Front. Comput. Sci., 2025, 19(5): 195903
https://doi.org/10.1007/s11704-024-31060-3

RESEARCH ARTICLE

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

Yanlin LI[1], Wantong JIAO[1], Ruihan LIU[1], Xuejin DENG[1], Feng ZHU (✉)[2], Weiwei XUE (✉)[1]

1 Chongqing Key Laboratory of Natural Product Synthesis and Drug Research, School of Pharmaceutical Sciences, Chongqing University, Chongqing 401331, China
2 College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China

Abstract Synthetic binding proteins (SBPs) with small size, marked solubility and stability, and high affinity are important for protein-based research, treatment, and diagnostics. Over the last several decades, site-directed mutagenesis and directed evolution of privileged protein scaffold make up the great majority of SBPs. The groundbreaking advancement of deep learning (DL) in recent years has revolutionized the problem of protein structure prediction and design. Here, for the first time, the cutting-edge DL framework ProteinMPNN was applied to fulfill the *de novo* design of 7,245 new synthetic proteins covering 55 different scaffolds based on the original SBPs collected in our SYNBIP database. Comprehensive bioinformatics analysis indicated that, in addition to the excellent performance of sequence recovery, the designed synthetic proteins have a significant improvement in solubility and thermal stability compared to the currently known SBPs. Meanwhile, 8 incredibly suitable protein scaffolds for ProteinMPNN have been identified, from which the designed synthetic proteins calculate displayed good performance on binding ability to their corresponding protein targets. Therefore, the DL-based framework shown great potential in target-directed *de novo* generation of synthetic protein library with high quality, which could assist experimental biologists to rational protein engineering to discover novel functional protein binders.

Keywords synthetic protein, deep learning, *de novo* protein design, solubility, stability

1 Introduction

Synthetic binding proteins (SBPs) with small size, marked solubility and stability, and high binding affinity are essential for protein-based research, treatment, and diagnostics [1]. SBPs were usually designed by site-directed mutagenesis and directed evolution techniques based on a privileged protein scaffold [2]. Over the past several decades, the number of SBPs has significantly increased. According to SYNBIP [2], a comprehensive database of synthetic binding proteins-related information, there are currently over 2,000 known SBPs have been discovered and with 74 SBPs have entered the clinic [2]. However, the classical techniques have limitations in expanding the protein space of SBPs [3]. Site-directed mutagenesis is highly dependent on the physiological properties and structure of parental protein, which limits the application of mutation results obtained from a single protein to other proteins [3], and it may also be hampered by the lack of high-resolution crystal structures [4]. While directed evolution explores only the sequence space regions around natural proteins, which means that this method only modifies certain amino acids in the protein sequence, and the quality of generated libraries is not guaranteed [5,6]. Protein *de novo* design uses physicochemical principles and molecular interactions to generate new protein variants, allowing for complete protein remodeling without stringent requirements on parent proteins. At the same time, the positions of the modified amino acids are relatively extensive [3]. Meanwhile, protein *de novo* design software based on physical methods relies on energy functions to predict the stability and folding behavior [3,7−12]. However, current energy functions may not be sufficient to accurately simulate the complex folding and interactions of proteins in nature [13], this has led to instances of design failures. Since deep learning (DL) has significant advantages in extracting features and integrating information from proteins [14], so it can uncover subtle patterns and correlations that traditional methods might overlook. Therefore, exploring synthetic binding protein design using DL-based *de novo* protein design is urgent and of great significance [15].

DL has been applied to multiple fields, including protein design. Modes such as ProteinSolver [16] and ProteinBERT [17] enable fast and flexible design based on real protein sequences and structures. In addition, deep networks that predict the structure of natural proteins are being inverted to design entirely new proteins [18], as well as creating entirely new luciferase enzymes by way of ‘full family illusions’ [19]. The advantage of *de novo* protein design based on DL is that

Received December 28, 2023; accepted April 17, 2024

E-mail: zhufeng@zju.edu.cn; xueww@cqu.edu.cn

Special Issue—Application of computational techniques in drug discovery

the computer learns a lot about the existing protein structure and fully understands what sequences are reasonable and popular [20−22], so the generated sequences are also more reasonable and likely to be popular [20].

More recently, DL-based *de novo* protein design has seen a breakthrough with the release of ProteinMPNN [23]. As an advanced DL framework, ProteinMPNN [23] achieves an unprecedented average sequence recovery rate of 52.4% and the design of a new protein containing 100 residues in 1.2 s on a single CPU [23]. For ProteinMPNN [23], the training 23,358 clusters (more than 500,000 crystal structures) were selected from the Protein Data Bank (PDB) [24]. Compared to other DL-based protein designs, ProteinMPNN not only has more model training data but also designs protein sequences quickly, with a high success rate and a wide range of applications. However, a thorough study of SYNBIP [2] revealed that only a very small number (245) of SBPs have crystal structures, and it is reasonable to speculate that there are few crystal structures for SBPs were used for ProteinMPNN training. Thus, to expand the number of SBPs by DL-based protein sequence design methods like ProteinMPNN [25] is of great necessity to extensively explore their applicability to SBPs with different scaffolds [26].

In this work, starting from the 3D structures of SBPs collected in SYNBIP database [2], ProteinMPNN [23] was first used for *de novo* design of synthetic protein sequences. Then, comprehensive bioinformatics analysis suggested that the new protein sequences have improved solubility and stability compared with the original SBPs. Furthermore, the differences between input structures in two forms including SBP (monomer) and the SBP bound to protein target (complex) were explored. It is found that the sequences generated based on monomer structure are better than that of complex structure in terms of solubility and stability. In contrast, the sequences designed based on the complex structure have better performance in calculated binding energy. Meanwhile, synthetic proteins designed by ProteinMPNN show superior performance compared to classical protein engineering methods. Overall, the DL-based protein sequence design framework ProteinMPNN is proven to be suitable for potential synthetic protein sequences generation. It shows great potential to overcome the limitation of protein engineering through the rapid generation of high-quality protein library, which provides scientists with new ideas to expand the sequence space of synthetic binding proteins.

2 Materials and methods

2.1 Input structures preparation

1,327 monomer structures were prepared including 1,100 trRosetta [27] predicted ones and 227 experimental ones were retrieved from the SYNBIP database [2]. The selected proteins are all single-chain SBPs with available structures from the SYNBIP database. The number of SBPs is consistent with the SYNBIP database. The reliability of predicted structures is discussed in the “Reliability of predicted structures” section of Supplementary Materials.

61 complex structures were prepared including 1 crystal structure [28] and 60 computational models [29]. In the preliminary stage, at least one representative SBP was selected for each protein scaffold. These SBPs were docked with their respective targets, resulting in 60 successfully docked complex structures. To facilitate comparative analysis across various sources of complex structures, additional one crystal complex structure was chosen from the successful docking complexes. Meanwhile, to control variables, additional 61 monomer structures were extracted from the complex ones and used for synthetic proteins design in the comparative analysis.

2.2 Generation of synthetic proteins using ProteinMPNN

Based on the prepared PDB files of 61 complex structures and 1,327 monomer structures, the DL-based framework of ProteinMPNN (See github.com/dauparas/ProteinMPNN/ website) [23] was used for new synthetic proteins generation. The training set for ProteinMPNN is protein assemblies in the PDB (as of August 2, 2021) with a resolution higher than 3.5 Å and less than 10,000 residues as determined by X-ray or Cryo-EM. Among them, the structures of 216 SBPs (95%) are included in the training set. Specific protein data for the training set can be downloaded from the website of files.ipd.uw.edu/pub/training_sets/pdb_2021aug02.tar.gz/. At different temperatures, SBPs in the monomer group were designed, and the sequence recovery exhibited an inverse trend with temperature (Fig. 1). As the sequence recovery is a crucial evaluation criterion for *de novo* protein design [30], so the temperature parameter of ProteinMPNN is set to 0.1 in this work. For each SBP, the number of amino acid sequences generated is set to 5.

2.3 Analysis of sequences similarity to proteins in protein data bank

The sequences similarity between the generated new synthetic proteins and proteins stored in PDB was analyzed by in-house Python script based on MMseqs2 algorithm [31]. The selection of sequence similarity threshold depends on the specific task. Generally, sequences with a similarity of less than 50% are considered non-homologous [32]. A lower threshold of 30% was chosen to ensure greater diversity in the generated sequences, thus verifying their non-homologous nature. MMseqs2 software [31] is employed for discovering similar protein sequences. In terms of sequence similarity search, MMseqs2 software outperforms common methods like BLAST [33] in terms of performance at comparable sensitivity levels.

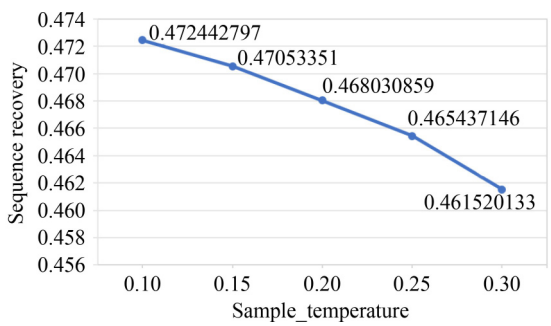


Fig. 1 Sequence recovery of protein sequences designed based on monomers at different temperatures

2.4 Properties analysis of designed synthetic proteins

2.4.1 Solubility analysis

Protein-Sol [34] was used for solubility calculation. The predicted solubility in Protein-Sol is a number from 0 to 1. When the value is higher than 0.45, it is considered that the solubility of this protein is better than the average soluble *E. coli* protein [35], that is, higher values mean better solubility. Out of the 6,635 amino acid sequences generated based on SBPs in monomer states, 65 were shorter than 20 in length, and 23 were identified as membrane proteins. For standardized statistics, these 88 amino acid sequences and other amino acid sequences generated from the same SBPs were not considered. Protein-Sol is a webserver and a program written by Python was used for automated calculation in this work.

2.4.2 Stability analysis

ProtParam method [36] integrated in TBtools [37] was used for thermal stability calculation. The instability coefficient is widely used to evaluate the instability of amino acid sequences *in vivo*. Protein with an instability index less than or equal to 40 is predicted to be stable, while protein with a value higher than 40 may be unstable. In other words, the smaller the value, the higher stability it is.

2.5 Analysis of binding energy between synthetic proteins and targets

To analyze the binding energy between synthetic proteins and targets, the new generated proteins were first docked to the binding site of protein target using PyMOL software [38]. Then, PDBePISA [39] was used for binding energy estimation. For five designs of each SBP, the one with the highest binding energy was selected for comparison.

2.6 Statistical analysis

Data analysis was performed using SPSS Statistics 29.0 software. To compare the solubility, stability, and binding energy between different design methods for the same SBP and determine if the differences were significant, a paired t-test was conducted under the assumption that the sample differences followed a normal distribution. Confirm in the Supplementary Materials that the sample differences follow a normal distribution. The null hypothesis was that the mean difference equals zero, with a significance level set at $P < 0.05$ to denote statistical significance. Specific calculation data can be found in the Supplementary Materials.

3 Results and discussion

3.1 7,245 new synthetic proteins designed by ProteinMPNN

Using the PDB files of 61 complex structures [29] and 61 monomer structures extracted from the complex structures (the details can be found in Section 2.1), and 1,327 monomer structures retrieved in SYNBIP database [2] as the inputs, new synthetic proteins were designed by the DL-based framework ProteinMPNN [23]. In this work, the temperature and the number of output sequences for each input SBP were set to 0.1 and 5, respectively. As a result, 7,245 synthetic binding proteins were successfully designed, which covering 55 scaffolds and 342 potential targets (**Appendix A**). Among them, 78.33% of the designed synthetic proteins have less than 200 amino acids. All the designed new synthetic protein sequences were used for further analysis.

3.2 Excellent sequence recovery of the designed synthetic proteins

Sequence recovery is recognized as an important criterion for *de novo* protein design evaluation [40]. Herein, the average sequence recovery for 6,635 sequences generated by SBP in monomer state is as high as 47.27%. In addition, there is no significant difference between the sequence recovery of 61 representative designs with SBPs in monomer and complex states (Fig. 2), which has the average value of 47.50% and 47.59%, respectively. The result indicated that although the decoding methods of ProteinMPNN are different for input structures, the accuracy of the algorithms for monomer structure at the sequence recovery is the same as that of and complex structure basically. Therefore, the ProteinMPNN-generated synthetic protein sequences basically share similar backbone structure with the parental SBPs (natural or artificial), which are characterized by sequence recovery and could be verified by protein 3D structure prediction [41].

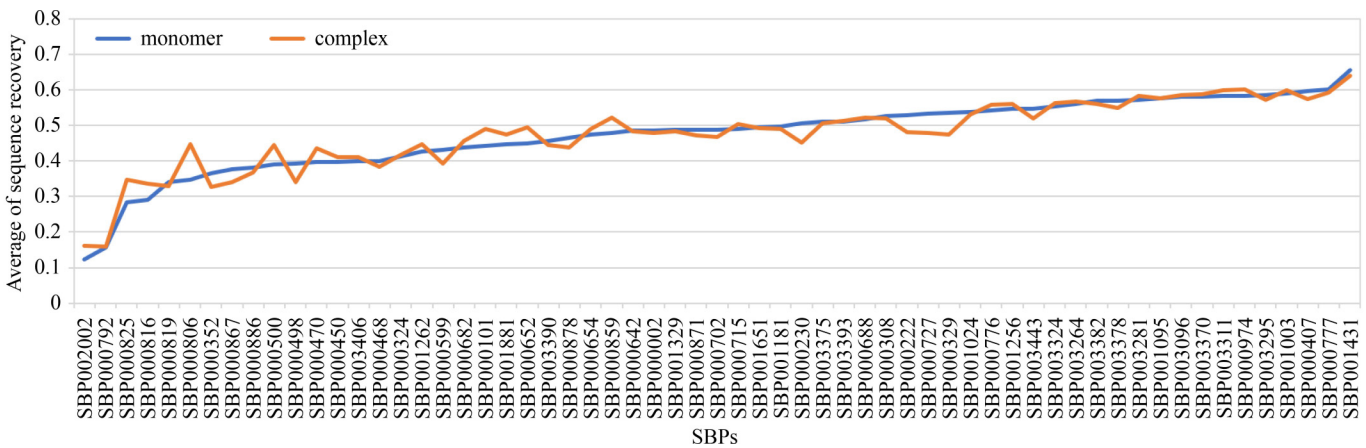


Fig. 2 Comparison of the average value of sequence recovery synthetic proteins designed by ProteinMPNN based on original SBPs in monomer and complex forms

3.3 Sequence diversity of synthetic proteins designed by ProteinMPNN

In addition to sequence recovery, it is important to analyze the diversity of new designed synthetic protein sequences [42]. In this work, a search was conducted on 6,635 sequences generated by SBP in monomer state. 516 amino acid sequences didn't receive search results. It can be concluded that these 516 amino acid sequences have no homologous proteins. In other words, the *de novo* designed proteins don't have direct relationship with any proteins present in nature. At the same time, 551 amino acid sequences with similarity less than or equal to 30%. This indicates that the ProteinMPNN design will result in novel amino acid sequences. Although most of the amino acid sequences designed by ProteinMPNN have over 30% sequence similarity, it also indirectly reflects the powerful ability and accuracy of ProteinMPNN to learn data from PDB databases.

Moreover, there are 20 repetitive sequences (0.30%) among the 6,635 ProteinMPNN-generated amino acid sequences in monomer state were identified in this work (Table 1). The repetitive protein sequences are mainly derived from proteins with shorter sequence lengths. At the same time, observing the 12 original synthetic binding proteins with amino acid sequence lengths less than 20, it was found that the similarity between the generated protein sequences is extremely high. It can be concluded that when designing short sequence synthetic binding proteins, their repeatability is relatively high. The reason for this situation may be that during the training process of deep learning ProteinMPNN, there are too few samples of short sequences. It was found that proteins with amino acid sequence lengths less than or equal to 20 accounted for about 6% of all proteins in PDB database, and most of them didn't meet the selection criteria of ProteinMPNN.

3.4 Improved protein solubility of the designed synthetic proteins

Solubility is a fundamental concept that is crucial in protein science [43,44]. As a protein trait determined by its primary amino acid sequence and environmental conditions, this concept is important for structural and biophysical research [45]. Here to explore the probability of improved solubility of generated amino acid sequences compared to original SBP, proportion of increased solubility (P_{SL}) is defined. P_{SL} is the number of generated amino acid sequences with solubility better than that of original SBP (N_b) divided by the total number of generated amino acid sequences (N_t = 5). For each original SBP, if the P_{SL} is equal or greater than 0.6, the solubility of at least 3 out of the 5 designed amino acid sequences is improved compared to the original SBPs. The predicted solubility data for each designed protein sequence and the original SBP can be found in **Appendix A**.

Table 1 List of 20 duplication among the ProteinMPNN-generated protein sequences

Sequence	Repetitions	Source of amino acid sequence
CTLPGYENNPEC	2	SBP000286
CHPLSTHPEC	3	SBP000288
CHPTSTHPLC	2	SBP000288
CHPDSTHPLC	2	SBP000289
CHPDSTHPDC	2	SBP000289
LVXEEAS [a]	3	SBP000293
SIPGTTL	2	SBP003419
GCTGPNCANSPGG	2	SBP003420
GCTGPNCANSAGP	2	SBP003420

[a] The "X" in the sequence means that amino acid at the position is unknown.

3.4.1 Solubility analysis of the sequences designed from monomer SBPs

Figure 3(a) shows the quantity distribution of different P_{SL}, and it can be concluded that most of the designed amino acid sequences (refers to 72% of the original SBPs) show improved solubility. Particularly, the solubility of the amino acid sequences designed from 46% of the original SBPs was all improved (Fig. 3(a)), demonstrating that ProteinMPNN [23] has significant advantages in improving the solubility of SBPs.

In addition, the distribution of SBPs with P_{SL} of 1 in the corresponding SBP scaffolds is also plotted. As shown in Fig. 3(b), when the proportion of those with P_{SL} of 1 in the total number of original SBPs corresponding to the same scaffolds exceeds 0.5, it means that the solubility of the amino acid sequences designed by most of the original SBPs to which the scaffold belong has been improved. It is believed that ProteinMPNN is extremely suitable for enhancing solubility in these 21 scaffolds (Neocarzinostatin based binder, scFv, Human VH dAb, Anticalin, OBody, Megabody, Nanobody, Diabody, CI2-based binder, Repebody, VL dAb, SH2 domain, Fab, Abdulin, Affilin, Alphabody, Evibody, Glubody, I-body, Transferrin based binder, and WW domain). Meanwhile, it should be noted that these 21 scaffolds only play a significant role in improving solubility, and does not represent poor solubility of amino acid sequences generated by other scaffolds. For example, the average solubility values of all amino acid sequences belonging to the 6 scaffolds Avimer, Cytochrome b562 based binder, dArmRP, GCN4 based binder, Im9 based binder, and PHD finger domain with a proportion of 0 in Fig. 3(b) are 0.77, 0.83, 0.80, 0.88, 0.76, and 0.71 inFig. 3(c), respectively, which still have a relatively high level of solubility.

In terms of average value and P_{SL}, ProteinMPNN performs well in both solubility and improving solubility. The reason for the good solubility of the designed amino acid sequence may be that most of the proteins used for DL are soluble ones. This is because the training ProteinMPNN protein structure was obtained from the PDB [24], and most protein structures in PDB resolved by NMR, X-ray, and (or) cryo-EM experiments [46] are soluble. Additionally, experimental validation has demonstrated that most proteins designed by ProteinMPNN are soluble [23].

3.4.2 Comparison of solubility between sequences designed from monomer and complex SBPs

In Fig. 4(a), the P_{SL} well demonstrated the difference in the number of amino acid sequences with enhanced solubility between sequences designed from monomer and complex SBPs. As shown, the amount of solubility improvement of amino acid sequences based on monomer design is slightly

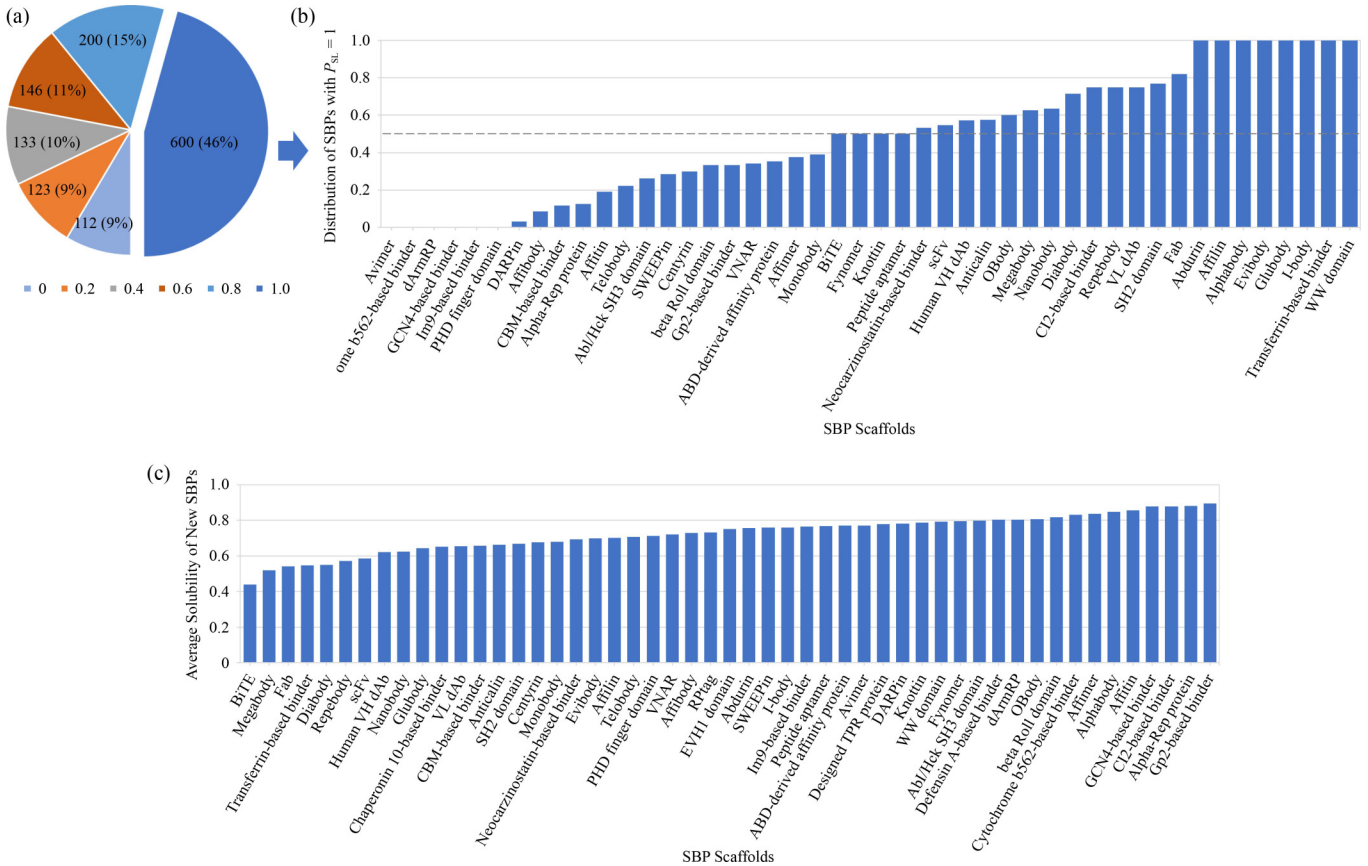


Fig. 3 Analysis of the solubility of synthetic proteins designed by ProteinMPNN. (a) Quantity distribution of different proportion of increased solubility (P_{SL}); (b) the distribution of SBPs with P_{SL} of 1 in the corresponding SBP scaffolds; (c) average solubility of all amino acid sequences under the scaffold

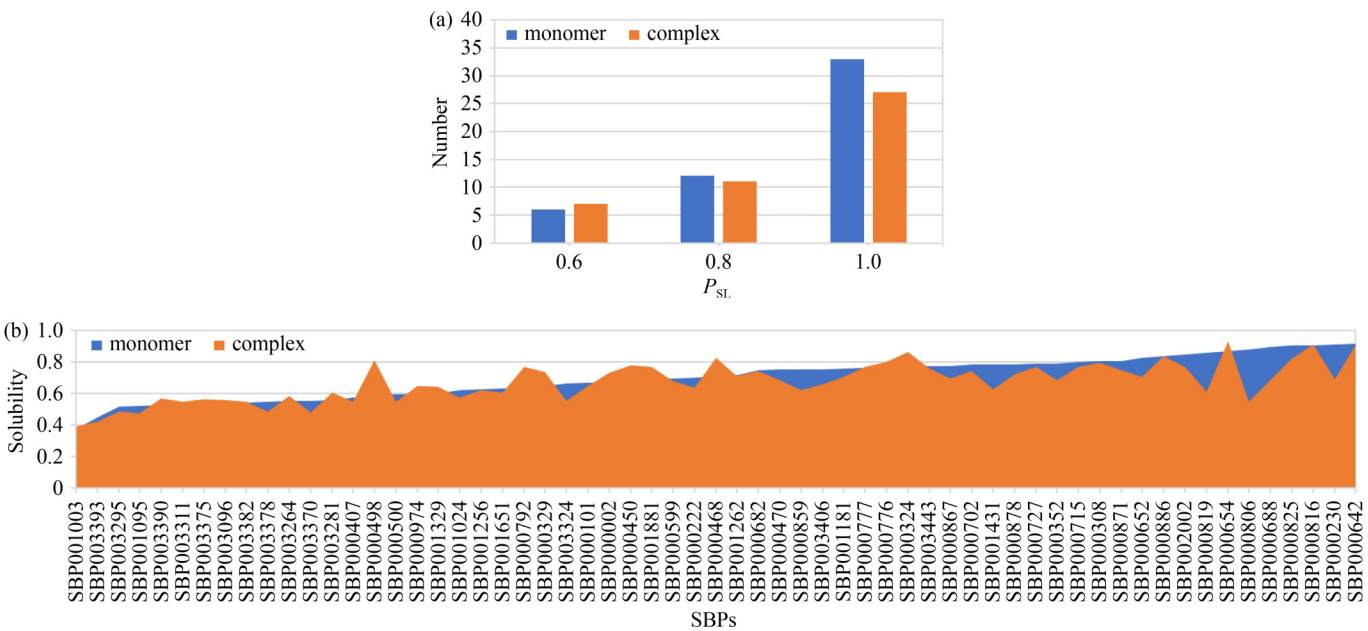


Fig. 4 Comparison of the solubility of synthetic proteins designed by ProteinMPNN original SBPs in monomer and complex forms. (a) The difference in the number of amino acid sequences with enhanced solubility between two designs; (b) comparison of average values of solubility between two designs

higher than that based on complex design when the value of P_{SL} > 0.6 (Fig. 4(a)). In addition, the value of average solubility of amino acid sequences obtained based on monomer design (62%) is also slightly higher than the average solubility of amino acid sequences obtained based on complex design (Fig. 4(b)), the differences are statistically significant (P-value = 0.031). One possible reason is that ProteinMPNN was trained on both single-chain and multi-chain proteins

[23]. When designing protein sequences, it considers the foldability of the sequences and their compatibility with the target structure. When designing a protein based on monomers, the model can focus more on optimizing the protein's own properties, including its solubility. However, in complex design, the model needs to consider both the interactions between the protein and the target, which may sacrifice some of the protein's solubility to ensure the overall stability of the complex.

3.5 Improved thermal stability of the designed synthetic proteins

Stability is another important property of proteins in production, and the derivation of new protein functions during evolution depends heavily on the stability of the proteins [47]. Consistent with solubility analysis, to explore the probability of improved stability of the designed amino acid sequence compared to the original SBPs, proportion of increased stability (P_{ST}) is defined here. P_{ST} is the number of design protein sequences with stability higher than that of the original SBPs (N_h) divided by the total number of design amino acid sequences (N_t = 5). For each original SBP, if the P_{ST} is equal or greater than 0.6, the stability of at least 3 out of the 5 designed amino acid sequences is improved compared to the original SBPs. The predicted instability index for each designed protein sequence and the original SBP can be found in **Appendix A**.

3.5.1 Stability analysis of the sequences designed from monomer SBPs

Figure 5(a) shows the quantity distribution of different P_{ST}, and it can be concluded that most of the designed amino acid sequences (refers to 62% of the original SBPs) show improved stability. Particularly, the stability of the amino acid sequences designed from 38% of the original SBPs was all improved (Fig. 5(a)), demonstrating that ProteinMPNN has significant advantages in improving the stability of SBPs. Because it has been experimentally validated that the majority of proteins generated by ProteinMPNN exhibit high thermal stability [23]. Meanwhile, ProteinMPNN-generated sequences are predicted to fold to native protein backbones more confidently and accurately than the original native sequences, while natural proteins are generally stable in structure, most sequences generated by ProteinMPNN have improved stability.

In addition, the distribution of SBPs with P_{ST} of 1 in the corresponding SBP scaffolds is also plotted. As shown in

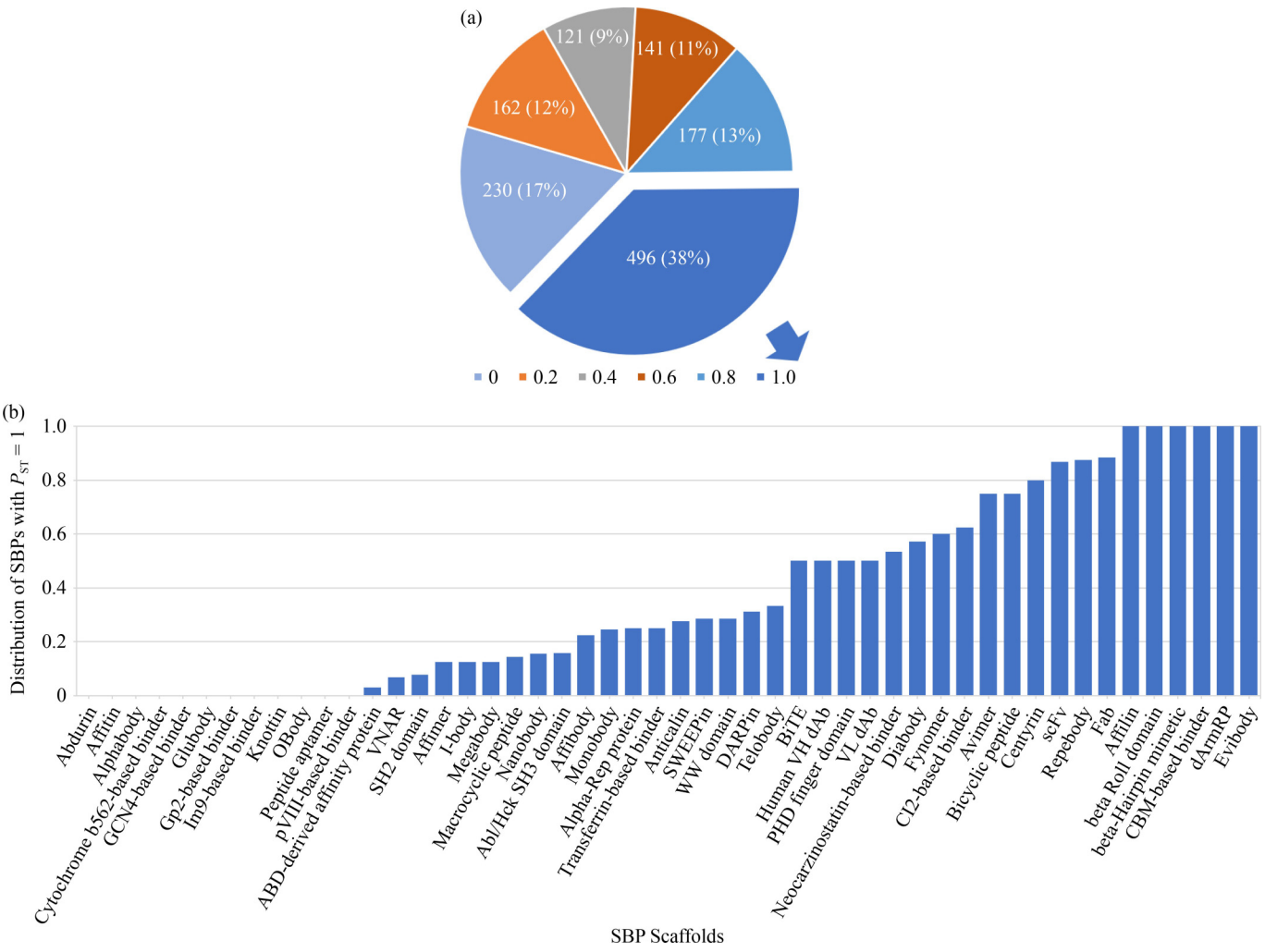


Fig. 5 Analysis of the stability of synthetic proteins designed by ProteinMPNN. (a) Quantity distribution of different proportion of increased stability (P_{ST}); (b) the distribution of SBPs with P_{ST} of 1 in the corresponding SBP scaffolds

Fig. 5(b), when the proportion of those with P_{ST} of 1 in the total number of original SBPs corresponding to the same scaffolds exceeds 0.5, it means that the stability of the amino acid sequences designed by most of the original SBPs to which the scaffold belong has been improved. It is believed that ProteinMPNN [23] is extremely suitable for enhancing stability in these 16 scaffolds (Neocarzinostatin based binder, Body, Fynomer, CI2 based binder, Avimer, Bicyclic peptide, Centyrin, scFv, Repebody, Fab, CBM based binder, Affilin, beta Roll domain, dArmRP, Evibody, beta Airpin mimetic).

Among the 12 scaffolds with a ratio of 0, 10 scaffolds only contain 2−5 original synthetic binding proteins, and contingency may be the reason why there is no proportion of increased stability of 1. The reason for other scaffold ratios of 0 or lower may be due to the important role of certain amino acids in protein stability, but the design of ProteinMPNN has modified it, such as the group IV WW domain, Pin1 in WW domain anti VEGFR-2 clone B1 [48] and disulfide bonds in Nanobody anti VSG cAbAn33-1-1-1 [49]. At the same time, after calculation, it was found that the variance of the stability of amino acid sequences in most groups is too large, indicating that the difference in stability between each group of amino acid sequences is too large. When using it, attention should be paid to selecting.

3.5.2 Comparison of stability between sequences designed from monomer and complex SBPs

In Fig. 6(a), the P_{ST} well demonstrated the difference in the number of amino acid sequences with enhanced stability between sequences designed from monomer and complex SBPs. As shown, the amount of stability improvement of amino acid sequences based on monomer design is slightly higher than that based on complex design when the value of $P_{ST} > 0.6$ (Fig. 6(a)). In addition, the value of average stability of amino acid sequences obtained based on monomer design (57%) is also slightly greater than the average stability of amino acid sequences obtained based on complex design (Fig. 6(b)), differences are statistically significant (P-value = 0.014). The reason for this may be similar to the relatively higher solubility observed.

3.6 Eight incredibly suitable protein scaffolds for ProteinMPNN have been identified

The scaffolds that show improvement in both solubility and stability are considered highly suitable for ProteinMPNN. They are Neocarzinostatin-based binder, Diabody, CI2-based binder, scFv, Repebody, Fab, Affilin, and Evibody. Among these, Diabody, scFv, and Fab belong to the antibody scaffolds, while the remaining five scaffolds belong to the non-antibody scaffolds [2].

Affilin scaffold comes in γ-B Crystallin or Ubiquitin [50]. Chymotrypsin inhibitor 2 (CI2) is a protease inhibitor naturally present in the seeds of barley [51]. Diabodies are small antibody fragments that have two antigen binding Fv domains [52]. Evibody is based on human cytotoxic-associated Antigen (CTLA-4) [53]. Fab refers to fragment antigen binding, which are smaller variants of monoclonal antibodies consisting of one VL and VH domain linked to their respective light and heavy constant domains [54]. Neocarzinostatin belongs to the family of bacterial chromoproteins [55]. Repebody is developed from Leucine-rich repeat (LRR) modules of variable lymphocyte receptors (VLRs) [56]. Single-chain variable fragment (scFv) consists of one VL and VH domain fused by a flexible linker [54]. Additional details of the 8 protein scaffolds are shown in Table 2.

3.7 New synthetic proteins designed from complex structures shown better binding energy

To investigate the potential of designed sequences binding to

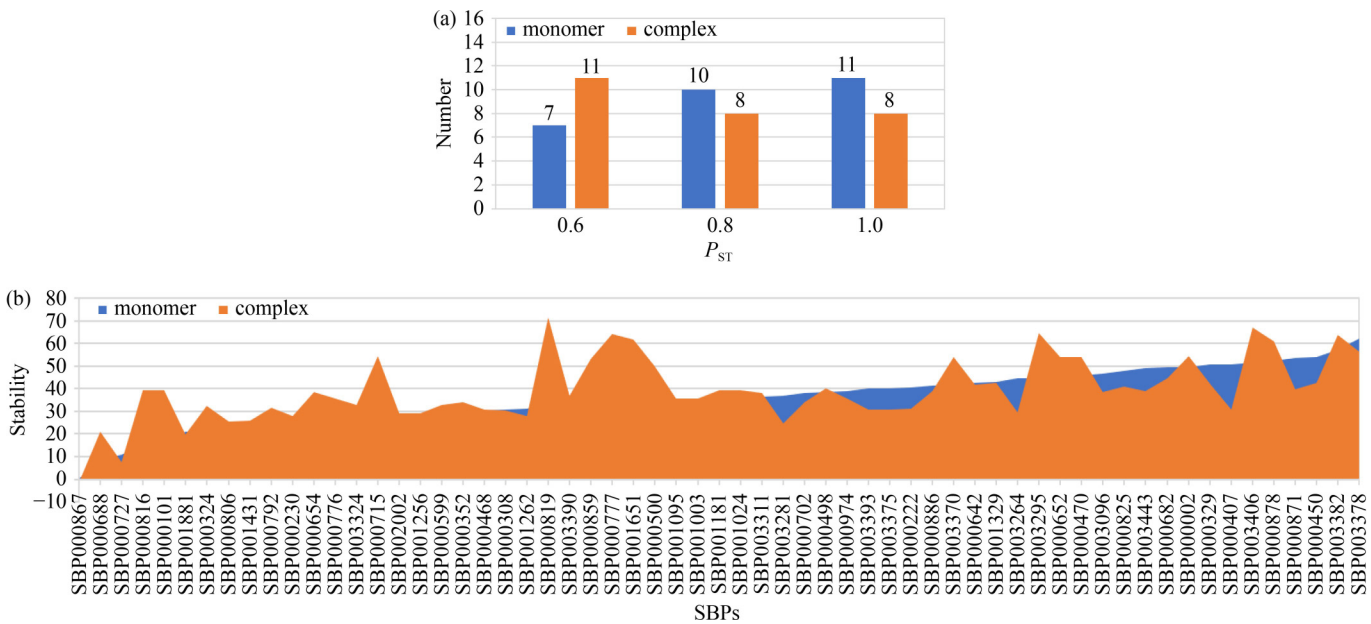


Fig. 6 Comparison of the stability of synthetic proteins designed by ProteinMPNN original SBPs in monomer and complex forms. (a) The difference in the number of amino acid sequences with enhanced stability between two designs; (b) comparison of average values of stability between two designs

Table 2 Details of 8 protein scaffolds applicable to ProteinMPNN from the SYNBIP database

Scaffold name	Thermal denaturation temperature range	Molecular weight range	Fold type
Affilin	56−72 ℃	17−22 kDa	Beta-Sheets + Beta-Turns + Loops
Diabody	25−55 ℃	21−60 kDa	Beta-Sheets + Loops
scFv	59−65 ℃	14−40 kDa	Beta-Sheets + Loops
Neocarzinostatin-based binder	51−57 ℃	10−14 kDa	Beta-Sheets + Loops
Evibody	90 ℃	12−16 kDa	Beta-Sheets + Loops
Fab	60 ℃	50−57 kDa	Beta-Sheets + Loops
Repebody	82 ℃	27−40 kDa	Alpha-Helices + Beta-Sheets + Loops
CI2-based binder	–	7 kDa	One Alpha-Helix + Beta-Sheets + Loops

the protein targets of their original SBPs, protein structure prediction and protein-protein interaction analysis were further conducted. First, based on 61 original SBPs in monomer and complex forms, 140 sequences from 14 sets of high-quality complex structures and 94 sequences with the lowest ProteinMPNN score (Lower score is better, the score represents model's uncertainty about the predictions) from ProteinMPNN design [23] were selected for protein structure prediction using AlphaFold2 [41]. Then, each new SBP was docked to the binding site on the protein target of the original SBP and with the binding energy estimated by PDBePISA [39].

The detailed information for protein-protein interaction analysis was provided in **Appendix B**. Compared to the original SBPs and their protein target interactions, 49.18% (30/61) of the estimated binding energy for new SBPs was increased in the complex based design group, while only 27.87% (17/61) of that in the monomer-based design group. Meanwhile, 67.21% (41/61) of the estimated binding energy for new SBPs in the complex-based design group was higher than that in the monomer-based design group, differences are statistically significant (P-value < 0.001). As shown in Fig. 7, the majority of the new SBPs in the complex based design group have potential hydrogen bond interactions (85%) and salt bridges (52%). Additionally, some complexes also exhibit potential disulfide bonds (3%). For example, in the interaction interface between SBP001262 and its target, hydrogen bond interactions, and salt bridges were observed (Fig. 8). Half of the estimated binding energy improvement of the complex based design group (49.18%) may be due to the consideration of protein-protein interaction in ProteinMPNN for protein function design [23].

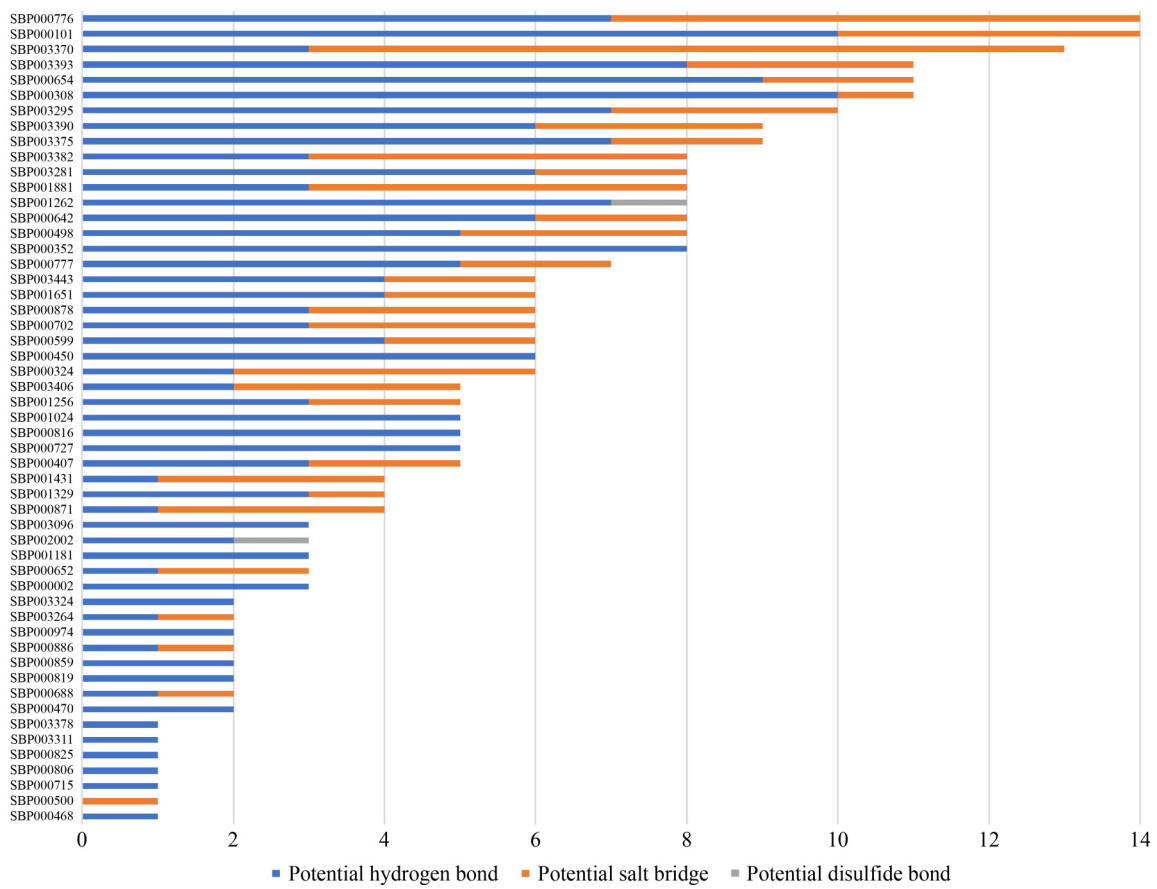


Fig. 7 Statistics of interaction types between SBPs based on complexes and their protein targets. The potential hydrogen bonds, salt bridges, and disulfide bonds are shown in blue, orange, and gray representations, respectively

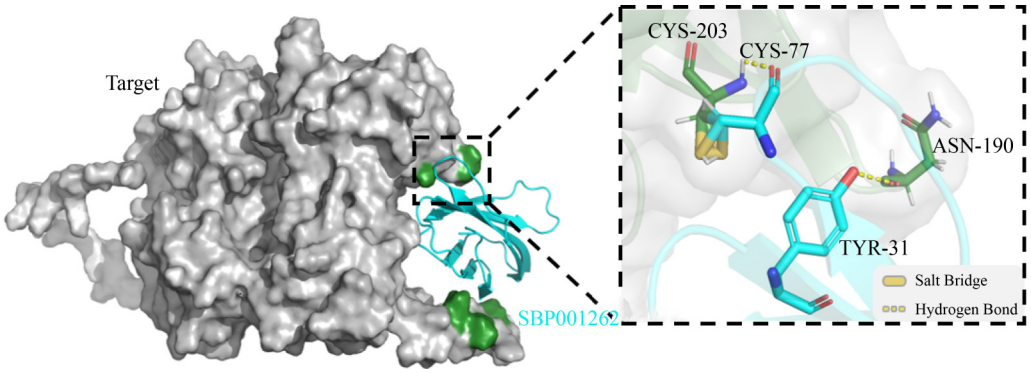


Fig. 8 Analysis of the interaction interface between SBP001262 and its target protein

In addition, a small part of the estimated binding energy of the monomer based design group (27.87%) is increased, which further verify the rationality of the development of new SBPs from privileged scaffolds [57]. ProteinMPNN has experimentally demonstrated that the designed proteins exhibit superior binding capacity [23], moreover, the enhanced binding energy of new proteins designed based on monomer and (or) complex structure of their original SBPs shown great potential of DL-based formwork like ProteinMPNN in target-directed *de novo* generation of SBP library with high quality, which could assist experimental biologists to rational protein engineering to discover functional binders.

3.8 ProteinMPNN demonstrates tremendous potential in designing SBPs compared to traditional protein engineering methods

Among the SBPs with crystal structures corresponding to their respective targets, Anticalins N7E and N9B [58] were selected as representatives of traditional protein engineering. They were designed from the same template protein, Lipocalin 2. For comparative analysis, ProteinMPNN was used to design the same template protein, following the same design and analysis process as with other SBPs.

From Table 3, it is evident to see that for the weaker affinity Anticalins N9B, 80% (4/5) the proteins designed by ProteinMPNN showed improved binding affinity. Compared to Anticalins N7E, one of the five designed proteins (ProntеinMPNN-5) exhibited better binding affinity. In contrast to the predicted stability, the solubility of the designed proteins is largely improved. Although ProntеinMPNN-5 demonstrated outstanding binding affinity, there was no improvement on the predicted solubility and stability. Therefore, it is still a great challenge for achieving good solubility and stability alongside strong binding affinity by ProteinMPNN. In addition, the average sequence recovery of 5 sequences designed from Lipocalin 2 is 46.4%. There are few sequences similar to the 5 sequences in the PDB database (**Appendix B**). More interestingly, a disulfide bond was discovered in the template protein Lipocalin 2. ProteinMPNN-5 not only folds into a structurally consistent form with substantial sequence changes but also retains the disulfide bond (Fig. 9).

It is evident that compared to traditional protein engineering methods which may only yield a few or a dozen active proteins from a large library, ProteinMPNN shows significant potential by exhibiting outstanding performance in the 5 designed proteins.

4 Conclusion

Using the advanced DL algorithm ProteinMPNN, 7,245 synthetic protein sequences were generated based on the original SBPs collected in SYNBIP. Post analysis of the designed sequences demonstrating excellent performance sequence recovery, solubility and stability. 8 scaffolds that have greatly improved solubility and stability include Neocarzinostatin based binder, Diabody, CI2 based binder, scFv, Repebody, Fab, Affilin, and Evibody. In addition, the sequence generated based on monomer is better than that generated based on complex in terms of solubility and stability, while it is the opposite for their binding energies to target. ProteinMPNN has some limitations in designing SBPs. For example, in the design of SBPs with short sequences (< 20), repeated sequences may appear among the 5 designed sequences. Additionally, there may be occurrences of long stretches of alanine (A), which is not consistent with expectations. This may be due to uncertainty in the model

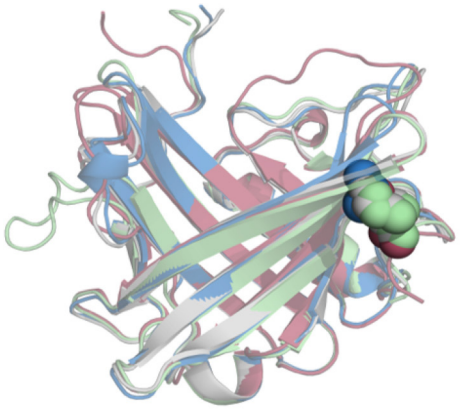
Table 3 Comparison of the properties of ProteinMPNN designed proteins with those obtained by conventional protein engineering

Protein name	K_d (nM)	Predected ΔiG	Predected solubility	Instability index
Anticalins N7E	7.18 ± 0.12	−12	0.535	33.65
Anticalins N9B	39.9 ± 1.0	−6.9	0.622	29.71
ProteinMPNN-1	–	−6	0.593*	35.54
ProteinMPNN-2	–	−9.5*	0.644**	31.78*
ProteinMPNN-3	–	−8*	0.678**	37.88
ProteinMPNN-4	–	−11.1*	0.681**	34.96
ProteinMPNN-5	–	−13.1**	0.527	44.85

* Represents the property exceeding that of a protein obtained by traditional methods.

** Represents the property exceeding that of two proteins obtained by traditional methods.

(a)



(b)

```
Lipocalin       QDSTSDLIPAPPLSKVPLQQNFQDNQFQGKWYVVGLAGNAILREDKDPQKMYATIYELKE 60
Anticalin N7E   QDSTSDLIPAPPLSKVPLQQNFQDNQFHGKWYVVGEAGNSLLREDKDPRKMYATIYELKE 60
Anticalin N9B   QDSTSDLIPAPPLSKVPLQQNFQDNQFHGKWYVVGRAGNMRLREDKDPAKMVATIYELKE 60
ProteinMPNN-5   ---SGSLIPAPPRSKVPLQPNFDPKKFEGKWYVVGLAGNTVFRDEENPLLMSSITFTLRP 57

Lipocalin       DKSYNVTSVLFRKKKCDYWIRTFVPGCQPGEFTLGNIKSYPGLTSYLVRVVSTNYNQHAM 120
Anticalin N7E   DKSYNVTSVRFRSKKCHYLIRTFVPGSQPGEFTLGLIKSKPGHTSFLVRVVSTNYNQHAM 120
Anticalin N9B   DKSYNVTKVMFQRKKCKYMINTFVPGSQPGEFTLGSIKSWPGLTSWLVRVVSTNYNQHAM 120
ProteinMPNN-5   DGTFDVVTRAWINGKCKVVPAVLVPGAEPGTFTLGGLDRYPGLTRYEVTVVETDYDNYAF 117

Lipocalin       VFFKKVSQNREYFKITLYGRTKELTSELKENFIRFSKSLGLPENHIVFPVPIDQCIDG   178
Anticalin N7E   VFFKTVAQNREYFFITLYGRTKELTSELKENFIRFSKSLGLPENHIVFPVPIDQCIDG   178
Anticalin N9B   VFFKRVYQNREFFGITLYGRTKELTSELKENFIRFSKSLGLPENHIVFPVPIDQCIDG   178
ProteinMPNN-5   VYERWVWGGREHWALLLLSKTQSVSEEVKNRFVELAKSLGLPEENIVFPIPITECVN-   174
```

Fig. 9 Analysis of Lipocalin 2, Anticalins N7E, N9B, and ProteinMPNN-5 in terms of both structure and sequence. (a) Structural comparison of four proteins. The RMSD values of Anticalins N7E (green), N9B (red), and ProteinMPNN-5 (blue) relative to Lipocalin 2 (gray) were 0.467 Å, 0.561 Å, and 0.470 Å, respectively. Disulfide bonds are demonstrated with spheres. The PDBIDs for Lipocalin 2, Anticalins N7E, and N9B are 1DFV, 5N47, and 5N48, respectively; (b) sequence comparison of the four proteins. The amino acids highlighted in red are different from Lipocalin 2, while the positions shaded in yellow indicate the locations of disulfide bonds

predictions or poor quality of input backbone sequences. To address this issue, negative alanine bias can be added. Overall, the result prove that ProteinMPNN can quickly generate high-quality libraries of synthetic proteins for experimental screen of specific target, which can provide scientists with a wide range of opportunities to explore the world of synthetic proteins.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 21505009), the Entrepreneurship and Innovation Support Plan for Chinese Overseas Students of Chongqing (cx2020127).

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

Electronic supplementary material Supplementary material is available in the online version of this article at journal.hep.com.cn and link.springer.com

Appendixes

Appendix A

Statistics of the 7,245 generated synthetic proteins according to protein scaffolds of their original SBPs.

Appendix B

Detailed information for protein-protein interaction analysis.

References

1. Gebauer M, Skerra A. Engineered protein scaffolds as next-generation therapeutics. Annual Review of Pharmacology and Toxicology, 2020, 60: 391−415
2. Wang X, Li F, Qiu W, Xu B, Li Y, Lian X, Yu H, Zhang Z, Wang J, Li Z, Xue W, Zhu F. SYNBIP: synthetic binding proteins for research, diagnosis and therapy. Nucleic Acids Research, 2022, 50(D1): D560−D570
3. Huang P S, Boyken S E, Baker D. The coming of age of *de novo* protein design. Nature, 2016, 537(7620): 320−327
4. Carpenter E P, Beis K, Cameron A D, Iwata S. Overcoming the challenges of membrane protein crystallography. Current Opinion in Structural Biology, 2008, 18(5): 581−586
5. Zeymer C, Hilvert D. Directed evolution of protein catalysts. Annual Review of Biochemistry, 2018, 87: 131−157
6. Engqvist M K M, Rabe K S. Applications of protein engineering and directed evolution in plant research. Plant Physiology, 2019, 179(3): 907−917
7. Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park J S, Jude K M, Marković I, Kadam R U, Verschueren K H G, Verstraete K, Walsh S T R, Bennett N, Phal A, Yang A, Kozodoy L, DeWitt M, Picton L, Miller L, Strauch E M, DeBouver N D, Pires A, Bera A K, Halabiya S, Hammerson B, Yang W, Bernard S, Stewart L, Wilson I A, Ruohola-Baker H, Schlessinger J, Lee S, Savvides S N, Garcia K C, Baker D. Design of protein-binding proteins from the target structure alone. Nature, 2022, 605(7910): 551−560
8. Baker D. What has de novo protein design taught us about protein folding and biophysics? Protein Science, 2019, 28(4): 678−683
9. Liang T, Jiang C, Yuan J, Othman Y, Xie X Q, Feng Z. Differential performance of RoseTTAFold in antibody modeling. Briefings in Bioinformatics, 2022, 23(5): bbac152

10. Chen W, Qian G, Wan Y, Chen D, Zhou X, Yuan W, Duan X. Mesokinetics as a tool bridging the microscopic-to-macroscopic transition to rationalize catalyst design. Accounts of Chemical Research, 2022, 55(22): 3230−3241
11. Chen W, Fu W, Duan X, Chen B, Qian G, Si R, Zhou X, Yuan W, Chen D. Taming electrons in Pt/C catalysts to boost the mesokinetics of hydrogen production. Engineering, 2022, 14: 124−133
12. Liang T, Chen H, Yuan J, Jiang C, Hao Y, Wang Y, Feng Z, Xie X Q. IsAb: a computational protocol for antibody design. Briefings in Bioinformatics, 2021, 22(5): bbab143
13. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology, 2019, 20(11): 681−697
14. Khakzad H, Igashov I, Schneuing A, Goverde C, Bronstein M, Correia B. A new age in protein design empowered by deep learning. Cell Systems, 2023, 14(11): 925−939
15. Wang F, Feng X, Kong R, Chang S. Generating new protein sequences by using dense network and attention mechanism. Mathematical Biosciences and Engineering, 2023, 20(2): 4178−4197
16. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim P M. Fast and flexible protein design using deep graph neural networks. Cell Systems, 2020, 11(4): 402−411.e4
17. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics, 2022, 38(8): 2102−2110
18. Anishchenko I, Pellock S J, Chidyausiku T M, Ramelot T A, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera A K, Dimaio F, Carter L, Chow C M, Montelione G T, Baker D. De novo protein design by deep network hallucination. Nature, 2021, 600(7889): 547−552
19. Yeh A H W, Norn C, Kipnis Y, Tischer D, Pellock S J, Evans D, Ma P, Lee G R, Zhang J Z, Anishchenko I, Coventry B, Cao L, Dauparas J, Halabiya S, DeWitt M, Carter L, Houk K N, Baker D. De novo design of luciferases using deep learning. Nature, 2023, 614(7949): 774−780
20. Ding W, Nakai K, Gong H. Protein design via deep learning. Briefings in Bioinformatics, 2022, 23(3): bbac102
21. Lin E, Lin C H, Lane H Y. De novo peptide and protein design using generative adversarial networks: an update. Journal of Chemical Information and Modeling, 2022, 62(4): 761−774
22. Yin R, Feng B Y, Varshney A, Pierce B G. Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants. Protein Science, 2022, 31(8): e4379
23. 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, Baker D. Robust deep learning–based protein sequence design using ProteinMPNN. Science, 2022, 378(6615): 49−56
24. Burley S K, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, Craig P A, Crichlow G V, Dalenberg K, Duarte J M, Dutta S, Fayazi M, Feng Z, Flatt J W, Ganesan S, Ghosh S, Goodsell D S, Green R K, Guranovic V, Henry J, Hudson B P, Khokhriakov I, Lawson C L, Liang Y, Lowe R, Peisach E, Persikova I, Piehl D W, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Vallat B, Voigt M, Webb B, Westbrook J D, Whetstone S, Young J Y, Zalevsky A, Zardecki C. RCSB protein data bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Research, 2023, 51(D1): D488−D508
25. Bennett N R, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, Peng Y P, Dauparas J, Baek M, Stewart L, Dimaio F, De Munck S, Savvides S N, Baker D. Improving de novo protein binder design with deep learning. Nature Communications, 2023, 14(1): 2625
26. Sequeiros-Borja C E, Surpeta B, Brezovsky J. Recent advances in user-friendly computational tools to engineer protein function. Briefings in Bioinformatics, 2021, 22(3): bbaa150
27. Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, Anishchenko I, Baker D, Yang J. The trRosetta server for fast and accurate protein structure prediction. Nature Protocols, 2021, 16(12): 5634−5651
28. Cortajarena A L, Kajander T, Pan W, Cocco M J, Regan L. Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. Protein Engineering, Design and Selection, 2004, 17(4): 399−409
29. Mijit A, Wang X, Li Y, Xu H, Chen Y, Xue W. Mapping synthetic binding proteins epitopes on diverse protein targets by protein structure prediction and protein-protein docking. Computers in Biology and Medicine, 2023, 163: 107183
30. Liu Y, Liu H. Protein sequence design on given backbones with deep learning. Protein Engineering, Design and Selection, 2024, 37: gzad024
31. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, 2017, 35(11): 1026−1028
32. Pierleoni A, Indio V, Savojardo C, Fariselli P, Martelli P L, Casadio R. MemPype: a pipeline for the annotation of eukaryotic membrane proteins. Nucleic Acids Research, 2011, 39(S2): W375−W380
33. Altschul S F, Gish W, Miller W, Myers E W, Lipman D J. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403−410
34. Hebditch M, Carballo-Amador M A, Charonis S, Curtis R, Warwicker J. Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics, 2017, 33(19): 3098−3100
35. Niwa T, Ying B W, Saito K, Jin W, Takada S, Ueda T, Taguchi H. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. Proceedings of the National Academy of Sciences of the United States of America, 2009, 106(11): 4201−4206
36. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M R, Appel R D, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker J M, ed. The Proteomics Protocols Handbook. Totowa: Humana, 2005, 571−607
37. Chen C, Chen H, Zhang Y, Thomas H R, Frank M H, He Y, Xia R. TBtools: an integrative toolkit developed for interactive analyses of big biological data. Molecular Plant, 2020, 13(8): 1194−1202
38. Lill M A, Danielson M L. Computer-aided drug design platform using PyMOL. Journal of Computer-Aided Molecular Design, 2011, 25(1): 13−19
39. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. Journal of Molecular Biology, 2007, 372(3): 774−797
40. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences of the United States of America, 2000, 97(19): 10383−10388
41. 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, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583−589
42. Wright C F, Teichmann S A, Clarke J, Dobson C M. The importance of sequence diversity in the aggregation and evolution of proteins. Nature, 2005, 438(7069): 878−881
43. Kramer R M, Shende V R, Motl N, Pace C N, Scholtz J M. Toward a molecular understanding of protein solubility: increased negative

surface charge correlates with increased solubility. Biophysical Journal, 2012, 102(8): 1907−1915
44. Navarro S, Ventura S. Computational re-design of protein structures to improve solubility. Expert Opinion on Drug Discovery, 2019, 14(10): 1077−1088
45. Smialowski P, Martin-Galiano A J, Mikolajka A, Girschick T, Holak T A, Frishman D. Protein solubility: sequence based prediction and experimental verification. Bioinformatics, 2007, 23(19): 2536−2542
46. Burley S K. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. Journal of Biological Chemistry, 2021, 296: 100559
47. Qing R, Hao S, Smorodina E, Jin D, Zalevsky A, Zhang S. Protein design: from the aspect of water solubility and stability. Chemical Reviews, 2022, 122(18): 14085−14179
48. Patel S, Mathonet P, Jaulent A M, Ullman C G. Selection of a high-affinity WW domain against the extracellular region of VEGF receptor isoform-2 from a combinatorial library using CIS display. Protein Engineering, Design and Selection, 2013, 26(4): 307−315
49. Saerens D, Conrath K, Govaert J, Muyldermans S. Disulfide bond introduction for general stabilization of immunoglobulin heavy-chain variable domains. Journal of Molecular Biology, 2008, 377(2): 478−488
50. Reverdatto S, Burz D S, Shekhtman A. Peptide aptamers: development and applications. Current Topics in Medicinal Chemistry, 2015, 15(12): 1082−1101
51. Karlsson G B, Jensen A, Stevenson L F, Woods Y L, Lane D P, Sørensen M S. Activation of p53 by scaffold-stabilised expression of Mdm2-binding peptides: visualisation of reporter gene induction at the single-cell level. British Journal of Cancer, 2004, 91(8): 1488−1494
52. Kwon N Y, Kim Y, Lee J O. Structural diversity and flexibility of diabodies. Methods, 2019, 154: 136−142
53. Hey T, Fiedler E, Rudolph R, Fiedler M. Artificial, non-antibody binding proteins for pharmaceutical and industrial applications. Trends in Biotechnology, 2005, 23(10): 514−522
54. Leenheer D, Ten Dijke P, Hipolito C J. A current perspective on applications of macrocyclic-peptide-based high-affinity ligands. Peptide Science, 2016, 106(6): 889−900
55. Nicaise M, Valerio-Lepiniec M, Minard P, Desmadril M. Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. Protein Science, 2004, 13(7): 1882−1891
56. Škrlec K, Štrukelj B, Berlec A. Non-immunoglobulin scaffolds: a focus on their targets. Trends in Biotechnology, 2015, 33(7): 408−418
57. Sandhya S, Mudgal R, Kumar G, Sowdhamini R, Srinivasan N. Protein sequence design and its applications. Current Opinion in Structural Biology, 2016, 37: 71−80
58. Gebauer M, Schiefner A, Matschiner G, Skerra A. Combinatorial design of an anticalin directed against the extra-domain b for the specific targeting of oncofetal fibronectin. Journal of Molecular Biology, 2013, 425(4): 780−802

Yanlin Li received her bachelor's degree in 2023 from the School of Pharmacy, Chongqing University, China. She is currently working toward a Master's degree at Chongqing University, China. Her research interests mainly include protein design and database construction.

Wantong Jiao is a class of 2021 undergraduate at the School of Pharmacy, Chongqing University, China. She is interested in the molecular mechanism of drug and target recognition and interaction as well as the research of new drug dosage forms.

Ruihan Liu is currently studying for a master's degree at Chongqing University in China. Her research interests mainly include database construction and structure-based drug design and screening.

Xuejin Deng received a bachelor's degree from Guizhou University, China. She is currently pursuing a master's degree at the School of Pharmacy, Chongqing University, China. Her research interest is protein design.

Feng Zhu is the Deputy Director of B&R International School of Medicine and a Distinguished Professor of Pharmaceutical Sciences at Zhejiang University, China. He obtained a bachelor's and master's degrees in Physics from Beijing Normal University, and a PhD in Pharmacy from the National University of Singapore. Based on artificial intelligence and OMIC (proteomics and metabolomics) technologies, their team conducts systematical exploration on the druggability and system profile of therapeutic targets, develops novel methods and online tools for target discovery, and further studies the mechanism underlying the interaction between drugs and their targets.

Weiwei Xue is an associate professor of Pharmaceutical Sciences at Chongqing University, China. He received a bachelor's degree in Chemistry (2009) and a PhD in Cheminformatics (2014) from Lanzhou University, China. He worked as a visiting scholar in the Institute for Protein Design at the University of Washington (2018−2019), USA. The research in Dr. Xue's Lab is focused on developing disease- and therapeutic-related bioinformatics databases and tools, and combing artificial intelligence and molecular modeling approaches to design innovative small molecules or protein binders against molecular targets of complex diseases, including psychiatric disorders, viral infection, and cancer. He has published more than 90 peer-reviewed papers in the area of bioinformatics and computational drug design.