MotifCraft: scalable functional protein binder design with AlphaFold2 hallucination

Anonymous Authors[1]

Abstract

Motif scaffolding is one promising direction in computational protein design where *de novo* protein scaffolds are designed to maintain a set of functional protein coordinates – a motif. Here we explore an AlphaFold-Multimer hallucination workflow, *MotifCraft*, for motif scaffolding and present approaches that improve efficiency while maintaining or improving protein generation accuracy. Our approach produces higher in silico success rates than existing diffusion-based scaffolding methods. Furthermore, we show that scaffolding a binder interface motif results in designs scoring as well, or better, according to established in silico interface metrics when we include the target protein in the workflow. Finally, we conclude that cropping the target protein to the minimal binding interface of a target protein is an effective method for scalable binder design against large proteins. *MotifCraft* thus enables fast, accurate, and scalable motif scaffolding as well as binder design against large protein targets.

1. Introduction

The field of *de novo* protein design has seen big advances in the past years, to a large extent tracking the successes in deep learning-based protein structure prediction (Khakzad et al., 2023). Generative approaches like RFdiffusion (Watson et al., 2023) and gradient-based methods like BindCraft (Pacesa et al., 2025), among many others, have emerged to solve the problem of conditionally and unconditionally generating a novel protein backbone and the corresponding amino acid sequence (Dauparas et al., 2022; 2025). However, design of functional *de novo* proteins does not consistently succeed (Albanese et al., 2025). One such area of functional design is motif scaffolding, where *de novo* protein scaffolds are generated to maintain a set of functional protein coordinates, such as enzyme active sites or interface residues.

BindCraft was shown to have exceptionally high experimental success rates for designed mini-protein binders, ranging from 10-100% (Pacesa et al., 2025). The paradigm of iterative protein backbone and sequence co-design seems to be very effective in generating mini-protein binders and allows for a high degree of freedom due to the modifiable losses. In this work, we exploit a similar AlphaFold2 (AF2)-backpropagation model for motif scaffolding. By initializing the network with the motif coordinates and performing gradient-descent using motif-specific losses, *MotifCraft* generates scaffolds both in the presence of and without a target protein.

One of the main known limitations of this approach is the computationally intensive backpropagation through AF2, which additionally scales poorly for large proteins (Evans et al., 2022; Bennett et al., 2023). Due to the complexity of the model, run time increases cubically with the length of the input protein sequence(s) (Jumper et al., 2021). A simple approach to reducing computational efforts would be to reduce the total number of amino acids in the protein (complex). Here we compare two effective methods for reducing the target protein to the essential binding interface to accelerate binder design through AF2-backpropagation.

Overall, we obtain high quality *de novo* scaffolds that meet established in silico designability criteria (Overath et al., 2025) across various benchmarks. We achieve an up to 9-fold acceleration of gradient-based design against large protein targets. Thus, using AF2-hallucination through backpropagation provides a novel, precise, and scalable approach to protein binder design and motif scaffolding.

2. AlphaFold2 backpropagation for motif scaffolding

To generate motif scaffolds, the ColabDesign implementation of AF2 is used, with inspiration from the BindCraft pipeline. The input consists of the structural template of the motif and the scaffold sequence, which is randomly initialized, but contains the motif placed at a random posi-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

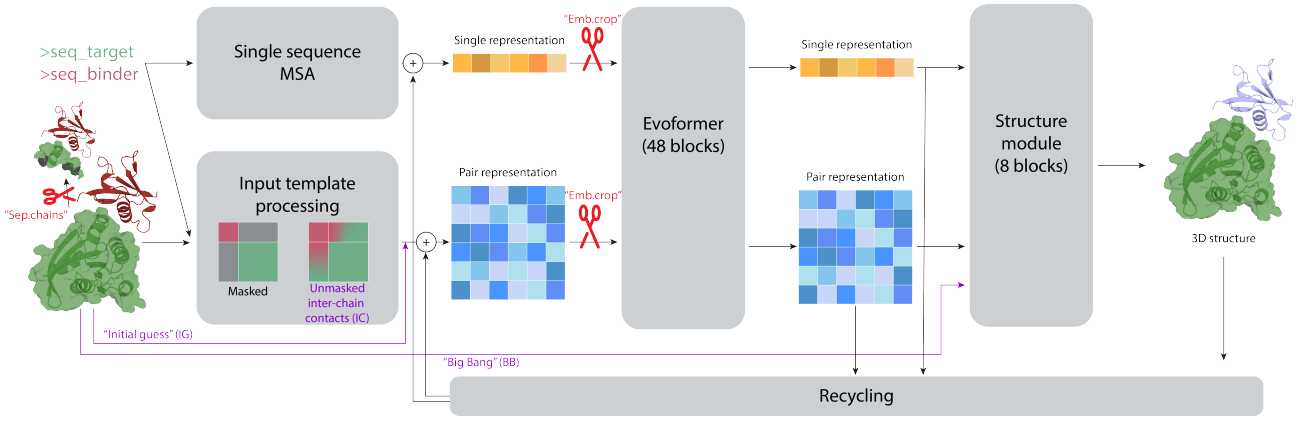

Figure 1. AlphaFold-Multimer architecture (Jumper et al., 2021; Evans et al., 2022), with implemented modifications in *MotifCraft* to the inputs (purple) and target cropping (red).

tion within the specified range (in the user-specified JSON file). The motif placement in the scaffold sequence is reselected at the beginning of each trajectory. These are passed through the AlphaFold-Multimer (AFM) network to obtain a 3D structure, which is used to compute the losses. The loss function is composed of several terms, with specific motif loss terms being added in *MotifCraft*, compared to the BindCraft and ColabDesign implementations (see Appendix B for full list). The position-specific errors are backpropagated through the AF2 network resulting in an L×20 gradient matrix (L=sequence length). This gradient is iteratively recomputed across multiple rounds of stochastic gradient descent, driving progressive refinement of the input binder sequence toward a minimized loss. As in BindCraft, to promote sequence robustness and prevent overfitting, one of the five AF2 models is selected at random at each iteration. The scaffold design is made up of four stages, starting from continuous logits to a discrete sequence (Pacesa et al., 2025). To improve the stability and solubility of the final design, eight sequences are generated with SolubleMPNN (Goverde et al., 2024), giving as input the predicted structure of the final iteration and fixing the motif residues. These sequences are then re-predicted using the two template-based models of AF2-monomer, with three recycles and in single sequence mode without templates.

We benchmark *MotifCraft*'s motif scaffolding capabilities using the RFdiffusion dataset (Watson et al., 2023) of 24 motifs. We generate 100 trajectories for each motif and consider a successful scaffold one that passes the in silico thresholds of predicted local distance difference test (pLDDT) ≥ 0.7, self-consistency root mean square deviation (scRMSD, calculated by comparing the Cα coordinates of the final trajectory step predicted by AFM and the validation: the SolubleMPNN sequence predicted with ESMFold (Lin et al., 2023), after structure alignment) ≤ 2Å and motifRMSD (calculated by comparing the backbone coordinates of the input motif and the corresponding atoms of the AF2-monomer-predicted validation, after structure alignment) ≤ 1Å. After obtaining all successful scaffolds for a motif, the final unique solutions are obtained from the cluster representatives using Foldseek's (van Kempen et al., 2024) easy-cluster method, as used in MotifBench (Zheng et al., 2025).

After adjusting the weights of the losses (tests were conducted on a single example task from the independent MotifBench dataset: PDB 1LDB), the method can generate unique scaffolds that pass the in silico scores by AF2-monomer (used for validation during development) validation (Supplementary Table C.2, MotifCraft_V1_AF2mono). Upon closer inspection of the more problematic motifs, we observed that the generated scaffolds had more trouble maintaining a low motifRMSD than the scRMSD (Supplementary Figure A.1). In other words, iterations led to proteins with high refolding success, but did not contain the correct motif structure. In order to counteract this known tendency to create over-stable designs, we implemented a two-step loss function weighting, emphasizing the motif structure conservation (motifRMSD) more in the first stage of design, and then allowing the scaffold to gain confidence in the later stages by lowering the motif losses (details in Appendix B).

The final version of *MotifCraft* shows a high success rate in generating distinct passing scaffolds, measured by the number of targets with at least one unique passing solution (using ESMFold as the structure prediction method) found in 100 trajectories (Figure 2A, Supplementary Tables C.1 and C.2), with improvements over both diffusion-based design methods (RFdiffusion (Watson et al., 2023), Proteina (Geffner et al., 2025), Genie2 (Lin et al., 2024), FrameFlow (Yim et al., 2023)), pLM-based (DayHoff Atlas (Yang et al., 2025)), as well as the single-step motif losses (MotifCraft_V1_AF2mono). Despite ESMFold typically reported to have higher success rates in motif scaffolding problems than AlphaFold2 (Zheng et al., 2025), in our case *MotifCraft* designs had higher success rates when using the

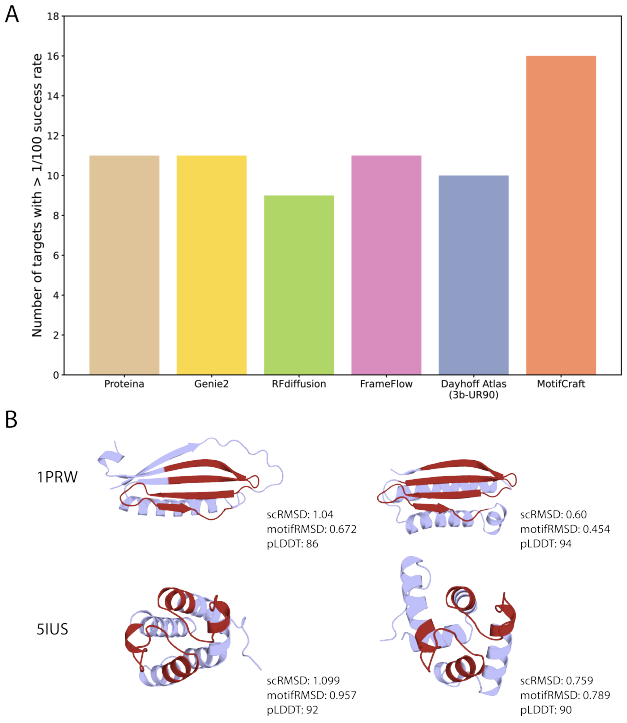

Figure 2. **A)** RFdiffusion motif scaffolding benchmark results. Number of motif scaffolding tasks with unique successful solutions found in $> 1/100$ trajectories (details in Appendix C.1). **B)** Examples of designs for two tasks from the RFdiffusion benchmark, PDB 1PRW and PDB 5IUS, for which *MotifCraft* found an in silico successful scaffold in 100 trajectories for the input motif (red), whereas benchmark methods did not.

AlphaFold2 structure prediction method compared to ESMFold (Supplementary Table C.2). Although our run time per trajectory is around 15 times longer than RFdiffusion for a protein 100 amino acids long, *MotifCraft* shows superior success rates in the same run time in 6 (selected examples in Figure 2B, Supplementary Table C.1) out of the 15 of the hardest tasks (defined as cases where the majority of the compared methods obtain $\leq 1/100$ unique successful trajectories). Thus, our proposed method generalizes to a diverse set of motif scaffolding problems, including active site, ion binding site, protein-binding interface, and RNA-binding interface motif scaffolding.

3. Motif scaffolding of binder-target interfaces

For the goal of scaffolding a motif that is part of a binder-target interface, we investigate the usefulness of the presence of the target protein in the design process. The optimal scaffold would not only pass the in silico motif scaffolding criteria of maintaining the motifRMSD and scRMSD, but also the established in silico protein binder interface metrics. These interface-specific scores include i_pTM, i_pAE and ipSAE (R. L. Dunbrack, 2025), which have been shown to correlate with the experimental success of designed binders (Pacesa et al., 2025; Overath et al., 2025; Cotet et al., 2025).

Our first observation after generating the initial designs is that validation of the protein structure output with AlphaFold2-monomer, across various loss parameters and design inputs, consistently places both proteins far from each other in space, a previously observed phenomenon when using this model with just the input template (Bennett et al., 2023). Therefore, we first analyze the validation itself. Running AF2-monomer with initial guess (IG) (Bennett et al., 2023) or "big bang" (BB) structure module initialization (Schweke et al., 2024) alone did not resolve the problem (Supplementary Figure A.2A,B). However, unmasking the inter-chain contacts from the template (IC (Mirabello et al., 2024), Figure 1), both with initial guess (or big bang) and without, places the two proteins in the intended proximity, indicated by a low RMSD to the original input motif structure (after aligning the target chains). ESMFold (Lin et al., 2023) does not correctly model the target protein itself, therefore also leading to high RMSDs (Supplementary Figure A.2A).

We benchmark the use of AF2-monomer (IC) with the experimental dataset of the Adaptyv Bio's Nipah binder design competition (Proteinbase, 2025), divided into "Not expressed", "Binding" and "No binding" categories. Given the Boltz-1 (Wohlwend et al., 2024) predicted structures as input templates, AF2-monomer validation is run in two modes, with and without IC, both with gapped-sequence (Roney & Ovchinnikov, 2022) template. The RMSDs in the analyses are computed by comparing the first 10 binder protein interface residues (to simulate a binder interface motif) of the re-predicted structure and the same residues in the input structure, after aligning by the target protein. We observe that, as in the case of the motif scaffolding examples, AF2-monomer is unable to place the Nipah binder designs in the same position as the input without IC (Supplementary Figure A.3A, RMSD compared to RMSD (IC)). Unsurprisingly, the confidence scores are also higher in the IC case. However, from the area under the ROC curve (AUC, Figure 4A, Supplementary Figure A.3B), as well as the area under the precision-recall curve (AP, Figure 4B), we can see that the IC-based scores are consistently more discriminative than the non-IC counterparts, although the discriminative power remains quite low. Since the primary objective of the validation and filtering step is obtaining a high precision, while allowing for low specificity, we compute the precision in the top 10% of ranked designs (by the respective score), and we can see that ipSAE, as well as the product of ipSAE and iLDDT (computed using OpenStructure v2.11.1 (Studer et al., 2026) with the `compare-structures` action) of the IC models result in the highest precision values (26% in the ranked top 10%) (Supplementary Figure A.3C). Therefore, we use AF2-monomer (IC) for the rest of the in silico validation analyses.

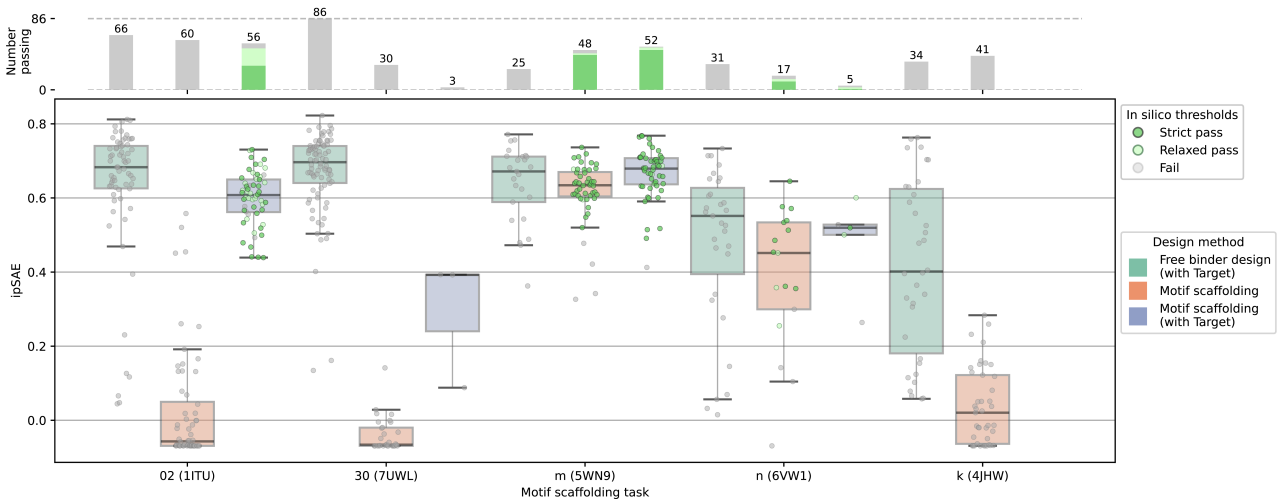

Figure 3. AF2-monomer-rescored (IC) trajectories of five design targets, with three different design methods: free binder design (green), motif scaffolding without the target (orange), and motif scaffolding with the target (purple). The y axis shows the ipSAE (R. L. Dunbrack, 2025) score for each trajectory sequence and the points are colored dark green if they pass the strict motifRMSD threshold of 1Å, light green if they pass the relaxed motifRMSD threshold of 2Å, or gray if they do not pass the thresholds. The number of passing trajectories is indicated as bar charts above the plot. For the free binder design (with target), which is not constrained by a motif, the RMSD is computed as the minimal distance to the input hotspots, thus the thresholds are set higher (strict: 7Å, relaxed: 10Å).

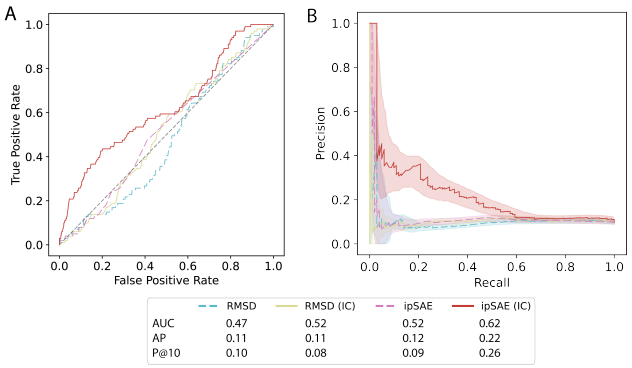

Figure 4. **A)** ROC curves and area under the ROC curves (AUC) of the RMSD and ipSAE scores for the binding and non-binding (non-expressed proteins were omitted) experimentally tested submissions to Adaptyv Bio's Nipah binder design competition (Proteinbase, 2025), rescored using AF2-monomer (IC). **B)** Precision-recall (PR) curves, area under the PR curves (AP) and precision in the top 10% of binder scores (P@10) for the RMSD and ipSAE scores of the same dataset.

We show five examples from the RFdiffusion (Watson et al., 2023) and MotifBench (Zheng et al., 2025) motif scaffolding benchmark datasets: MotifBench dataset: 02 (PDB 1ITU) and 30 (PDB 7UWL), RFdiffusion dataset: m (PDB 5WN9), n (PDB 6VW1) and k (PDB 4JHW). These tasks are selected because their original PDB entries contain a target protein interacting at the motif site, and they had few or no successful scaffolds when designed without the target. For these five examples, we run the optimized motif scaffolding workflow for 100 trajectories, (1) giving only the motif coordinates from the benchmark datasets as input and (2) giving both the motif coordinates and the corresponding target protein from the same PDB entry (in this workflow, the interface-specific losses 7+8 from Appendix B were additionally included). For k (4JHW) and m (5WN9) from the RFdiffusion dataset, the antibody target is cropped to the variable regions to reduce run time, creating a single-chain variable fragment as the target protein.

We observe that including the target protein with interface-specific losses often rescued these cases and allowed obtaining successful binders containing the motif of interest. In some examples we observe a decrease in the number of designs reaching the final stage of iterative design (number of points in Figure 3), failing due to low confidence in the initial stages. In the case of k_4JHW, no trajectory passed the final stage thresholds, but we suppose that with more extensive sampling, passing scaffolds could be found also for this motif.

Secondly, the presence of the target protein improved or maintained the correct placement of the design upon refolding and aligning on the target protein chains (Figure 3, fraction of green (motifRMSD $\leq$ 1 or 2Å) / gray (motifRMSD > 2Å) points) and improved or maintained the interface-specific ipSAE scores (Figure 3) (R. L. Dunbrack, 2025). The other common interface-specific scores (i_pAE (Supplementary Figure A.2C), i_pTM (Supplementary Figure A.2D)) show the same trend of improvement when designing with the target protein. For some motif-target pairs, the confidence scores were significantly worse than in the Nipah binder design competition, suggesting that the in silico scoring thresholds might have to be adjusted on a case-by-case basis. In conclusion, more consistent high interface

confidences can be found in the case that the scaffold has been designed with the intended target.

4. Protein design with AF2-hallucination can be accelerated through target truncation

One main limitation of iterative generation of *de novo* designs through AlphaFold2 backpropagation is the high computational resources required. The run time increases cubically with the protein sequence length L due to the triangle attention in the Evoformer blocks (Jumper et al., 2021). Therefore, a drastic reduction in run time for each trajectory can be achieved by cropping the target down to the essential binding site, which would then serve as input to design a *de novo* protein binder or to scaffold a protein-binder interface motif. We implement two strategies to explore this: (1) crop the target before the whole workflow and input the now discontinuous target interface as several separate chains (Sep.chains), or (2) crop the target after the AlphaFold2 embedding of the input data (Emb.crop), such that the heavy pass through the Evoformer blocks is reduced to the interface residues, but the embeddings contain information from the full protein context (Figure 1, red scissors). In both cases, the SolubleMPNN and validation steps are run using the full target protein.

The default number of one recycle in the initial stages works without modifications for the Sep.chains mode. However, in the Emb.crop mode, recycling requires the single and pair representations obtained before the Structure Module, as well as the predicted output coordinates. Since the dimensions of these tensors are not equivalent to the cropped representations they are merged into, we explore two possible solutions. First, recycling is omitted during the design process (0 recycles). Alternatively, after the first pass, the inputs are cropped to the same dimension (same input therefore as Sep.chains) and then merged with the outputs of the first pass. This case of Emb.crop 1recycle therefore sees the full protein in the first pass through the AF2 network and then refines the structure with the cropped target in the second pass. All subsequent iterations start with the full protein and updated sequence from the backpropagation. A trajectory is considered successful if it passes all in silico thresholds after AF2-monomer validation, which are equivalent to the BindCraft default thresholds (Pacesa et al., 2025), except without the filters for Rosetta interface scores and hydrogen bonds at the interface.

Our first application is the target protein PBP2a (PDB 1VQQ), a protein of 494 amino acids known to confer resistance to methicillin-resistant staphylococci (MRSA) (Chambers, 1999). PBP2a's transpeptidase activity surface can be drastically cropped to 270 amino acids, to which an inhibitor of PBP2a could be designed. 25 trajectories (binder length: 85-95 amino acids) are run for each combination of the two cropping methods and two AlphaFold2 options: initial guess (IG) and big bang (BB) (Supplementary Figure A.4). From this example we conclude that both cropping strategies are up to an order of magnitude faster and produce the same or more successful trajectories. The embedding-cropping (Emb.crop) mode with 1 recycle, initial guess (IG) and big bang (BB) shows an advantage in terms of number of passing designs, however, the equivalent separate chains (Sep.chains) input mode is faster, due to more optimized GPU usage and less overhead. In general, 1 recycle is necessary to obtain good trajectories (a higher proportion of trajectories fail after 50 steps with 0 recycles), and the speedup of performing 0 recycles is not significant (Supplementary Figure A.4).

These conclusions can also be observed in two other test cases (same binder length of 85-95 amino acids), Derf7 (PDB 9HAE), the largest protein target in the BindCraft dataset (Pacesa et al., 2025), and the Nipah Glycoprotein G (PDB 2VSM) from Adaptyv Bio's protein design competition (Proteinbase, 2025), with a varying degree of success (Figure 5). Derf7, already being a much smaller protein with 190 amino acids, still benefits from cropping in terms of run time, though the full target input does just as well in terms of trajectory successes, whereas the Emb.crop mode fails to produce successful trajectories. The Nipah protein, being cropped from 416 amino acids to 166 amino acids around the binding site to the human cell receptor ephrinB2, also shows a considerable decrease in run time. For this target, the Sep.chains method showed better success rates and faster run times per step. We conclude that the optimal target cropping method is task-dependent, but in any case is very successful at rendering binder design using AF2-hallucination much more time-efficient.

5. Discussion

Our results demonstrate that *MotifCraft*, an AlphaFold-Multimer backpropagation-based hallucination workflow, can be successfully applied to the protein design task of motif scaffolding. The high flexibility of the design process through adaptations of the loss function allows us to guide the designs in the desired direction, resulting in protein scaffolds that upon validation with an independent model (AlphaFold2-monomer), are confident, and keep the motif structure intact. *MotifCraft* has a high in silico success rate in the RFdiffusion benchmark dataset, notably generating successful scaffolds for motifs that are considered difficult test cases.

We show that to obtain *de novo* designed scaffolds containing protein-binding motifs, including the target protein in the design process can lead to more in silico successful designs. An important observation is the failure of the commonly used model for protein design scoring, AlphaFold2-

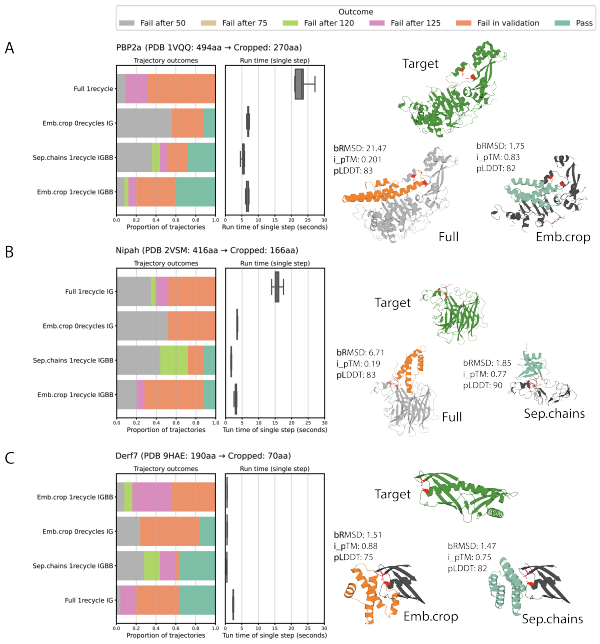

Figure 5. Results of the different cropping methods on three example targets of varying lengths. Cropping methods: Full input protein (with initial guess (IG) initialization), Sep.chains cropping (with initial guess (IG) and big bang (BB) initializations) and Emb.crop (with 0 or 1 recycles, and initial guess (IG) and big bang (BB) initializations), illustrations for each modification can be found in Figure 1. Each method is run for 25 trajectories with a binder sequence length of 85-95 amino acids. Target examples: **A)** PBP2a (PDB 1VQQ, 494 amino acids → 270 amino acids), **B)** Nipah Glycoprotein G (PDB 2VSM, 416 amino acids → 166 amino acids) and **C)** Derf7 (PDB 9HAE, 190 amino acids → 70 amino acids). Bar charts on the left correspond to the proportion of trajectories that fail, colored by the iteration at which it fails, or pass all filters (light green). On the right, the box plots of the average run times of a single step in each trajectory are plotted. Some examples of protein structures from the final trajectory step are shown on the right of the plots, with the PDB target in dark green, the designed binders in light green (passing all filters) and orange (failing at the validation step), the full protein targets in light gray and the cropped protein target interfaces in dark gray, with the defined hotspots in red sticks. The metrics of the structure re-predicted with AF2-monomer (default mode) are shown next to the binder: binder Cα RMSD after alignment on the target protein (bRMSD, ="Hotspot RMSD" in BindCraft), i_pTM and pLDDT (of the binder protein).

monomer in single-sequence mode with template - to distinguish between binders and non-binders. We partially compensate for these shortcomings by using unmasked interchain contacts (IC) AF2-monomer in single sequence mode with the protein binder and target template to re-predict and filter our designs. However, there remains ample room for improvement in correctly distinguishing protein binders.

Due to repeated passes through AlphaFold2, a limitation of motif scaffolding with this method is its relatively slow computational speed. Our results show that cropping the target protein down to the essential binding surface is a simple, yet effective method to reduce run time. Our designs exhibit good in silico scores similar to validated *de novo* binder design results, indicating a strong likelihood of solubility and binding. There are several promising directions for future development, including further optimized and specialized losses, as well as improved scoring through new co-folding methods.

Ultimately, we show with *MotifCraft* that the AF2 backpropagation framework can be used for motif scaffolding, as well as the benefits of including the target protein for scaffolding protein-protein interface motifs. The inclusion of a target protein results in longer run times, for which we propose an efficient speed-up using target truncation. Our analyses thus demonstrate the advantages of having such a versatile platform for *de novo* protein design. In the future, these designs will also need to be validated experimentally to verify successful motif conservation and binding. Overall, the ability to guide designed proteins to specific characteristics represents a generalizable approach that takes the field one step closer to the design of functional protein binders, enzymes, and DNA-binding proteins.

Data availability

The data and source code supporting the findings of this study will be made available in a public repository upon acceptance.

Impact Statement

This paper presents work whose goal is to advance the field of computational protein design. There are many potential societal consequences of our work, including addressing bottlenecks and missing solutions in healthcare, agriculture and ecology.

References

Albanese, K. I., Barbe, S., Tagami, S., Woolfson, D. N., and Schiex, T. Computational protein design. *Nature Reviews Methods Primers*, 5(1), 2025.

Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y. P., Dauparas, J., Baek, M., Stewart, L., DiMaio, F., Munck, S. D., Savvides, S. N., and Baker, D. Improving de novo protein binder design with deep learning. *Nat Commun.*, 614(2625), 2023.

Chambers, H. F. Penicillin-binding protein-mediated resistance in pneumococci and staphylococci. *The Journal of Infectious Diseases*, 179(Supplement$_2$) : $S353 - -S359$, 1999.

Cotet, T.-S., Krawczuk, I., Stocco, F., Ferruz, N., Gitter, A.,

Kurumida, Y., de Almeida Machado, L., Paesani, F., Calia, C. N., Challacombe, C. A., Haas, N., Qamar, A., Correia, B. E., Pacesa, M., Nickel, L., Subr, K., Castorina, L. V., Campbell, M. J., Ferragu, C., Kidger, P., Hallee, L., Wood, C. W., Stam, M. J., Kluonis, T., Ünal, S. M., Belot, E., Naka, A., and Organizers, A. C. Crowdsourced protein design: Lessons from the adaptyv egfr binder competition. *bioRxiv*, 2025.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning-based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

Dauparas, J., Lee, G. R., Pecoraro, R., An, L., Anishchenko, I., Glasscock, C., and Baker, D. Atomic context-conditioned protein sequence design using ligandmpnn. *Nat Methods*, 22(4):717–723, 2025.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022.

Geffner, T., Didi, K., Zhang, Z., Reidenbach, D., Cao, Z., Yim, J., Geiger, M., Dallago, C., Kucukbenli, E., Vahdat, A., and Kreis, K. Proteina: Scaling flow-based protein structure generative models. *arXiv*, 2025.

Goverde, C. A., Pacesa, M., Goldbach, N., Dornfeld, L. J., Balbi, P. E. M., Georgeon, S., Rosset, S., Kapoor, S., Choudhury, J., Dauparas, J., Schellhaas, C., Kozlov, S., Baker, D., Ovchinnikov, S., Vecchio, A. J., and Correia, B. E. Computational design of soluble and functional membrane protein analogues. *Nature*, 631(8020):449–458, 2024.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583—589, 2021.

Khakzad, H., Igashov, I., Schneuing, A., Goverde, C., Bronstein, M., and Correia, B. A new age in protein design empowered by deep learning. *Cell Syst.*, 14(1):925–939, 2023.

Lin, Y., Lee, M., Zhang, Z., and AlQuraishi, M. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2. *arXiv*, 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Mirabello, C., Wallner, B., Nystedt, B., Azinas, S., and Carroni, M. Unmasking alphafold to integrate experiments and predictions in multimeric complexes. *Nat Commun.*, 15 (8724), 2024.

Overath, M. D., Rygaard, A. S. H., Jacobsen, C. P., Brasas, V., Morell, O., Sormanni, P., and Jenkins, T. P. Predicting experimental success in de novo binder design: A meta-analysis of 3,766 experimentally characterised binders. *bioRxiv*, 2025.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., Cho, Y., Ghamary, K. H., Vinué, L., Yachnin, B. J., Wollacott, A. M., Buckley, S., Westphal, A. H., Lindhoud, S., Georgeon, S., Goverde, C. A., Hatzopoulos, G. N., Gönczy, P., Muller, Y. D., Schwank, G., Swarts, D. C., Vecchio, A. J., Schneider, B. L., Ovchinnikov, S., and Correia, B. E. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

Proteinbase. Nipah competition results, 2025. URL https://proteinbase.com/collections/nipah-binder-competition-results. Accessed: 2026-04-29.

R. L. Dunbrack, J. Rēs ipsae loquuntur: What's wrong with alphafold's iptm score and how to fix it. *bioRxiv*, 2025.

Roney, J. P. and Ovchinnikov, S. State-of-the-art estimation of protein model accuracy using alphafold. *Phys. Rev. Lett.*, 129(23):238101, 2022.

Schweke, H., Pacesa, M., Levin, T., Goverde, C. A., Kumar, P., Duhoo, Y., Dornfeld, L. J., Dubreuil, B., Georgeon, S., Ovchinnikov, S., Woolfson, D. N., Correia, B. E., Dey, S., and Levy, E. D. An atlas of protein homo-oligomerization across domains of life. *Cell*, 187(4):999–1010.e15, 2024.

Studer, G., Robin, X., Bienert, S., Durairaj, J., Skrinjar, P., Tauriello, G., Waterhouse, A. M., and Schwede, T. A fully automated benchmarking suite to compare macromolecular complexes. *Nature Methods*, 23(2):387–394, 2026.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nat Biotechnol*, 42(2):243–246, 2024.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., Bortoli, V. D., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv*, 2024.

Yang, K. K., Alamdari, S., Lee, A. J., Kaymak-Loveless, K., Char, S., Brixi, G., Domingo-Enrich, C., Wang, C., Lyu, S., Fusi, N., Tenenholtz, N., and Amini, A. P. The dayhoff atlas: scaling sequence diversity for improved protein generation. *bioRxiv*, 2025.

Yim, J., Campbell, A., Foong, A. Y. K., Gastegger, M., Jiménez-Luna, J., Lewis, S., Satorras, V. G., Veeling, B. S., Barzilay, R., Jaakkola, T., and Noé, F. Fast protein backbone generation with se(3) flow matching. *arXiv*, 2023.

Zheng, Y., Zhang, B., Didi, K., Yang, K. K., Yim, J., Watson, J. L., Chen, H., and Trippe, B. L. Motifbench: A standardized protein design benchmark for motif-scaffolding problems. *arXiv*, 2025.

A. Supplementary Figures

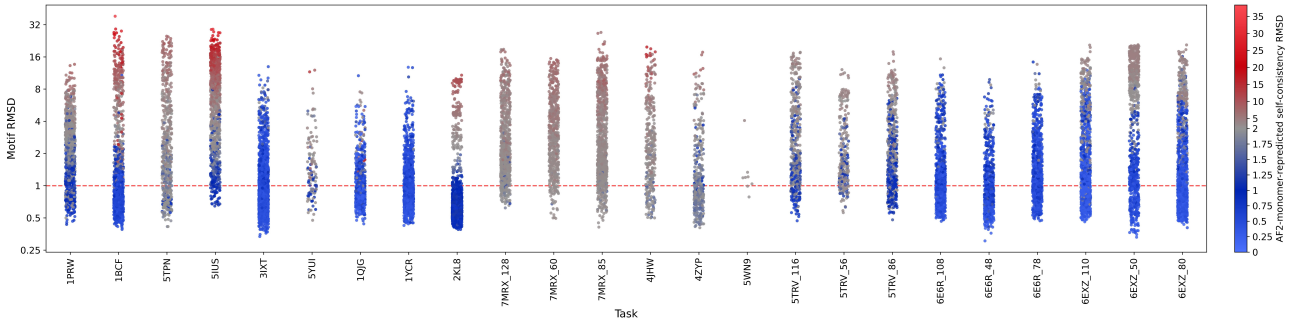

Figure A.1. RFdiffusion motif scaffolding benchmark dataset. Individual trajectory sequences (outputs of SolubleMPNN) are re-predicted using AF2-monomer in single sequence mode without templates and plotted on the y-axis at the Cα motifRMSD to the input motif coordinates and colored by Cα self-consistency RMSD (scRMSD) of the full scaffold to the final trajectory step structure.

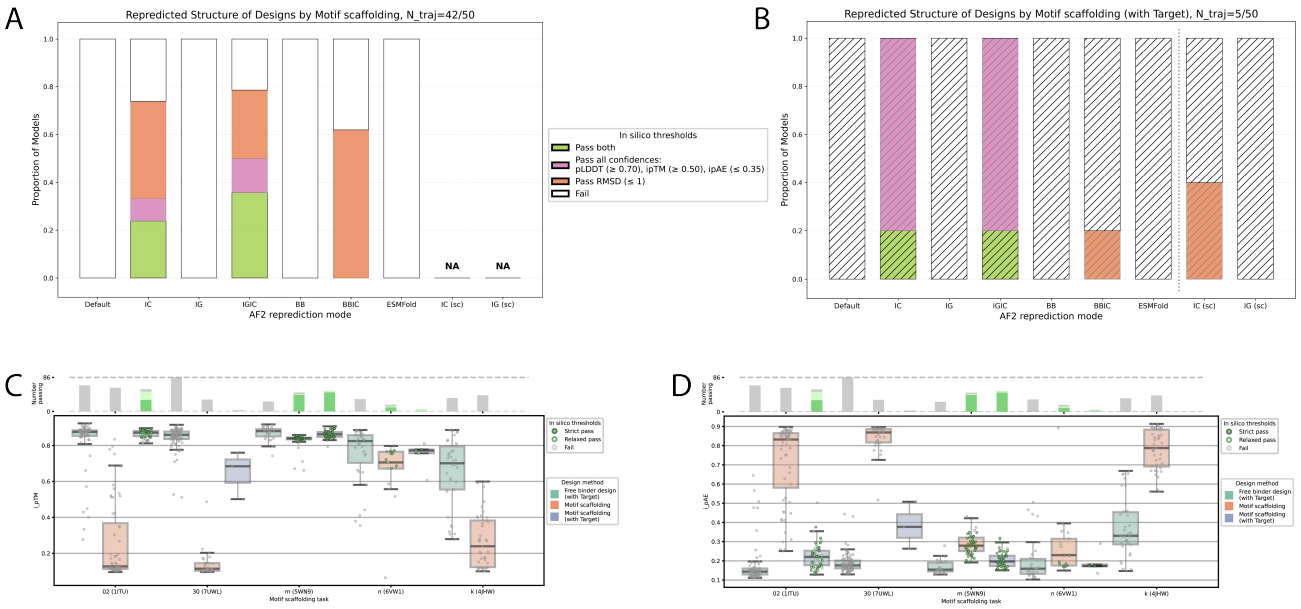

Figure A.2. Results of in silico validation metrics of the test case n (PDB 6VW1) from the RFdiffusion benchmark dataset. The motif scaffolding method is *MotifCraft* either without target (**A**, solid) or design in the presence of the target protein (**B**, hatched). For each trajectory reaching the final step (=**N**, out of 50 trajectories run), 20 MPNN sequences were generated and the top-ranked sequence was rescored using AF2-monomer, with the different modes on the x axis. The y axis values correspond to the proportion of sequences passing either the motifRMSD threshold of 1Å to the input (orange), the in silico confidence score thresholds of pLDDT ≥ 0.7, i_pTM ≥ 0.5 and i_pAE ≤ 0.35 (pink), or both (green). Comparison of the re-predicted Cα motifRMSD to the final trajectory structure, after aligning on the target protein, is plotted on the two right-most columns. The designs run without target protein do not have this metric, since the structure of the final step does not contain the target to align to. **C,D)** AF2-monomer-rescored (IC) trajectories of five design targets, with three different design methods: free binder design (green), motif scaffolding without the target (orange) and motif scaffolding with the target (purple). The y axis shows the i_pAE (C) and i_pTM (D) score, respectively, for each trajectory sequence and the points are colored dark green if they pass the strict motifRMSD threshold of 1Å, light green if they pass the relaxed motifRMSD threshold of 2Å, or gray if they do not pass the thresholds. The fraction of passing trajectories is indicated as bar charts above the plot. For the free binder design (with target), which is not constrained by a motif, the RMSD is computed as the minimal distance to the input target hotspots, thus the thresholds are set higher (strict: 7Å, relaxed: 10Å).

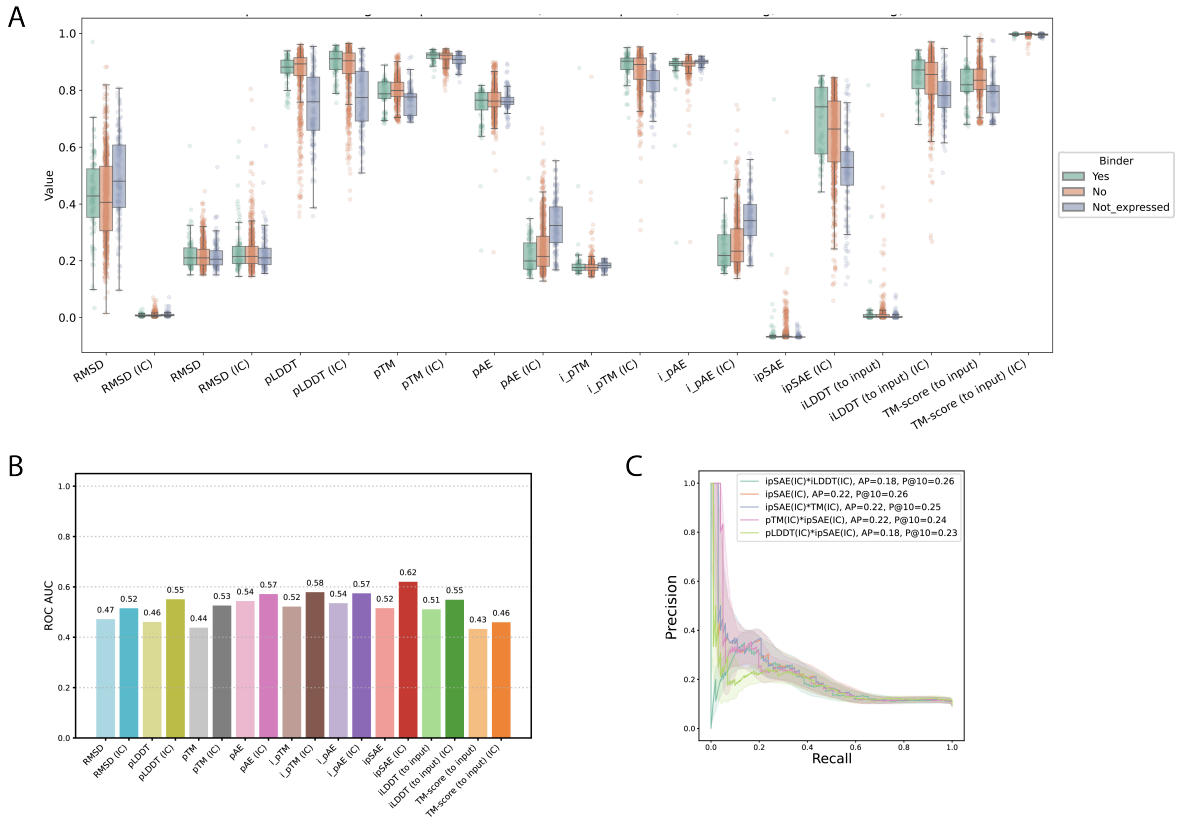

Figure A.3. Results of in silico scoring metrics of the Nipah protein design competition binding and non-binding complexes (Proteinbase, 2025), predicted using default AF2 parameters, versus with IC, both with the Boltz-1 (Wohlwend et al., 2024) predicted structures as templates. **A)** Absolute values of the individual scores, separated by experimental success (binding, non-binding, and non-expressed). **B)** Area under the ROC curve scores (AUC) of the individual scores for the binding and non-binding (non-expressed proteins were omitted) experimentally tested submissions. **C)** For the same binder dataset, precision-recall (PR) curves, area under the PR curves (AP) and precision in the top 10% (P@10) of the top-ranked scores (or product of scores), ranked by the precision in the top 10%.

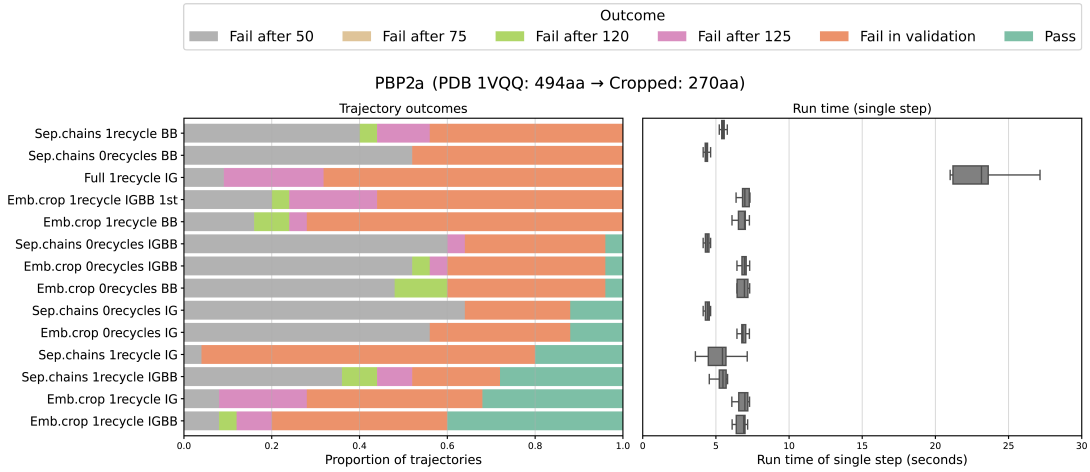

Figure A.4. Results of different methods for accelerated protein design through target cropping, evaluated on the target protein PBP2a (PDB 1VQQ, 494 amino acids → 270 amino acids). Different AlphaFold2 initialization options, initial guess (IG) and big bang (BB), were tested alone and in combination with each other. Each method is run for 25 trajectories with a binder sequence length of 85-95 amino acids. Bar charts on the left correspond to the proportion of trajectories that fail, colored by the iteration at which it fails, or pass all filters (green). On the right, the box plots of the average run times of a single step in each trajectory are plotted.

B. Motif scaffolding loss composition

The final optimized motif scaffolding loss is composed of the following terms, with the italic points only applying to the case where the target is also given as input (1-3: full binder protein losses, 4-6: motif losses, *7-8: interface losses*, 9-10: other full binder protein losses). The default BindCraft and *MotifCraft* (for both the single-step and two-step models, if only one value is indicated, then it is valid for both models) weights are indicated in brackets.

1. full protein confidence pLDDT (BindCraft default=0.1, MotifCraft weight=0.1)
2. normalized predicted alignment error (pAE) within the full protein (BindCraft default=0.4, MotifCraft weight=0.4)
3. residue contact loss within the full protein (BindCraft default=1.0, MotifCraft weight=0.5)
4. RMSD of motif Cα atoms to input structure (NA in BindCraft, MotifCraft default one-step design weight=5.0, two-step design: weight=6.0 in the first stage, then 4.0)
5. weighted geometric mean of the motif residues pAE (4.0, weight_mean=1.0) and RMSD (5.0, smoothened with sigmoid scale 3.0, weight_mean=2.0) losses (NA in BindCraft, MotifCraft default one-step design weight=15.0, two-step design: weight=20.0 in the first stage, then 10.0)
6. residue contact loss between the motif and scaffold (NA in BindCraft, MotifCraft weight=0.5)
7. *normalized predicted alignment error (pAE) between the full binder protein and target (BindCraft default=0.1, MotifCraft weight=0.1)*
8. *interface confidence i_pTM (BindCraft default=0.05, MotifCraft weight=0.05)*
9. radius of gyration of the full protein (BindCraft default=0.3, MotifCraft weight=0.1)
10. 'helicity loss': penalize or promote backbone contacts every one in a three-residue offset to promote the hallucination of helical or non-helical designs (BindCraft default=-0.3, MotifCraft weight=-0.3)

C. Motif scaffolding RFdiffusion benchmark results

ID	Task	Proteina	Genie2	RFdiffusion	FrameFlow	Dayhoff Atlas (3b-UR90)	MotifCraft
p	6E6R_long	71.3	41.5	38.1	11.0	3.0	42.0
p	6E6R_medium	41.7	27.2	15.1	9.9	4.0	27.0
p*	6E6R_short	5.6	2.6	2.3	2.5	14.0	20.0
q	6EXZ_long	29.0	32.6	16.7	40.3	0.0	31.0
q	6EXZ_med	4.3	5.4	2.5	11.0	0.0	39.0
q	6EXZ_short	0.3	0.2	0.1	0.3	0.0	29.0
o	5TRV_long	17.9	9.7	2.3	7.7	0.0	10.0
o	5TRV_med	2.2	2.3	1.0	2.1	0.0	9.0
o	5TRV_short	0.1	0.3	0.1	0.1	0.0	0.0
j	7MRX_long	5.1	2.7	6.6	3.5	42.0	0.0
j	7MRX_med	3.1	2.3	1.3	2.2	19.0	0.0
j	7MRX_short	0.2	0.5	0.1	0.1	0.0	0.0
h*	1YCR	24.9	13.4	0.7	14.9	2.0	52.0
e*	3IXT	0.8	1.4	0.3	0.8	18.0	26.0
c	5TPN	0.4	0.8	0.5	0.6	0.0	3.0
g*	1QJG	0.3	0.5	0.1	1.8	0.0	0.0
f	5YUI	0.5	0.3	0.1	0.1	0.0	0.0
l	4ZYP	1.1	0.3	0.6	0.4	0.0	2.0
a*	1PRW	0.1	0.1	0.1	0.1	94.0	14.0
d*	5IUS	0.1	0.1	0.1	0.0	0.0	8.0
b	1BCF	0.1	0.1	0.1	0.1	10.0	4.0
m	5WN9	0.2	0.1	0.0	0.3	0.0	0.0
i	2KL8	0.1	0.1	0.1	0.1	1.0	1.0
k	4JHW	0.0	0.0	0.0	0.0	0.0	0.0

Table C.1. Number of unique successes, per 100 backbones generated and validated using ESMFold structure prediction, for each task in the RFdiffusion motif scaffolding benchmark dataset. Reference method results are obtained from previously published sources and divided by 10 in the cases where the 1000 trajectories were run (Geffner et al., 2025; Yang et al., 2025). Note that the clustering was done slightly differently (hierarchical instead of Foldseek clustering) for Proteina, Genie2, RFdiffusion and FrameFlow. *Tasks with *MotifCraft* showing superior success rate than RFdiffusion in the same run time.

ID	Task	MotifCraft_V1 (AF2-monomer)	MotifCraft (AF2-monomer)	MotifCraft (ESMFold)
p	6E6R_long	43.0	37.0	42.0
p	6E6R_medium	38.0	32.0	27.0
p*	6E6R_short	38.0	30.0	20.0
q	6EXZ_long	29.0	23.0	31.0
q	6EXZ_med	26.0	26.0	39.0
q	6EXZ_short	8.0	13.0	29.0
o	5TRV_long	17.0	11.0	10.0
o	5TRV_med	4.0	11.0	9.0
o	5TRV_short	0.0	1.0	0.0
j	7MRX_long	1.0	0.0	0.0
j	7MRX_med	0.0	0.0	0.0
j	7MRX_short	0.0	0.0	0.0
h*	1YCR	65.0	63.0	52.0
e*	3IXT	34.0	31.0	26.0
c	5TPN	0.0	2.0	3.0
g*	1QJG	15.0	32.0	0.0
f	5YUI	1.0	0.0	0.0
l	4ZYP	0.0	2.0	2.0
a*	1PRW	18.0	12.0	14.0
d*	5IUS	15.0	10.0	8.0
b	1BCF	6.0	4.0	4.0
m	5WN9	0.0	0.0	0.0
i	2KL8	1.0	1.0	1.0
k	4JHW	0.0	0.0	0.0

Table C.2. Number of unique successes, per 100 backbones generated, for *MotifCraft* variants on the RFdiffusion motif scaffolding benchmark dataset. *Tasks with *MotifCraft* showing superior success rate than RFdiffusion in the same run time.