Supplementary Materials for

Expanding the sequence spaces of synthetic binding protein using deep learning-based framework ProteinMPNN

1. Reliability of predicted structures

Combined algorithms of deep residual network and Rosetta-constrained energy minimization, trRosetta has been widely used to the rapid and accurate prediction of protein structure [1, 2]. Meanwhile, all predicted structures in the SYNBIP database have confidence estimation scores (TM-scores) above 0.7, indicating a correctly modeled topology [1].

To further validate the accuracy of structure prediction, four experimentally characterized SBPs (SBP000066, SBP000118, SBP000710, and P001886) from different scaffolds were randomly selected. These four SBPs cover the four main protein secondary structures (α-helix, β-sheet, β-turn, and random coil). trRosetta was used to predict the structures of these four SBPs and analyzed the differences between the predicted and experimental structures. The results and **Fig. S1** show that the predicted structures are generally consistent with the experimental structures, demonstrating the reliability of the structure-based design.

2 Statistical analysis

The solubility and stability of the five sequences designed by each SBP were averaged as a subsequent comparative analysis. In the analysis of binding energy, three SBPs, SBP000682, SBP000792, and SBP001003, were excluded due to the lack of detected interactions with the target.

2.1 Normality test for sample differences

Based on the comprehensive analysis including the Kolmogorov-Smirnov test (**Table S1**), skewness (absolute value less than 3), and kurtosis (absolute value less than 10) (**Table S1**), as well as histograms (**Fig. S2**, **Fig. S4**, and **Fig. S6**) and Quantile-Quantile plots (Q-Q plot) (**Fig. S3**, **Fig. S5**, and **Fig. S7**), it can be reasonably concluded that the differences in solubility, instability index and binding energy conform to a normal distribution.

2.2 Paired t-test analysis

In the case where the differences in samples follow a normal distribution, paired t-test analysis was conducted, and the specific calculation results are shown in **Table S2**.

Table S1. Skewness, Kurtosis, as well as results of Kolmogorov-Smirnov test and Shapiro-Wilk test for sample differences

Monomer – Complex	Skewness	Kurtosis	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
			Statistic	df	Sig.	Statistic	df	Sig.
Solubility	0.627	2.013	0.103	61	0.177	0.958	61	0.037
Instability Index	-0.657	0.439	0.100	61	0.200	0.969	61	0.124
Binding Energy	0.631	1.230	0.093	58	0.200	0.964	58	0.079

[a] Lilliefors Significance Correction

Table S2. Paired t-test specific data for proteins based on monomer design and complex design.

Monomer – Complex	Paired Differences					Significance
	Mean	Std. Deviation	Std. Error Mean	t	df	Two-Sided p
Solubility	0.026	0.092	0.012	2.214	60	0.031
Instability Index	-3.850	11.860	1.519	-2.535	60	0.014
Binding Energy	2.693	4.831	0.634	4.245	57	<0.001

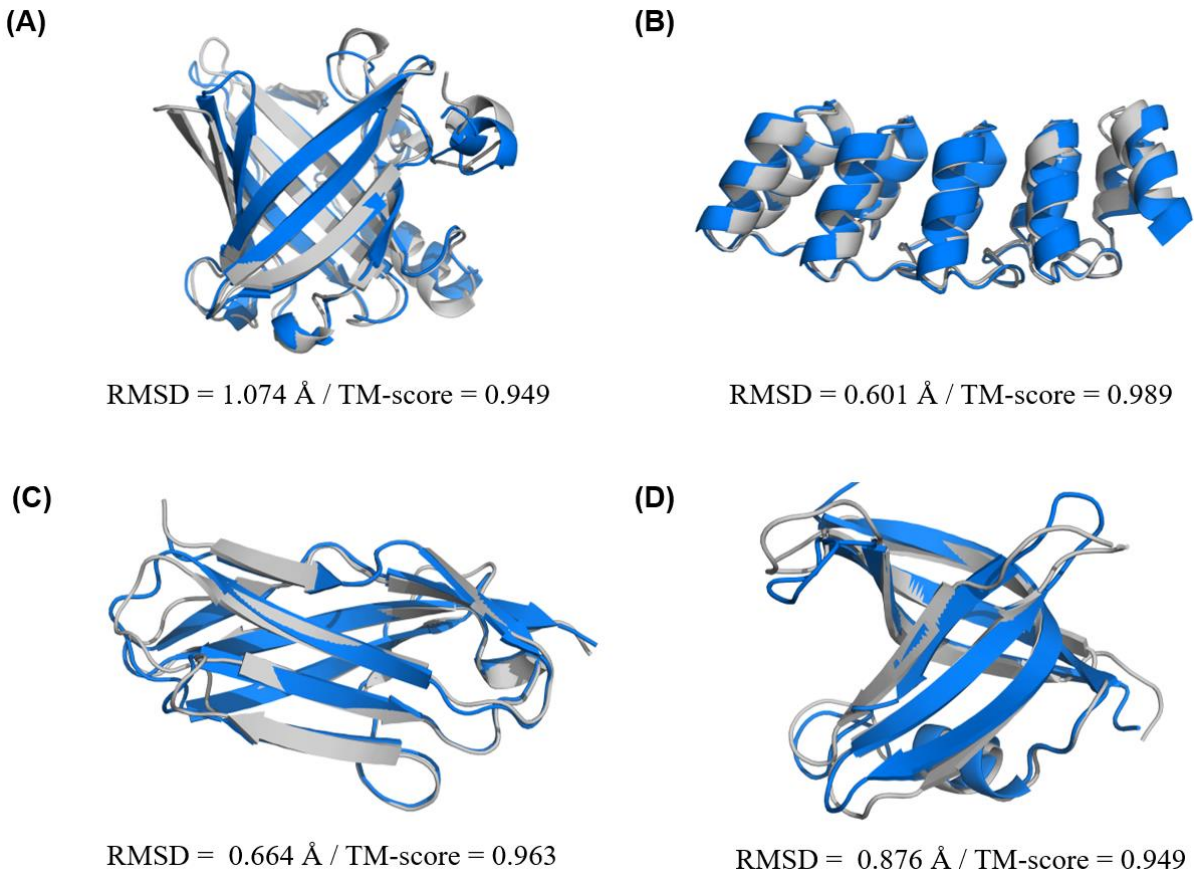

Fig. S1. Structural alignment of the predicted structure (blue) and crystal structure (gray) of (A) SBP000066, (B) SBP000118, (C) SBP000710, and (D) SBP001886.

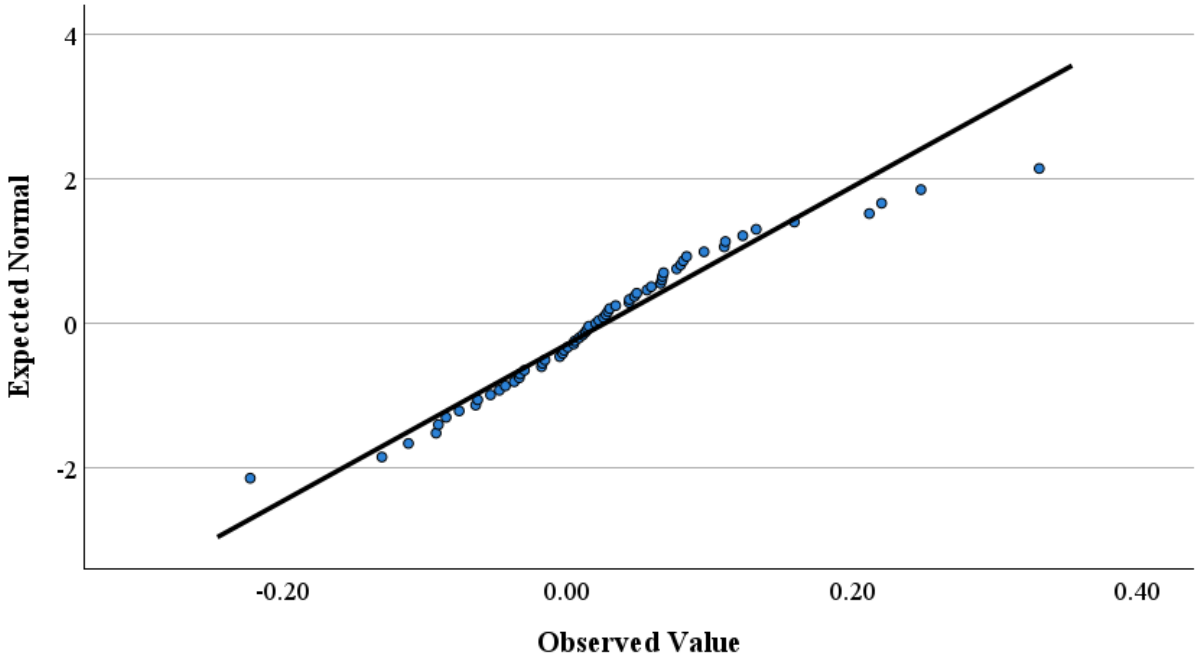

Fig. S2. Normal Q-Q Plot of differences in solubility.

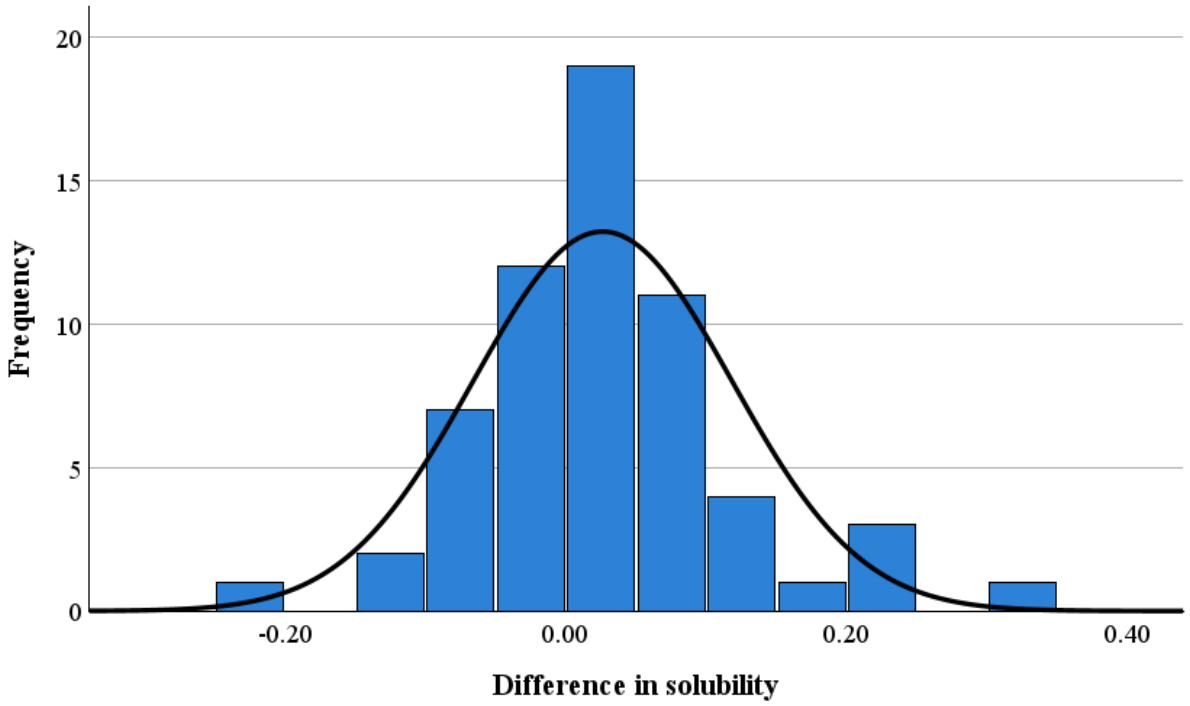

Fig. S3. Histogram of differences in solubility.

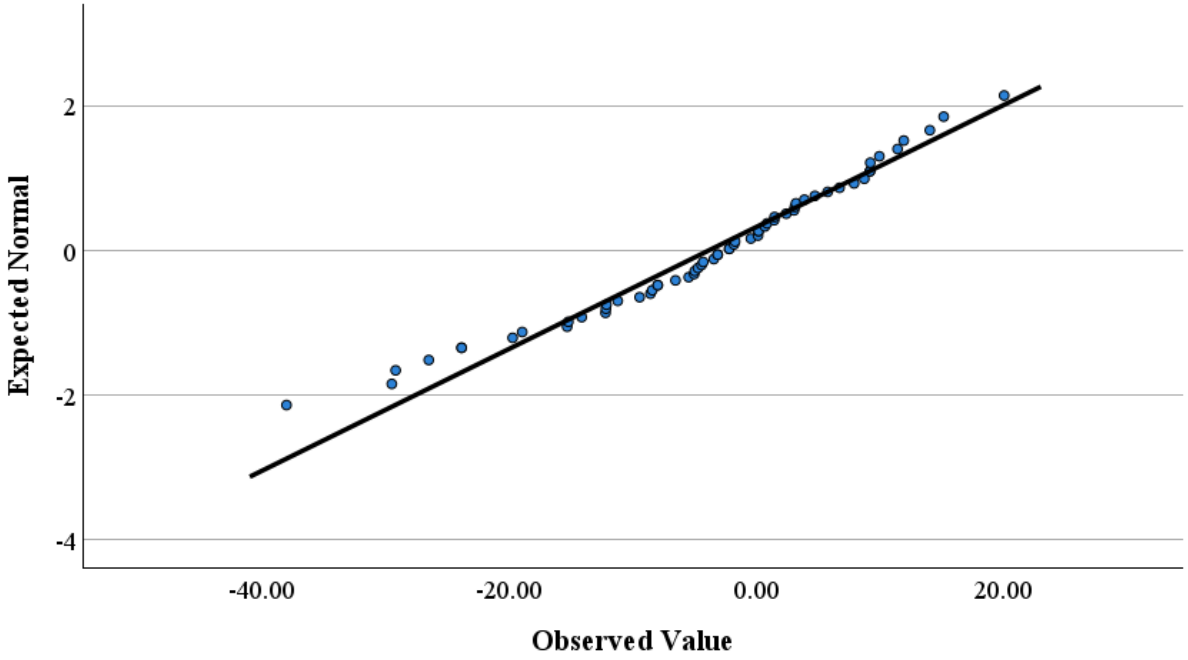

Fig. S4. Normal Q-Q Plot of differences in instability index.

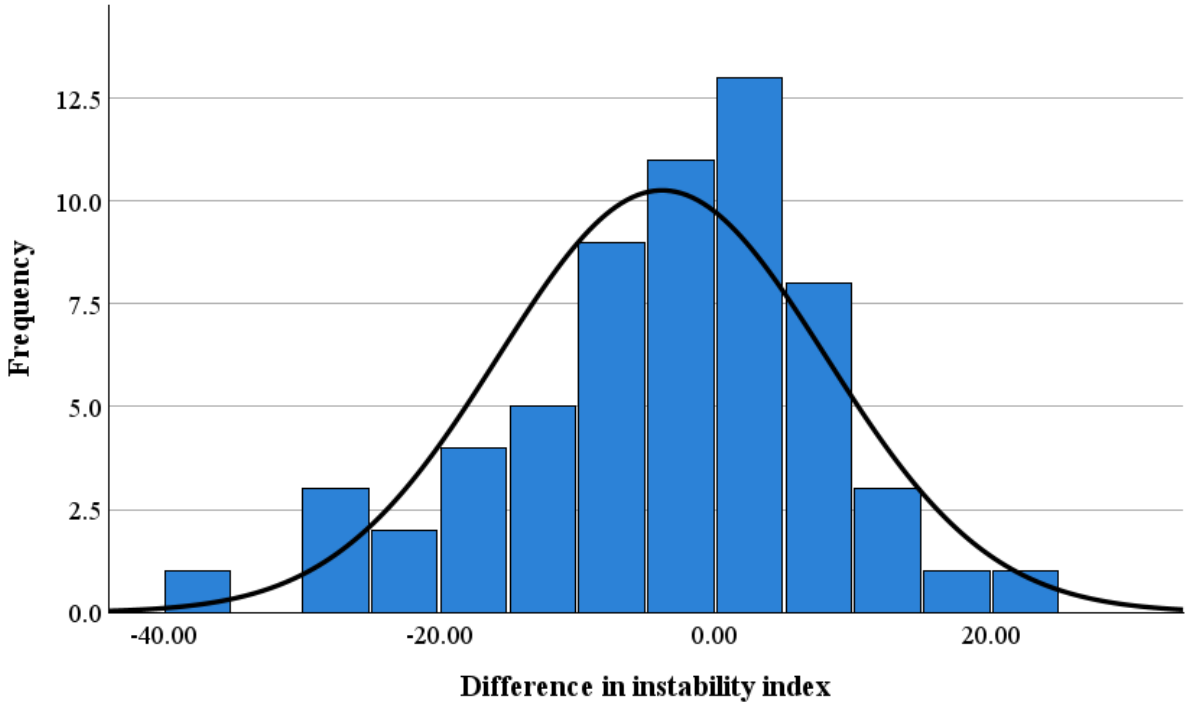

Fig. S5. Histogram of differences in instability index.

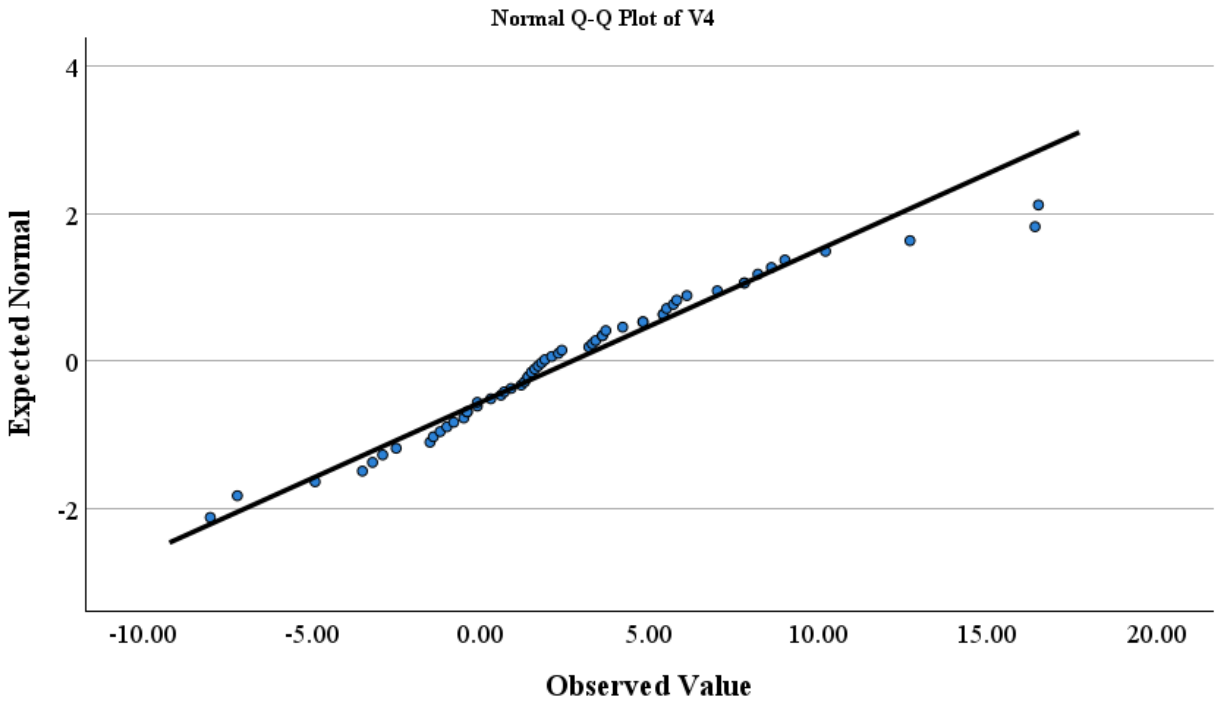

Fig. S6. Normal Q-Q Plot of differences in binding energy.

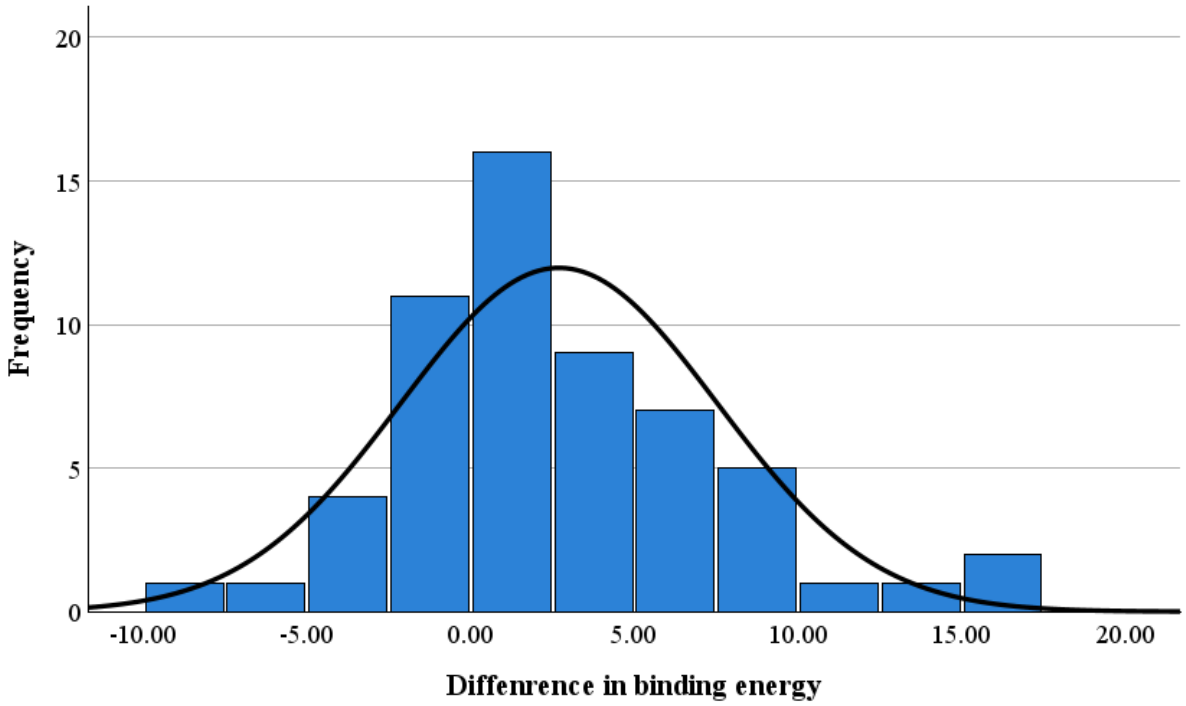

Fig. S7. Histogram of differences in binding energy.

References

1. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A, 2020, 117(3): 1496-1503
2. Abriata L A, Dal Peraro M. State-of-the-art web services for de novo protein structure prediction. Brief Bioinform, 2021, 22(3)