with 15 percentage point higher precision. The advantage is particularly pronounced for hard targets.

For the generative step, we apply the CycleCritic to two different hallucination approaches (ESMFold2 and Mosaic). We design for seven targets with both the VHH and scFV format resulting in a total of 14 design tasks. Our approach improves the top-50 designs on $13/14$ and $10/14$ design tasks for Mosaic and ESMFold respectively.

Briefly, our main contributions are:

- We analyse the dynamics of AF3-like models during recycling, show that the trunk is the main determining factor in folding performance which can often be predicted at early recycles.
- We propose a CycleCritic that can directly predict confidence and structural compatibility from the pair representation even at early recycle.
- For screening, we deploy the CycleCritic for dynamic allocation of the recycles. This drastically reduces the cost of passing designs since the model no longer needs to spend unnecessary recycles on designs that are easy to score.
- For hallucination-based generation, we use this CycleCritic to construct more economical confidence losses. We demonstrate that this leads to better design outcomes.

2 Related Work

Interpreting the trunk representation. Migliorini et al. (2026) extract interpretable structural concepts directly from the pair representation of a co-folding model. Similarly, Feldman & Skolnick (2026) probe AlphaFold 3's internal representations across the forward pass and find that the Pairformer compresses its inputs into a compact latent geometry in which complex biophysical features become linearly decodable. Lin et al. (2023) and Candido et al. (2026) investigate the ESMFold latent space and show that real structural information is carried in the pair representation. These works provide the foundation for our work by pointing out the richness of the trunk representation.

Screening with trunk representation. Verburgt et al. (2025) refine AlphaFold's pLDDT head with equivariant graph neural networks over the predicted C_α coordinates emitted by the structure module. Liu et al. (2025) adapt AlphaFold 3 to a score-only model by feeding input coordinates directly into the confidence head, bypassing thus the diffusion structure module. Both approaches require a structure to exist, while our method can predict confidences without a given structure. (Candido et al., 2026) proposes a distogram ipTM proxy for ESMFold2. It has no learned parameters and provides a valuable baseline for our approach.

Generation with trunk representation. Liu et al. (2026) derive inter-chain contact probabilities from the Pairformer trunk and use them as a blueprint constraining generation before diffusion, pruning docking modes and reducing sampling roughly tenfold; they further note that these probabilities indicate prediction difficulty and sampling utility. ESMFold-based binder design Candido et al. (2026) and EasyNano (Hu et al., 2026) optimize against trunk outputs alone, computing their objective from the distogram or pair representation with no structure module in the loop. Our contribution work extends on this body of literature by not only skipping diffusion but also dynamically allocating the recycle budget of the trunk. Jang et al. (2026) improve folding models by learning to transfer the trunk representations of different models to each other.

Scoring and pose selection. Confidence-based pose selection is a standard filter applied in co-folding-based workflows to improve hit-rates despite well-characterised limitations (Fromm et al., 2026; Thüning, 2026). Recent works have focussed on developing structure-based antibody pose scorers trained with decoys. Xu et al. (2026) train an equivariant GNN for scoring antibody-antigen poses derived from classical docking methods. Grandguillaume et al. (2026) learn a scoring function from pLM embeddings, and force field and structure-derived features over docking decoys. Tadiello et al. (2026) train a Deep Set ranking model over an ensemble of AlphaFold 3 co-folding samples.

3 Motivation: Dynamics of Recycling

AlphaFold3-type folding models are generative models and therefore stochastic. We recall this stochasticity is not only because of the diffusion module but also due to the trunk. Specifically, recycles in the trunk are stochastic due to MSA subsampling and possible inference dropout. Our work is motivated by the observation that most of the variance of state-of-the-art AlphaFold 3-like models can be attributed to the trunk and not, as one may naively expect, the diffusion (Figure 2a). We further observe that some components of the pair representation correlate closely with confidence metrics, even early during recycling, see Figure 2b.

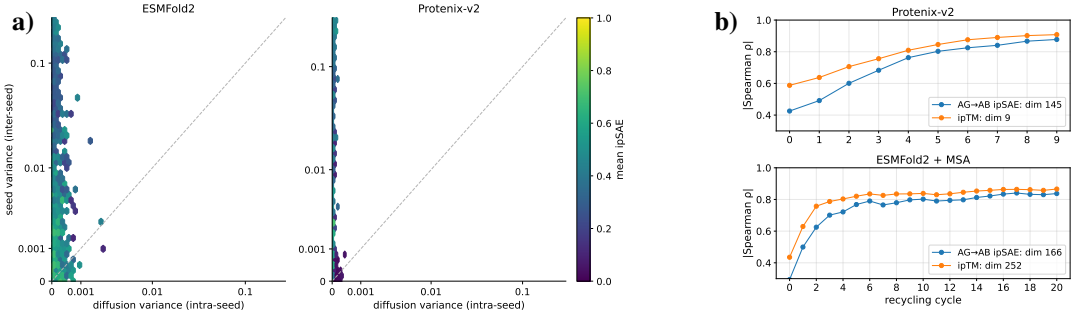


Figure 2: **Model variance and latent dimension predictability. a) Seed and diffusion variance.** How much does the ipSAE vary between different seeds and diffusion samples. Computed with 3 seeds and diffusion samples each. **b) Latent dimensions correlate with final confidence metrics.** We average-pool the latent dimensions over all pairs in the pair representations and take the highest correlated dimension. We see a high Spearman correlation with the final metric outputs even at early recycles. a) 2,500 b) 350 samples; both on Sabdab.

To assess the dependence of confidence on the supplied structure, we fix trunk outputs and re-evaluate the Protenix-v2 and ESMFold2 confidence heads on perturbed coordinates for 2,500 an-

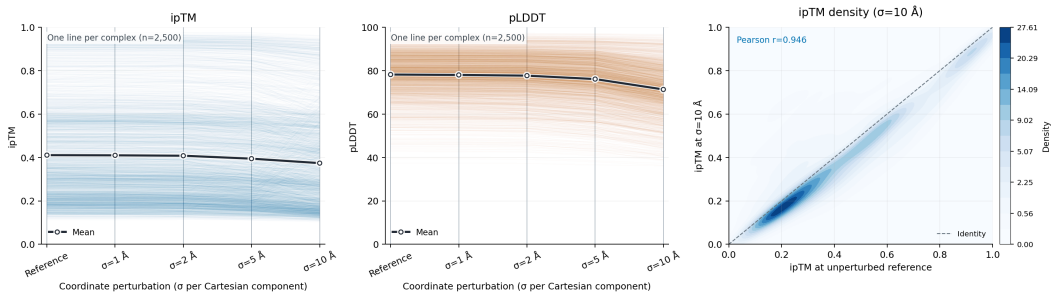


Figure 3: **Protenix-v2 Confidence metrics are largely determined by trunk representations.** Trunk representations were fixed while the confidence head scored perturbed structures for 2,500 antibody–antigen complexes randomly sampled from SAbDab. Perturbations comprised independent Gaussian noise with per-coordinate standard deviation $\sigma \in \{1, 2, 5, 10\}$ Å. **Left:** ipTM across perturbations; traces denote individual complexes and black line denotes mean. **Centre:** pLDDT across perturbations; traces denote individual complexes and black line denotes mean. **Right:** Unperturbed versus perturbed ipTM for $\sigma = 10$ Å; dashed diagonal indicates perfect agreement.

tibody–antigen complexes. For Protenix-v2, adding Gaussian noise with σ =10 Å modestly reduces ipTM and plDDT, with strong correlation between perturbed and unperturbed scores (Pearson $r = 0.946$) (Figure 3). Input ablations localize the dominant contributor to confidence scores to the trunk's pair representation. Removing the pair input collapses ipTM, while removing other inputs has limited effect (Figure A.2a). Together, these observations indicate that confidence metrics are largely determined by the pair representation and only weakly adjusted by the supplied coordinates. We observe different, co-ordinate dependent, behaviour for ESMFold2, discussed in Appendix A.

4 CycleCritic

The analysis of the previous section strongly suggests that one should be able to predict confidences as well as structural consistency directly from the pair representation even at early recycles. This then allows to dynamically allocate the recycling budget with the CycleCritic deciding the number of recyclings. In this section, we will outline our concrete implementation of this.

We use a simple DeepSets-style architecture (Zaheer et al., 2017). The architecture then consists of a shared per-pair encoder (no cross-attention), a permutation-invariant weighted pooling step, and a small head on the pooled vector. We refer to Figure 4 for an overview of the architecture.

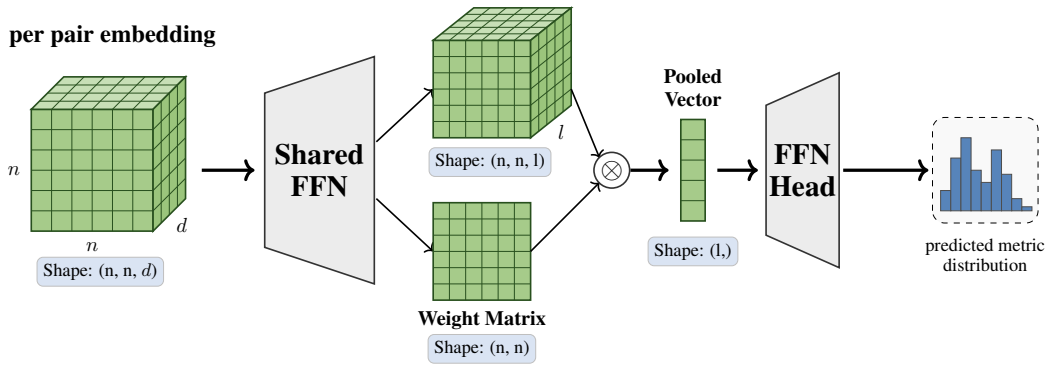


Figure 4: **Schematic architecture of the learned CycleCritics.** Both apply a shared feedforward network to a latent representation per pair representation *individually* before aggregating by weighted mean pooling. The structure critic additionally receives the pair distances of the designed backbone; the predicted metric can be e.g. ipSAE or ipTM for the confidence critic and self-consistency RMSD for the structure critic.

For each recycle, the CycleCritic is trained to predict the final metric after a fixed number of reference recycles. The number of reference recycles is chosen for each folding model such that it leads to high folding fidelity. This is motivated by the observation that prediction quality typically plateaus for a sufficiently large number of recycles.

Self-consistency can only be predicted with respect to a given reference design structure while for confidence only the sequence of the design is sufficient. As a result, we train slightly different CycleCritic for structural consistency and confidence.

Confidence CycleCritic. This critic takes the pair representation $z_{ij}^{(c)}$ of the current recycle as input as well as the recycle number c and chain-identity. It then predicts the ipSAE, ipTM, pTM or pLDDT of the reference refold. The final head emits a categorical distribution $P_c \in \Delta^K$ over K bins tiling $[0, 1]$, with scalar prediction $s_c = \sum_k P_{c,k}\, b_k$ for bin centres b_k.

Structure CycleCritic. This critic takes the pair representation $z_{ij}^{(c)}$ of the current recycle, the recycle number c, and the C_β distance matrix of the design as input. Instead of directly using the pair representation for the per-pair embeddings, we found it advantageous to first transform it into the distogram with the frozen small distogram head. This head is part of the standard AlphaFold3 architecture and is used throughout the folding model training for the distogram loss. The resulting per-pair embeddings are then processed by the DeepSet-style architecture outlined above. The final head predicts the scRMSD discretized by 40 bins uniformly covering 0-20 Å.

4.1 Analytic Alternative

We also propose analytic CycleCritics that require no learning, lowering the barrier to adoption at the cost of reduced performance.

Structure CycleCritic. We propose to leverage the distogram p_{ij}^c predicted at cycle c directly. We then compute the mean negative log likelihood of the design's distance matrix D_{ij} across the binder-target interface, defined by $C_\alpha < 12$Å on the designed backbone. We refer to Algorithm 1 for complete details. The log likelihood thresholds corresponding to the corresponding scRMSD thresholds can then be calibrated on a validation set. Alternatively, one can impose a log likelihood screening threshold directly.

Confidence CycleCritic. For ipTM, a proxy was proposed in Candido et al. (2026) (Algorithm 15). Its basic idea is to use the entropy of the interface distogram as a proxy for ipTM. This is well motivated as the structure module produces the residue positions conditioned on the trunk. The distogram encodes how constraining this conditioning information is. Highly constrained structures come with lower uncertainty. We propose to use this proxy as an analytic Confidence CycleCritic. A downside of this approach is that it is restricted to ipTM.

As we discuss in Appendix E.1, these analytic alternatives are underperforming relative to their learned counterparts but still show meaningful performance.

5 Experiments

In the following, we demonstrate the power of the CycleCritic for screening and hallucination-based design. For both, we generate designs for seven benchmark targets: CD45, CTLA-4, EGFR, IL7Ra, INSR, PDGFRβ, and PD-L1.

5.1 Efficient Screening with CycleCritic

The CycleCritic can significantly speed up screening. To demonstrate this, we generated 126,000 VHH designs using BoltzGen Stärk et al. (2025). We use the seven benchmark targets described above and four VHH frameworks with PDB IDs `8coh`, `8z8v`, `7eow`, and `7xl0`. This yields 4500 designs per target–framework pair and 18,000 designs per target.

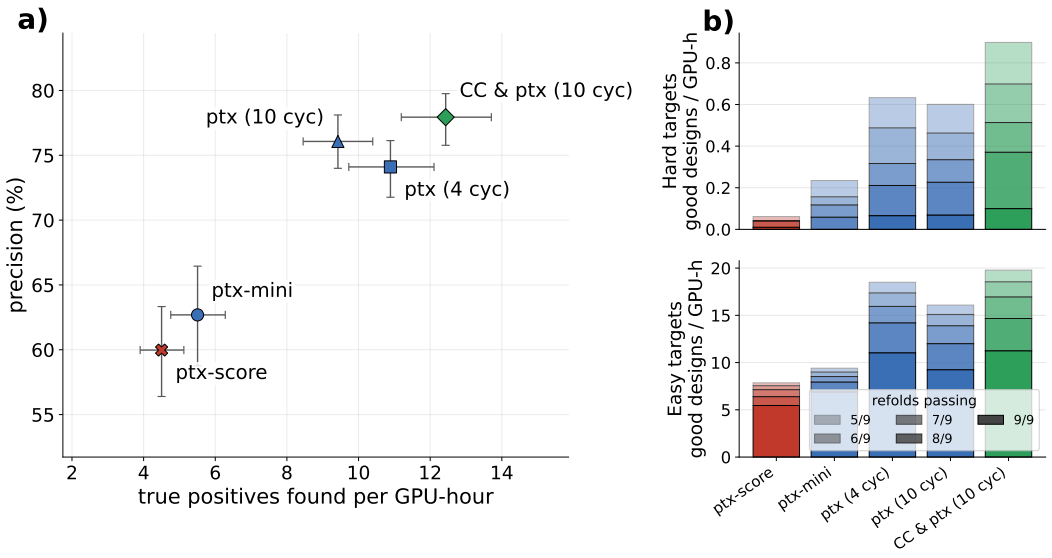


Figure 5: **CycleCritic outperforms all baselines. a)** Precision and true positive throughput. Error bars represent 95% percentile bootstrap confidence intervals (2.5th–97.5th percentiles) computed across 2,000 draws by resampling 50-design chunks with replacement within each target. **b)** Hard vs. rest mean yield. Pooled mean good-design yield per GPU-hour across hard targets (< 100 positives: CD45, CTLA-4, PDGFRβ) and remaining targets (EGFR, IL7RA, PD-L1, INSR). Stacks show 5–$9/9$ labels.

These designs are then filtered. Protenix-v2 with 10 recycles is used as a reference, which is one of the leading folding models for antibody-antigen complexes Zhang et al. (2026). We aim to reproduce the fidelity of Protenix-v2 with 10 recycles at significantly lower computational cost.

We compare with three relevant baselines:

- Protenix (4 cycles) Chen et al. (2025): the same Protenix-v2 model with 4 instead of 10 recycles.
- Protenix-mini Gong et al. (2025): a lightweight version of Protenix that is optimized for high inference speed at the cost of reduced folding fidelity.
- ProtenixScore Ring (2024): an implementation of the AF3Score architecture Liu et al. (2025) for Protenix. This approach turns Protenix into a scorer. The trunk is templated on the input structure and the diffusion rollout is skipped. The confidence head takes as input the trunk embeddings as well as the design structure and produces the resulting confidence score. Note that this model cannot evaluate structural self-consistency.

Since AlphaFold3-like folding models are stochastic, we refold the same complex 9 times with Protenix-v2 at 10 recycles. Positives are designs that pass more than half the refolds with the thresholds ipSAE > 0.5 and scRMSD < 3Å. We then fold the design with all methods, check whether it passes the filters, and compare to the label obtained by the reference method.

Ideally, a method can identify a high number of positives at drastically reduced runtime. Crucially, this needs to be achieved at high precision, i.e. the number of false positives needs to be low. Note that even the reference method will not perfectly agree with the reference labels due to the stochasticity of the refold. The independent repeat still serves as a reference for label agreement and screening performance. We quantify this by refolding the designs one more time with the reference method.

Fig. 5 demonstrates that the CycleCritic significantly outperforms all baselines. Impressively, it also matches the precision of the reference method within error bars. The CycleCritic is particularly beneficial for harder targets. We note that its performance can likely be improved by a more involved calibration approach. For the experiments, we prioritize convenience over maximal performance by

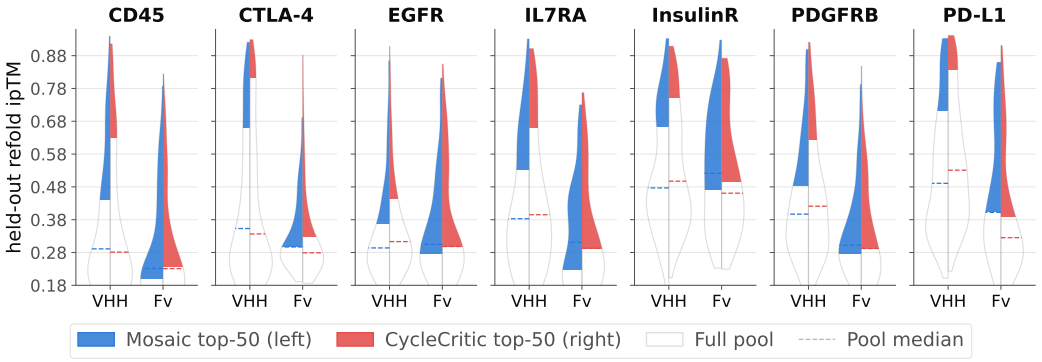


Figure 6: **For Mosaic, CycleCritic significantly improves top-**50 **designs.** Filled areas denote top-50 designs by refold ipTM and dashed lines are median over all produced designs. Both baseline and CycleCritic used the same 6 B200 compute hour budget per design task.

adopting a very simple calibration method. Specifically, we chose the validation set of the cycle critic and adjusted the gating thresholds of the CycleCritic such that full recall is achieved, i.e., all positives in this set are recovered. We refer to Appendix E for more details on the gating experiments.

5.2 Hallucination-based Design with CycleCritic

In hallucination-based design Pacesa et al. (2025); Mirdita et al. (2022), the amino acid sequence of the designed binder is optimized by gradient descent through the folding model. The optimization objective typically involves terms that try to maximize the confidence of the design. For this reason, CycleCritic can be used in these terms to effectively approximate the confidence scores that would be obtained from a large number of recycles. In contrast, previous methods often use a low number of recycles or even detach some of them from the computational graph. We demonstrate that the CycleCritic approach improves antibody design for two state-of-the-art frameworks: Mosaic (Boyd et al., 2025) with Protenix-v2 and the design strategy of ESMFold2 (Candido et al., 2026). We design antibodies both VHH and scFvs format for the seven benchmark targets.

Interestingly, we observe better performance when averaging the CycleCritic-predicted terms across recycling loops, indicating that the CycleCritic provides useful information even in early recycling loops.

Mosaic. We replace all confidence-head-based losses with alternatives derived from the CycleCritic. This allows us to skip the diffusion and confidence modules, generating nearly 60% more samples within the same wall-clock time. We also replaced the pAE-based losses with analogous distogram-based losses. For fair comparison, the final evaluations for both methods were performed using Protenix-v2 at 10 recycles for the refold.

In protein design, one is typically only interested in the distribution over the best designs since wet-lab capacity is scarce. We therefore also compare the design strategies on the top-50 designs. Figure 6 demonstrates that the CycleCritic leads to substantial improvements. The top-50 mean is improved for 13 out of 14 design tasks (7 targets and 2 formats) with a mean gain of 6 p.p.. We refer to Appendix G for more details.

ESMFold2. This approach already skips the diffusion and confidence heads during hallucination, as it relies on distogram-based losses. Nevertheless, our experiments establish that optimization can strongly benefit from the CycleCritic. We attribute this to CycleCritic's ability to effectively predict confidence at large number of recycles, allowing us to estimate design confidence at higher fidelity at essentially the same computational cost.

We add to the baseline optimization objective two terms based on the predicted ipTM and $\mathrm{ipSAE}_{\mathrm{AG}\to\mathrm{AB}}$. We refer to Appendix F for more details.

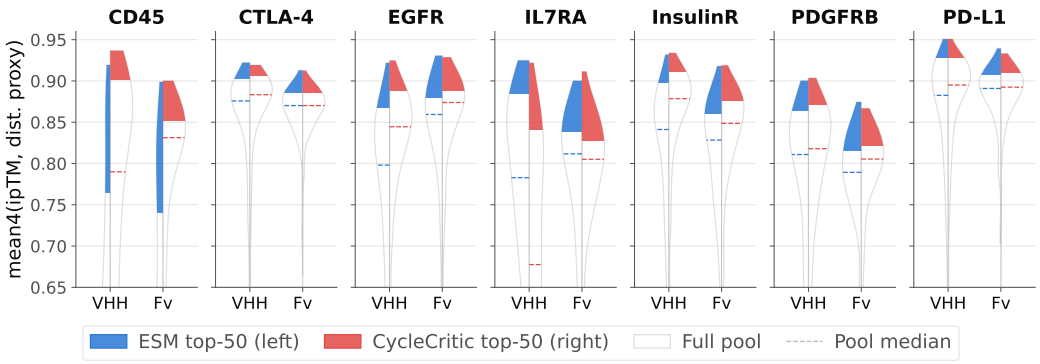


Figure 7: **For ESMFold2, CycleCritic raises per-design quality at matched compute.** We use the same ranking metric as in the ESMFold2 publication for comparison, i.e. the mean over ipTM and distogram-proxy ipTM averaged over 4 refolds with different checkpoints. Filled area is top-50 subset by the ranking metric. Dashed lines are medians over all produced designs.

We compare with the baseline for the same compute budget. Figure 7 demonstrates that the CycleCritic meaningfully improves over the baseline. Specifically, the CycleCritic improves the best designs in 10 out of the 14 design tasks by 1%. As was to be expected, performance can vary by target. CD45 sees a drastic improvement while IL7Ra leads to a small reduction.

Design Quality. To test whether CycleCritic degrades or collapses the designs, we measure a range of diversity and quality metrics. For both ESMFold2- and Mosaic-based design pipelines, CycleCritic designs mirror the diversity and novelty of their baselines. As summarized in Figure 8, we observe comparable CDR diversity, novelty with respect to human repertoire, humanness, and interface diversity. More details can be found in the Appendix H.

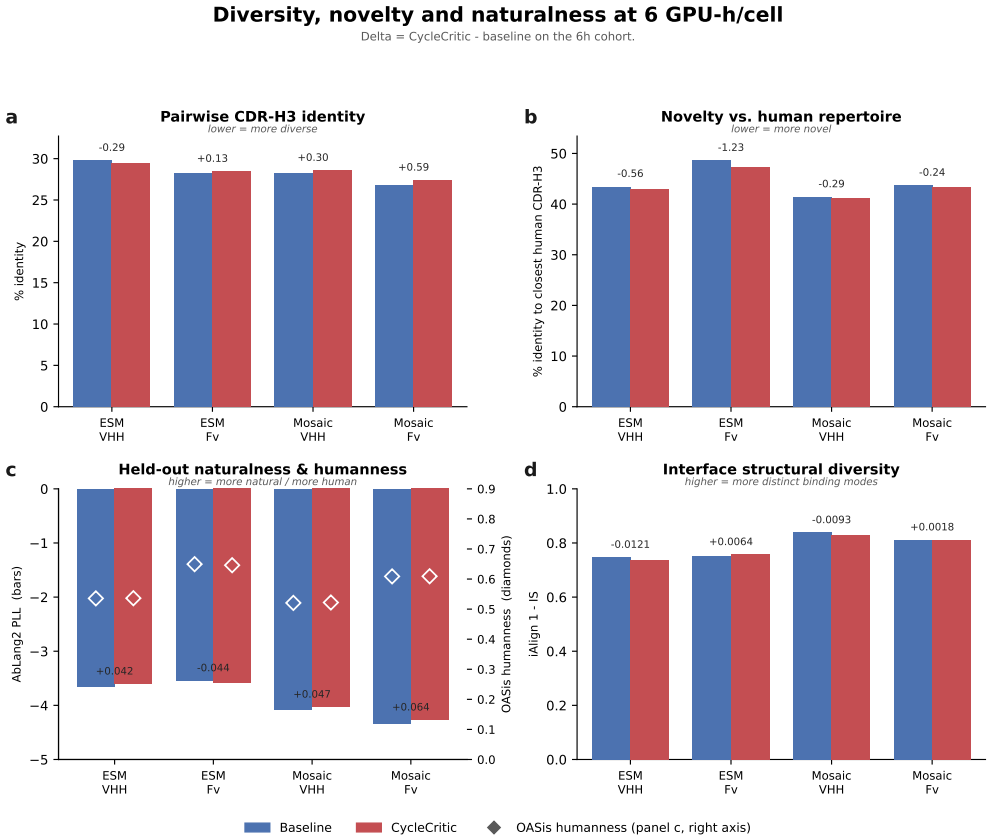


Figure 8: **Diversity, novelty and naturalness of designed binders.** Baseline (blue) versus CycleCritic (red) on the six-hour design cohorts. Annotations show CycleCritic minus baseline. **(a)** Mean pairwise CDR-H3 identity within target–framework cells. **(b)** Identity to the closest human CDR-H3 among 2.3M unique heavy-chain loops from OAS. **(c)** AbLang2 pseudo-log-likelihood (bars, left axis) and OASis humanness (diamonds, right axis) over designed CDR positions. **(d)** Interface structural diversity measured by iAlign $1 - \text{IS}$ within target–framework–arm cells. The iAlign comparison describes the available structural subset, whose coverage differs between arms, rather than the complete design cohorts.

6 Conclusion

A simple CycleCritic uses trunk representations from early recycles to predict the confidence and structural consistency metrics that would be obtained after a large number of recycles. This can speed up and improve both the design and screening steps of binder design pipelines. Specifically, the CycleCritic allows to dynamically allocate recycles depending on the difficulty of the protein complex at hand. This is somewhat reminiscent of reasoning where compute budget is dynamically allocated by a language model.

Reproducibility Statement

Appendix C details dataset and split construction. Appendix B and D details the training procedure for the CycleCritic and structure CycleCritic. Appendix F and G describes the design protocol with ESMFold2 and Mosaic. Appendix E details the use of the CycleCritic in gating binder designs. Appendix H describes the analysis procedure for binder designs.

AI USE STATEMENT

Generative AI was used in preparation of figure code, targeted re-phrasing of drafted text and assistance in implementation of experiments. The authors are responsible for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

In detail: Generate synthetic data sets ×; develop theoretical models or conceptual frameworks ×; formulate mathematical claims ×; provide ingredients for proving claims ×; assist in writing proofs ×; propose or refine hypotheses ×; design or provide feedback on research methodology or experiments ✓; implement methods ✓; assist with translation ×; clean and reformat dataset ✓; support qualitative and thematic data analysis ✓; interpret results ✓. Formulate survey or interview questions ×; create or modify scientific figures or images ✓; suggest experimental parameters ✓; create or edit software code ✓; create artifacts ×; draft parts of a research paper ✓; transcribe research recordings ✓; summarize or analyse existing literature ✓; discover research topics or identify gaps ✓; brainstorming ×; source or search for information ✓; edit a research paper for readability ✓; identify relevant literature ✓; format references ✓; suggest a research-paper structure ×; propose a title or keywords ×.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, Sebastian Bodenstein, David A Evans, Chia-Chun Hung, Michael O'Neill, Martin Reiman, Kathryn Tunyasuvunakool, Zachary Higgins, Augustin Žídek, Michal Green, Demis Hassabis, and John M Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Nathaniel R Bennett, Joseph L Watson, Robert J Ragotte, Andrew J Borst, DéJenaé L See, Connor Weidle, Riti Biswas, Yutong Yu, Ellen L Shrock, Russell Ault, et al. Atomically accurate de novo design of antibodies with rfdiffusion. *Nature*, 649(8095):183–193, 2026.

N. Boyd, S. Guns, and Escalante Bio. Mosaic. `https://github.com/escalante-bio/mosaic`, 2025. URL `https://github.com/escalante-bio/mosaic`.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 89–96, 2005. doi: 10.1145/1102351.1102363.

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

Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, Shenghao Wu, Kuangqi Zhou, Yanping Yang, Zhenyu Liu, Lan Wang, Bo Shi, Shaochen Shi, and Wenzhi Xiao. Protenix - advancing structure prediction through a comprehensive alphafold3 reproduction. *bioRxiv*, 2025. doi: 10.1101/2025.01.08.631967.

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.

Jonathan Feldman and Jeffrey Skolnick. Alphainterp: Probing alphafold 3's internal representations reveals evolutionary determinants of predicted structure and confidence. *bioRxiv*, 2026. doi: 10.64898/2026.04.22.720175.

Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody–antigen complexes. *Bioinformatics*, 42(4):btag136, 04 2026. ISSN 1367-4811. doi: 10.1093/bioinformatics/btag136. URL https://doi.org/10.1093/bioinformatics/btag136.

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

Ilyas Grandguillaume, Fernando Luis Barroso da Silva, and Catherine Etchebest. Arid-sf: A physics-informed deep learning scoring function to improve antibody–antigen docking model ranking. *Journal of Chemical Information and Modeling*, 66(15):9540–9553, 07 2026. ISSN 1549-9596. doi: 10.1021/acs.jcim.6c00836. URL https://doi.org/10.1021/acs.jcim.6c00836.

Yue Hu, Wanyu Cheng, Junqing Wang, and Yingchao Liu. Easynano: rapid epitope-targeted nanobody cdr design via differentiable distogram optimization with esmfold2. *arXiv preprint arXiv:2606.12772*, 2026.

Isomorphic Labs Team. Accurate predictions of novel biomolecular interactions with IsoDDE. Technical report, Isomorphic Laboratories, February 2026. URL https://storage.googleapis.com/isomorphiclabs-website-public-artifacts/isodde_technical_report.pdf.

Hyosoon Jang, Taewon Kim, and Sungsoo Ahn. Synthesizing state-of-the-art structure predictions from soup of co-folding models. *arXiv preprint arXiv:2609.15552*, 2026.

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.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan Dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Lihang Liu, Yang Liu, Xianbin Ye, Shanzhuo Zhang, Yuxin Li, Kunrui Zhu, Yang Xue, Xiaonan Zhang, and Xiaomin Fang. Reshaping biomolecular structure prediction through strategic conformational exploration with helixfold-s1. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01264-2.

Yu Liu, Qilin Yu, Di Wang, and Mingchen Chen. Af3score: A score-only adaptation of alphafold3 for biomolecular structure evaluation. *Journal of Chemical Information and Modeling*, 65(15):8207–8214, 2025. doi: 10.1021/acs.jcim.5c01144.

Giosue Migliorini, Aristofanis Rontogiannis, Grigori Guitchounts, Nicholas Franklin, Axel Elaldi, and Olivia Viessmann. Pairsae: Mechanistic interpretability from pair representations in protein co-folding. *arXiv preprint arXiv:2606.27440*, 2026.

Luis S Mille-Fragoso, Claudia L Driscoll, John N Wang, Haoyu Dai, Talal Widatalla, Jim L Zhang, Xiaowei Zhang, Bing Rao, Liang Feng, Brian L Hie, et al. Efficient generation of epitope-targeted antibodies with germinal. *Nature biotechnology*, pp. 1–10, 2026.

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.

Aaron Ring. Protenixscore: Score existing protein structures (pdb or cif) with the protenix confidence head, without running diffusion. https://github.com/cytokineking/ProtenixScore, 2024.

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

Hannes Stärk, Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Tommi Jaakkola, and Regina Barzilay. Boltzgen: Toward universal binder design across biomolecular modalities. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494.

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

Matteo Tadiello, Marko Ludaic, Vsevolod Viliuga, and Arne Elofsson. Abag-rank: Improving model selection of alphafold antibody–antigen complexes by learning to rank. *Bioinformatics*, September 2026. ISSN 1367-4811. doi: 10.1093/bioinformatics/btag663. URL http://dx.doi.org/10.1093/bioinformatics/btag663.

Moritz Thüning. Abag-xm: 335,360 dockq-labelled antibody-antigen structure predictions from four models, 2026. URL https://huggingface.co/datasets/Tenstorrent/abag-xm. Analysis: https://moritztng.github.io/abag-scaling/.

Jacob Verburgt, Zicong Zhang, and Daisuke Kihara. Alphafold model quality self-assessment improvement via deep graph learning. *Protein Science*, 34(9):e70274, 2025. doi: 10.1002/pro.70274.

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.

Xiaotong Xu, Ilaria Coratella, Victor Reys, and Alexandre MJJ Bonvin. Deeprank-ab: a scoring function for antibody-antigen complexes based on geometric deep learning. *Communications Biology*, 9(1), June 2026. ISSN 2399-3642. doi: 10.1038/s42003-026-10408-4. URL http://dx.doi.org/10.1038/s42003-026-10408-4.

Lian Yan, Robert Dodier, Michael C. Mozer, and Richard Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pp. 848–855, 2003.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Yuxuan Zhang, Chengyue Gong, Jinyuan Sun, Jiaqi Guan, Milong Ren, Song Xue, Hanyu Zhang, Wenzhi Ma, Zhenyu Liu, Xinshi Chen, and Wenzhi Xiao. Protenix-v2: Broadening the reach of structure prediction and biomolecular design. *bioRxiv*, 2026. doi: 10.64898/2026.04.10.717613.

A DYNAMICS DURING RECYCLING

Figure A.1 shows that trunk outputs carry information about the final DockQ score even in early cycles. We ran diffusion and confidence head once after every recycle and looked at the predictiveness for the final structure being acceptable (DockQ > 0.23).

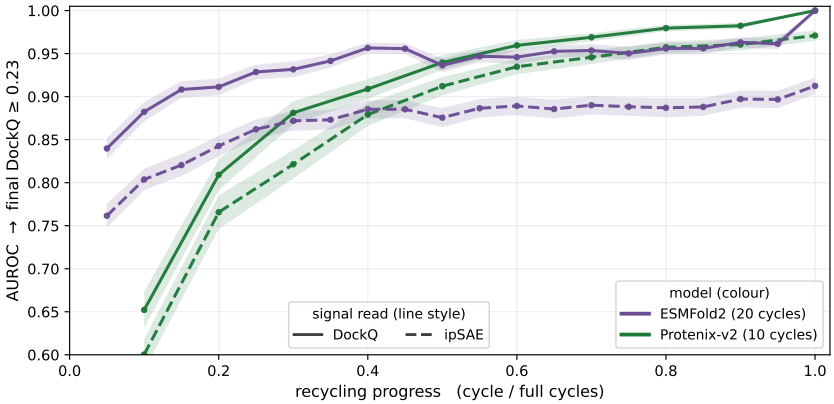


Figure A.1: intermediate AG→AB ipSAE and DockQ during recycling as predictive of final DockQ

Figure A.2a shows the confidence metrics of Protenix-v2 and ESMFold2 samples under different input ablations. For Protenix-v2, we observe limited impact for ablation of trunk single representations. While we see an effect on ablating coordinates entirely, when accounting for our observation of the limited effect of noising the coordinates on confidence scores in Figure 3, we hypothesise this is attributable to the ablation being a severe perturbation which is out of distribution for the confidence head. Ablating the trunk pair representation entirely collapses the ipTM. For the pLDDT prediction, the single representation is the primary input, which is unsurprising given the architectural design of confidence heads which primarily operate on this input.

Figure A.2b shows the same experiment performed on the ESMFold2 confidence head inputs. Here, unlike for Protenix-v2, we observe strong co-ordinate dependence with ipTM prediction collapsing for co-ordinate ablation. Similarly, in Figure A.3 we observe monotonic degradation in ipTM and pLDDT confidence scores with increased noising of the structural input. However, we note ESMFold2 trains the confidence head with 20% dropout on the pair representation to deliberately reduce the dependence on the pair (Candido et al., 2026) so this difference is unsurprising.

B TRAINING THE CONFIDENCE CYCLECRITIC

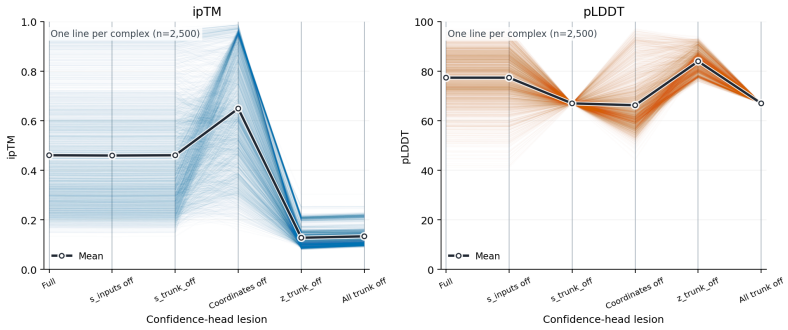
The host folding model is *frozen* throughout. For each training structure we run the model once through its complete inference schedule and record two things: the pair representation $z_c \in \mathbb{R}^{N\times N\times d_{\text{pair}}}$ at *every* recycle c, and the *final* structural metric obtained after the full schedule completes — ipSAE, ipTM, pTM or pLDDT computed on the output of the confidence head. The metric after all recycling loops is the target; the per-cycle representations are the inputs.

The CC consumes $z_c \in \mathbb{R}^{N\times N\times d_{\text{pair}}}$ at recycle c together with learned chain-identity and recycle embeddings and emits a categorical distribution $P_c \in \Delta^K$ over K bins tiling $[0, 1]$, with scalar prediction $s_c = \sum_k P_{c,k}\, b_k$ for bin centres b_k. Three terms are combined, each evaluated independently at every recycle and then averaged:

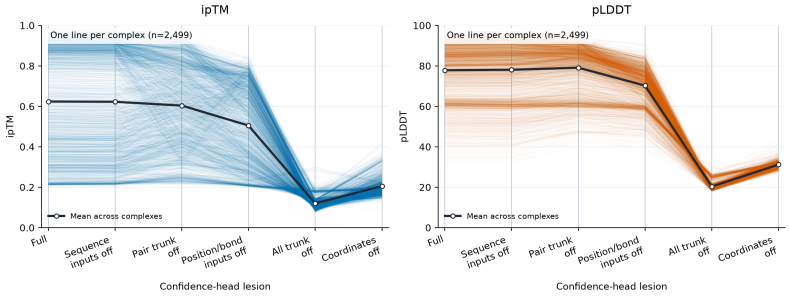
$$\mathcal{L} \;=\; \lambda_{\text{nll}}\,\mathcal{L}_{\text{NLL}} \;+\; \lambda_{\text{rank}}\,\mathcal{L}_{\text{rank}} \;+\; \lambda_{\text{auroc}}\,\mathcal{L}_{\text{AUROC}}. \tag{1}$$

$\mathcal{L}_{\text{NLL}}$ is the cross-entropy against the bin containing the target metric. $\mathcal{L}_{\text{rank}}$ is a RankNet (Burges et al., 2005) term over a mini-batch of B structures,

$$\mathcal{L}_{\text{rank}} \;=\; \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \operatorname{softplus}\!\left(-\frac{s_i - s_j}{\tau}\right), \qquad \mathcal{P} = \{(i,j) : y_i > y_j\}, \tag{2}$$



(a) Protenix-v2 confidence head input ablations



(b) ESMFold2 confidence head input ablations.

Figure A.2: **Ablations of inputs to Protenix-v2 and ESMFold2 confidence heads.** The confidence head was re-evaluated on 32 sampled structures for each of 2,500 antibody–antigen complexes randomly selected from SAbDab after selectively ablating input embeddings (s_{inputs}), trunk single representations (s_{trunk}), trunk pair representations (z_{trunk}), or the complete trunk or coordinate pathways. **Left:** ipTM under each ablation; colored lines show per-complex behavior with points denoting a mean across complexes. **Right:** pLDDT under each ablation. These results identify the trunk pair representation as the dominant input supporting interface confidence for Protenix-v2.

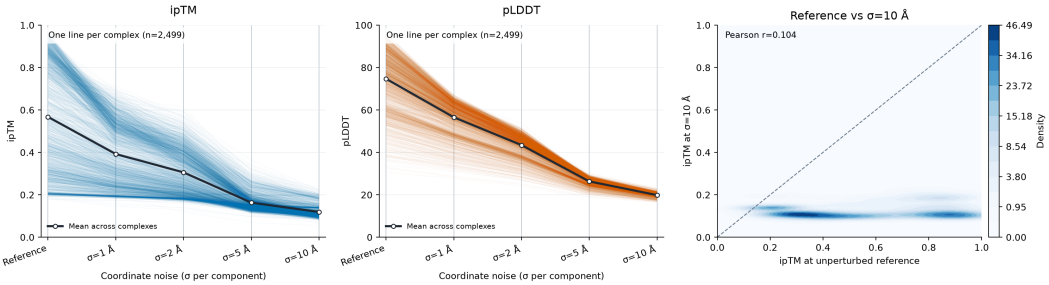


Figure A.3: **Perturbation of structural input to ESMFold2 confidence head.** Trunk representations were fixed while the confidence head scored perturbed structures for 2,499 antibody–antigen complexes randomly sampled from SAbDab. Perturbations comprised independent Gaussian noise with per-coordinate standard deviation $\sigma \in \{1, 2, 5, 10\}$ Å. **Left:** ipTM across perturbations; traces denote individual complexes and black line denotes mean. **Centre:** pLDDT across perturbations; traces denote individual complexes and black line denotes mean. **Right:** Unperturbed versus perturbed ipTM for $\sigma = 10$ Å; dashed diagonal indicates perfect agreement.

a smooth surrogate for Spearman correlation over the strictly ordered pairs $\mathcal{P}$ and $\tau = 0.1$. $\mathcal{L}_{\text{AUROC}}$ is the Wilcoxon–Mann–Whitney surrogate (Burges et al., 2005; Yan et al., 2003) at a quality bar $\theta = 0.5$, applied to the exceedance mass $\sum_{b_k > \theta} P_{i,k}$ rather than to s_i, so that it optimises the separability the gate actually thresholds on.

Table B.1: **CycleCritic hyperparameter search space.** Both substrates were searched with Bayesian optimisation under Hyperband early termination (min_iter $= 8$, $\eta = 3$), maximising the mean Spearman correlation across recycles on a held-out validation set. The Protenix column applies identically to the gating sweep and all four design sweeps; the ESMFold2 column describes the experimental-fast design sweep.

Hyperparameter	Protenix-v2 (stock)	ESMFold2 exp.-fast
Searched		
Architecture	`ffn`, `swiglu`	
d_{hid}	16, 32, 64, 128, 256	64, 128, 256
d_{lat}	16, 32, 64, 128, 256, 384	64, 128, 256
n_{hid}	1, 2, 3	
Bins K	16, 32, 64	16, 32, 64, 128
p_{drop}	0.0, 0.1	0.0, 0.1, 0.2
Learning rate	log-$\mathcal{U}[4\times10^{-4}, 6\times10^{-3}]$	log-$\mathcal{U}[10^{-4}, 3\times10^{-3}]$
λ_{NLL}	1	1
λ_{rank}	log-$\mathcal{U}[0.1, 12]$	
λ_{auroc}	0, 0.3, 1	0, 0.3, 1, 3
`rank_loss`	true, false	
`weighted_pool`	true, false	
`nll_over_samples`	true, false	
`head_on_pool`	false	true, false
Fixed		
Optimiser	AdamW, $\beta = (0.9, 0.98)$, weight decay 10^{-3}	
Schedule	cosine annealing over all steps, gradient clip 1.0	
Epochs $\times$ steps	30×100	
λ_{nll}, τ, θ	1.0, 0.1, 0.5	
Rank batch B	4	12
Run budget	300 (gating), 200 (design)	800

Table B.2: **CycleCritic head configurations, grouped by regime and substrate.** Gating heads score native sequences over the host model's full recycling schedule; design heads score CDR-noised sequences at four cycles. d_{pair} is the trunk pair-representation width, d_{asym} the asymmetric chain-pair embedding, and *w-pool* / *h-pool* the `weighted_pool` and `head_on_pool` flags.

Target	Arch.	d_{pair}	Cyc.	d_{hid}	d_{lat}	n_{hid}	Bins	d_{asym}	p_{drop}	w-pool	h-pool	Params
Gating												
Protenix-v2 (stock trunk), 10 cycles												
ipSAE	`ffn`	256	10	256	–	3	16	4	0.1	✓	×	205,457
Design												
Protenix-v2 (stock trunk), 4 cycles												
ipSAE	`swiglu`	256	4	32	128	1	32	4	0.1	✓	×	74,793
ipTM	`swiglu`	256	4	64	384	2	32	4	0.1	✓	×	238,473
pTM	`swiglu`	256	4	16	128	1	32	4	0.1	✓	×	72,217
pLDDT	`swiglu`	256	4	16	384	1	16	4	0.1	✓	×	214,025
ESMFold2 experimental-fast-cutoff2021/2025, 4 cycles												
ipTM	`ffn`	256	4	256	–	1	16	4	0.2	✓	✓	73,849
ipSAE	`ffn`	256	4	128	–	1	32	4	0.1	✓	✓	39,305

Applying the pairwise terms per recycle trains the head to rank candidates correctly at every point along the trajectory, not only once recycling has converged.

We train for up to 30 epochs of 100 optimiser steps, each coupling B structures, using AdamW ($\beta = (0.9, 0.98)$, weight decay 10^{-3}) with cosine annealing and gradient clipping at 1.0. Checkpoints are selected on the mean-over-recycles validation Spearman correlation.

The hyperparameters λ_{rank} and λ_{AUROC} are found via hyperparameter optimization with the search space as defined in Table B.1. The objective of the sweep was mean spearman correlation of the CC prediction to the target metric.

Target	lr	λ_{rank}	λ_{auroc}	b_{rank}
Gating				
Protenix-v2				
ipSAE	3.46×10^{-3}	0	0	-
Design				
Protenix-v2				
ipSAE	1.67×10^{-3}	1.431	1	4
ipTM	3.36×10^{-3}	4.665	1	4
pTM	5.62×10^{-3}	4.818	0	4
pLDDT	5.20×10^{-3}	2.387	1	4
ESMFold2 exp.-fast-cutoff2021/2025				
ipTM	2.14×10^{-3}	6.839	1	12
ipSAE	2.19×10^{-3}	11.562	0.3	12

Table B.3: CycleCritic training hyperparameters for deployed models the optimal parameters were also found via Bayesian optimization.

B.1 Architecture details for the Confidence CycleCritic

For each trunk pair representation $z_{ij} \in \mathbb{R}^{d_{\text{pair}}}$, learned row-chain, column-chain, and recycling-cycle embeddings are concatenated and LayerNorm is applied:

$$x_{ij} = \text{LayerNorm}\big([z_{ij}; e_i^{\text{chain}}; e_j^{\text{chain}}; e^{\text{cycle}}]\big). \tag{3}$$

Thus, $x_{ij} \in \mathbb{R}^{d_{\text{in}}}$, where $d_{\text{in}} = d_{\text{pair}} + 3d_{\text{asym}}$. The same FFN or SwiGLU operation E_{pair} is applied independently to every x_{ij}, producing one latent representation per pair. The pair axis is reduced by a permutation-invariant mean, optionally weighted by a learned confidence, and the output head is applied either before or after pooling. The head maps its input to K logits with a final linear layer and applies softmax, producing a categorical distribution P over K equal-width bins of the target in $[0, 1]$. The scalar prediction is the expected bin center, $\hat{y} = \sum_{k=1}^{K} P_k b_k$, where b_k is the center of bin k.

Per-pair activation. The FFN uses n_{hid} linear–GELU layers. Its final per-pair latent has width d_{hid}:

$$h_{ij}^{(0)} = x_{ij}, \qquad h_{ij}^{(r)} = \text{GELU}\Big(W_r h_{ij}^{(r-1)} + b_r\Big), \quad r = 1, \ldots, n_{\text{hid}}. \tag{4}$$

Thus, $E_{\text{pair}}(x_{ij}) = h_{ij}^{(n_{\text{hid}})} \in \mathbb{R}^{d_{\text{hid}}}$. Its output head is

$$\text{Head}_{\text{FFN}}(h) = \text{softmax}(W_{\text{out}} h + b_{\text{out}}), \qquad W_{\text{out}} \in \mathbb{R}^{K \times d_{\text{hid}}}. \tag{5}$$

The SwiGLU encoder produces a per-pair latent of width d_{lat} through a gated projection:

$$E_{\text{pair}}(x_{ij}) = (W_{\text{up}} x_{ij} + b_{\text{up}}) \odot \text{SiLU}(W_{\text{gate}} x_{ij} + b_{\text{gate}}), \qquad W_{\text{up}}, W_{\text{gate}} \in \mathbb{R}^{d_{\text{lat}} \times d_{\text{in}}}. \tag{6}$$

Its output head applies an n_{hid}-layer SiLU MLP of width d_{hid}, followed by a linear map to K logits:

$$\text{Head}_{\text{SwiGLU}}(h) = \text{softmax}\Big(W_{\text{out}}\, \text{MLP}_{\text{SiLU}}^{(n_{\text{hid}})}(h) + b_{\text{out}}\Big), \qquad W_{\text{out}} \in \mathbb{R}^{K \times d_{\text{hid}}}. \tag{7}$$

The MLP maps $\mathbb{R}^{d_{\text{lat}}}$ to $\mathbb{R}^{d_{\text{hid}}}$; its first weight matrix has shape $d_{\text{hid}} \times d_{\text{lat}}$, and subsequent matrices have shape $d_{\text{hid}} \times d_{\text{hid}}$. Thus, both architectures end in a linear map to K logits, while the SwiGLU head includes the preceding SiLU MLP.

Pooling operation. The pooling operation is either a learned weighted mean or a uniform mean. With weighted pooling, a linear confidence head predicts c_{ij} from each pair latent and assigns weight $w_{ij} = \sigma(c_{ij})$; without weighting, $w_{ij} = 1$. These weights are used in either aggregation order below.

Table C.1: **Parameters used to generate the Protenix-v2 gating dataset.** Each native complex is folded once using the full recycling schedule, with pair representations captured after every recycle. Antigen and antibody-framework MSAs are provided.

Parameter	Value
Host model	Protenix-v2
MSA	Yes
Recycles, C	10
N_{step}	200
N_{decode}	5
Token cap (train / val)	600 / 1,000
Complexes (train / val)	3,498 / 500

Head placement. `head_on_pool` determines the ordering of the head and pooling operations. When it is false, the head is applied to each pair before pooling; when it is true, the pair representations are pooled before applying the head:

- `False`: $\text{Pool}(\text{Head}(E_{\text{pair}}(x_{ij})))$.
- `True`: $\text{Head}(\text{Pool}(E_{\text{pair}}(x_{ij})))$.

C Dataset construction

All datasets were generated by running a folding model over antibody–antigen complexes from SAbDab and recording the per-recycle pair representation $z_{\text{cycles}} \in \mathbb{R}^{C\times N\times N\times D}$, together with `asym_id` (per-token chain index), `is_ab` (antibody mask), and decoded confidence labels. Each complex is decoded N_{decode} times under independent diffusion noise; reported labels are the mean over decodes, with the per-decode values retained.

Two regimes are used. *Gating* datasets fold native sequences and are used to predict refold quality from the trunk state. *Design* datasets fold CDR-corrupted sequences and are used to guide hallucination-based design trajectories.

For splitting, we use the epitope similarity proposed in Isomorphic Labs Team (2026). We use mmseqs clustering at 40 percent similarity. The resulting clusters are split into validation and training with a 1:7 ratio.

C.1 Gating datasets

The parameters for generating the gating dataset for Protenix-v2 are summarized in Table C.1.

C.2 Design datasets

For training the cCC for design, we explicitly also noise the CDRs in order for the CC to not be out of distribution at the start of design. We observed that this had a great positive effect on generation.

CDR noise model A soft position-specific scoring matrix replaces the one-hot residue encoding (`restype`) at CDR positions. For CDR position i with native one-hot $e_i \in \mathbb{R}^{20}$,

$$\mathbf{v}_i = 10\, e_i + \sigma\, \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, I_{20}) \tag{8}$$

$$\text{PSSM}_i = \text{softmax}(\mathbf{v}_i). \tag{9}$$

Framework and antigen positions retain their one-hot encoding. For all $\sigma > 0$ the cysteine logit is set to -10^6 prior to the softmax. Each complex contributes one $\sigma = 0$ anchor and $k_{\text{noised}} = 4$ noised draws, giving five items per complex.

Inspired by Karras et al. (2022), the noise scale is drawn from a clipped log-normal, $\sigma = \min(\exp(u),\, \sigma_{\max})$ with $u \sim \mathcal{N}(P_{\text{mean}},\, P_{\text{std}})$.

Parameter	Value
Model	Protenix-v2
Recycles C	4
Diffusion steps N_{step}	200
Decodes N_{decode}	5
$(P_{\text{mean}}, P_{\text{std}}, \sigma_{\max})$	$(1.15,\ 0.45,\ 12.0)$
Median σ	$\approx$3.2
Cropping	epitope-centered, pad 30
Max tokens	600

C.2.1 PROTENIX/ MOSAIC

The soft PSSM is written into the `restype` feature. The `profile` feature receives the rank-one query-row correction $\Delta\text{profile}_i = (\text{PSSM}_i - \text{old}_i)/N_{\text{msa}}$; the `msa` feature is unmodified. Columns 20+ of `restype` are zeroed.

Schedule constants were calibrated on 50 complexes (23 SAbDab, 27 VHH) whose $\sigma = 0$ anchors reached ipTM ≥ 0.70. Both regimes are flat to $\sigma = 2$; VHH then falls to 0.65 of anchor by $\sigma = 3$, while SAbDab decays to a plateau near 0.66.

We sample $u \sim \mathcal{N}(1.15, 0.45^2)$ and set $\sigma = \min(\exp(u), 12)$. The resulting median noise scale is approximately 3.2, with the central one-standard-deviation interval approximately $[2.0, 5.0]$.

Each complex contributes five folds: one $\sigma = 0$ anchor and four draws from the schedule. Training uses 7,089 complexes (35,445 folds) and validation 100 complexes (500 folds).

C.2.2 ESMFOLD2

Parameter	Value
Model	`ESMFold2-Experimental-Fast-cutoff2021/2025`
`lm_dropout`	0.5
`num_loops`	3 (4 carry cycles)
`num_sampling_steps`	20
Recycles C captured	4
$(P_{\text{mean}}, P_{\text{std}}, \sigma_{\max})$	$(0.9,\ 0.7,\ 12.0)$
Median σ	≈ 2.5
Cropping	epitope-centered, pad 30
Max tokens	600

Binder construction. The binder is a single suffix chain, VH + `GGGSGGGSGGGSGGGS` + VL, or VH alone for VHH/nanobody. Complexes are folded as `antigen | binder`.

Filtering. Retained complexes have a heavy chain, exactly one antigen chain, and a defined epitope. The antigen is cropped to $[\min(\text{epitope}) - 30,\ \max(\text{epitope}) + 30]$, and complexes whose cropped target + binder exceeds 600 tokens are dropped.

Calibration on this model gives ipTM collapse across $\sigma \in [1, 4]$, from 0.8–0.9 at $\sigma = 0$ to ≈ 0.2 for $\sigma \geq 4$.

Splits. Train = SAbDab train + VHH train; val = SAbDab val + VHH val; reusing the same epitope-cluster splits. Val are 200 complex-folds drawn with `WeightedSampler` (`weights_epitope_then_cdr`, without replacement,) Training uses all valid complexes.

D STRUCTURE CYCLECRITIC

The structure cycle critic (sCC) scores a candidate at recycle c from two quantities that are already available before diffusion is run: the trunk's current belief about pairwise geometry, and the designed

Algorithm 1 Interface distogram NLL at recycle c

Require: $z^{(c)} \in \mathbb{R}^{N\times N\times d}$, trunk pair representation at recycle c
Require: $D \in \mathbb{R}^{N\times N}$, pairwise distances between the representative atoms of the *design* backbone
Require: chain roles $\mathrm{r}(i) \in \{\text{binder}, \text{target}\}$; interface cutoff τ (12 Å); bin boundaries $e_1 < \cdots < e_{B-1}$ of the trunk's distogram head
Ensure: $\mathrm{nll}_{\text{iface}}$, the mean negative log-likelihood of the design's interface distances under the trunk's own belief

$\ell \leftarrow \mathrm{DistogramHead}(z^{(c)})$ ▷ The distogram head, exactly as the trunk runs it, frozen;
$\log p \leftarrow \text{log-softmax}(\ell)$ ▷ symmetrised internally over the bin axis.

$b_{ij} \leftarrow \sum_{k=1}^{B-1} \mathbf{1}[D_{ij} > e_k]$ ▷ Bin the observed backbone distances.

▷ Interface pairs: cross-role, in contact, both tokens resolved.
$M \leftarrow \{(i,j) : \mathrm{r}(i) \neq \mathrm{r}(j) \ \wedge\ i \neq j \ \wedge\ D_{ij} < \tau \ \wedge\ i, j \text{ resolved}\}$
if $M = \varnothing$ **then**
  return $\varnothing$ ▷ no interface: the gate falls back or abstains
end if

$\mathrm{nll}_{\text{iface}} \leftarrow -\dfrac{1}{|M|} \displaystyle\sum_{(i,j)\in M} \log p_{ij}\big[b_{ij}\big]$ ▷ Average the surprise over the interface.

return $\mathrm{nll}_{\text{iface}}$

backbone the refold is supposed to reproduce. Both are expressed on the same distance basis, which is what allows them to be compared pair by pair.

D.1 Analytic baseline

As a zero-parameter baseline we use the analytic distogram criterion (Algorithm 1), which based on the predicted distogram and the designed backbone score a refold on a simple statistic: the mean negative log-likelihood, under p_{ij}, of the bin containing the reference distance, taken over cross-role pairs closer than 12 Å. A candidate is discarded when this statistic exceeds a per-recycle threshold, subject to a minimum-recycle guard that suppresses kills before the pose has had a chance to form.

D.2 Learned sCC: Architecture and training

In Fig. D.1 the architecture of the learned sCC is illustrated. A more detailed description is outlined below.

Inputs. In addition to the predicted distogram p_{ij}, learned sCC from the designed backbone takes the C_β distance d_{ij} of the same token pair (C_β constructed from N, C_α and C, so the reference is backbone-only and independent of side chains) and encodes it on *those same bins* as a soft code q_{ij}. Two flags accompany each pair: whether both tokens are resolved in the reference, and whether d_{ij} fell in the open-ended far bin, where the head cannot distinguish 22 Å from 80 Å.

Pair selection. Scoring all N^2 pairs is unnecessary and, for a pooled statistic, actively harmful: interface contacts are well under 1% of the grid, so a uniform draw would contain almost none of them. We therefore keep a stratified budget of $10{,}240$ pairs in three bands. Every cross-role pair within 18 Å is retained, so the *interface* band ($<$ 12 Å) and the *halo* band (12–18 Å) are stored complete rather than sampled; the remaining budget is drawn uniformly from the rest of the grid, so the sampled *background* band's mean is an unbiased estimate of the mean over the remaining pairs.

Per-pair encoding. Each retained pair is described by the concatenation of p_{ij}, q_{ij}, the two flags, a learned embedding of the role pair (binder–binder, binder–target, target–target) and a learned embedding of the recycle index, together with the scalar agreement $a_{ij} = -\log(p_{ij} \cdot q_{ij})$, the surprise of the reference distance under the trunk's own prediction. Averaging a_{ij} over the interface

only would reproduce the analytic baseline, the distogram NLL. This gives a 139-dimensional token, layer-normalized, which a single SwiGLU shared across all pairs maps to a 128-dimensional latent.

Pooling and readout. Latents are averaged within each band, each normalized by its own occupancy so that a band contributes equally regardless of how many pairs it holds. Alongside them we append the mean, maximum and 95th percentile of a_{ij} within each band, computed directly as a_{ij} above and bypassing the SwiGLU, so that a purely analytic summary of interface agreement remains a linear readout of the pooled vector. The concatenation (393 dimensions) passes through a 128-unit layer with SiLU activation and dropout 0.1, and a final linear layer with a softmax over 40 bins uniformly covering 0–20 Å yields a predicted distribution over sc-RMSD. Gating uses the tail mass $P(\text{sc-RMSD} < 3\ \text{Å})$, the sum of the first six bins; 3 Å is chosen to fall exactly on a bin edge. The head has $96{,}946$ parameters, negligible against a single trunk recycle.

Training. For each item we record the sc-RMSD of five diffusion decodes at full depth. At every recycle the sCC predicts the distribution of that same full-depth outcome. The loss is the negative log-likelihood of those five outcomes under the prediction, averaged over decodes and summed over recycles with weights decaying linearly from 3 at recycle 0 to 1 at recycle 9, which concentrates capacity on the early recycles where a rejection is worth making.

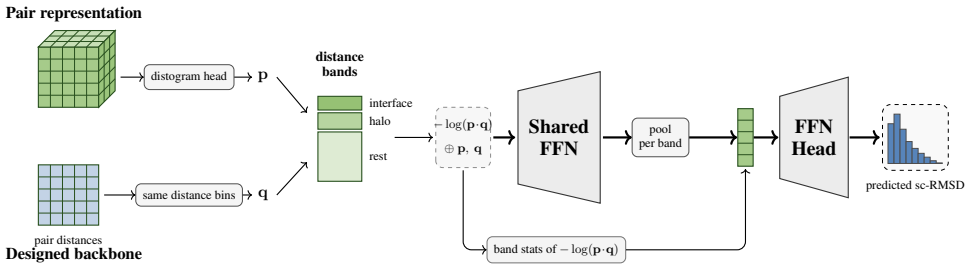


Figure D.1: **Structure CycleCritic.** The trunk's predicted distogram $\mathbf{p}$ and the designed backbone's pair distances $\mathbf{q}$ are encoded on the same distance bins, so their per-pair agreement $-\log(\mathbf{p}\cdot\mathbf{q})$ is the analytic gate. Pairs are banded by distance — interface (< 12 Å), halo (12–18 Å) and a subsample of the rest — and pooled within bands. Band statistics of the agreement bypass the shared network, so the analytic gate stays a linear readout. The head predicts the distribution of self-consistency RMSD.

D.3 Dataset

We train and evaluate the learned sCC on a curated set of VHH complexes derived from the Protein Data Bank. The corpus comprises 1444 distinct PDB entries, decomposed into 2635 VHH-antigen structures. Train/validation partitioning is epitope-cluster-based. The split targets $90/10$ at the cluster level: 574 training and 64 validation clusters, with no cluster shared between sides. This yielded 2196 training and 439 validation VHH-antigen pairs, or 16.7% by structure count rather than 10% due to the validation clusters being arbitrarily larger than those in the training set. A maximum sequence threshold of 1000 tokens per complex was enforced, yielding 1994 unique structures and 282 held-out validation structures.

E Gating binder designs

E.1 Prediction fidelity and analytic CycleCritic

We evaluate the CycleCritic on VHH designs generated as described in Section 5.1.

We first compare the Confidence CycleCritics to the following baselines: 1) The latent dimension in the pair representation that best predicts ipTM or ipSAE and 2) the ipTM proxy as proposed in Candido et al. (2026) (Algorithm 15) for ipTM. In Figure E.1, we evaluate the Protenix-based gating CC and the ESM-based design model. We can see that already early on in recycling the CC is more

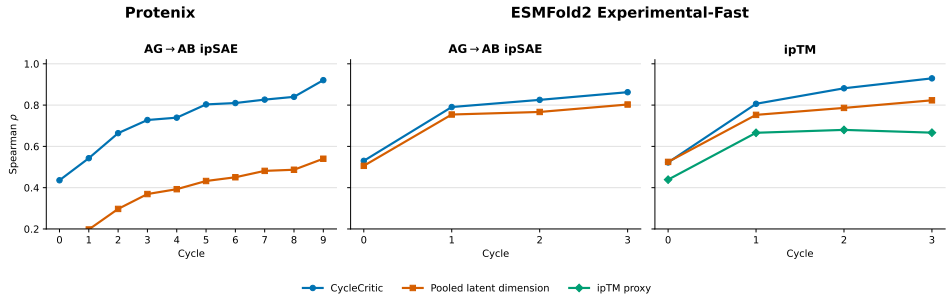


Figure E.1: **Confidence CycleCritic predictions strongly correlate with confidence metrics even at early recycle.** This holds both for Protenix-v2 and ESMFold2. Pooled latent is the baseline described in Figure 2b. The baseline ipTM proxy was proposed in Candido et al. (2026).

predictive of the final metric, where gains are most valuable because they allow us to directly apply hallucination or early gating.

For the Structure CycleCritic (sCC), we include two analytic critics as baselines: a variant that aggregates likelihood over all inter-chain distances and not only the interface and the one that aggregates only over the interface. In Fig. E.2 we show correlation of the sCC scores with the self-consistency RMSD of the binder as well as its predictive power (AUROC) per cycle. The learned sCC exceeds the analytic baseline at every recycle in both rank agreement and AUROC. The operating point is where the two separate: sweeping a per-cycle schedule to a matched recall of 0.96, the learned critic saves 0.284 of the recycling-plus-diffusion budget against 0.209 for the analytic score. Both critics are considerably stronger at late recycles, but the compute saved by a gate accrues early, so the early-cycle behavior dominates deployment value.

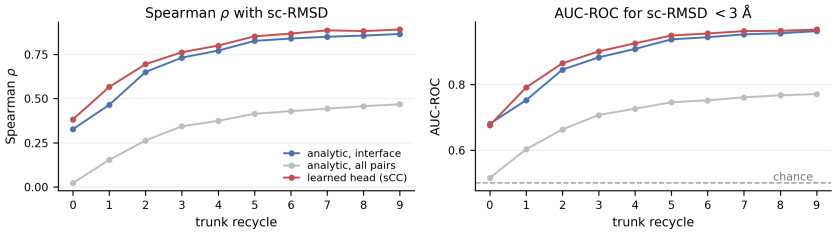


Figure E.2: **Structure CycleCritic (sCC) discriminates self-consistency from the second recycle onward. Left:** Spearman ρ between the cycle-c sCC scores and the pose's sc-RMSD **Right:** AUC-ROC of sCC scores for the binary call sc-RMSD < 3 Å.

E.2 Refolding filters

We refold approximately $126{,}000$ designs with Protenix-v2, using MSAs for the antigen and antibody framework, 10 recycles, 10 sampling steps of the deterministic probability-flow ODE, 5 diffusion samples per trunk pass, and ten independent trunk seeds per design. For each trunk seed, we select the diffusion sample with the highest ipTM among the five samples. The seed passes when this selected sample satisfies both filters. Two cheaper refolding baselines are evaluated on the same designs: Protenix-v2 with 4 recycles and Protenix-mini (Gong et al., 2025).

A refold seed *passes* when ipSAE > 0.5 and antigen-superposed self-consistency RMSD < 3 Å. Repeat 0 is reserved for independent method evaluation and excluded from the reproducibility label. Each design receives a $k/9$ label, where $k \in \{0, \ldots, 9\}$ is the number of passing refolds among the remaining nine seeds. Across the campaign, 1.0% of designs pass on all nine reference seeds, while 4.3% pass on at least one.

	recycle									
	0	1	2	3	4	5	6	7	8	9
sCC kill	0.130	0.095	0.067	0.035	0.018	0.053	0.028	0.019	0.041	0.055
cCC kill	0.297	0.271	0.239	0.130	0.102	0.044	0.055	0.034	0.098	0.368

Table E.1: Deployed per-recycle gating thresholds. The sCC rejects when its predicted tail mass $P(\text{sc-RMSD} < 3\,\text{Å})$ falls below the bar. The cCC rejects when $P(\text{ipSAE} > 0.5)$ falls below its kill bar; Both gates use $\texttt{min_cycles} = 1$, permitting a first rejection at recycle 0.

E.3 ProtenixScore baseline

As a baseline for lightweight, diffusion-free scoring we opted for ProtenixScore (Ring, 2024) which is an implementation of the AF3Score (Liu et al., 2025) approach for Protenix. ProtenixScore can be used to score a given structure without running the full structure prediction of Protenix models. It runs the trunk for all ten recycles but does not proceed to predict the structure using the diffusion Structure Module. Instead, it passes the structure we are scoring, in this case the designed structure as output by BoltzGen (Stark et al., 2025) into the confidence head which produces its scores. The trunk sees the antigen MSA, a framework-realigned binder MSA, and is templated on the input structure. This templating is per-chain, so no inter-chain distance enter the trunk and the pose of the structure is only given to the confidence head.

ProtenixScore runs the trunk on the design's own coordinates and skips diffusion, producing confidence scores but no structure. Designs are retained at a fixed threshold of $\texttt{ipsae_max} > 0.9$. The arm is priced as a single pass over the whole cohort: 227.03 GPU-hours over 126,000 designs, or 6.51 s each. Against the nine-seed reproducibility label, the 1,704 declarations contain 1,022 true positives at the $\geq$ 5-of-9 threshold with precision 0.600, recall 0.374, delivering 4.50 reproducible designs per GPU-hour.

E.4 Gating campaign

Gates. Two early-exit gates operate together. If either gate fires, the refold stops after the current trunk recycle, before diffusion and confidence-head evaluation:

- **cCC** is a learned head on the trunk pair representation that predicts a binned final-depth ipSAE distribution at every recycle. It terminates the refold when the predicted probability $P(\text{ipSAE} > 0.5)$ falls below a per-cycle threshold.
- **sCC** is a learned head that uses the trunk's distogram and the design backbone's pair distances at every recycle to predict the distribution of self-consistency RMSD at full recycling depth. It terminates the refold when $P(\text{sc-RMSD} < 3\,\text{Å})$ falls below a per-cycle threshold.

The gates are kill-only: designs that survive both complete the full refold and are assessed with the downstream ipSAE and sc-RMSD filters. The per-cycle thresholds for both critics were calibrated on held-out SAbDab refolds (282 validation complexes from Sec. D.3) to retain all passing refolds in the calibration set (i.e. 100% recall). Threshold details are shown in Table E.1.

E.5 Results

The gate terminates 107,478 designs (85.6% of the evaluated campaign) before diffusion; 18,057 reach the sampler. The sCC accounts for 85,182 terminations at a mean zero-based exit index of 3.6, and the cCC for 22,296 at 8.8. Mean trunk depth, including designs that survive, falls from 10 to 6.31 passes (36.9% fewer trunk passes).

Of the 5,401 designs passing at least one of nine reference refolds, 4,816 survive the gate (89.2% pre-diffusion retention); 172 are terminated by cCC and 413 by sCC. Under the majority criterion ($\geq 5/9$), 2,658 of 2,732 positives survive (97.3% retention).

Gating survival is not a positive declaration: after diffusion, 2,874 survivors pass both direct ipSAE and sc-RMSD filters. Against the $\geq 1/9$ label used for precision and recall in Table E.2, 2,769 declarations are true positives, giving precision 0.96 and recall 0.51.

		Against the $\geq 1/9$ label		**GPU-h per confirmed passer**		
Method	Accepted	Precision	Recall	$\geq 1/9$	$\geq 5/9$	$9/9$
ptx-mini	1,056	0.78 ± 0.03	0.15 ± 0.01	0.146	0.182	0.252
ptx@4	2,618	0.92 ± 0.01	0.45 ± 0.02	0.074	0.092	0.157
ptx@10	2,925	0.95 ± 0.01	$\mathbf{0.52 \pm 0.02}$	0.085	0.106	0.189
ptx_score	1,704	0.73 ± 0.03	0.23 ± 0.02	0.183	0.222	0.321
cc+sccc gated	2,874	$\mathbf{0.96 \pm 0.01}$	0.51 ± 0.02	**0.065**	**0.080**	**0.145**

Table E.2: Method comparison on the 126,000-design campaign. *Accepted* is the number of designs a method declares a passer; precision and recall score that declaration against a design that reproduces on at least one of nine held-out Protenix@10 seeds ($n = 5{,}401$), with $\pm$ half the width of a 95% chunk-clustered bootstrap interval over designs (50-design chunks resampled within target; designs in a chunk share a backbone and task and are not independent). The right-hand block divides the same measured cost by the passers confirmed at three progressively stricter labels: reproducing on at least one seed, at least five, and all nine. Only the label moves across those columns - a method's verdict is its own pose against its own bars and does not depend on the label. Best value per column in bold.

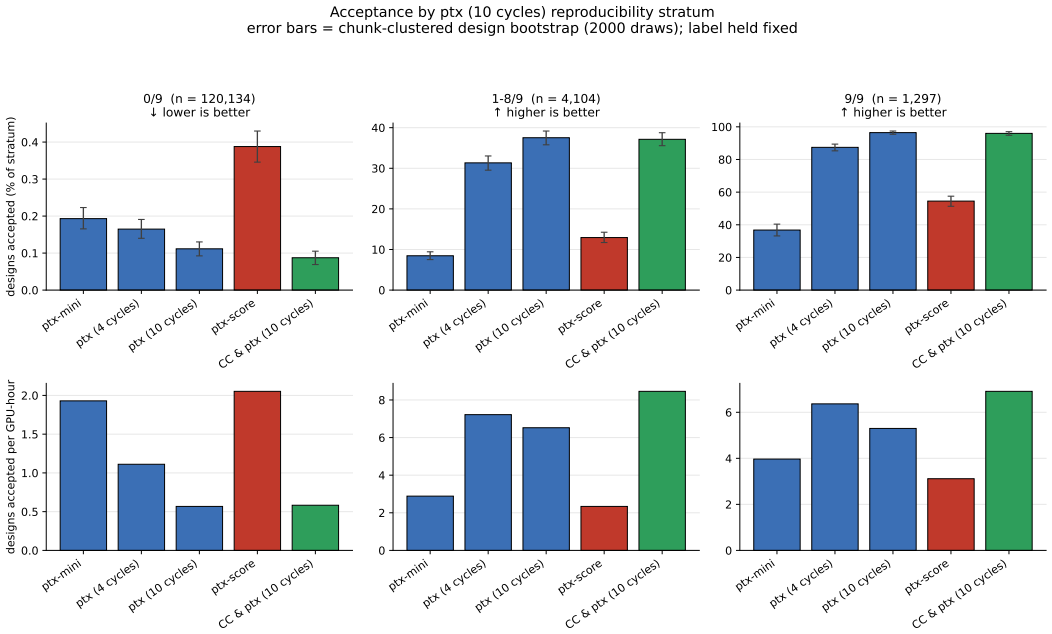


Figure E.3: Acceptance by reproducibility stratum. Each design is classified by the number of passing refolds among nine held-out seeds (102–110): $0/9$ ($n = 120{,}134$), 1–$8/9$ ($n = 4{,}104$), or $9/9$ ($n = 1{,}297$). Methods declare independently on seed 101. **Top:** fraction accepted within each stratum. **Bottom:** accepted designs per GPU-hour within each stratum. The gate accepts 0.087% of $0/9$ designs while retaining 95.99% of $9/9$ designs; ptx@10 accepts 0.112% and 96.45%, respectively. Error bars are 95% confidence intervals from 2,000 bootstrap draws of 50-design chunks resampled within target.

E.6 Simulating gating performance on a smaller set of generated designs

We evaluated the critics on a separate set of 14,000 generated designs, with 2,000 per target. Sweeping their per-cycle thresholds gives the trade-off between pre-diffusion retention of designs passing

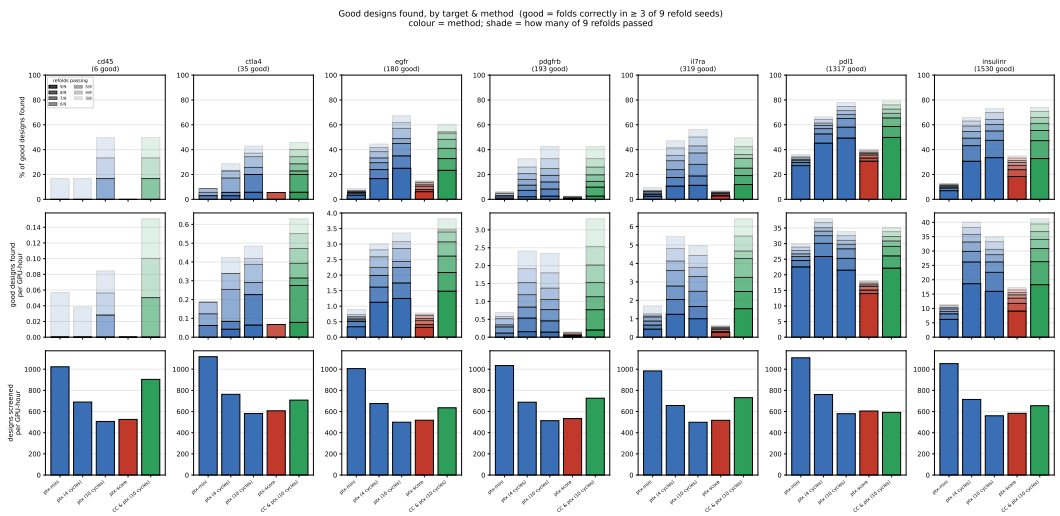


Figure E.4: Delivery of reproducible designs by target. A design is *good* if it passes at least three of nine held-out refolds; seed 101 is reserved for each method's declaration. Targets are ordered by their number of good designs. Colour identifies the method and shade the number of passing jury seeds. **Top:** recall among good designs. **Middle:** good designs delivered per GPU-hour. **Bottom:** designs screened per GPU-hour. The gate delivers more good designs per GPU-hour than ptx@10 on all seven targets, with gains from 3.5% (PD-L1) to 78.5% (CD45), despite lower recall on EGFR and IL7RA.

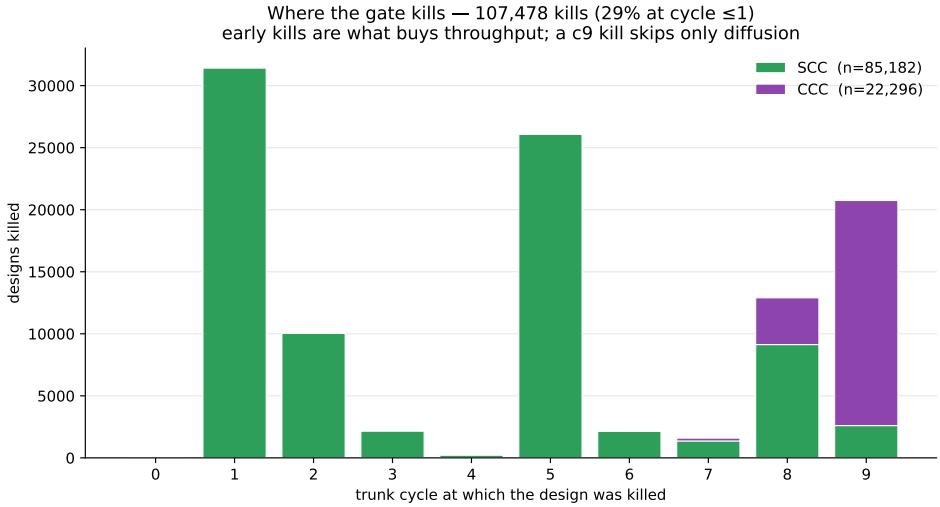


Figure E.5: Exit decisions of the union gate. Bars show terminations by zero-based trunk-cycle index and the gate that fired. Of approximately 126,000 designs, 107,478 (85.6%) are terminated: 85,182 by sCC and 22,296 by cCC. Their median exit indices are 3 and 9, corresponding to four and ten completed trunk passes.

at least three of nine held-out refolds and screening throughput. Note that the deployed thresholds in the full experiment were selected independently of this evaluation set. Figure E.6 shows the resulting frontier and the deployed operating point.

Cost model. For a design, modeled runtime in seconds is

$$\text{cost} = F_0 + cC + sD,$$

Gate	Critics	Designs/GPU-hour	Recall ($\geq 3/9$)	Kill rate
No gate	—	514.2	1.0000	0.000
Learned (deployed)	cCC ∪ sCC	**736.0**	**0.8966**	0.848
Analytic (deploy-like)	ipTM proxy ∪ dCC	721.6	0.8549	0.946

Table E.3: Simulated prescreening on 14,000 VHH designs. Recall is the fraction of $\geq 3/9$ designs retained before diffusion; it is not final declaration recall. Critics use their deployed per-cycle thresholds and minimum-cycle guards. Seed 101 is excluded from the nine-seed reference label.

where F_0 is a fixed per-design cost (2.76 s gated, 2.23 s ungated), c is the number of completed trunk passes, $C = 0.30$ s is the cost per pass, s indicates survival to diffusion, and $D = 1.74$ s is the diffusion cost. A design terminated at zero-based cycle index k has $c = k + 1$; a survivor has $c = 10$. Throughput is $3600/\overline{\text{cost}}$ designs per GPU-hour. The printed constants give an ungated cost of approximately 7 s per design, or approximately 514 designs per GPU-hour using the unrounded fitted values.

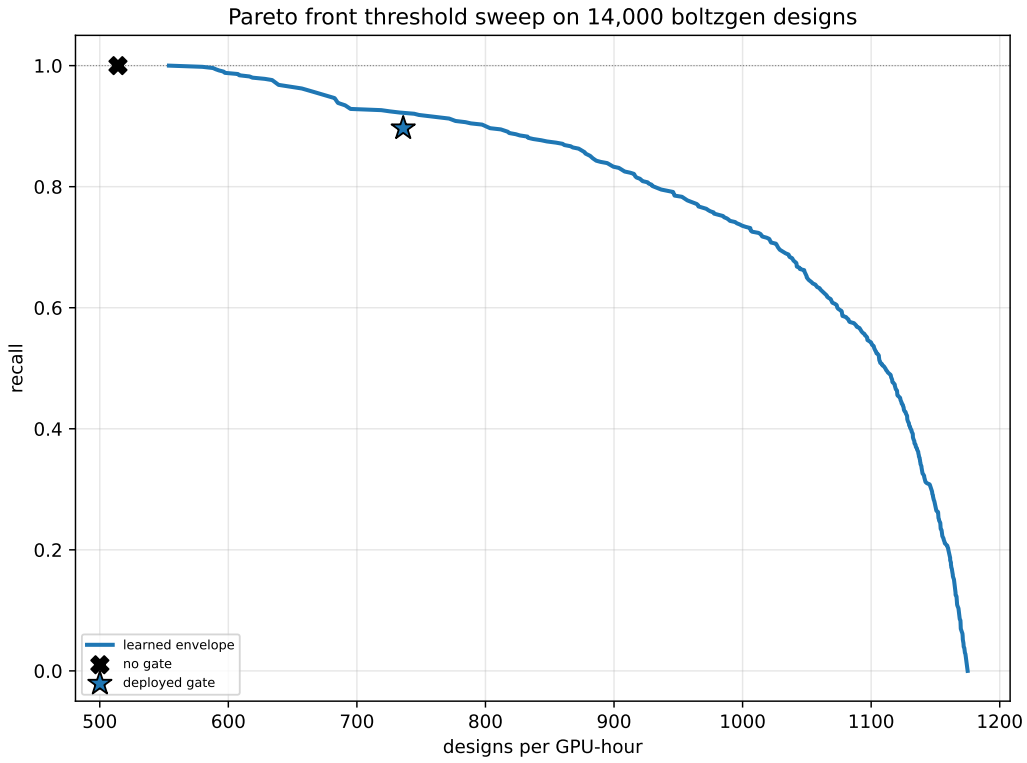


Figure E.6: Simulated trade-off between pre-diffusion recall and screening throughput on 14,000 generated designs. Recall uses the $\geq 3/9$ held-out-refold label. The star marks the deployed thresholds, which were selected independently of this evaluation set.

F Design with ESMFold2

F.1 Base ESMFold2 design objective

We build on the gradient-guided binder design procedure of the ESMFold2 paper, using their two folding trunks, `ESMFold2-Experimental-Fast` (cutoff 2021) and `ESMFold2-Experimental-Fast-Cutoff2025`, with one trunk randomly sampled each op-

timization step. The base structure loss $\mathcal{L}_{\text{struct}}$ is computed entirely from the predicted distogram,

$$\mathcal{L}_{\text{struct}} = 0.5\,\mathcal{L}_{\text{intra}} + 0.5\,\mathcal{L}_{\text{inter}} + 0.2\,\mathcal{L}_{\text{glob}}, \tag{10}$$

where $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$ are distogram-based intra-binder and binder–target contact terms, and $\mathcal{L}_{\text{glob}}$ is a radius-of-gyration globularity term. A sequence-plausibility regularizer (ESM-C pseudo-perplexity) is applied identically in the base and CycleCritic runs and combined at the gradient level, normalized per design, with weight 0.05. Notably, the base objective contains no differentiable structure of confidence-head based term: for design they also skip diffusion and confidence head.

F.2 CycleCritic-augmented objective

Since the original ESM-design randomly picks one of the two experimental models per step for generation, we trained different CCs to predict each of their pair representations.

Further we compute two CycleCritic-based losses

$$\mathcal{L}^{(t)}_{\text{CC-ipTM}} = 1 - \widehat{\text{ipTM}}_{\text{CC}^{(t)}}, \tag{11}$$

$$\mathcal{L}^{(t)}_{\text{CC-ipSAE}} = 1 - \widehat{\text{ipSAE}}_{\text{CC}^{(t)}} \tag{12}$$

for $ipTM$ and $\text{ipSAE}_{\text{AG}\to\text{AB}}$. Here, each head $CC^{(t)}$ only ever sees z from the trunk t it was trained on. The full objective is

$$\mathcal{L}_{\text{struct,CC}} = \mathcal{L}_{\text{struct}} + 0.5\left(\mathcal{L}^{(t)}_{\text{CC-ipTM}} + \mathcal{L}^{(t)}_{\text{CC-ipSAE}}\right) \tag{13}$$

F.3 Evaluation/Selection metric.

Like in Candido et al. (2026), each design is refolded once with each of four checkpoints: `ESMFold2-Experimental-Fast`, `ESMFold2-Experimental-Fast-Cutoff2025`, `ESMFold2-Experimental`, and `ESMFold2-Experimental-Cutoff2025`. Refolds use three recycling loops and 200 sampling steps, with the same folding seed shared across checkpoints for each design. We define the selection score as

$$S_i = \frac{1}{8}\sum_{m=1}^{4}\left(\text{ipTM}_{im} + \text{ipTM}^{\text{proxy}}_{im}\right), \tag{14}$$

where $\text{ipTM}^{\text{proxy}}$ is the distogram-derived ipTM surrogate. All designs have complete scores from all four checkpoints.

The two Fast checkpoints also serve as generation trunks; the two non-Fast checkpoints are held out from generation.

F.4 Guidance parameter study for ESMFold2 binder generation

We investigate design-time guidance on four targets (CD45, EGFR, PDGFRβ, and PD-L1) and three antibody frameworks (trastuzumab, atezolizumab, and ocankitug), giving 12 target–framework cells. All nine arms use the same generator and differ only in their guidance settings. Each arm contains 255 designs per cell, or 3,060 designs. The eight guided arms are seed-paired.

Reported errors are within-cell SEs; comparisons use nominal paired $t(11)$ statistics across the 12 cells

F.4.1 The parameters

We vary four parameters of the guidance objective and select, axis by axis, the setting that maximises selection:

parameter	candidate settings
guidance source	none (base), training-free ipTM proxy, CycleCritic
critic heads	ipTM, ipTM + $\mathrm{ipSAE}_{\mathrm{AG}\to\mathrm{AB}}$ (dual)
recycling schedule	all cycles, last cycle only
guidance weight	0.50, 0.85, 1.00, 1.50

The Cycle-Critic carries per-trunk-matched heads (a critic head trained for the folding trunk it scores). Unless a comparison holds a parameter fixed to isolate another, arms are compared at the eventual recommended setting (dual head, all cycles, weight as noted).

F.4.2 Guidance source — the learned critic beats the training-free proxy

Both guidance sources improve substantially over unguided design, and the learned critic beats the training-free surrogate:

guidance source (all cycles, w 0.85)	selection	Δ vs base
none (base)	0.7832	—
ipTM proxy	0.8061	$+0.0229$ ($t(11) = 3.4$, 12/12)
Cycle-Critic (ipTM)	0.8175	$+0.0343$ ($t(11) = 3.3$, 12/12)

The critic beats the proxy by a paired $+\mathbf{0.0114}$ (SE 0.0043, $t(11) = 2.6$; 10/12 cells), with the tiers cleanly separated (the weakest critic arm sits above the strongest proxy arm). Because generator and launcher are held fixed, this is attributable to the guidance objective alone.

F.4.3 Critic heads — the dual head beats ipTM alone

Adding a second CC head on top of the ipTM head, holding routing and weight fixed:

critic heads (all cycles, w 0.85)	selection	Δ
ipTM	0.8175	—
ipTM + ipSAE	0.8267	$+0.0092$ ($t(11) = 1.8$, 8/12)

F.4.4 Recycling schedule — apply the critic on all cycles

Averaging the guidance over all recycling carry cycles vs. applying it only on the final cycle has a much larger effect for the learned critic than for the proxy:

guidance source	all $-$ end (Δ selection)	SE	$t(11)$	cells won
learned critic (dual, w 0.85)	$+\mathbf{0.0157}$	0.0048	$\mathbf{3.3}$	**12/12**
distogram proxy (w 0.85)	$+0.0021$	0.0020	1.0	9/12

All-cycle averaging is worth $\sim 7\times$ more to the learned critic, a clean 12/12 sweep; the end-cycle dual critic (0.8111) even falls *below* the all-cycle ipTM-only critic (0.8175). The learned critic evidently extracts usable interface signal from the intermediate recycling trajectory that the training-free surrogate cannot.

F.4.5 Guidance weight — saturates by about 1.0

Sweeping the weight for the dual critic on all cycles (all four points seed-paired):

weight	selection	Δ vs previous	$t(11)$
0.50	0.8202	—	—
0.85	0.8267	$+0.0065$	1.7 (9/12)
1.00	**0.8292**	$+0.0025$	1.6 (7/12)
1.50	0.8308	$+0.0015$	1.2 (8/12)

The curve rises monotonically but saturates: the step to 1.00 still adds a measurable increment, whereas the further step to 1.50 is negligible. This can be explained by the fact that the gradients of the structure loss get normalized to unit length. Since the gains after 1.0 are marginal and we want to avoid breaking optimization we choose weight 1.0.

F.4.6 RECOMMENDED CONFIGURATION

The parameter study converges on
Cycle-Critic, dual head (ipTM + ipSAE), all recycling cycles, guidance weight 1.0,
which reaches 0.8292 selection ($+5.88\%$ over unguided base). Full macro-averaged leaderboard over the 12 shared cells (± 1 SE = within-cell sampling):

arm	selection	± 1 SE	Δ vs base
Cycle-Critic, dual, all cycles, weight 1.50	0.8308	± 0.0012	$+6.08\%$
Cycle-Critic, dual, all cycles, weight 1.00 — recommended	**0.8292**	$\mathbf{\pm 0.0013}$	$\mathbf{+5.88\%}$
Cycle-Critic, dual, all cycles, weight 0.85	0.8267	± 0.0013	$+5.56\%$
Cycle-Critic, dual, all cycles, weight 0.50	0.8202	± 0.0014	$+4.73\%$
Cycle-Critic, ipTM, all cycles, weight 0.85	0.8175	± 0.0014	$+4.39\%$
Cycle-Critic, dual, end cycle, weight 0.85	0.8111	± 0.0014	$+3.56\%$
iptm-proxy, all cycles, weight 0.85	0.8061	± 0.0016	$+2.93\%$
iptm-proxy, end cycle, weight 0.85	0.8040	± 0.0016	$+2.66\%$
base (no guidance)	0.7832	± 0.0018	—

Summarizing, while the iptm-proxy already improves hallucination-based design ($+2.93\%$ over base), the Cycle-Critic wins out for two reasons: it is flexible enough to model other confidence metrics (the added ipSAE head), and its early-cycle predictions are far more informative than the proxy's — guiding on all recycling cycles rather than only the last helps the critic roughly seven times as much as it helps the proxy. The recommended configuration reaches $+5.88\%$ over base, roughly double the proxy's gain.

F.5 RESULTS

While the additional compute of the two loss terms across the two CycleCritic executions leads to a modest increase of 0.4% in time per design (Table F.2), the optimization metric used by ESMFold2-design favors the CC-based optimization in 6 out of 7 targets for both VHHs and Fvs (Table F.1). The optimization metric is the mean of the ipTM and distogram proxy predicted by four different critics refolding the designed sequence.

G HALLUCINATION WITH MOSAIC

We compare Mosaic's (Boyd et al., 2025) standard differentiable sequence design with full Protenix against a trunk-only CycleCritic objective. The purpose of the CycleCritic objective is to replace the expensive diffusion and confidence-head losses during optimization. Final selected designs were still evaluated by full Protenix refolding model.

G.1 ORIGINAL MOSAIC OBJECTIVE

The standard Mosaic objective we used in the baseline design runs is

$$\begin{aligned}\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{contact}} + 0.05\,\mathcal{L}_{\text{target}\to\text{binder PAE}} + 0.05\,\mathcal{L}_{\text{binder}\to\text{target PAE}} \\ + 0.025\,\mathcal{L}_{\text{iPTM}} + 0.4\,\mathcal{L}_{\text{binder PAE}} + 0.025\,\mathcal{L}_{\text{pTM-energy}} + 0.1\,\mathcal{L}_{\text{pLDDT}}.\end{aligned} \tag{15}$$

Here, $\mathcal{L}_{\text{contact}}$ is a distogram-based binder-target contact term. The two interface PAE terms penalize predicted aligned error from target to binder and binder to target. The binder PAE term penalizes predicted aligned error within the designed binder chain. The ipTM and pLDDT terms come from the full Protenix confidence machinery, while $\mathcal{L}_{\text{pTM-energy}}$ is a pTM-like energy computed from the PAE logits.

Table F.1: Design-suite ESM results pooled by framework class and target, at equal 6 GPU-h per cell. Metric is mean4(ipTM, ipTM distogram proxy). Top-K columns report the mean of a cell's K best designs. The budget is drawn against the critic: the measured $+0.4\%$ per-design overhead is applied, so CycleCritic is allotted slightly fewer designs than the baseline. Larger Base/CC entries are bold; the summary row gives the format totals and the mean across targets.

Format	Target	n (Base/CC)	Mean		Median		Top-5		Top-10		Top-25		Top-50	
			Base	CC	Base	CC	Base	CC	Base	CC	Base	CC	Base	CC
VHH	CD45	403/401	.522	**.739**	.477	**.790**	.914	**.929**	.908	**.925**	.895	**.919**	.855	**.913**
	CTLA-4	600/596	.846	**.860**	.876	**.883**	**.919**	.917	**.917**	.916	.912	**.913**	.908	**.911**
	EGFR	392/392	.762	**.817**	.798	**.845**	.913	**.918**	.905	**.914**	.894	**.908**	.883	**.900**
	IL7RA	388/388	**.748**	.669	**.783**	.678	**.920**	.905	**.915**	.899	**.909**	.888	**.900**	.873
	INSR	477/476	.812	**.856**	.841	**.878**	.928	**.932**	.924	**.929**	.916	**.925**	.909	**.920**
	PDGFRβ	400/400	.762	**.792**	.811	**.818**	.896	**.900**	.893	**.897**	.887	**.890**	.878	**.883**
	PD-L1	604/600	.855	**.875**	.883	**.895**	**.947**	.945	**.944**	.940	**.939**	.936	**.935**	.933
	mean	3264/3253	.758	**.801**	.781	**.827**	.919	**.921**	.915	**.917**	.907	**.911**	.896	**.905**
Fv	CD45	159/158	.631	**.793**	.616	**.831**	.875	**.890**	.869	**.884**	.851	**.876**	.821	**.867**
	CTLA-4	222/221	.860	**.864**	.870	**.870**	**.910**	.907	**.906**	.904	**.899**	.899	.894	**.894**
	EGFR	156/156	.829	**.852**	.859	**.874**	**.921**	.921	.916	**.918**	.906	**.910**	.897	**.902**
	IL7RA	156/155	.775	**.778**	**.812**	.805	**.893**	.892	**.886**	.883	**.874**	.869	**.862**	.854
	INSR	183/181	.806	**.832**	.828	**.849**	.907	**.910**	.899	**.904**	.888	**.896**	.879	**.889**
	PDGFRβ	159/159	.780	**.791**	.789	**.805**	**.869**	.862	**.862**	.859	.851	**.852**	.838	**.841**
	PD-L1	225/225	.879	**.882**	.891	**.892**	**.934**	.929	**.930**	.926	**.923**	.921	**.917**	.917
	mean	1260/1255	.794	**.828**	.809	**.847**	.901	**.902**	.895	**.897**	.885	**.889**	.873	**.881**

Table F.2: Cycle-Critic (dual ipTM+ipSAE) generation-time for ESM hallucination. The per-design row is the 149-step optimization loop

Quantity	Format	Base	CC	Δ
Per design	VHH	130.8 s	131.3 s	$+0.4\%$
	Fv	265.9 s	267.0 s	$+0.4\%$

G.2 CycleCritic replacement objective

In the CycleCritic runs, we replaced the confidence-dependent terms by lightweight CycleCritic predictions computed from the pair representations. Our objective then was

$$\begin{aligned}\mathcal{L}_{\mathrm{CC}} = \mathcal{L}_{\mathrm{contact}} &+ 0.05\,\mathcal{L}_{\mathrm{target}\to\mathrm{binder\ dist}} + 0.05\,\mathcal{L}_{\mathrm{binder}\to\mathrm{target\ dist}} \\ &+ 0.025\,\mathcal{L}_{\text{CC-iPTM}} + 0.4\,\mathcal{L}_{\mathrm{binder\ dist}} + 0.025\,\mathcal{L}_{\text{CC-pTM}} + 0.1\,\mathcal{L}_{\text{CC-pLDDT}}. \end{aligned} \tag{16}$$

The pTM-energy term was replaced by the pTM CycleCritic prediction in the same weighted slot.

The CycleCritic metric losses are defined as

$$\mathcal{L}_{\text{CC-iPTM}} = -\widehat{\mathrm{iPTM}}_{\mathrm{CC}}, \tag{17}$$

$$\mathcal{L}_{\text{CC-pTM}} = -\widehat{\mathrm{pTM}}_{\mathrm{CC}}, \tag{18}$$

$$\mathcal{L}_{\text{CC-pLDDT}} = -\widehat{\mathrm{pLDDT}}_{\mathrm{CC}}. \tag{19}$$

Each CycleCritic head predicts a distribution over bins for the corresponding confidence metric, and the scalar prediction is the expected value of that distribution:

$$\widehat{m}_{\mathrm{CC}} = \sum_b p_{\mathrm{CC}}(b)\, c_b, \qquad m \in \{\mathrm{iPTM}, \mathrm{pTM}, \mathrm{pLDDT}\}. \tag{20}$$

For runs using all trunk cycles, the CycleCritic predictions were averaged over cycles:

$$\widehat{m}_{\mathrm{CC}} = \frac{1}{C}\sum_{c=1}^{C} \widehat{m}^{(c)}_{\mathrm{CC}}. \tag{21}$$

The PAE terms in the original full-Protenix objective were replaced by distogram expected-distance terms. For each residue pair, the expected distance was computed from the distogram distribution:

$$\mathbb{E}[d_{ij}] = \sum_b p_{ij}(b)\, d_b. \tag{22}$$

The binder-target distance term averages the nearest expected distance to the target over binder residues:

$$\mathcal{L}_{\text{binder}\to\text{target dist}} = \frac{1}{N_{\text{binder}}} \sum_{i\in\text{binder}} \min_{j\in\text{target}} \mathbb{E}[d_{ij}]. \tag{23}$$

The binder-internal distance term averages expected distances within the binder:

$$\mathcal{L}_{\text{binder dist}} = \frac{1}{N_{\text{binder}}^2} \sum_{i,j\in\text{binder}} \mathbb{E}[d_{ij}]. \tag{24}$$

We did not apply hyperparameter optimization to fine-tune the loss weights.

G.3 Design objectives and optimization settings

Both objectives used identical optimization settings. We used APGM with a three-stage schedule totaling $50 + 30 + 30 = 110$ optimization steps:

Stage	Steps	Step size	Momentum	Scale	Log-space
1	50	$1.5\sqrt{N_{\text{design}}}$	0.2	1.0	yes
2	30	$0.5\sqrt{N_{\text{design}}}$	0.0	1.1	no
3	30	$0.25\sqrt{N_{\text{design}}}$	0.0	1.1	yes

The production runs used batch size 1, two Protenix recycling steps during optimization, one final sample per seed, and full Protenix-v2 refolding with 10 cycles and 200 diffusion steps for endpoint evaluation.

We ran the optimization for 6h per target $\times$ framework and loss, so 49 combinations to compare on.

G.4 Results

The results can be seen in Table G.1. The CC speeds up sampling substantially, while yielding a similar metric distribution.

H Analysis of binder designs

Analyses were run over the full generated pool of both campaigns, across two formats (VHH, scFv) and seven antigens. Comparisons are within a campaign and format. All quantities are reported as rates or per-design means rather than raw totals, as the equal-time budget for the different methods leave the baseline and CycleCritic arms with unequal pool sizes.

H.1 Diversity and novelty

Within-sample diversity. Designs are compared on their concatenated designed CDRs. The exact-duplicate rate is zero everywhere, with a unique fraction of 1.000 in all cells in both baseline and CycleCritic arms. Pairwise CDR-H3 identity is computed with a Levenshtein-indel normalised similarity within each cell and sits at $\sim$27–30$\%$ in every condition, with a CycleCritic-versus-baseline difference of at most 0.6% (Figure 8a). Greedy clustering of the same identity matrix at 70, 80 and 90% leaves approximately one design per cluster throughout, so no near-duplicate families form in either arm.

Table G.1: Design-suite Mosaic results pooled by framework class and target, at equal GPU-time. Metric is held-out full-Protenix refold ipTM. Top-K columns report the mean of a cell's K best designs, which is rank-comparable across arms holding unequal numbers of designs. Larger Base/CC entries are bold. n is the number of designs each arm produced for that target within the shared budget; the summary row gives the format totals and the mean across targets.

Format	Target	n (Base/CC)	Mean		Median		Top-5		Top-10		Top-25		Top-50	
			Base	CC	Base	CC	Base	CC	Base	CC	Base	CC	Base	CC
VHH	CD45	175/282	.363	**.372**	**.291**	.281	.873	**.905**	.819	**.888**	.738	**.821**	.624	**.744**
	CTLA-4	253/497	**.431**	.427	**.353**	.337	.899	**.917**	.884	**.911**	.843	**.891**	.784	**.860**
	EGFR	167/253	.339	**.353**	.294	**.313**	.814	**.837**	.743	**.783**	.620	**.685**	.518	**.589**
	IL7RA	165/263	.428	**.446**	.383	**.395**	.887	**.891**	.841	**.880**	.759	**.841**	.678	**.780**
	INSR	191/313	.521	**.532**	.477	**.497**	**.911**	.904	.883	**.890**	.822	**.863**	.767	**.826**
	PDGFRβ	168/269	.432	**.448**	.397	**.421**	.882	**.898**	.843	**.866**	.742	**.805**	.647	**.739**
	PD-L1	241/440	.526	**.557**	.491	**.531**	.930	**.934**	.920	**.930**	.891	**.918**	.827	**.893**
	mean	1360/2317	.434	**.448**	.384	**.397**	.885	**.898**	.848	**.878**	.773	**.832**	.692	**.776**
FV	CD45	82/106	**.295**	.294	**.231**	.231	.667	**.725**	.617	**.678**	.506	**.561**	.376	**.426**
	CTLA-4	107/154	**.335**	.329	**.297**	.279	.630	**.787**	.578	**.713**	.501	**.588**	.420	**.478**
	EGFR	81/98	.353	**.366**	**.305**	.298	.746	**.758**	.659	**.716**	.539	**.622**	.431	**.489**
	IL7RA	78/100	**.337**	.336	**.311**	.293	.675	**.713**	.625	**.674**	.523	**.581**	.423	**.459**
	INSR	85/115	**.526**	.500	**.521**	.461	.839	**.854**	.803	**.837**	.736	**.773**	.645	**.663**
	PDGFRβ	78/100	**.342**	.336	**.302**	.292	.684	**.708**	.619	**.647**	.501	**.558**	.407	**.452**
	PD-L1	101/142	**.458**	.392	**.402**	.325	.837	**.861**	**.814**	.814	**.743**	.725	**.614**	.597
	mean	612/815	**.378**	.365	**.339**	.311	.725	**.772**	.674	**.725**	.578	**.630**	.474	**.509**

Table G.2: Measured per-design wall-clock generation time for the Mosaic campaign at equal GPU-time (six hours per cell and arm). Duration is the gap between consecutive design start-times within a worker, so it captures designs per unit GPU-time including per-design overhead. Δ is the change of CycleCritic relative to the baseline.

Format	Mean (s)			Median (s)		
	Base	CC	Δ	Base	CC	Δ
VHH	453.2	261.4	$\mathbf{-42.3\%}$	465.0	262.0	$\mathbf{-43.7\%}$
FV	790.7	588.5	$\mathbf{-25.6\%}$	795.0	582.5	$\mathbf{-26.7\%}$
All	552.1	341.6	$\mathbf{-38.1\%}$	515.5	308.0	$\mathbf{-40.3\%}$

Novelty against the human repertoire. For every design we locate its closest natural counterpart among 2,295,779 unique human heavy CDR-H3 loops drawn from 3.0M paired antibodies across 19 OAS studies, matching on length and scoring by Hamming identity. No design reproduces a natural CDR-H3 exactly (0 of 14,136), and the closest human loop is only $\sim$41–48% identical (Figure 8b). CycleCritic is equal or more novel than baseline in every cell. Importantly, since paired OAS contains no camelid sequences, nanobody CDR-H3s are scored against human heavy chains, inflating their apparent novelty by construction; both arms are scored against the same reference, however, so the comparison between them is unaffected.

Antibody language-model likelihood and humanness. We assess the designs' naturalness and humanness with two models that are not part of the design loop, AbLang2 pseudo-log-likelihood and the OASis humanness score. By equivalence testing, CycleCritic is equal or better in all eight comparisons (Figure 8c), with uniformly small effect sizes, suggesting CycleCritic does not degrade designs naturalness.

Interface diversity. Sequence diversity does not preclude designs converging on a common binding mode, so we compare interfaces directly with iAlign, scoring all pairs within each (framework, target, arm) cell of both campaigns. Mean $1 - \mathrm{IS}$ is 0.753 for baseline and 0.748 for CycleCritic in the ESMFold2 campaign, and 0.831 against 0.821 in Mosaic, indicating low similarity (Figure 8d).

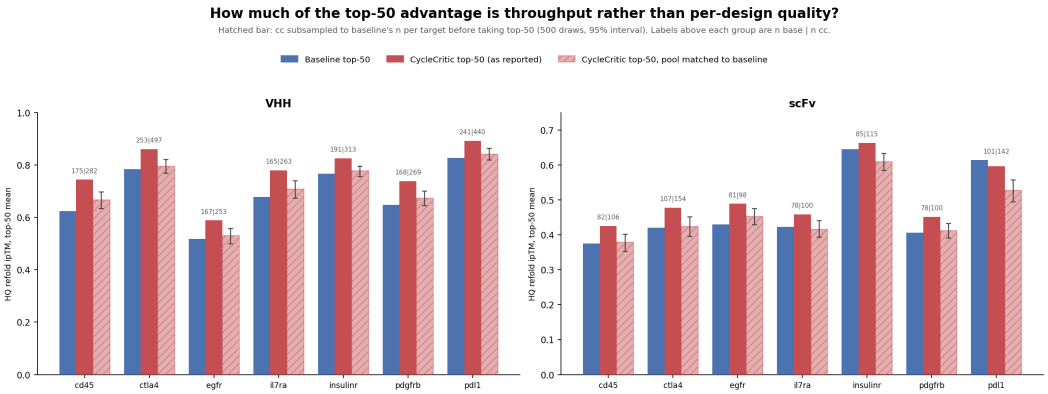


Figure H.1: **How much of the Mosaic top-50 advantage is throughput rather than per-design quality.** Per antigen, baseline top-50 (blue), CycleCritic top-50 as reported (red), and CycleCritic top-50 after subsampling its pool to the baseline's n before taking the top-50 (hatched, 500 draws, 95% interval). Labels above each group are n baseline | n CycleCritic. Where the hatched bar falls back toward the baseline the top-50 edge is largely a more-shots effect; where it stays near the solid red bar the edge survives pool-matching.

Throughput and the top-K advantage. In the ESMFold2 campaign CycleCritic costs $+0.4\%$ per design and lifts the whole distribution. In Mosaic its per-design optimisation is $\sim 1.6\times$ faster, so the same 6 GPU-h/cell budget yields $1{,}972$ baseline against $3{,}132$ CycleCritic designs, and the advantage is concentrated in the selected top-K rather than the full-set mean. Subsampling the CycleCritic pool to the baseline's n before taking each arm's top-50 (Figure H.1) shows how much of that edge is due to throughput. This further supports that the underlying methods are not degraded by CycleCritic; instead, CycleCritic provides a speedup that allows us to produce more designs in less time, improving the best-of-K selection under a fixed time budget.