Backprop-based Motif Scaffolding Beats Generative Models

Anisha Parsan[1]
aparsan@mit.edu

Bowen Jing[1]
bjing@mit.edu

Bonnie Berger[1,2]
bab@mit.edu

[1] CSAIL, Massachusetts Institute of Technology
[2] Department of Mathematics, Massachusetts Institute of Technology

Abstract

Designing protein backbones that scaffold functional motifs is a central task in *de novo* protein design. Current state-of-the-art methods are based on generative models, yet they require extensive sampling and filtering of structures, typically resulting in low *in silico* success rates. Here, we revisit backprop-based design and systematically evaluate ColabDesign's protocol for motif scaffolding. Remarkably, we find that this protocol, which we call *MotifCraft*, outperforms all known methods on the standard 24-motif RFDiffusion benchmark, ranking first on 17 motifs, improving the median unique success rate 8.5x from 0.7% to 5.9%, and substantially exceeding recent generative models like Proteína. Detailed error analysis reveals that, for the most difficult motifs, MotifCraft attains an even higher success rate when considering only design trajectories that terminate in structures containing the motif. These results establish backprop-based optimization as the state-of-the-art strategy for motif scaffolding and challenge the prevailing assumption that generative models are necessary for this task.

1 Introduction

Designing a scaffold that preserves a fixed structural motif is one of the core tasks in protein design, with myriad applications from enzyme design to therapeutics. The prevailing paradigm for motif scaffolding involves generative models that take the motif's 3D coordinates as input and sample compatible backbone structures (Geffner et al., 2025; Lin et al., 2024; Watson et al., 2023; Yim et al., 2023), followed by sequence design with inverse folding models like ProteinMPNN (Dauparas et al., 2022). The design is considered successful if the motif is preserved when the sequence is refolded with a structure prediction model. Despite substantial progress from many generative modeling frameworks, *in silico* success rates remain very low, with the leading generative model, Proteína, attaining a >1% unique success rate on only 11 out of 24 benchmarks motifs. Evidently, folding models impose complex constraints that are difficult to satisfy with structure generative models alone.

We reasoned that directly using a structure prediction model in the design process should increase the proportion of designs passing filters defined by such models. Thus, we revisit a simpler paradigm for protein design: backprop-based optimization, in which a structure predictor is treated as differentiable to make iterative updates to an input sequence, as popularized in ColabDesign (Ovchinnikov, 2022). This approach has proven highly effective for *de novo* design of binders and antibodies (Pacesa et al., 2025; Cho et al., 2025; Mille-Fragoso et al., 2025), exceeding success rates for generative models. Although motif scaffolding with similar protocols has also been demonstrated (Frank et al., 2024), the performance on standard benchmarks compared to recent generative models has not been established.

In this brief report, we systematically evaluate the motif scaffolding protocol implemented in ColabDesign (Ovchinnikov, 2022; Frank et al., 2024), where motif residues are fixed and the surrounding

6th Workshop on Machine Learning in Structural Biology (MLSB 2025), San Diego, CA.

backbone sequence is optimized with losses based on geometry and fold confidence. For convenience, we refer to this protocol as *MotifCraft*. Surprisingly, we find that MOTIFCRAFT ranks first on 17 out of 24 benchmark motifs when compared against leading generative models, attaining a $> 1\%$ unique success rate on 17 motifs. Error analysis demonstrates that the most difficult motifs fail primarily by loss of motif-geometry preservation at end of the structure trajectory rather than by failure to preserve the motif during refolding. This suggests that successful trajectories from MOTIFCRAFT produce much higher-quality structures than generative models, whose outputs contain the motif by construction but often fail to preserve it during sequence design and refolding.

2 Methods

ColabDesign's Scaffolding Procedure We use ColabDesign's scaffolding pipeline and initialize their optimization procedure with three of their losses, at equal weight: contact, categorical cross entropy of the distogram and motif RMSD (each weighted $= 1.0$). Let $M_{\text{fix}} \subset \{1, \ldots, L\}$ denote residues corresponding to the motif, and and M_{free} the designable positions. The procedure maintains amino-acid logits $\theta \in R^{L\times 20}$ freeze θ_i for $i \in M_{\text{fix}}$, and update only $\theta_{M_{\text{free}}}$.

- **Contact loss** ($\mathcal{L}_{contact}$): encourages at least one strong contact per position.
- **Distogram loss** ($\mathcal{L}_{dgram}$): categorical cross-entropy (CCE) between the predicted distance map and the motif reference (from PDB).
- **Motif RMSD loss** ($\mathcal{L}_{mRMSD}$): penalizes structural deviation of designed motifs from the target backbone motif.

The total objective is a weighted sum of these terms:

$$\mathcal{J} = \alpha\, \mathcal{L}_{contact} + \beta\, \mathcal{L}_{dgram} + \gamma\, \mathcal{L}_{mRMSD}. \tag{1}$$

Algorithm 1 Gradient-based Scaffold Optimization (ColabDesign's Scaffolding Procedure)

Require: template structure S (motif coords), masks $M_{\text{fix}}, M_{\text{free}}$, length L, steps N

Initialize logits $\theta \in \mathbb{R}^{L\times 20}$; freeze θ_i for all $i \in M_{\text{fix}}$
for $i \leftarrow 1$ **to** N **do**
  $p \leftarrow \mathrm{softmax}(\theta)$
  $(\hat{X}, \hat{D}) \leftarrow \textsc{Predict}(p, S)$
  $\mathcal{L}_{\text{contact}} \leftarrow \textsc{ContactLoss}(\hat{X};\ M_{\text{free}})$
  $\mathcal{L}_{\text{dgram}} \leftarrow \mathrm{CCE}(\hat{D}, D^{\text{ref}}(S);\ M_{\text{fix}} \times M_{\text{fix}})$
  $\mathcal{L}_{\text{mRMSD}} \leftarrow \mathrm{RMSD}(\hat{X}, S;\ M_{\text{fix}})$
  $\mathcal{J} \leftarrow \mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{dgram}} + \mathcal{L}_{\text{mRMSD}}$ ▷ all weights $= 1.0$
  $\theta_{M_{\text{free}}} \leftarrow \theta_{M_{\text{free}}} - \eta \nabla_{\theta_{M_{\text{free}}}} \mathcal{J}$ ▷ no updates on M_{fix}
end for
return final sequence AA_seq where $\text{AA\_seq}_i = \arg\max p_i$ if $i \in M_{\text{free}}$; otherwise residue identity from S

Motif Scaffolding Task Set We evaluate on the functional-site scaffolding benchmark curated by RFdiffusion (Watson et al., 2023), consisting of 25 motif-scaffolding problems, including viral epitopes, enzymatic active sites, and binding interfaces. Subsequent works (e.g. Genie2 and Proteína (Geffner et al., 2025; Lin et al., 2024) report results on a 24-task single-motif subset that excludes 6VW1 due to its multi-chain motif, which some methods do not support. This standardized suite consists of discontinuous vs. contiguous segments, varied residue counts, and diverse functional contexts. Examples of problems and task specifications are provided in Fig. 6.

Evaluation Pipeline For each motif, we sample 1000 specifications of scaffold lengths and layouts following the sampling procedure in Lin et al. (2024). For each specification, we run MOTIFCRAFT to generate an initial sequence and corresponding structure and redesign seven additional sequences with ProteinMPNN, producing 8 sequences per backbone scaffold. We refold all candidates with ESMFold (Lin et al., 2022; Dauparas et al., 2022) and mark a scaffold successful if any of its 8 sequences meet

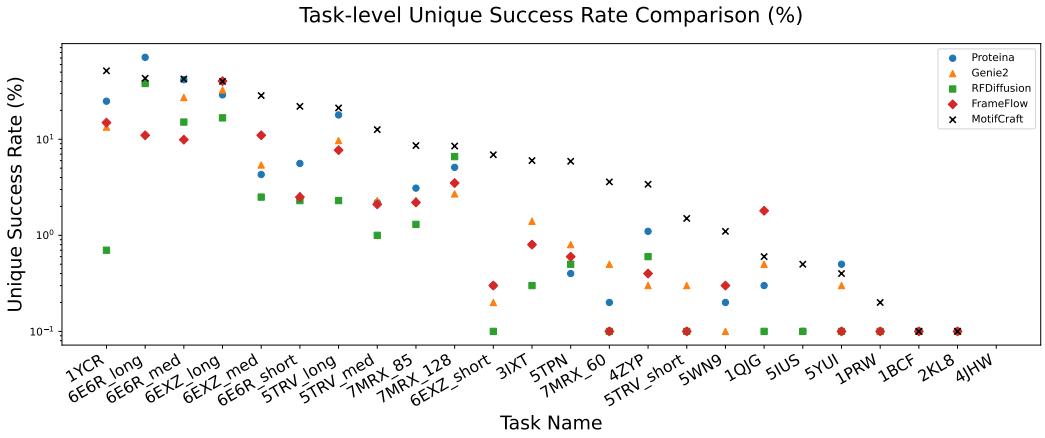

Figure 1: **MOTIFCRAFT performance across tasks.** Unique success rate of 1000 samples for each problem in dataset (after FoldSeek clustering). Baseline (Proteína, Genie2, RFDiffusion, FrameFlow) numbers from Geffner et al. (2025).

all of the following criteria: motif RMSD (backbone) ≤ 1.0 Å, self-consistency RMSD (C_α only) ≤ 2.0 Å, PAE ≤ 5 Å, and pLDDT ≥ 70. Finally, we cluster the successful scaffolds using Foldseek with a TM-score threshold of 0.6 and report the number of clusters as unique successes.

3 Results

Figure 1 shows the unique success rate of MOTIFCRAFT compared to recent generative model baselines on all 24 motifs. MOTIFCRAFT ranks as the top method on 17 of the 24 motifs (Fig. 2), significantly outperforming Proteína (2) and FrameFlow (2). On a head-to-head comparison, MOTIFCRAFT outperforms both Proteína and FrameFlow on 19 motifs, and both Genie2 and RFDiffusion on 21 motifs. In aggregate, the median unique success rate of MOTIFCRAFT is 5.9%, a factor of 8.5 times higher than the best baseline methods, FrameFlow. In particular, this method increases the number of motifs with unique success rates of $> 1\%$ from 11 to 17.

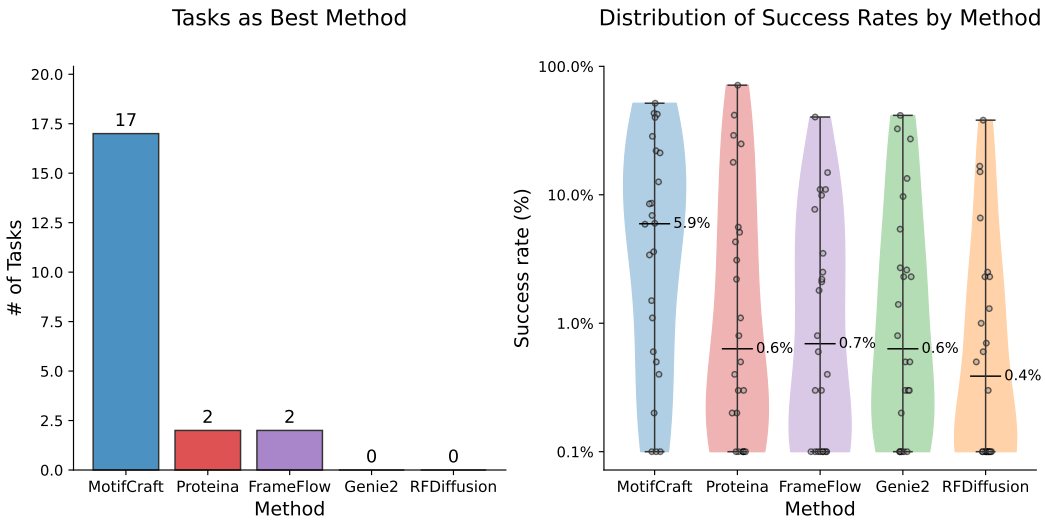

Figure 2: **Task Success Comparisons.** (Left) The number of tasks in which each method has the highest unique success rate. (Right) Distribution of unique success rates for each method. The numerical median values are reported prior to the log-transform.

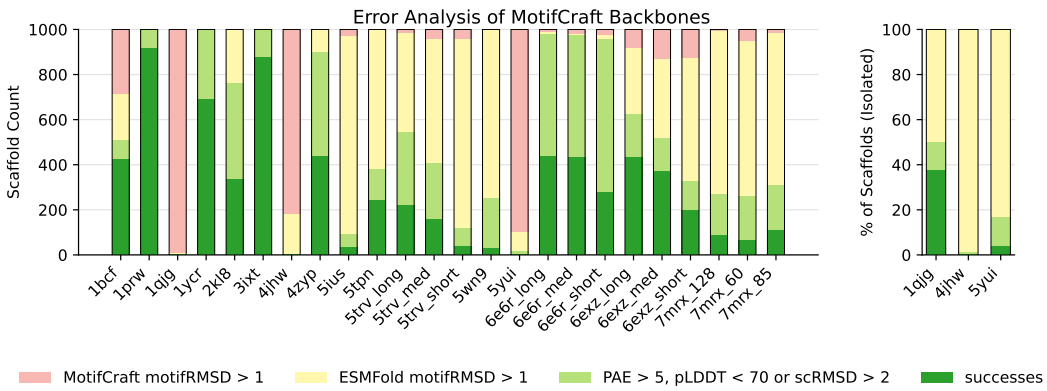

Figure 3: **Breakdown of failure modes via sequential filtering.** At each stage, we filter the subset from the prior stage using the denoted criterion. Counts are prior to FoldSeek clustering.

Error Analysis For each of 24 motifs, we provide a breakdown of the failure or success modes across 1000 scaffolds in Figure 3, left. In particular, we examine failures across chronological stages of the pipeline (before uniqueness clustering) using cumulative gates: MOTIFCRAFT design motifRMSD$>$ 1, then ESMFold refolds where none of the eight designs pass motifRMSD, and finally failure on the remaining confidence or self-consistency metrics. Most motifs exhibit significant overall pre-clustering success rates (dark green), with failures roughly equally distributed between insufficiently confident or self-consistent designs (light green) and disruption of motif geometry upon refolding (yellow). The most difficult motifs (1qjg, 4jhw, 5yui) are characterized by a majority of design trajectories converging to structures that do not contain the target motif (red). Filtering these cases out, a substantial fraction of designs are successful for two out of three motifs (Figure 3, right).

Figure 4 shows more detailed success/failure breakdowns for all 192k sequences across 24 motifs. Overall, the metric pass rates are as follows: motifRMSD < 1 Å (51.7%), scRMSD < 2 Å (64.7%), PAE < 5 Å (42.3%), and pLDDT > 70 (83.1%); success across all thresholds is 28.8%. The most common failure combinations of a given scaffold are: all four criteria failing, the triple {PAE, scRMSD, motifRMSD}, and {PAE} alone. This indicates that while sequence foldability is typically sufficient (pLDDT passes) while long-range confidence and consistency are bottlenecks. We note that in generative models, the pass rate for motifRMSD is 100% by construction, suggesting higher failure rates on the other criteria.

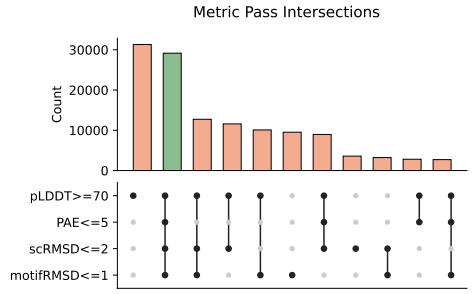

Figure 4: Frequency of sequences passing most common combinations of criteria.

Secondary-structure diversity An analysis of secondary-structure content reveals that MOTIFCRAFT generates a wide variety of structural compositions in successful designs for a given problem. See Fig.7 for per-task secondary-structure analysis (helix, sheet, loop fractions). As a case study, we analyzed designs for 6E6R_short. On average, successful scaffolds contained 54.6% α-helix, 10.6% β-sheet, and 34.8% loop. Individual designs span a wide range: one helix-dominated scaffold reached 85% α, another was enriched for β at 54%, and a third adopted a loop-heavy architecture with 62% coil. This diversity in successful designs indicates that the pipeline explores a variety of secondary-structure realizations while preserving the target motif (see Fig. 5 for representative helix- and sheet-rich examples).

3.1 Ablations

ProteinMPNN Redesign. We compare success rates using only the original sequence versus after generating seven redesigns with ProteinMPNN and refolding all eight candidates with ESMFold. Fig. 8 reports per-problem success before vs. after ProteinMPNN; the table below summarizes average

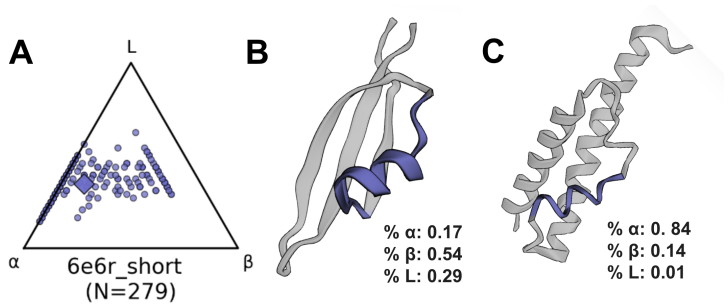

Figure 5: **Secondary Structure Case Study.** (A) Percentage of α helix, β sheet, and loop in `6E6R_short` successful samples. (B) Successful sample with the most β percentage. (C) Successful sample with the greatest α percentage

pass rates and deltas across criteria. Overall, the redesigns improve success rates over the original on a majority of problems (17 out of 24). Notably, ProteinMPNN unlocks solutions on tasks that were previously unsolved with the original sequence (e.g., 5IUS and 1QJG), indicating that sequence-level adjustments can rescue borderline backbones. We observe that ProteinMPNN improves model confidence (pLDDT and PAE) but reduces geometric fidelity (drops in motifRMSD/scRMSD), yielding a modest net gain in end-to-end success. This pattern suggests that while ProteinMPNN sequences are more readily interpretable by the structure predictor, they do not always translate into better geometric alignment with the target motif.

Template Guidance. We assessed the effect of conditioning on a template structure (the motif segment) during optimization. With template guidance (the default setting), we achieve equivalent or better performance on 16 out of 24 problems, but no additional problems are solved overall. On average, template guidance improves motif fidelity but results in reductions to confidence metrics and scRMSD. Despite these differences, overall success is higher with template guidance (considering per-problem improvements), suggesting that structural conditioning enforces motif alignment but can constrain solutions in a way that reduces model confidence.

Table 1: Average pass rates across criteria for ablation studies: before/after PMPNN and with/without template guidance. Overall = fraction of scaffolds meeting all criteria.

Criteria	Original Sequence	After PMPNN	With Template	Without Template
motifRMSD ≤ 1	36.6%	35.3%	35.5%	31.6%
scRMSD ≤ 2	40.2%	35.9%	36.4%	42.7%
PAE ≤ 5	16.6%	23.9%	23.0%	25.0%
pLDDT ≥ 70	44.6%	59.4%	57.5%	62.4%
Overall	13.1%	15.6%	15.3%	12.6%

4 Discussion

Our results show that ColabDesign's backprop-based optimization method outperforms top generative models on a majority of motifs, challenging the prevailing paradigm that generative models are necessary or well-suited for motif scaffolding. These results present several promising avenues for further investigation. During optimization, one can prune or terminate trajectories early when intermediate metrics are poor, or adopt sequential Monte Carlo-like resampling schemes to prioritize promising trajectories. Such strategies could improve success rates even further. Another compelling direction is to explore similar strategies for all-atom motif scaffolding—i.e. motifs specified only with functional groups or sidechain coordinates. This will require the development of differentiable scoring functions that measure the distance between a partially optimized sequence and a desired atomic motif, which are currently not yet available. However, if the dramatic improvement in success rate carries over to such settings, this could unlock many high-precision applications, such as enzyme active sites, metal coordination, and high-specificity binding interfaces.

References

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

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

Christopher Frank, Ali Khoshouei, Lara Fuß, Dominik Schiwietz, Dominik Putz, Lara Weber, Zhixuan Zhao, Motoyuki Hattori, Shihao Feng, Yosta de Stigter, Sergey Ovchinnikov, and Hendrik Dietz. Scalable protein design using optimization in a relaxed sequence space. *Science*, 386(6720): 439–445, October 25 2024. doi: 10.1126/science.adq1741.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint*, arXiv:2503.00710, 2025. arXiv, https://arxiv.org/abs/2503.00710.

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2. *arXiv preprint*, arXiv:2405.15489, 2024. arXiv, https://arxiv.org/abs/2405.15489.

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

Luis Santiago Mille-Fragoso, John N. Wang, Claudia L. Driscoll, Haoyu Dai, Talal M. Widatalla, Xiaowe Zhang, Brian L. Hie, and Xiaojing J. Gao. Efficient generation of epitope-targeted de novo antibodies with germinal. *bioRxiv preprint*, bioRxiv:2025.09.19.677421, 2025. arXiv, https://www.biorxiv.org/content/10.1101/2025.09.19.677421v1.

Sergey Ovchinnikov. Colabdesign, 2022. URL https://github.com/sokrypton/ColabDesign.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. One-shot design of functional protein binders with bindcraft. *Nature*, August 27 2025. doi: 10.1038/s41586-025-09429-6.

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, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, August 31 2023. doi: 10.1038/s41586-023-06415-8.

Jason Yim, Andrew Campbell, Andrew Y. K. Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S. Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noé. Fast protein backbone generation with se(3) flow matching. *arXiv preprint*, arXiv:2310.05297, 2023. arXiv, https://arxiv.org/abs/2310.05297.

5 Appendix

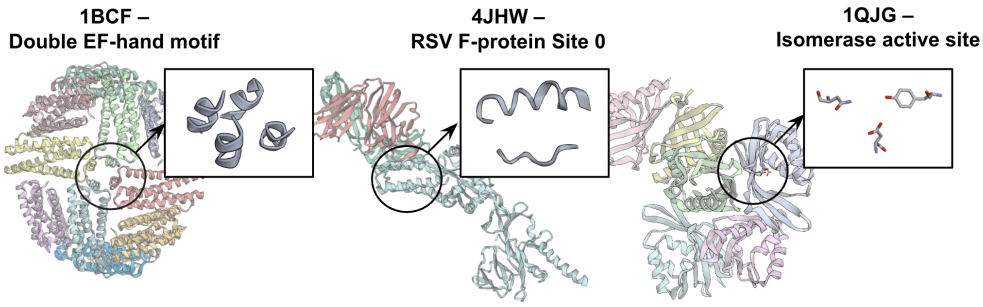

Figure 6: **Representative benchmark tasks.** Example problems from the functional-site scaffolding benchmark illustrate the design setup: a subset of residues corresponding to a functional motif is fixed (boxed), and the protein must be redesigned subject to task-specific constraints such as allowable N- and C-terminal lengths or total sequence length. Shown are three representative cases: `1BCF` (di-iron binding motif), `4JHW` (RSV F-protein epitope, Site 0), and `1QJG` (Δ^5-3-ketosteroid isomerase active site).

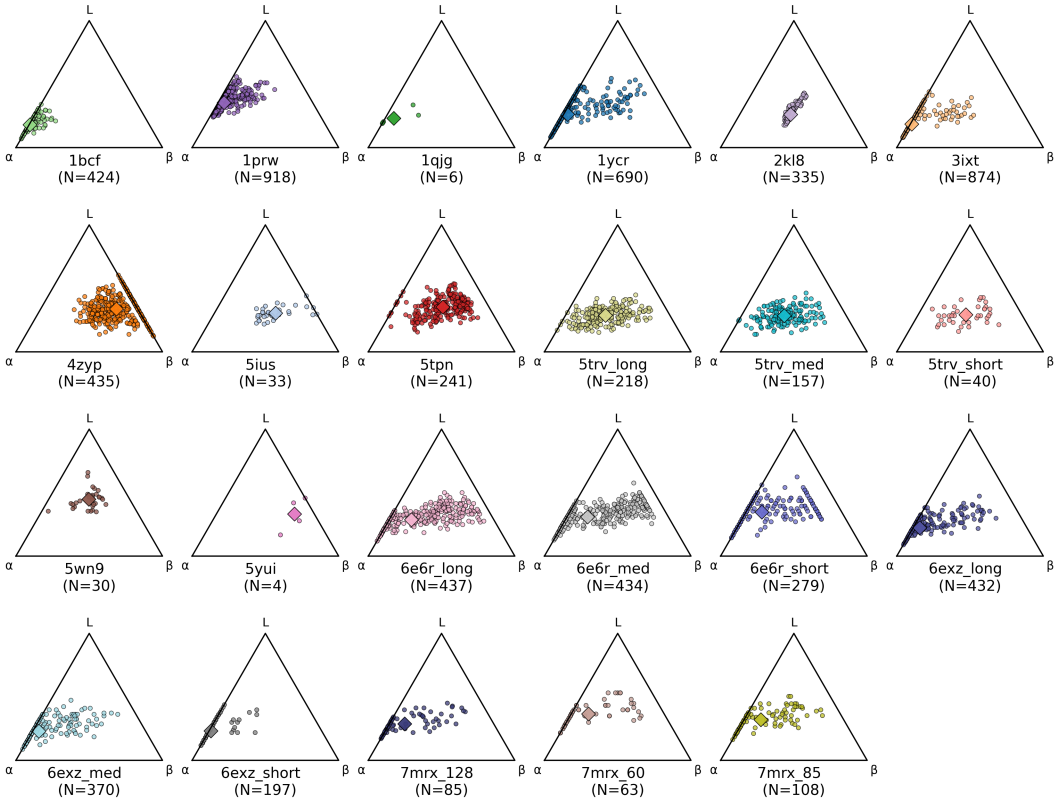

Figure 7: **Secondary Structure Breakdown.** Distribution of α-helix, 10.6% β-sheet, and loops in total successful designs for each problem (not clustered for uniqueness).

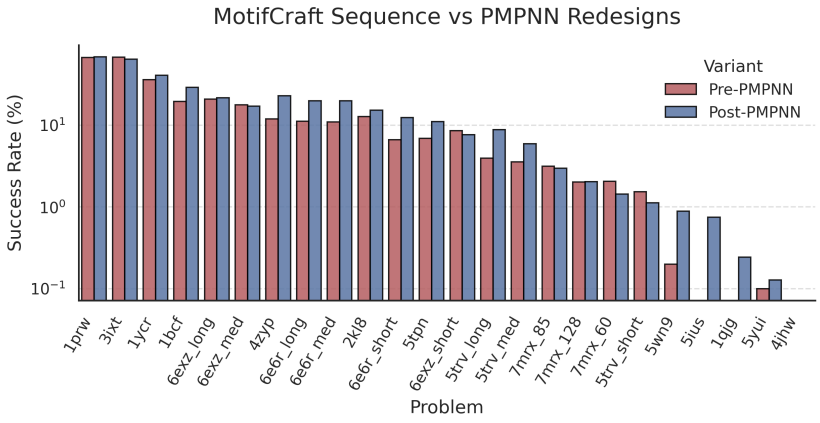

Figure 8: **Effects of PMPNN on success rates** Use of ProteinMPNN to redesign sequences boosts success on majority of samples and produces successes for problems that were unsolved under the original MotifCraft sequence. Numbers are reported pre-clustering.

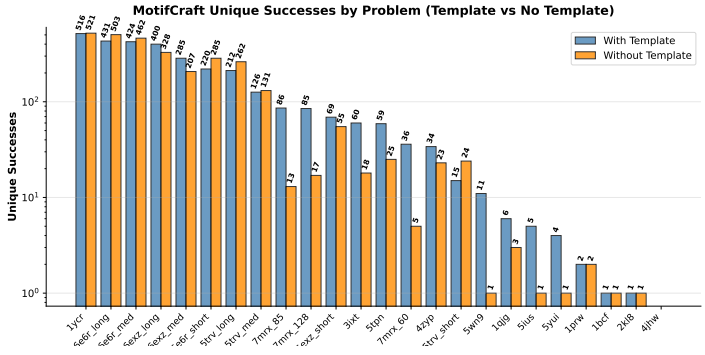

Figure 9: **Effects of template ablation on success rates.** Addition of template improves success rates on cases in which unique successes are lower. Numbers are reported after FoldSeek clustering.

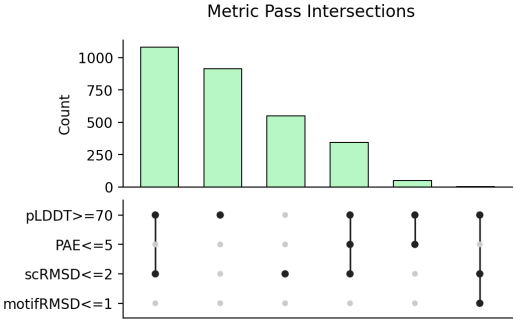

Figure 10: **Success modes of `4JHW`.** The most stringent failure mode is motifRMSD, followed by PAE. No designs pass all four criteria for this method. Passes are computed on all 8000 structures (8 refolds per original scaffold design)

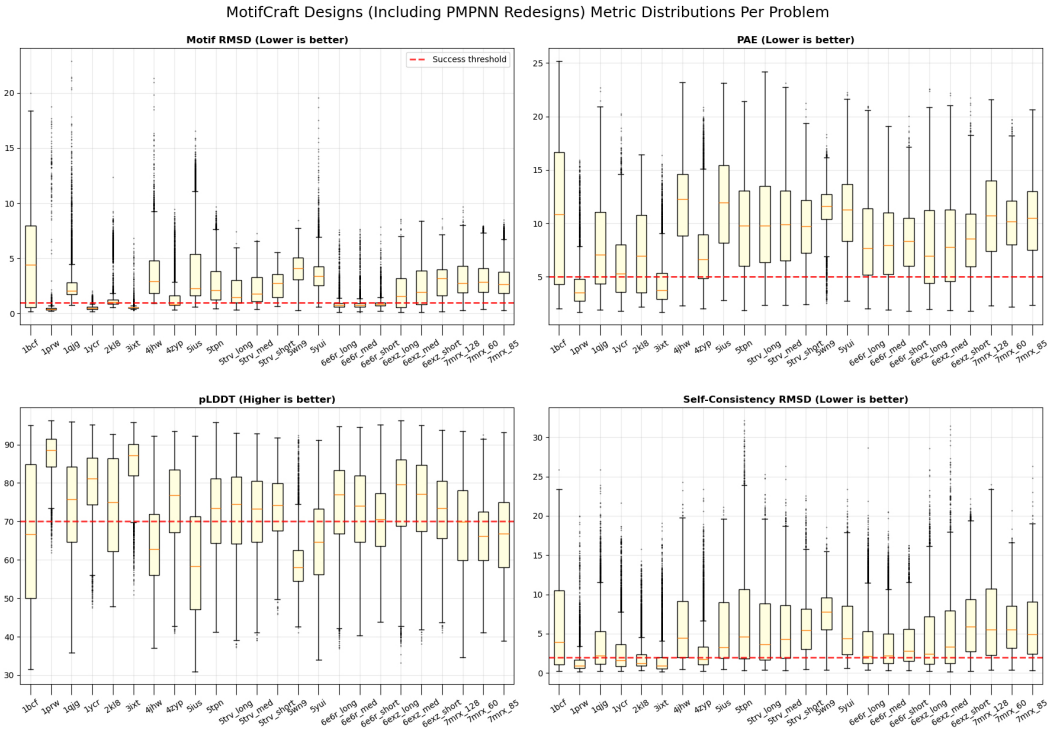

Figure 11: **Boxplots of metrics.** Including all refolds for each particular scaffold.

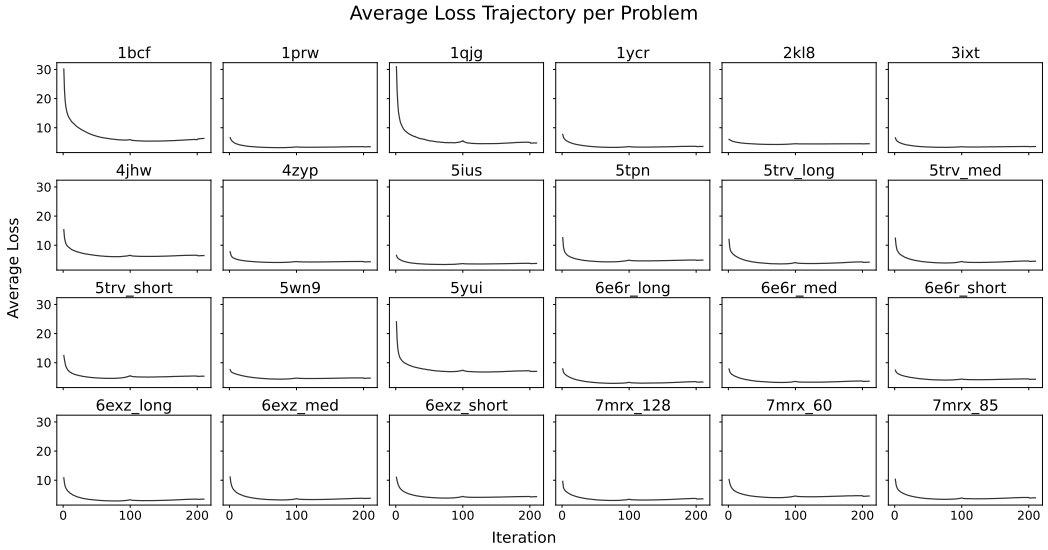

Figure 12: **Average trajectories over designs.** Plot of ColabDesign's loss trajectory; for each sample, the composite loss is recorded at each iteration and averaged aross a problem.