

Figure 6: Correlation between pLDDT and C α RMSD of designed sequences from ResiDPO.

A APPENDIX

A.1 PLDDT IS A GOOD DESIGNABILITY PROXY.

Protein designability quantifies how well a designed sequence folds into a target structure. Common metrics for assessing designability rely on structural similarity between the predicted and target structures, such as root-mean-square deviation (RMSD). For instance, in experimental protein design, only sequences exhibiting low C α RMSD are typically synthesized Kim et al. (2024).

While RMSD is calculated per residue, its value is highly sensitive to protein alignment. Even minor changes in dihedral angles can disproportionately inflate RMSD, particularly for residues located distally. This sensitivity makes RMSD suboptimal as a local quality metric. The Local Distance Difference Test (LDDT) addresses this limitation by providing a superposition-free measure of local distance differences. Notably, the predicted LDDT (pLDDT) from AlphaFold2 has been shown to correlate well with RMSD (Fig. 6). Given this correlation and its capacity for per-residue assessment, we adopt pLDDT as a proxy for protein designability in this work.

A.2 ANALYSIS OF PLDDT DISTRIBUTIONS.

In the left panel of Fig. 7, we compare the pLDDT score distributions of all methods. Notably, ResiDPO shifts many of the lower-scoring models into the 80–90 range, creating a distinct peak in this region that largely drives its higher success rate. In the right panel, we report independent two-sample t-tests between ResiDPO and both LigandMPNN and DPO. Four asterisks (****) indicate $p < 1e-4$, demonstrating that the improvements achieved by ResiDPO are highly statistically significant.

A.3 ANALYSIS OF PAE DISTRIBUTIONS.

Figure 8 compares the Prediction Angle Error (PAE) distributions for three methods, where lower PAE values indicate better performance. While DPO reduces the occurrence of high PAE outcomes, it does not substantially increase the proportion of low PAE results. In contrast, ResiDPO exhibits superior generalization, leading to a significant improvement in low PAE results and a concurrent reduction in high PAE instances. Nevertheless, a considerable number of results still present high PAE values, underscoring the need for future research focused on more targeted PAE optimization.

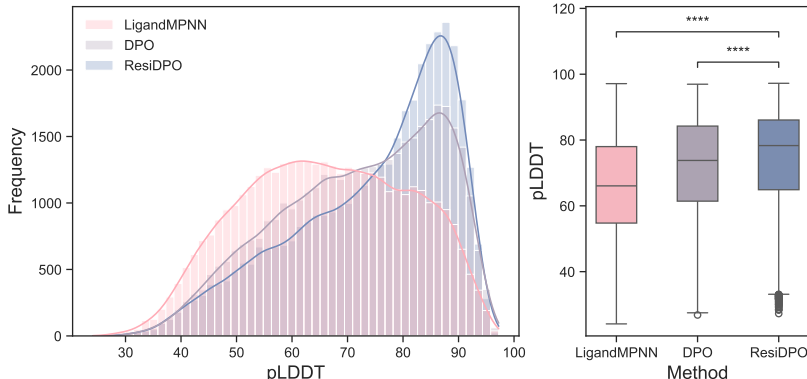


Figure 7: Comparison of the pLDDT distribution of predicted sequence from different methods on the enzyme design benchmark.

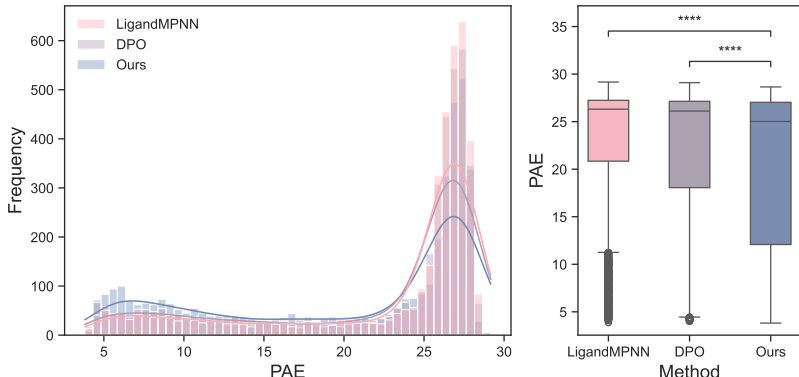


Figure 8: Comparison of the PAE distribution of predicted sequence from different methods on the binder design benchmark.

A.4 VISUALIZING DESIGN ACCURACY: CASE STUDIES

To qualitatively assess the structural fidelity of designs generated by various methods, we present two illustrative case studies from the enzyme (EC4) and binder (TrkA) design benchmarks. For each case, the target backbone (generated by RFDiffusion2 and RFDiffusion) is superimposed with the AlphaFold2-predicted structures of sequences designed by LigandMPNN, and fine-tuned ones with DPO and our proposed ResiDPO.

As depicted in Figure 9a, the enzyme designs from both LigandMPNN and DPO exhibit significant structural deviations. Specifically, sequences from these baseline methods result in disordered loop conformations within the catalytic pocket, leading to numerous steric clashes. In stark contrast, the ResiDPO-designed sequence accurately recapitulates the target backbone, folding with high precision into the desired catalytic architecture.

A similar outcome is observed in the binder design challenge (Figure 9b). LigandMPNN and DPO fail to produce viable binding candidates; the predicted structures of their designed sequences are disengaged from the target protein, indicating a loss of intended interaction. ResiDPO, however, successfully generates a sequence predicted to adopt the target conformation and maintain the crucial binding interface, underscoring its superior ability to enforce structural and functional constraints.

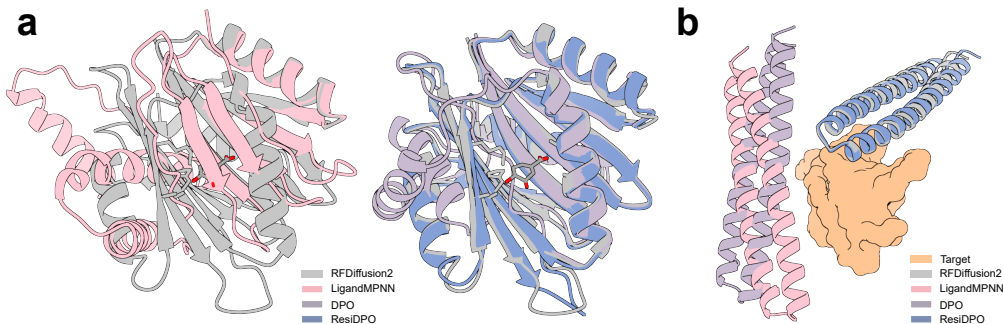


Figure 9: Comparative structural analysis of designed proteins. AlphaFold2 predictions of sequences designed by LigandMPNN, DPO, and ResiDPO for (a) an enzyme and (b) a protein binder, superimposed onto the designed backbone (grey). (a) ResiDPO (blue) achieves high backbone fidelity, accurately forming the catalytic pocket, while LigandMPNN (pink) and DPO (violet) designs exhibit disordered loops and steric clashes. (b) ResiDPO’s binder design maintains the intended binding pose and interface with the target (orange), whereas LigandMPNN and DPO designs fail to bind, becoming dissociated from the target. These visualizations highlight ResiDPO’s enhanced capability in generating sequences that precisely adhere to complex structural and functional requirements.

Table 2: Evaluation with ESMFold.

Method	Seq. Succ. Rate	Backbone Succ. Rate
LigandMPNN	11.83	30.36
ResiDPO	22.28	47.48

A.5 EVALUATION WITH ESMFOLD

Since reward hacking is a common issue for reinforcement learning methods, to evaluate whether our method is overfitting to the reward model (AlphaFold2), we also evaluate our model using another folding method, ESMFold. The result is shown in Tab. 2, our method’s significant performance gains (around 2x sequence success rate) hold when evaluated by this non-AlphaFold oracle. This provides strong evidence that ResiDPO learns generalizable principles of designability that are not specific to the biases of a single model family.

A.6 BENCHMARK DETAILS

To rigorously evaluate the performance of EnhancedMPNN against baseline models, we established two challenging de novo protein design benchmarks: enzyme design and binder design. These benchmarks were specifically constructed to assess ”designability”—the capacity of a designed sequence to fold into its target structure with high fidelity.

A.6.1 ENZYME DESIGN BENCHMARK

The enzyme design benchmark utilizes five distinct catalytic motifs. These motifs were sourced from the Atomic Motif Ensemble (AME) dataset, as previously characterized in the RFDiffusion2 paper Ahern et al. (2025). Each motif corresponds to an enzyme from a different top-level Enzyme Commission (EC) number classification. Specific details for each motif, including originating PDB ID, EC number, and constituent residues, are provided in Supplementary Tab. 3.

Following the RFDiffusion2 paper, we generated 1,000 backbone structures for each of the five selected catalytic motifs based on the motif atoms. For each of the 1,000 generated backbones, 8 protein sequences were designed using the Protein Sequence Design (PSD) models under evaluation (e.g., the baseline ligandMPNN and our proposed EnhancedMPNN). The amino acid identities of the catalytic motif residues were held fixed during sequence design, and the side-chain conformations

Table 3: Motif information for the enzyme design benchmark.

EC No.	Enzyme name	PDB ID	Motif Atoms	Ligands
EC1	UDP-glucose 6-dehydrogenase	1DLI	A118: [N, CA, C, CB]; A145: [N, CA]; A204: [NZ, CE, CD]; A208: [OD1, CG, CB, ND2]; A260: [SG, CB, CA]; A264: [OD1, CG, CB, OD2];	UDX, NAD
EC2	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	1Q0N	A82: [NE, CD, CZ]; A92: [NH2, CZ, NE, NH1]; A95: [OD1, CG]; A97: [OD2, CG];	APC, PH2, MG
EC3	dCTP deaminase	1XS1	A124: [O, C]; A126: [O, C]; A138: [OE1, CD, CG, OE2]; C111: [N, CA, C, CB]; C115: [NH2, CZ, NE, NH1];	DUT
EC4	3-dehydroquinatase (type I)	1QFE	A86: [O, C, CA]; A143: [NE2, CD2, CE1, CG, ND1]; A170: [NZ, CE, CD];	DHS
EC5	delta 5-3-ketosteroid isomerase	1E3V	A16: [OH, CZ, CE1, CE2]; A40: [OD2, CG]; A100: [N, CA, C, CB]; A103: [OD2, CG];	DXC

Table 4: Target information of the binder design benchmark.

Target	PDB ID	Hotspot
PD-L1	5O45	A56, A115, A123
IL7 Receptor subtype Alpha	3DI3	B58, B80, B139
Insulin Receptor	4ZXB	E64, E88, E96
TrkA Receptor	1WWW	X294, X296, X333
Influenza Hemagglutinin	5VLI	B521, B545, B552

of these fixed motif residues were provided as structural context to the PSD models. Cysteine residues were omitted from the vocabulary of possible amino acids, and others were sampled with a temperature of 0.1.

The designability of each generated sequence was assessed by predicting its structure using AlphaFold2 Jumper et al. (2021). To accurately simulate *de novo* design evaluation, no Multiple Sequence Alignments (MSAs) or external structural templates were utilized during AlphaFold2 predictions. A design was classified as successful if its pLDDT \geq 80 and C α RMSD \leq 1.5Å.

A.6.2 BINDER DESIGN BENCHMARK

The binder design benchmark was adapted from the protocols detailed in the original RFDiffusion study Watson et al. (2023). This benchmark comprises five distinct protein targets against which binders are designed. Detailed information for each target protein, including PDB ID and relevant chain information, is provided in Supplementary Tab. 4.

For each of the five target proteins, 100 unique *de novo* binder backbone structures were generated *de novo* using RFDiffusion. The generation of these binder backbones was guided by predefined "hotspot" residues, which specify the desired interaction interface on the target protein. For each of the 100 generated binder backbones per target, 8 candidate sequences were designed using the PSD models with the target chain fixed. The temperature is set to 0.1.

Designed binder sequences were evaluated for their ability to fold correctly and bind to their respective targets using AlphaFold2 Jumper et al. (2021). During this structural prediction step, the native structure of the target protein chain was provided as a template to AlphaFold2 to ensure its accurate representation in the complex. A binder design was considered successful if it satisfied all three of the following criteria: the binder sequence pLDDT \geq 80, inter-chain PAE \leq 10, and the overall C α RMSD \leq 1Å.

Table 5: Comparison of ResiDPO with DPO variants on the validation set.

Method	Seq. Recovery	pLDDT Acc.
LigandMPNN Dauparas et al. (2025)	57.63	57.71
DPO Rafailov et al. (2023)	57.03	62.11
Robust DPO Chowdhury et al. (2024)	56.96	62.56
RSO Liu et al. (2023)	56.91	61.23
KTO Ethayarajh et al. (2024)	57.03	61.23
NCA Chen et al. (2024)	56.96	59.09
SPPO Wu et al. (2025)	56.92	59.91
ResiDPO	55.56	66.08

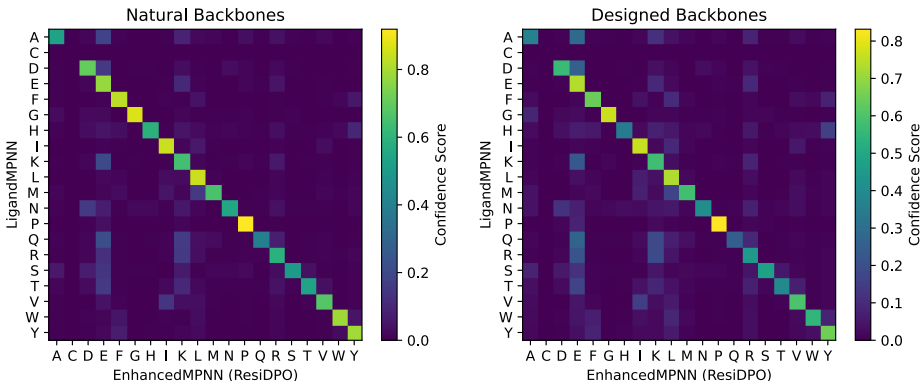


Figure 10: Confusion matrices showing residue substitutions induced by EnhancedMPNN on natural backbones (left) and designed enzyme backbones (right).

A.7 EVALUATION ON ATOMIC MOTIF ENZYME (AME) BENCHMARK WITH 41 ENZYMES

To evaluate our methods in a broad range of protein targets, we extended our benchmark to the recently released Atomic Motif Enzyme (AME) Benchmark Ahern et al. (2025) with 41 enzymes.

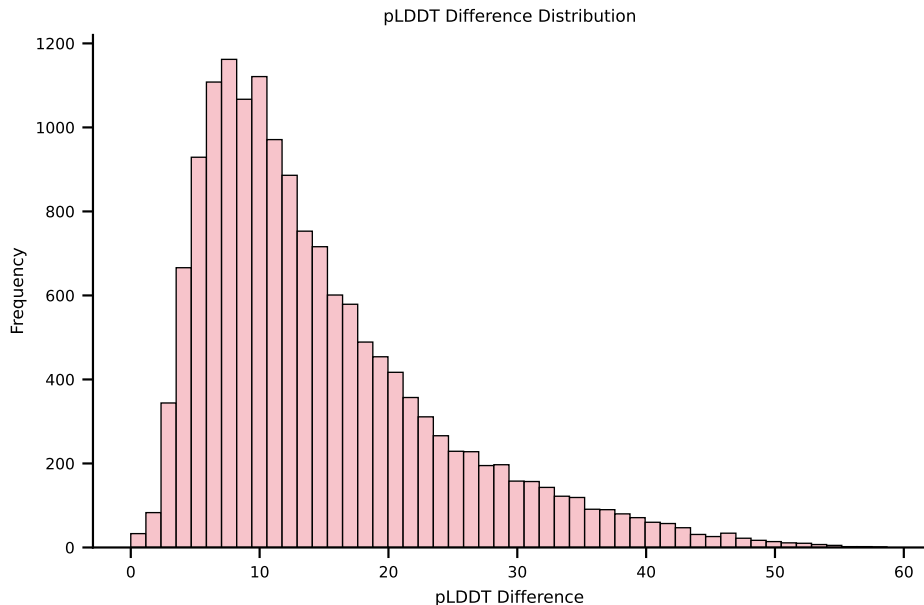


Figure 11: Distribution of pLDDT difference (δ) for the PDB-D training set. Most structures exhibit a pLDDT difference ranging from 5 to 15.

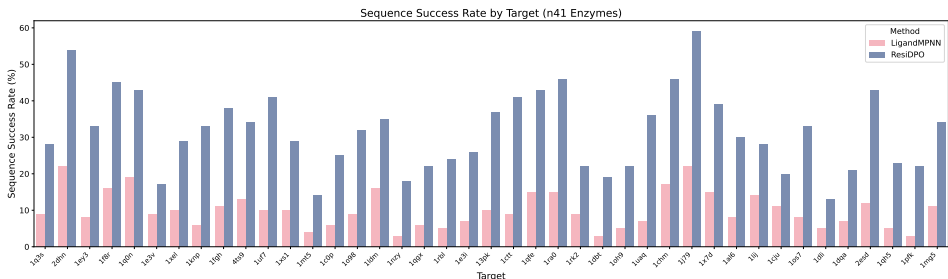


Figure 12: Evaluation on the 41-enzyme Atomic Motif Enzyme (AME) benchmark.