Nucleic Acids Research, 2025, **53**, D595–D603
https://doi.org/10.1093/nar/gkae893
Advance access publication date: 16 October 2024
Database issue

SYNBIP 2.0: epitopes mapping, sequence expansion and scaffolds discovery for synthetic binding protein innovation

Yanlin Li[1,†], Fengcheng Li[2,3,†], Zixin Duan[1], Ruihan Liu[1], Wantong Jiao[1], Haibo Wu[4], Feng Zhu[3,*] and Weiwei Xue[1,*]

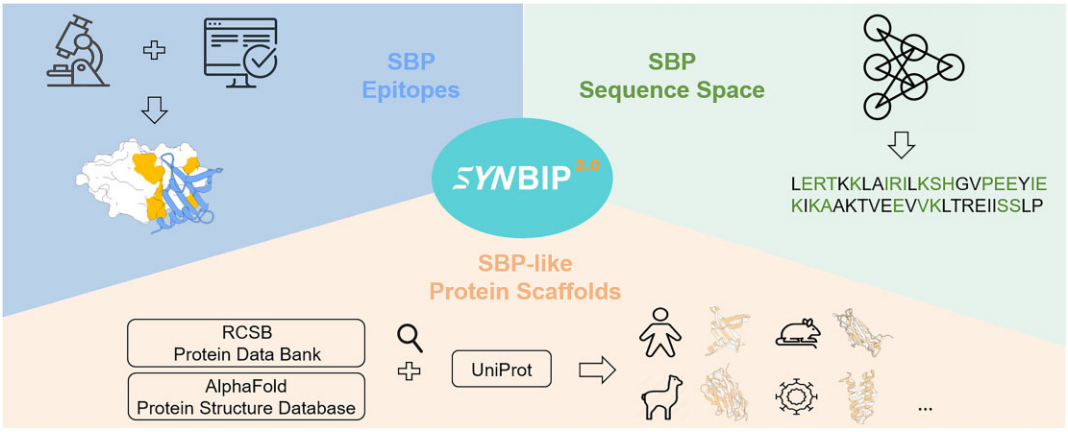
[1]Chongqing Key Laboratory of Natural Product Synthesis and Drug Research, School of Pharmaceutical Sciences, Chongqing University, No. 55 South University Town Road, High-tech Zone, Chongqing 401331, China
[2]Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center for Child Health, 3333 Binsheng Road, Hangzhou, Zhejiang 310052, China
[3]College of Pharmaceutical Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou, Zhejiang 310058, China
[4]School of Life Sciences, Chongqing University, No. 55 South University Town Road, High-tech Zone, Chongqing 401331, China

[*]To whom correspondence should be addressed. Tel: +86 187 0236 4293; Fax: +86 023 6567 8450; Email: xueww@cqu.edu.cn
Correspondence may also be addressed to Feng Zhu. Email: zhufeng@zju.edu.cn
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Synthetic binding proteins (SBPs) represent a pivotal class of artificially engineered proteins, meticulously crafted to exhibit targeted binding properties and specific functions. Here, the SYNBIP database, a comprehensive resource for SBPs, has been significantly updated. These enhancements include (i) featuring 3D structures of 899 SBP–target complexes to illustrate the binding epitopes of SBPs, (ii) using the structures of SBPs in the monomer or complex forms with target proteins, their sequence space has been expanded five times to 12 025 by integrating a structure-based protein generation framework and a protein property prediction tool, (iii) offering detailed information on 78 473 newly identified SBP-like scaffolds from the RCSB Protein Data Bank, and an additional 16 401 555 ones from the AlphaFold Protein Structure Database, and (iv) the database is regularly updated, incorporating 153 new SBPs. Furthermore, the structural models of all SBPs have been enhanced through the application of the AlphaFold2, with their clinical statuses concurrently refreshed. Additionally, the design methods employed for each SBP are now prominently featured in the database. In sum, SYNBIP 2.0 is designed to provide researchers with essential SBP data, facilitating their innovation in research, diagnosis and therapy. SYNBIP 2.0 is now freely accessible at https://idrblab.org/synbip/.

Graphical abstract



Introduction

Synthetic binding proteins (SBPs) are a class of artificially designed protein binders with specific functions, which have broad applications in research, diagnosis and therapy. Compared with classical antibodies, the SBPs are smaller, and most of them are more stable, less immunogenic and better of tissue penetration (1). As an alternative to classical antibodies, the key data of SBP (e.g. protein scaffold type, sequence information and biophysical property) are essential for the next-generation protein development (2). The SYNBIP database

Received: August 19, 2024. **Revised:** September 18, 2024. **Editorial Decision:** September 25, 2024. **Accepted:** September 26, 2024

(1) containing 68 diverse SBP scaffolds was, therefore, constructed and has rapidly emerged as an indispensable complement to other SBP-related databases (3–11) in aiding protein engineering (12–16), protein structure prediction (17–20), protein function annotation (21–27), proteomic research (28–32) and so on (33–39).

In recent years, both physics-based method (40) and deep learning (41,42) have achieved notable advancements in the design of SBPs (43–49), while the ongoing challenge of acquiring high-quality, diverse and representative datasets persists, thereby limiting the full potential of those methods (50). First of all, the structural understanding of SBP in complex with target protein, which encompasses vital details of the interaction interface. This knowledge is crucial for enhancing the precision and efficiency of SBP design (44,51–53). Additionally, the concept of expanding the sequence space of SBPs has emerged as a cutting-edge trend in synthetic biology, aimed at uncovering novel protein binders with tailored functional capabilities (54–57). It is imperative to effectively construct high-quality *in silico* protein libraries, characterized by excellent solubility and stability, for subsequent experimental validation against target proteins (58–61). Furthermore, the remarkable progress of structure-based protein comparison algorithm (62) has made it technically viable to investigate SBP-like scaffolds across the scale of the known protein universe (63). This exploration can serve as a vital data repository for the development of SBPs, offering valuable starting points for protein engineering endeavors (64) and providing extensive datasets for the training of artificial intelligence (AI) models to generate innovative scaffolds (65).

To date, several active online repositories have been established to freely provide SBP-related data. The majority of them focused on providing comprehensive information on either sequences (e.g. ABCD (3) and sdAb-DB (6)), or structures (e.g. Thera-SAbDab (5) and PyIgClassify (7)), or both (e.g. Yvis (4)) of antibody and nanobody. Although esteemed databases such as STRING (8), BioGRID (9) and PDBbind (66) contain extensive records of biological interactions including protein–protein interactions, an initial analysis of the data indicates that only a small subset (∼8%) of SBPs in complex with target proteins are actually available. For the sequence space, it exhibits considerable variation in SYNBIP, with the number of SBPs underlying each scaffold varying from 1 (e.g. Chaperonin 10-based binder) to 271 (e.g. nanobody) (1). While the conventional method of directed evolution, which is typically employed to uncover new functions or enhance protein properties (65), appears to be inadequate for efficiently expanding the sequence space of SBPs such as Chaperonin 10-based binder. Furthermore, the number of scaffolds (e.g. 68 collected in SYNBIP (1)) is insufficient for training advanced AI models. In summary, all the existing online resources do not systematically describe those important data, which require for a major update of SYNBIP to provide the comprehensive information on describing the binding epitopes, expanded sequence space and new scaffolds of SBPs.

To fill in those gaps, a major update of SYNBIP was therefore performed in this work. (i) The binding epitopes of 870 representative SBPs covering 86.8% scaffolds and 74.1% target proteins collected in SYNBIP were mapped at atomic level by providing 899 structures of SBP–target complexes. (ii) The sequence space of SBPs was expanded by five times (12 025 sequences) through integrating the structure-based protein generation framework ProteinMPNN (41) and a protein property prediction tool based on monomer or complex structures. (iii) 78 473 and 16 401 555 protein domains with SBP-like scaffolds come from 54 736 different organisms were identified by Foldseek (62) from the RCSB Protein Data Bank (RCSB PDB) (67) and AlphaFold Protein Structure Database (AlphaFold DB) (63), respectively. (iv) Routine updates were performed on this database, with 153 new SBPs added, covering 25% (17/68) of the scaffolds, and all SBP structures have been meticulously updated utilizing AlphaFold2, complemented by the latest information on their clinical status and a showcase of their design methodologies. Therefore, the systematical data refer to this majority update of SYNBIP laid a solid foundation for rational protein engineering and design, which will facilitate discovery of novel functional-specific protein binders used in research, diagnosis and therapy. SYNBIP 2.0 is now freely accessible without any login requirement at: https://idrblab.org/synbip/.

Factual content and data retrieval

The binding epitopes of SBPs in complex with their target proteins

The binding epitopes determine the potency and specificity of the interactions between SBPs and their target proteins, thereby exerting a direct influence on their functional characteristics (68–72). In SYNBIP 2.0, we have curated a dataset of 899 SBP–target complex structures to present a comprehensive overview of binding epitope information between SBPs and their target proteins. Of the 899 complexes, 221 were experimentally determined SBP–target complexes obtained from the RCSB PDB (67), encompassing 27 distinct scaffolds and 116 different target proteins, respectively. Despite this, the information of SBPs' binding epitopes of remains insufficient, necessitating the application of molecular modeling techniques to construct more SBP–target complexes. While the experimental complexes can provide valuable information for the precise selection of binding regions, these data are essential during the complex's modeling process.

Accordingly, the remaining 678 SBP–target complexes covering 41 scaffolds, which were provided as a valuable supplementary resource for the experimental structures, were constructed through a rigorously validated computational framework (17). Details of this computational framework and flowchart (Supplementary Figure S1) are given in the Supplementary Data. Four representative crystal complexes with SBP from different scaffolds and binding targets were selected for redocking study in the Supplementary Data, demonstrating the applicability of this method in modeling SBP–target complexes (Supplementary Figure S2). The structural qualities of the 678 computationally modeled SBP–target complexes were thoroughly assessed using the RosettaDock criteria (73) as described in the Supplementary Data. Figure 1A illustrates that the models exhibit an overall satisfactory quality, with the distribution of high-quality, medium-quality and acceptable-quality complexes being 296 (44%), 314 (46%) and 68 (10%), respectively. This categorization underscores the robustness of our computational framework in generating reliable models for further analysis. Using the high-quality structure of Affibody anti-CTX-M-15 A5888 (SBP) in complex with β-lactamase (target) as an example (Figure 1B), the I_rmsd values for the top 5 models are 1.044, 0.993, 1.002, 1.013 and 0.982 Å (green plots in Figure 1B).

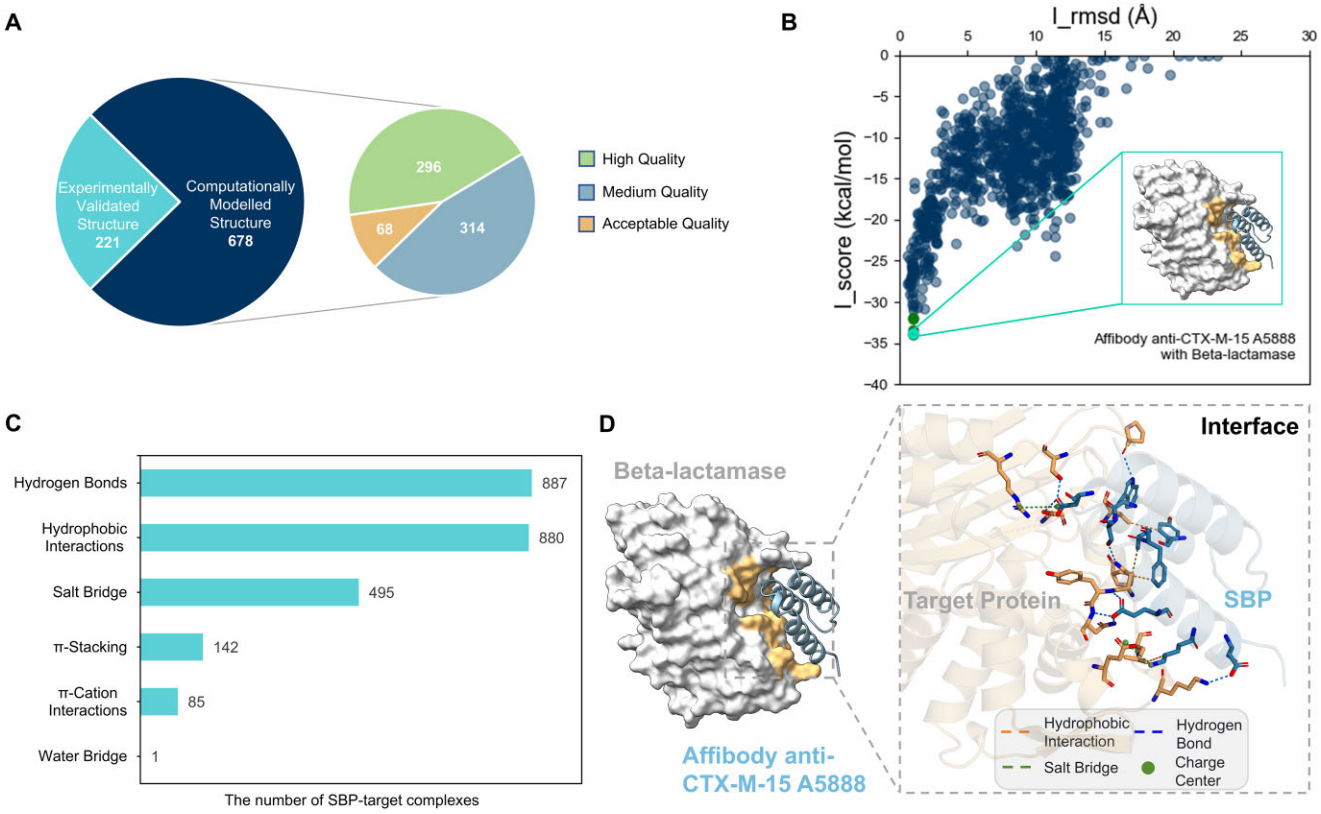


Figure 1. A statistical analysis of the binding epitopes of SBPs, along with an illustrative example of SBP–target interaction. (**A**) The distribution of SBP–target complexes from experiment and computational framework, as well as the quality distribution of computationally modeled complexes. (**B**) The Rosetta docking funnels of the high-quality computationally modeled complex (e.g. Affibody anti-CTX-M-15 A5888 in complex with β-lactamase). The plots on green color have the lowest docking interface score with an I_rmsd ≤ 4 Å. (**C**) The number of SBP–target complexes with corresponding interaction types. They represent the number of complexes possessing the corresponding interaction, with each complex counted as 1 regardless of the presence of multiple such interactions within it. (**D**) The interaction analysis of the high-quality computationally modeled complex (e.g. Affibody anti-CTX-M-15 A5888 in complex with β-lactamase).

The final selected model has an I_rmsd value of 0.993 Å, which is considered high-quality according to RosettaDock criteria.

In addition, the interaction interfaces of these complexes were analyzed using the Protein–Ligand Interaction Profiler (PLIP) tool (74). Summary of the 899 analyzed complexes revealed that the primary interactions between SBPs and their targets are hydrogen bonds, hydrophobic interactions, salt bridges, π-stacking and π–cation interactions, with the numbers of 887, 880, 495, 142 and 85, respectively, as depicted in Figure 1C. For instance, in the Affibody anti-CTX-M-15 A5888 bound to the β-lactamase complex, several key interactions as mentioned above are observed (Figure 1D). The residues Tyr25, Ala29, Phe30 and Lys50 on Affibody anti-CTX-M-15 A5888 engage in hydrophobic interactions with Leu105, Asn107, Pro170 and Thr174 on β-lactamase. Additionally, Trp32, Asp36, Glu47 and Asp53 on the Affibody form hydrogen bonds with Lys102, Asn107, Tyr108, Asn109, Ser240, Pro271 and Ser275 on β-lactamase. Furthermore, Asp36 and Lys50 on the Affibody establish salt bridges with Glu113 and Arg277 on β-lactamase. The detailed information at the interaction interface can be utilized to guide the rational design of new SBPs with enhanced potency and specificity. Moreover, researchers have the option to download analysis files for each complex from the website of the database, enabling more intuitive visualization.

Expanding the sequence space of SBPs and the concurrent prediction of their properties

Currently, the main methods of SBPs within SYNBIP are developed through conventional protein design techniques, including site-directed mutagenesis and directed evolution (1). There are limited variety of available SBPs under each scaffold, exemplified by the presence of only one SBP in the chaperonin 10-based binder scaffold. This suggests that the traditional approaches have inherent constraints when it comes to engineering SBPs for specific protein scaffolds. This limitation may stem from the fact that conventional protein design lacks the capability to exhaustively explore the entire sequence space (75). More recently, deep learning-based methods in *de novo* protein design have seen a breakthrough (41). Among them, the state-of-the-art ProteinMPNN framework outperforms in rapidly designing functional proteins, such as SBPs, with a remarkably high success rate (41,45). Furthermore, it demonstrates the capability to enhance various protein attributes, including solubility and stability (60,76).

In SYNBIP 2.0, the sequence space of SBPs was expanded by five times to 12 025 using ProteinMPNN (41) based on 1527 SBP monomers and 878 SBP–target complexes. Notably, the generated proteins have good sequence recovery (average sequence recovery is 49.2%) (77). Furthermore, the solubility and stability of those potential SBPs were predicted by Protein-Sol (78) and Instability Index (II) (79), respectively.

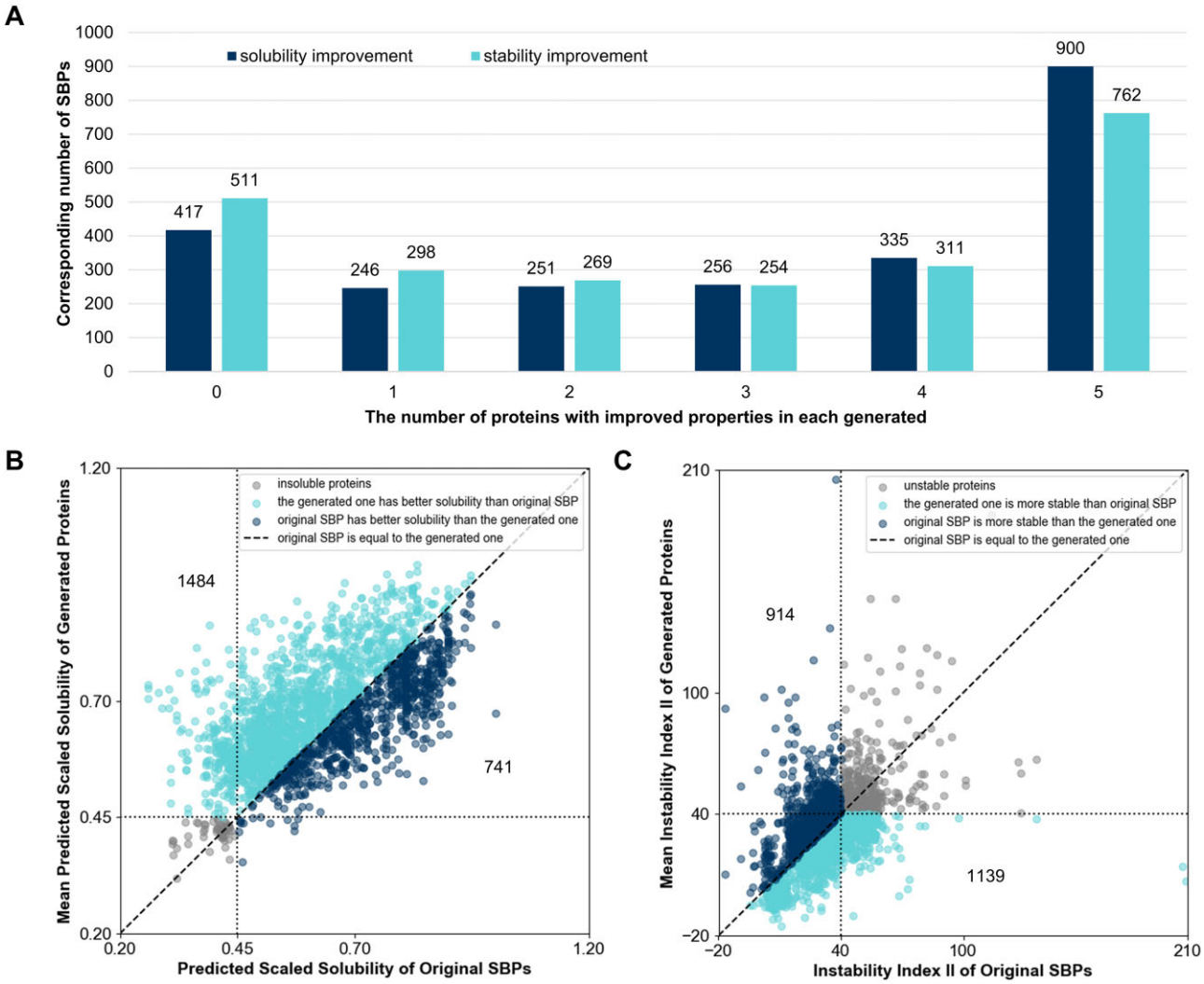


Figure 2. Statistics and comparison of properties between expanded sequences and original SBPs. (**A**) The number of proteins with improved properties (solubility or stability) after advanced generative protein design. In this work, the sequence space of each original SBP was expanded by five times. Taking the number of horizontal axes as 4 and the number of vertical axes as 335 as an example, it indicates that for 335 original SBPs, four out of the five generated sequences have better properties than the original one. (**B**) The horizontal axis of each data point represents the solubility of the original SBP, while the vertical axis indicates the average solubility of the five generated proteins for that SBP. A predicted scaled solubility score above 0.45 suggests a higher probability of solubility compared to the average soluble *E. coli* protein, with higher scores correlating to better solubility. The diagonal dashed line ($y = x$) signifies that the average solubility of the generated proteins matches that of the original SBP. The cyan points denote the proteins with improved solubilities. (**C**) The horizontal axis for each point signifies the stability of the original SBP, and the vertical axis shows the average stability of the five generated proteins under that SBP. An Instability Index (II) below 40 is considered indicative of a stable protein structure, with lower scores indicating greater stability. The diagonal dashed line ($y = x$) indicates that the average stability of the generated proteins is equivalent to that of the original SBP. The cyan points highlight the proteins with enhanced stabilities.

The workflow for ProteinMPNN (Supplementary Figure S3) and the codes for calculating the proteins' solubility and stability in this work were provided in Supplementary Data. The two properties are highlighted due to their substantial importance and relevance in the fields of biotechnology and biochemistry. They play a pivotal role in processes such as protein expression and purification, which are crucial for the advancement of drug development (80). To illustrate the similarities and differences among the designed sequences, both sequence logos and multiple sequence alignment results are presented in the 'Expanding Sequence Space for This Complex' section in the database.

Statistics in Figure 2A demonstrate that, on average, the majority of the generated proteins exhibit superior properties than the original SBPs. Regarding the solubility parameter, a predicted scaled solubility score exceeding 0.45 is indicative of a higher likelihood of solubility compared to the average soluble *Escherichia coli* protein (Figure 2B), as derived from empirical solubility data. In this context, higher scores are associated with enhanced solubility (78). As for the stability metric, an Instability Index (II) below 40 denotes a stable protein structure (Figure 2C), with lower scores reflecting superior stability (79). These generated proteins offer scientists novel insights for broadening the sequence space of SBPs. Researchers can now discern and select appropriate sequences for enhanced design. Alternatively, they can leverage the sequence generation and property prediction algorithms included in the Supplementary Data to execute batch designs. This approach facilitates the rapid creation of high-quality protein libraries for efficient binders screening.

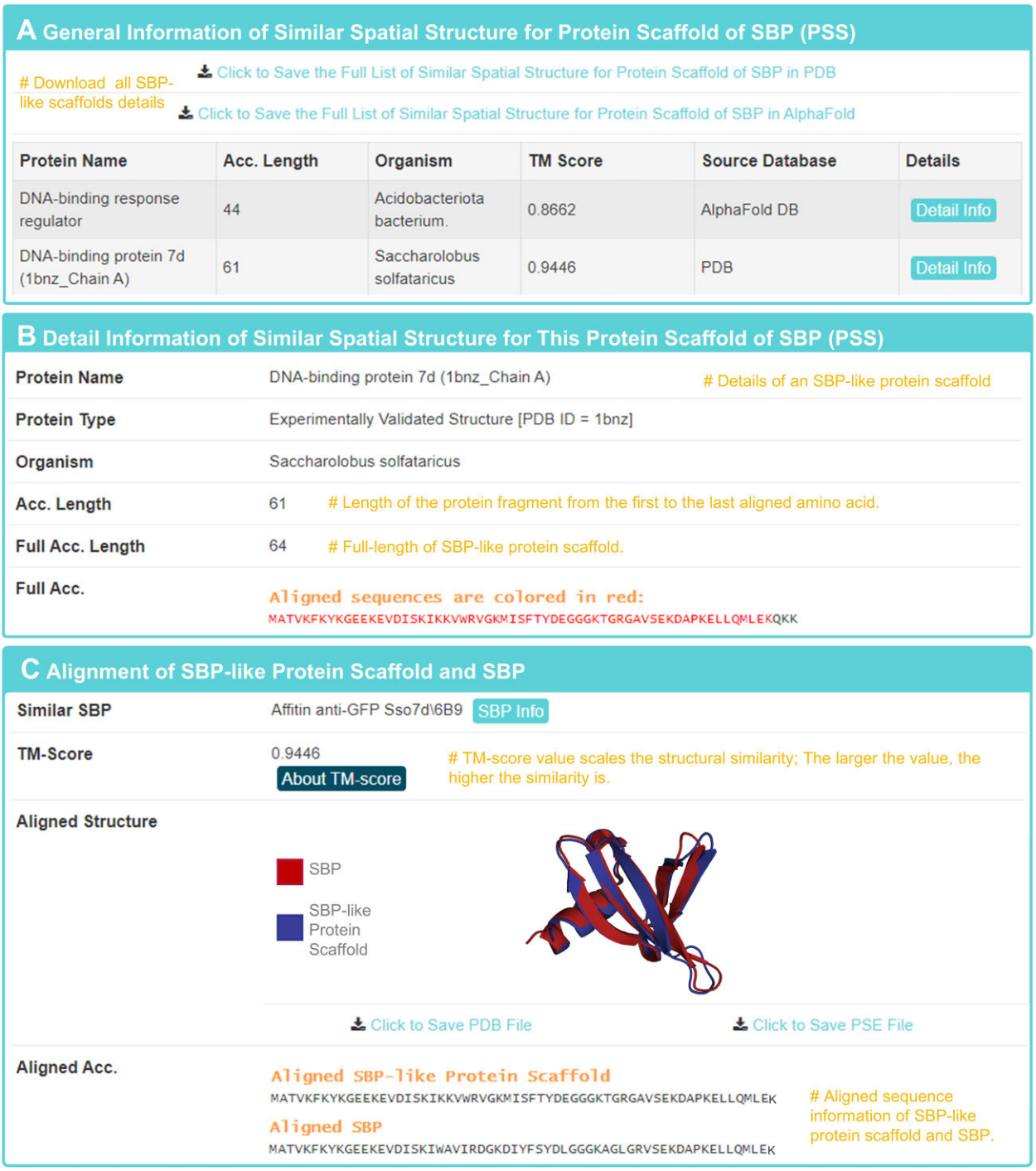


Figure 3. The typical page of SBP-like protein scaffolds. (**A**) General information of similar spatial structure for protein scaffold of SBP, including CSV files of all SBP-like scaffolds details under this scaffold, and representative SBP-like scaffolds. (**B**) Detail information of similar spatial structure for this protein scaffold of SBP, including protein name, protein type, PDB ID, length of alignment sequence and aligned sequence. (**C**) Structural alignment of SBP-like scaffold and corresponding SBP, including similar SBP name, TM-score, aligned structures and sequences.

Expansive SBP-like scaffolds data: insights from entire proteomes

The parent scaffold is a pivotal starting point in protein engineering. The 68 scaffolds cataloged in the SYNBIP database have largely been discovered by knowledge or designed from scratch (1). Currently, the insufficiency of SBP scaffolds is a primary impediment to the development of SBPs, underscoring the urgent need to identify SBP-like scaffolds (50). By comparing the structural similarity of SBPs under various scaffolds collected in SYNBIP, it can be observed that some scaffolds exhibit a certain degree of similarity (Supplementary Figure S4). The results suggest that structural similarity serves as a robust criterion for the identification of SBP-like scaffolds from natural sources, presenting a wealth of potential candidates for future research and practical applications. To this end, we have systematically mined SBP-like scaffolds across diverse organisms' proteomes through structure-based similarity analysis (12) leveraging the extensive protein databases includ-

Table 1. SYNBIP 2.0 vs. SYNBIP 1.0: a detailed comparative analysis of database enhancements

	SYNBIP 2.0	SYNBIP 1.0	Enhancements
The number of SBPs	2264	2111	153 new SBPs
SBP structures	1640	1329	311 new SBP structures
The number of targets	609	476	133 new targets
SBP epitopes	899	–	899 experimentally and computationally SBP–Target complex structures to interpret the mechanism of SBP action and guide the rational design of SBPs
SBP sequence space	12 025	–	12 025 variants of SBP were designed utilized the deep-learning algorithm and predicted the solubility and stability of the designed sequences
SBP-like scaffolds	78 473 from RCSB PDB and 16 401 555 from AlphaFold DB	–	16 480 028 SBP-like scaffolds were identified across the entire proteome, supporting diverse design and application possibilities

ing RCSB PDB (67), AlphaFold DB (63) and UniProt (81). It promises to yield a unique and valuable dataset for SBP-related research. The details of our methodology are thoroughly outlined in the Supplementary Data.

In addition, the accuracy of the method is demonstrated through two compelling examples. First, it can discern complex structures, including those with multiple disulfide bonds, as illustrated in Supplementary Figure S5. Second, it can successfully identify the human neural cell adhesion molecule 1 using the original vNAR scaffold (82) as a template. Human neural cell adhesion molecule 1 has been advanced into the development of i-body drugs (83), highlighting the method's capacity for uncovering novel scaffolds.

As a result, 78 473 SBP-like scaffolds from the RCSB PDB and an additional 16 401 555 from the AlphaFold DB were identified. These scaffolds span across 54 736 species, including *Homo sapiens*, *Ruminococcus albus*, *Bacteroides fragilis* and others. Since those scaffolds were identified from different kinds of species, they could be developed for different applications. For instance, humanized SBP-like scaffolds could advance the humanization of SBPs, facilitating the development of superior drugs and diagnostic tools (12,84). Meanwhile, scaffolds from other species, particularly those thriving in extreme environments like microorganisms, favor the creation of environmental remediation reagents (85) and may provide ideas for optimizing the physicochemical properties of SBPs (86). On the other hand, the identified scaffolds encompass a diverse array of functions. For example, Basiliximab is instrumental in preventing organ rejection in kidney transplants (87), while Motavizumab safeguards high-risk infants against respiratory syncytial virus infections (88). These applications not only underscore the versatility of the scaffolds but also pave the way for the innovation of additional functionalities for SBPs.

Figure 3 illustrates the information of identified Affitin scaffolds, including the SBP names, aligned protein names, species details, sequence ranges for the aligned proteins and TM-scores, along with their corresponding PDB IDs. Those data are categorized by scaffolds and available for download. For each scaffold, the webpage prominently features the SBP-like scaffolds with the highest TM-scores, ensuring that at least one is from a human species to emphasize the importance of humanization in SBP development.

Regular updates on SBPs and related information

The integration of newly emerged SBPs and targets to SYNBIP 2.0 was also routinely conducted in this update. First, 153 new SBPs which represent 25% of the known scaffolds (17 out of 68) were added from recent literature, the company's pipeline reports, and other sources (89–91). For an in-depth understanding of our collection procedures, please refer to the SYNBIP literature (1). To maintain data precision and thoroughness, SYNBIP includes only those sequences from patents that have been corroborated by publications in peer-reviewed journals. Furthermore, our database not only compiles the most effective SBPs reported in scholarly articles but also encompasses other SBPs that have demonstrated affinity for their targets, as documented in the literature. In addition, all SBP structures were updated with AlphaFold2 (92), adding 201 previously unpredictable structures. Users can download the structures and view the confidence score of each amino acid (Supplementary Figure S6) by entering the corresponding code in the PyMOL (see Supplementary Data). Third, the clinical status of SBPs was updated through multiple channels, including well-known clinical databases (ClinicalTrials.gov, ChiCTR) and pharmaceutical company websites. The design methodologies for all SBPs have been meticulously classified into two primary approaches: conventional techniques such as site-directed mutagenesis and directed evolution, as well as the innovative *de novo* protein design. The number of SBPs designed by traditional and *de novo* approaches are 2228 and 36, respectively.

Conclusion and perspectives

SYNBIP 2.0 is a significant update to the SYNBIP database, providing an extensive suite of resources designed to facilitate the development of SBPs (Table 1). This update mainly focuses on three key sections in SYNBIP 2.0: 'SBP epitopes', 'SBP sequence space' and 'SBP-like scaffolds'. The display of binding epitopes presents crucial structural insights into the interaction interface between SBPs and their target proteins, which are vital for deciphering the mechanisms of SBP action and guiding the rational design of these proteins. Expanding the SBP sequence space introduces a curated collection of high-quality, *de novo* designed sequences. These sequences are readily available for direct application or can be integrated into SBP libraries for enhanced protein screening and design. The SBP-like scaffolds mined from entire proteome may serve multiple purposes, including developing potential scaffolds, humanized protein design, property optimization, new functional SBP design, evolutionary analysis and so on. In addition, SYNBIP 2.0 is committed to regular updates, ensuring that the database remains at the forefront with the most re-

cent SBP data. These features collectively solidify SYNBIP 2.0 as an indispensable platform, offering innovative perspectives to guide the rational design of next-generation SBPs.

Data availability

SYNBIP2.0 is freely accessible to all users without any login requirement at: https://idrblab.org/synbip/.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We would like to thank the users who using SYNBIP and providing us with valuable suggestions for further improvement of this databse.

Authors' contribution: W.X. and F.Z. designed the research. Y.L., F.L., and Z.D. performed the research. Y.L., F.L., Z.D., R.L., W.J., and H.W. analyzed the data. Y.L. and W.X. wrote the manuscript. All authors reviewed the manuscript.

Funding

Natural Science Foundation of Chongqing [2023NSCQ-MSX0140]; Technology Innovation and Application Demonstration Project of Chongqing [cstc2018jscx-msybX0287]; Entrepreneurship and Innovation Support Plan for Chinese Overseas Students of Chongqing [cx2020127]. Funding for open access charge: Natural Science Foundation of Chongqing [2023NSCQ-MSX0140]; Technology Innovation and Application Demonstration Project of Chongqing [cstc2018jscx-msybX0287]; Entrepreneurship and Innovation Support Plan for Chinese Overseas Students of Chongqing [cx2020127].

Conflict of interest statement

None declared.

References

1. Wang,X., Li,F., Qiu,W., Xu,B., Li,Y., Lian,X., Yu,H., Zhang,Z., Wang,J., Li,Z., *et al.* (2022) SYNBIP: synthetic binding proteins for research, diagnosis and therapy. *Nucleic Acids Res.*, **50**, D560–D570.
2. Crook,Z., Nairn,N. and Olson,J. (2020) Miniproteins as a powerful modality in drug development. *Trends Biochem. Sci.*, **45**, 332–346.
3. Lima,W.C., Gasteiger,E., Marcatili,P., Duek,P., Bairoch,A. and Cosson,P. (2020) The ABCD database: a repository for chemically defined antibodies. *Nucleic Acids Res.*, **48**, D261–D264.
4. Carvalho,M.B., Molina,F. and Felicori,L.F. (2019) Yvis: antibody high-density alignment visualization and analysis platform with an integrated database. *Nucleic Acids Res.*, **47**, W490–W495.
5. Raybould,M., Marks,C., Lewis,A., Shi,J., Bujotzek,A., Taddese,B. and Deane,C. (2020) Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Res.*, **48**, D383–D388.
6. Wilton,E., Opyr,M., Kailasam,S., Kothe,R. and Wieden,H. (2018) sdAb-DB: the single domain antibody database. *ACS Synth. Biol.*, **7**, 2480–2484.
7. Adolf-Bryfogle,J., Xu,Q., North,B., Lehmann,A. and Dunbrack,R.L. Jr. (2015) PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res.*, **43**, D432–D438.
8. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P., *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
9. Oughtred,R., Stark,C., Breitkreutz,B.J., Rust,J., Boucher,L., Chang,C., Kolas,N., O'Donnell,L., Leung,G., McAdam,R., *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47**, D529–D541.
10. Zhang,Z., Li,F., Duan,Z., Shi,C., Wang,X., Zhu,F. and Xue,W. (2024) OPTICS: an interactive online platform for photosensory and bio-functional proteins in optogenetic systems. *Comput. Biol. Med.*, **178**, 108687.
11. Zhang,Y., Zhou,Y., Zhou,Y., Yu,X., Shen,X., Hong,Y., Zhang,Y., Wang,S., Mou,M., Zhang,J., *et al.* (2024) TheMarker: a comprehensive database of therapeutic biomarkers. *Nucleic Acids Res.*, **52**, D1450–D1464.
12. Wang,X., Zhang,Y., Li,Z., Duan,Z., Guo,M., Wang,Z., Zhu,F. and Xue,W. (2024) PROSCA: an online platform for humanized scaffold mining facilitating rational protein engineering. *Nucleic Acids Res.*, **52**, W272–W279.
13. Gomes,D.E.B., Yang,B., Vanella,R., Nash,M.A. and Bernardi,R.C. (2024) Integrating dynamic network analysis with AI for enhanced epitope prediction in PD-L1:affibody interactions. *J. Am. Chem. Soc.*, **146**, 23842–23853.
14. Yang,Z., Shao,Q., Jiang,Y., Jurich,C., Ran,X., Juarez,R., Yan,B., Stull,S., Gollu,A. and Ding,N. (2023) Mutexa: a computational ecosystem for intelligent protein engineering. *J. Chem. Theory Comput.*, **19**, 7459–7477.
15. Mao,M., Ahrens,L., Luka,J., Contreras,F., Kurkina,T., Bienstein,M., Sárria Pereira de Passos,M., Schirinzi,G., Mehn,D., Valsesia,A., *et al.* (2024) Material-specific binding peptides empower sustainable innovations in plant health, biocatalysis, medicine and microplastic quantification. *Chem. Soc. Rev.*, **53**, 6445–6510.
16. Yang,J., Zhang,Z., Yang,F., Zhang,H., Wu,H., Zhu,F. and Xue,W. (2021) Computational design and modeling of nanobodies toward SARS-CoV-2 receptor binding domain. *Chem. Biol. Drug Design*, **98**, 1–18.
17. Mijit,A., Wang,X., Li,Y., Xu,H., Chen,Y. and Xue,W. (2023) Mapping synthetic binding proteins epitopes on diverse protein targets by protein structure prediction and protein-protein docking. *Comput. Biol. Med.*, **163**, 107183.
18. Liang,T., Jiang,C., Yuan,J., Othman,Y., Xie,X.Q. and Feng,Z. (2022) Differential performance of RoseTTAFold in antibody modeling. *Brief. Bioinform*, **23**, bbac152.
19. Sun,J., Liu,X., Zhang,S., Li,M., Zhang,Q. and Chen,J. (2023) Molecular insights and optimization strategies for the competitive binding of engineered ACE2 proteins: a multiple replica molecular dynamics study. *Phys. Chem. Chem. Phys.*, **25**, 28479–28496.
20. Zheng,L., Shi,S., Sun,X., Lu,M., Liao,Y., Zhu,S., Zhang,H., Pan,Z., Fang,P., Zeng,Z., *et al.* (2024) MoDAFold: a strategy for predicting the structure of missense mutant protein based on AlphaFold2 and molecular dynamics. *Brief. Bioinform*, **25**, bbae006.
21. Liu,Y., Zhang,Y., Chen,Z. and Peng,J. (2024) POLAT: Protein function prediction based on soft mask graph network and residue-Label ATtention. *Comput. Biol. Chem.*, **110**, 108064.
22. Zhao,L., Chen,M., Wang,X., Kang,S., Xue,W. and Li,Z. (2022) Identification of Anti-TNFα VNAR single domain antibodies from Whitespotted Bambooshark (Chiloscyllium plagiosum). *Mar. Drugs*, **20**, 307.
23. Qiu,X.Y., Wu,H. and Shao,J. (2022) TALE-cmap: Protein function prediction based on a TALE-based architecture and the structure information from contact map. *Comput. Biol. Med.*, **149**, 105938.
24. Fu,J., Zhang,Y., Wang,Y., Zhang,H., Liu,J., Tang,J., Yang,Q., Sun,H., Qiu,W., Ma,Y., *et al.* (2022) Optimization of metabolomic data processing using NOREVA. *Nat. Protoc.*, **17**, 129–151.

25. Yang,Q., Li,B., Chen,S., Tang,J., Li,Y., Li,Y., Zhang,S., Shi,C., Zhang,Y., Mou,M., *et al.* (2021) MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J. Proteomics*, **232**, 104023.
26. Yang,Q., Li,B., Tang,J., Cui,X., Wang,Y., Li,X., Hu,J., Chen,Y., Xue,W., Lou,Y., *et al.* (2020) Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief. Bioinf.*, **21**, 1058–1068.
27. Amahong,K., Zhang,W., Zhou,Y., Zhang,S., Yin,J., Li,F., Xu,H., Yan,T., Yue,Z., Liu,Y., *et al.* (2023) CovInter: interaction data between coronavirus RNAs and host proteins. *Nucleic Acids Res.*, **51**, D546–D556.
28. Lian,X., Zhang,Y., Zhou,Y., Sun,X., Huang,S., Dai,H., Han,L. and Zhu,F. (2024) SingPro: A knowledge base providing single-cell proteomic data. *Nucleic Acids Res.*, **52**, D552–D561.
29. Zhang,Y., Sun,H., Lian,X., Tang,J. and Zhu,F. (2023) ANPELA: significantly enhanced quantification tool for cytometry-based single-cell proteomics. *Adv. Sci.*, **10**, e2207061.
30. Li,F., Yin,J., Lu,M., Yang,Q., Zeng,Z., Zhang,B., Li,Z., Qiu,Y., Dai,H., Chen,Y., *et al.* (2022) ConSIG: consistent discovery of molecular signature from OMIC data. *Brief. Bioinf.*, **23**, bbac253.
31. Li,F., Zhou,Y., Zhang,Y., Yin,J., Qiu,Y., Gao,J. and Zhu,F. (2022) POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability. *Brief. Bioinf.*, **23**, bbac040.
32. Li,B., Tang,J., Yang,Q., Li,S., Cui,X., Li,Y., Chen,Y., Xue,W., Li,X. and Zhu,F. (2017) NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res.*, **45**, W162–W170.
33. Li,F., Yin,J., Lu,M., Mou,M., Li,Z., Zeng,Z., Tan,Y., Wang,S., Chu,X., Dai,H., *et al.* (2023) DrugMAP: molecular atlas and pharma-information of all drugs. *Nucleic Acids Res.*, **51**, D1288–D1299.
34. Hosseininejad-Chafi,M., Kianmehr,Z., Pooshang-Bagheri,K., Kazemi-Lomedasht,F. and Behdani,M. (2023) Development of a functional nanobody targeting programmed cell death protein-1 as immune checkpoint inhibitor. *Curr. Pharm. Des.*, **29**, 2336–2344.
35. Singh,J.K., Anand,S. and Srivastava,S.K. (2023) Is BF.7 more infectious than other Omicron subtypes: insights from structural and simulation studies of BF.7 spike RBD variant. *Int. J. Biol. Macromol.*, **238**, 124154.
36. Tu,G., Fu,T., Zheng,G., Xu,B., Gou,R., Luo,D., Wang,P. and Xue,W. (2024) Computational chemistry in structure-based solute carrier transporter drug design: recent advances and future perspectives. *J. Chem. Inf. Model.*, **64**, 1433–1455.
37. Li,F., Mou,M., Li,X., Xu,W., Yin,J., Zhang,Y. and Zhu,F. (2024) DrugMAP 2.0: molecular atlas and pharma-information of all drugs. *Nucleic Acids Res.*, https://doi.org/10.1093/nar/gkae791.
38. Tu,G., Xu,B., Luo,D., Liu,J., Liu,Z., Chen,G. and Xue,W. (2023) Multi-state model-based identification of cryptic allosteric sites on human serotonin transporter. *ACS Chem. Neurosci.*, **14**, 1686–1694.
39. Xue,W., Fu,T., Deng,S., Yang,F., Yang,J. and Zhu,F. (2022) Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram. *ACS Chem. Neurosci.*, **13**, 340–351.
40. 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., *et al.* (2022) Design of protein-binding proteins from the target structure alone. *Nature*, **605**, 551–560.
41. Wang,J., Lisanza,S., Juergens,D., Tischer,D., Watson,J.L., Castro,K.M., Ragotte,R., Saragovi,A., Milles,L.F., Baek,M., *et al.* (2022) Scaffolding protein functional sites using deep learning. *Science*, **377**, 387–394.
42. Xu,B., Chen,Y. and Xue,W. (2024) Computational protein design - where it goes? *Curr. Med. Chem.*, **31**, 2841–2854.
43. Silva,D.A., Yu,S., Ulge,U.Y., Spangler,J.B., Jude,K.M., Labão-Almeida,C., Ali,L.R., Quijano-Rubio,A., Ruterbusch,M., Leung,I., *et al.* (2019) De novo design of potent and selective mimics of IL-2 and IL-15. *Nature*, **565**, 186–191.
44. Cao,L., Goreshnik,I., Coventry,B., Case,J.B., Miller,L., Kozodoy,L., Chen,R.E., Carter,L., Walls,A.C., Park,Y.J., *et al.* (2020) De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science*, **370**, 426–431.
45. Bennett,N.R., Coventry,B., Goreshnik,I., Huang,B., Allen,A., Vafeados,D., Peng,Y.P., Dauparas,J., Baek,M., Stewart,L., *et al.* (2023) Improving de novo protein binder design with deep learning. *Nat. Commun.*, **14**, 2625.
46. Gainza,P., Wehrle,S., Van Hall-Beauvais,A., Marchand,A., Scheck,A., Harteveld,Z., Buckley,S., Ni,D., Tan,S., Sverrisson,F., *et al.* (2023) *De novo* design of protein interactions with learned surface fingerprints. *Nature*, **617**, 176–184.
47. Kang,S., Davidsen,K., Gomez-Castillo,L., Jiang,H., Fu,X., Li,Z., Liang,Y., Jahn,M., Moussa,M., DiMaio,F., *et al.* (2019) COMBINES-CID: an efficient method for de novo engineering of highly specific chemically induced protein dimerization systems. *J. Am. Chem. Soc.*, **141**, 10948–10952.
48. Zambaldi,V., La,D., Chu,A.E., Patani,H., Danson,A.E., Kwan,T.O.C., Frerix,T., Schneider,R.G., Saxton,D., Thillaisundaram,A., *et al.* (2024) *De novo* design of high-affinity protein binders with AlphaProteo. arXiv doi: https://arxiv.org/abs/2409.08022v1, 12 September 2024, preprint: not peer reviewed.
49. Lu,M., Yin,J., Zhu,Q., Lin,G., Mou,M., Liu,F., Pan,Z., You,N., Lian,X., Li,F., *et al.* (2023) Artificial intelligence in pharmaceutical sciences. *Engineering-Prc*, **27**, 37–69.
50. Kortemme,T. (2024) *De novo* protein design-From new structures to programmable functions. *Cell*, **187**, 526–544.
51. Mou,M., Pan,Z., Zhou,Z., Zheng,L., Zhang,H., Shi,S., Li,F., Sun,X. and Zhu,F. (2023) A transformer-based ensemble framework for the prediction of protein-protein interaction sites. *Research*, **6**, 0240.
52. Zhang,S., Amahong,K., Zhang,C., Li,F., Gao,J., Qiu,Y. and Zhu,F. (2021) RNA-RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection. *Brief. Bioinf.*, **23**, bbab397.
53. Wang,Y., Chen,Z., Pan,Z., Huang,S., Liu,J., Xia,W., Zhang,H., Zheng,M., Li,H., Hou,T., *et al.* (2023) RNAincoder: a deep learning-based encoder for RNA and RNA-associated interaction. *Nucleic Acids Res.*, **51**, W509–W519.
54. Huang,P.S., Boyken,S.E. and Baker,D. (2016) The coming of age of de novo protein design. *Nature*, **537**, 320–327.
55. Freschlin,C.R., Fahlberg,S.A. and Romero,P.A. (2022) Machine learning to navigate fitness landscapes for protein engineering. *Curr. Opin. Biotechnol.*, **75**, 102713.
56. Fu,T., Li,F., Zhang,Y., Yin,J., Qiu,W., Li,X., Liu,X., Xin,W., Wang,C., Yu,L., *et al.* (2022) VARIDT 2.0: structural variability of drug transporter. *Nucleic Acids Res.*, **50**, D1417–D1431.
57. Sun,X., Zhang,Y., Li,H., Zhou,Y., Shi,S., Chen,Z., He,X., Zhang,H., Li,F., Yin,J., *et al.* (2023) DRESIS: the first comprehensive landscape of drug resistance information. *Nucleic Acids Res.*, **51**, D1263–D1275.
58. Kim,D.E., Jensen,D.R., Feldman,D., Tischer,D., Saleem,A., Chow,C.M., Li,X., Carter,L., Milles,L., Nguyen,H., *et al.* (2023) *De novo* design of small beta barrel proteins. *Proc. Natl Acad. Sci. USA*, **120**, e2207974120.
59. Shin,J.E., Riesselman,A.J., Kollasch,A.W., McMahon,C., Simon,E., Sander,C., Manglik,A., Kruse,A.C. and Marks,D.S. (2021) Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, **12**, 2403.
60. Sumida,K.H., Núñez-Franco,R., Kalvet,I., Pellock,S.J., Wicky,B.I.M., Milles,L.F., Dauparas,J., Wang,J., Kipnis,Y., Jameson,N., *et al.* (2024) Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.*, **146**, 2054–2061.
61. Tian,P., Lemaire,A., Sénéchal,F., Habrylo,O., Antonietti,V., Sonnet,P., Lefebvre,V., Isa Marin,F., Best,R.B., Pelloux,J., *et al.*

(2022) Design of a protein with improved thermal stability by an evolution-based generative model. *Angew. Chem. Int. Ed Engl.*, **61**, e202202711.
62. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Söding,J. and Steinegger,M. (2024) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
63. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
64. Sakuma,K., Kobayashi,N., Sugiki,T., Nagashima,T., Fujiwara,T., Suzuki,K., Kobayashi,N., Murata,T., Kosugi,T., Tatsumi-Koga,R., *et al.* (2024) Design of complicated all-α protein structures. *Nat. Struct. Mol. Biol.*, **31**, 275–282.
65. Eisenstein,M. (2023) AI-enhanced protein design makes proteins that have never existed. *Nat. Biotechnol.*, **41**, 303–305.
66. Liu,Z., Li,Y., Han,L., Li,J., Liu,J., Zhao,Z., Nie,W., Liu,Y. and Wang,R. (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.
67. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chen,L., Crichlow,G.V., Christie,C.H., Dalenberg,K., Di Costanzo,L., Duarte,J.M., *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
68. Ascher,D.B., Jubb,H.C., Pires,D.E.V., Ochi,T., Higueruelo,A. and Blundell,T.L. (2015) Springer Netherlands, Dordrecht, pp. 141–163.
69. Persch,E., Dumele,O. and Diederich,F. (2015) Molecular recognition in chemical and biological systems. *Angew. Chem. Int. Ed Engl.*, **54**, 3290–3327.
70. Zheng,L., Shi,S., Lu,M., Fang,P., Pan,Z., Zhang,H., Zhou,Z., Zhang,H., Mou,M., Huang,S., *et al.* (2024) AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding. *Genome Biol.*, **25**, 41.
71. Xue,W., Yang,F., Wang,P., Zheng,G., Chen,Y., Yao,X. and Zhu,F. (2018) What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci.*, **9**, 1128–1140.
72. Wang,Y., Pan,Z., Mou,M., Xia,W., Zhang,H., Zhang,H., Liu,J., Zheng,L., Luo,Y., Zheng,H., *et al.* (2023) A task-specific encoding algorithm for RNAs and RNA-associated interactions based on convolutional autoencoder. *Nucleic Acids Res.*, **51**, e110.
73. Chaudhury,S., Berrondo,M., Weitzner,B.D., Muthu,P., Bergman,H. and Gray,J.J. (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One*, **6**, e22477.
74. Adasme,M.F., Linnemann,K.L., Bolz,S.N., Kaiser,F., Salentin,S., Haupt,V.J. and Schroeder,M. (2021) PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. *Nucleic Acids Res.*, **49**, W530–W534.
75. Zeymer,C. and Hilvert,D. (2018) Directed evolution of protein catalysts. *Annu. Rev. Biochem.*, **87**, 131–157.
76. Li,Y., Jiao,W., Liu,R., Deng,X., Zhu,F. and Xue,W. (2025) Expanding the sequence spaces of synthetic binding protein using deep learning-based framework ProteinMPNN. *Front. Comput. Sci.*, **19**, 195903.
77. Liu,Y. and Liu,H. (2024) Protein sequence design on given backbones with deep learning. *Protein Eng. Des. Sel.*, **37**, gzad024.
78. Hebditch,M., Carballo-Amador,M.A., Charonis,S., Curtis,R. and Warwicker,J. (2017) Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, **33**, 3098–3100.
79. Guruprasad,K., Reddy,B.V. and Pandit,M.W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Protein Eng.*, **4**, 155–161.
80. Qing,R., Hao,S., Smorodina,E., Jin,D., Zalevsky,A. and Zhang,S. (2022) Protein design: from the aspect of water solubility and stability. *Chem. Rev.*, **122**, 14085–14179.
81. UniProt Consortium (2022) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
82. Streltsov,V.A., Varghese,J.N., Carmichael,J.A., Irving,R.A., Hudson,P.J. and Nuttall,S.D. (2004) Structural evidence for evolution of shark Ig new antigen receptor variable domain antibodies from a cell-surface receptor. *Proc. Natl Acad. Sci. USA*, **101**, 12444–12449.
83. Griffiths,K., Dolezal,O., Cao,B., Nilsson,S.K., See,H.B., Pfleger,K.D.G., Roche,M., Gorry,P.R., Pow,A., Viduka,K., *et al.* (2016) i-bodies, Human Single Domain Antibodies That Antagonize Chemokine Receptor CXCR4. *J. Biol. Chem.*, **291**, 12641–12657.
84. Yin,J., Chen,Z., You,N., Li,F., Zhang,H., Xue,J., Ma,H., Zhao,Q., Yu,L., Zeng,S., *et al.* (2024) VARIDT 3.0: the phenotypic and regulatory variability of drug transporter. *Nucleic Acids Res.*, **52**, D1490–D1502.
85. Możejko-Ciesielska,J., Ray,S. and Sankhyan,S. (2023) Recent challenges and trends of polyhydroxyalkanoate production by extremophilic bacteria using renewable feedstocks. *Polymers (Basel)*, **15**, 4385.
86. Karan,R., Mathew,S., Muhammad,R., Bautista,D.B., Vogler,M., Eppinger,J., Oliva,R., Cavallo,L., Arold,S.T. and Rueping,M. (2020) Understanding high-salt and cold adaptation of a polyextremophilic enzyme. *Microorganisms*, **8**, 1594.
87. Onrust,S.V. and Wiseman,L.R. (1999) Basiliximab. *Drugs*, **57**, 207–213.
88. Cingoz,O. (2009) Motavizumab. *MAbs*, **1**, 439–442.
89. Zhou,Y., Zhang,Y., Zhao,D., Yu,X., Shen,X., Zhou,Y., Wang,S., Qiu,Y., Chen,Y. and Zhu,F. (2024) TTD: therapeutic target database describing target druggability information. *Nucleic Acids Res.*, **52**, D1465–D1477.
90. Li,Y., Li,X., Hong,J., Wang,Y., Fu,J., Yang,H., Yu,C., Li,F., Hu,J., Xue,W., *et al.* (2020) Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief. Bioinf.*, **21**, 649–662.
91. Shen,L., Sun,X., Chen,Z., Guo,Y., Shen,Z., Song,Y., Xin,W., Ding,H., Ma,X., Xu,W., *et al.* (2024) ADCdb: the database of antibody-drug conjugates. *Nucleic Acids Res.*, **52**, D1097–D1109.
92. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Received: August 19, 2024. **Revised:** September 18, 2024. **Editorial Decision:** September 25, 2024. **Accepted:** September 26, 2024