

# High-accuracy protein complex structure modeling based on sequence-derived structure complementarity

Received: 10 March 2025

Accepted: 30 September 2025

Published online: 19 November 2025

 Check for updatesMinghua Hou<sup>1,3</sup>, Yuhao Xia<sup>2,3</sup>, Pengcheng Wang<sup>1</sup>, Zexin Lv<sup>1</sup>, Dongliang Hou<sup>1</sup>, Xiaogen Zhou<sup>1</sup>, Jianyang Zeng<sup>2</sup>✉ & Guijun Zhang<sup>1</sup>✉

In living organisms, proteins perform key functions required for life activities by interacting to form complexes. Determining the protein complex structure is crucial for understanding and mastering biological functions. Although AlphaFold2 makes a revolutionary breakthrough in predicting protein monomeric structures, accurately capturing inter-chain interaction signals and modeling the structures of protein complexes remain a formidable challenge. In this work, we report DeepSCFold, a pipeline for improving protein complex structure modeling. DeepSCFold uses sequence-based deep learning models to predict protein-protein structural similarity and interaction probability, providing a foundation for identifying interaction partners and constructing deep paired multiple-sequence alignments (MSAs) for protein complex structure prediction. Benchmark results show that DeepSCFold significantly increases the accuracy of protein complex structure prediction compared with state-of-the-art methods. For multimer targets from CASP15, DeepSCFold achieves an improvement of 11.6% and 10.3% in TM-score compared to AlphaFold-Multimer and AlphaFold3, respectively. Furthermore, when applied to antibody-antigen complexes from the SAbDab database, DeepSCFold enhances the prediction success rate for antibody-antigen binding interfaces by 24.7% and 12.4% over AlphaFold-Multimer and AlphaFold3, respectively. These results demonstrate that DeepSCFold effectively captures intrinsic and conserved protein-protein interaction patterns through sequence-derived structure-aware information, rather than relying solely on sequence-level co-evolutionary signals.

Interacting proteins play pivotal roles in cellular processes by forming functional multi-protein complexes (i.e., multimers or assemblies) that are essential for executing biological processes such as signal transduction, transport, and metabolism<sup>1–4</sup>. To fully understand these functions, determining the structures of these complexes is crucial. However, this often remains a challenge for existing experimental methods, such as X-ray crystallography, nuclear magnetic resonance

(NMR), and cryo-electron microscopy (cryo-EM)<sup>5</sup>. Consequently, computational methods for obtaining complex structures have become an indispensable and important complement to experimental techniques in structural biology. Nevertheless, predicting the quaternary structure of a protein complex is significantly more challenging than predicting the tertiary structure of a single protein monomer, as it necessitates the accurate modeling of both intra-chain

<sup>1</sup>College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. <sup>2</sup>School of Engineering, Westlake University, Hangzhou, China.

<sup>3</sup>These authors contributed equally: Minghua Hou, Yuhao Xia. ✉e-mail: [zengjy@westlake.edu.cn](mailto:zengjy@westlake.edu.cn); [zgj@zjut.edu.cn](mailto:zgj@zjut.edu.cn)

and inter-chain residue-residue interactions among multiple protein chains<sup>6</sup>.

Traditionally, the prediction of protein complex structures relies on template-based homology modeling or docking-based prediction methods<sup>6,7</sup>. While template-based homology modeling is effective when high-quality templates are available, its applicability is often limited by the difficulty in obtaining suitable templates for most target complexes<sup>8</sup>. In protein-protein docking methods, monomer structures are assembled into a complex based on the “lock-and-key”<sup>9</sup> and “induced fit”<sup>10</sup> principles, as exemplified by tools such as ZDOCK<sup>11–13</sup>, HADDOCK<sup>14–16</sup>, and HDOCK<sup>17–19</sup>. The docking process aims to identify the optimal binding mode that satisfies the spatial shape complementarity through energy minimization. However, docking methods still face challenges in predicting complex structures due to the complexity of conformational sampling, the inaccuracy of energy functions, and the inherent flexibility of proteins in the interface regions<sup>20,21</sup>.

With the remarkable advancements achieved by AlphaFold2<sup>22</sup> in protein monomer structure prediction, deep learning methods have increasingly been applied to predict protein multimer structures, including AlphaFold-Multimer<sup>23</sup>, DMFold-Multimer<sup>24,25</sup>, MULTICOM3<sup>6</sup>, and the recently released AlphaFold3<sup>26</sup>. Among these, AlphaFold-Multimer, an extension of AlphaFold2 specifically tailored for protein multimer structure prediction, has significantly improved the accuracy of complex predictions. Nevertheless, the accuracy of multimer structure predictions using AlphaFold-Multimer remains considerably lower than that of AlphaFold2 for monomer structures<sup>6</sup>. To address this limitation, the research community has actively developed strategies based on AlphaFold-Multimer to further enhance performance. These strategies include extensive sampling through variations in multiple sequence alignments (MSAs) construction, the use of multiple seeds, increased recycling, and extensive network dropout<sup>27</sup>. Results from CASP15 highlight the effectiveness of these methods, demonstrating superior performance in predicting protein multimers compared to the baseline AlphaFold-Multimer (i.e., the NBIS-AF2-standard group)<sup>28</sup>. Notable examples include the Zheng group (i.e., DMFold)<sup>25</sup>, Venclovas group<sup>29</sup>, Wallner group<sup>30</sup>, Yang-Multimer group<sup>31</sup> and MULTICOM\_human group (i.e., MULTICOM3)<sup>6</sup>, etc. Among these advancements, the quality of MSAs plays a critical role in structure prediction, as their implicit co-evolutionary information is often essential for locating an approximate global minimum in the protein conformation space<sup>28,32</sup>. This importance is further amplified in the context of protein complexes, where accurately capturing the binding modes of protein chains may significantly benefit from the paired MSAs of the complex.

In protein complex structure prediction, monomeric MSAs derived from individual chains are systematically paired across different chains to generate comprehensive paired MSAs. This pairing strategy enables the identification of inter-chain co-evolutionary signals between interacting partners, providing valuable insights into the dynamic behavior and stability of molecular interactions within the protein complex. However, popular sequence search tools such as HHblits<sup>33,34</sup>, Jackhammer<sup>35</sup>, and MMseqs<sup>36,37</sup> are primarily designed for constructing monomeric MSAs and cannot be directly applied to paired MSAs construction. This limitation may compromise the accuracy and generality of protein complex structure predictions, particularly for tightly intertwined complexes or highly flexible interactions, such as those observed in antibody-antigen systems<sup>38</sup>. Recently, several methods have been proposed to construct paired MSAs, aiming to enhance the ability to capture inter-chain interactions. Examples include DeepMSA<sup>24</sup>, MULTICOM3<sup>6</sup>, DiffPALM<sup>38</sup>, ESMPair<sup>39</sup>. DeepMSA2 improves monomeric MSAs and constructs paired MSAs by performing iterative alignment searches across genomic and metagenomic sequence databases, followed by filtering using AlphaFold2/AlphaFold-Multimer. MULTICOM3 generates a diverse set of paired

MSAs by concatenating the MSAs of subunits, leveraging potential protein-protein interactions extracted from multiple sources. ESMPair ranks monomeric MSAs using ESM-MSA-Ib<sup>40</sup> and integrates species information to construct paired MSAs. DiffPALM employs an MSA transformer to estimate amino acid probabilities, creating a permutation matrix to pair protein sequences. While these methods effectively capture inter-chain co-evolutionary information through paired MSAs construction based on sequence similarity or species-related information, they may face limitations when applied to complexes lacking clear co-evolutionary signals at the sequence level. For instance, virus-host and antibody-antigen systems often do not exhibit inter-chain co-evolution, as identifying orthologs between host and pathogenic proteins is challenging due to the absence of species overlap.

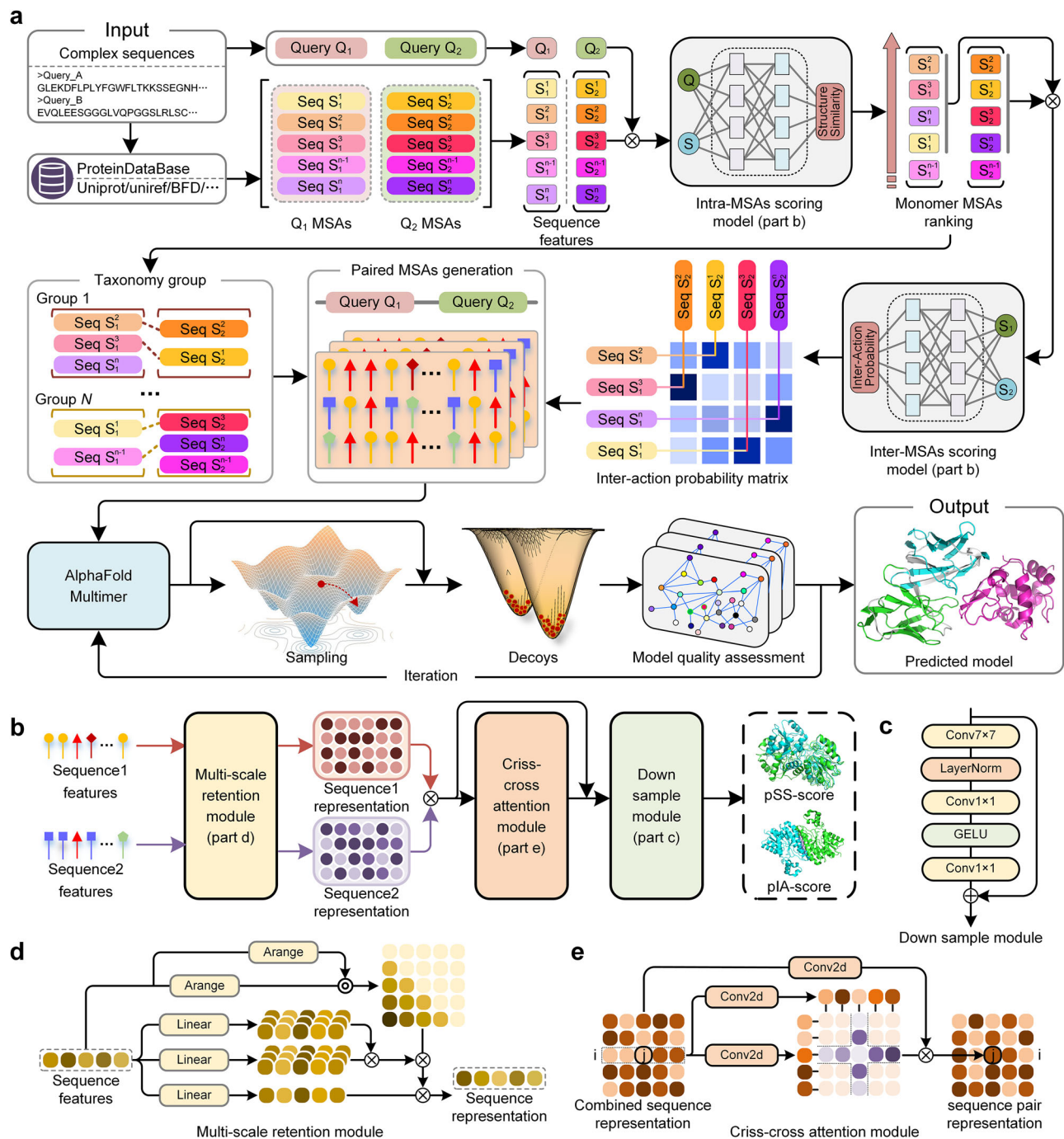
Protein structures are generally more functionally conserved than their corresponding sequences due to their direct involvement in mediating biological processes. This evolutionary conservation is particularly evident at the structural level of protein-protein interactions (PPIs), where interaction interfaces tend to be more conserved than sequence motifs. Extensive experimental evidence suggests that the repertoire of protein interaction modes in nature is remarkably limited<sup>41,42</sup>, with similar structural binding patterns observed across diverse PPIs<sup>43,44</sup>. In this case, leveraging deep learning architectures to directly extract spatial conformational complementarity information from monomeric protein sequences could provide structural insights, potentially enhancing the precision of protein complex structure prediction<sup>45</sup>.

In this work, we present DeepSCFold, which combines protein sequence embedding with physicochemical and statistical features through a deep learning framework to systematically capture structural complementarity<sup>46,47</sup> between protein chains and accurately predict complex quaternary structures. Benchmark evaluations on the CASP15 protein complex dataset demonstrate that DeepSCFold outperforms existing state-of-the-art methods in both global and local interface accuracy. To further validate the robustness of our method, we performed an evaluation using antibody-antigen complexes from the SABDab<sup>48</sup> database, focusing on challenging cases that may lack inter-chain co-evolution signals. Results demonstrate that structural complementarity-based paired MSAs can effectively compensate for the absence of co-evolutionary information by providing reliable inter-chain interaction signals.

## Results

### Overview of the method

DeepSCFold is a computational protocol specifically developed for high-accuracy prediction of protein complex structures. The method constructs paired multiple sequence alignments (pMSAs) by integrating two key components: (1) assessing structural similarity between monomeric query sequences and their corresponding homologs within individual MSAs, and (2) identifying potential interaction patterns among sequences across distinct monomeric MSAs. This dual-strategy approach enables the systematic generation of high-quality pMSAs, which serve as the foundation for accurate protein complex modeling. At the core of DeepSCFold's paired-MSA construction are two advanced sequence-based deep learning models: one predicts protein-protein structural similarity (pSS-score) purely from sequence information, while the other estimates interaction probability (pIA-score) based solely on sequence-level features. These models enable the inference of structural and interaction properties without relying on prior structural knowledge, making DeepSCFold uniquely capable of modeling complex interactions from sequence data alone. Figure 1 shows an overview of the DeepSCFold protocol. Starting from the input protein complex sequences, DeepSCFold first generates monomeric multiple sequence alignments (MSAs) from multiple sequence databases (UniRef30<sup>49</sup>, UniRef90<sup>50</sup>, UniProt<sup>50</sup>, Metaclust<sup>51</sup>, BFD<sup>51,52</sup>,



**Fig. 1 | Pipeline of DeepSCFold.** **a** DeepSCFold takes multiple query sequences of a protein complex as input, with each component sequence being independently searched against protein databases to generate monomeric multiple sequence alignments (MSAs). A deep learning model is designed to assess structural similarity between each sequence homolog in the MSA and the query sequence, enabling the ranking and selection of appropriate homologs. The pairing of homologs from different monomeric MSAs is accomplished through two distinct strategies: one leveraging species information or manually curated annotations, and another utilizing a probability matrix derived from an interaction probability prediction model that scores potential homolog pairs. These paired MSAs are

subsequently processed by AlphaFold-Multimer to generate potential complex structures. The resulting models undergo quality assessment, with the top-ranked model serving as a template for a single iteration of refinement, ultimately yielding the final predicted structure of the protein complex. **b** Deep learning model architecture: The features of a pair of sequences are input into a multi-scale retention module (**d**) to generate sequence representations, which are then fed into a criss-cross attention module (**e**) to generate paired representations, followed by a down sample module (**c**) to generate the final predicted scores (pSS or pIA). Source data are provided as a Source Data file.

MGnify<sup>53–55</sup>, and the ColabFold DB<sup>56</sup>). Then, the predicted pSS-score, which quantifies the structural similarity between the input sequence and its corresponding homologs in the monomeric MSAs, was employed as a complementary metric to traditional sequence similarity, thereby enhancing the ranking and selection process of

monomeric MSAs. Subsequently, the developed deep learning model predicts the pIA-scores for each potential pair of sequence homologs derived from distinct subunit MSAs. These interaction probabilities are then utilized to systematically concatenate monomeric homologs and construct paired MSAs, enabling the identification of biologically

relevant interaction patterns. Additionally, we integrate multi-source biological information, including species annotations, UniProt accession numbers, and experimentally determined protein complexes from the Protein Data Bank (PDB), to construct additional paired MSAs with enhanced biological relevance. Finally, DeepSCFold uses the series of paired MSAs constructed above to perform complex structure predictions through AlphaFold-Multimer. The top-1 model is selected based on our in-house complex model quality assessment method DeepUMQA-X and then used as the input template of AlphaFold-Multimer for one iteration to generate the final output structure.

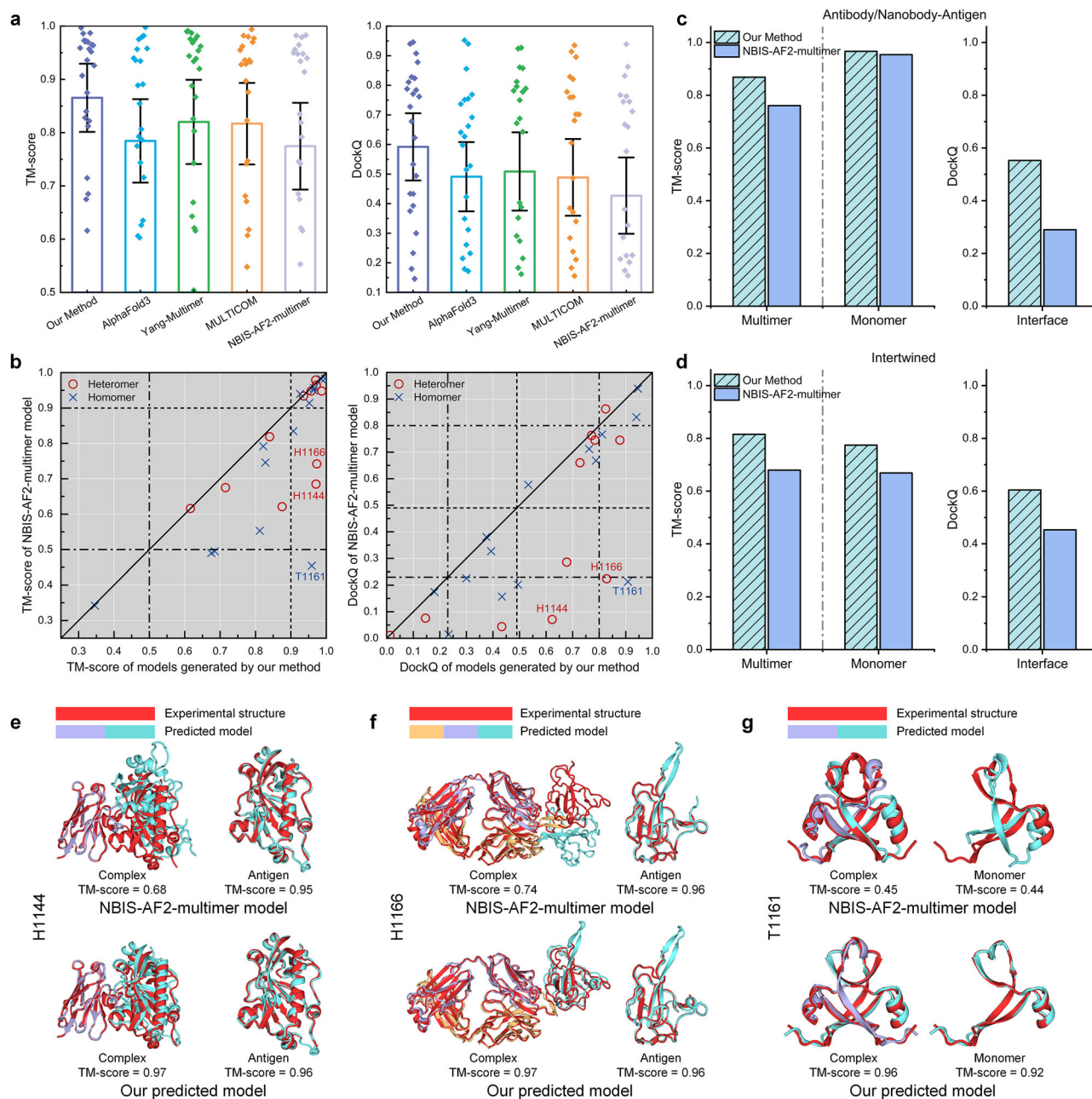
### Improvements of protein complex structure prediction by DeepSCFold

To evaluate the performance of DeepSCFold in predicting protein complex structures, we systematically tested the method on a benchmark set of multimeric targets from the CASP15 competition. For each target, complex models were generated using protein sequence databases available up to May 2022, ensuring a temporally unbiased assessment of our method's predictive capabilities. The predictions were then compared with those from several state-of-the-art methods, including AlphaFold3, Yang-Multimer, MULTICOM, and NBIS-AF2-multimer. The prediction results of AlphaFold3 models were generated using its online server (<https://golgi.sandbox.google.com/>), while the last three methods were retrieved from the CASP15 official website. To assess the accuracy of the predicted models, we employed two complementary metrics: TM-score, which quantifies the global topological similarity to the experimentally determined structure, and DockQ, which specifically evaluates the quality of protein-protein docking interfaces. A detailed description of these metrics is provided in Supplementary Note 1.

Figure 2a shows the TM-score and DockQ of the predicted multimer protein models with respect to the experimental structures. The detailed evaluation results are listed in Supplementary Table 7. DeepSCFold demonstrates superior performance in protein complex structure prediction, achieving an average TM-score of 0.87 and DockQ score of 0.59. This represents a significant improvement over current state-of-the-art methods, including AlphaFold3 (TM-score = 0.78, DockQ = 0.49), Yang-Multimer (TM-score = 0.82, DockQ = 0.51), MULTICOM (TM-score = 0.82, DockQ = 0.49), and NBIS-AF2-multimer (TM-score = 0.78, DockQ = 0.43). The consistent enhancement across both evaluation metrics highlights DeepSCFold's advanced capabilities in modeling protein complexes with greater accuracy and reliability. These findings indicate that the integration of sequence-derived structural complementarity in DeepSCFold likely enhances its ability to generate protein multimer models with improved global topological accuracy and biologically meaningful docking modes, as evidenced by its higher TM-score and DockQ. Given that NBIS-AF2-multimer serves as the implementation of AlphaFold-Multimer in CASP15, we performed a direct comparative analysis between DeepSCFold and NBIS-AF2-multimer to evaluate the effectiveness of our enhanced input strategy in improving the accuracy of multimer structure predictions. Figure 2b provides a detailed comparison of prediction accuracy between DeepSCFold and NBIS-AF2-multimer for each individual target. The results demonstrate that DeepSCFold outperforms NBIS-AF2-multimer, achieving superior TM-scores in 92% of the test cases and higher DockQ scores in 88% of the cases. To provide a more detailed evaluation, we categorized the results based on different performance thresholds for TM-score (0.5 and 0.9) and DockQ (0.23, 0.49, and 0.8), allowing for a nuanced comparison across various quality levels. Several cases show notable improvements in both metrics, with models improving from unreliable to good or high-quality levels, such as intertwined complexes (e.g., T1161), nanobody-antigen complexes (e.g., H1144), and antibody-antigen complexes (e.g., H1166).

To further elucidate the underlying mechanisms of these improvements, we conducted an in-depth analysis of these complex types, as illustrated in Fig. 2c, d. This focused investigation provides valuable insights into the specific structural features and interactions contributing to the enhanced predictive performance. Figure 2c shows a comparison of the results for the antibody/nanobody-antigen complex, showing the average multimer TM-score, monomer TM-score, and interface DockQ. Both methods demonstrate robust performance in monomer prediction, with DeepSCFold attaining a monomer TM-score of 0.97 and NBIS-AF2-multimer achieving 0.95. However, the critical distinction lies in their interface prediction capabilities. DeepSCFold achieves a substantially higher interface DockQ score of 0.55, outperforming NBIS-AF2-multimer's score of 0.29. This marked enhancement in interface prediction accuracy directly contributes to DeepSCFold's superior multimer TM-score of 0.87, compared to 0.76 for NBIS-AF2-multimer, highlighting the importance of precise interface modeling in multimer structure prediction. As a representative case study, in Fig. 2e we present an example from H1144, which is a nanobody-antigen complex containing two protein chains with 341 residues in total. Although NBIS-AF2-multimer generates a high-quality monomer model with a TM-score of 0.95, the quaternary orientation of the complex is completely wrong resulting in a poor complex TM-score of 0.68. In contrast, our method provides more accurate inter-chain signals, leading to a significantly improved complex model with a TM-score of 0.97. Figure 2f shows another illustrative case from H1166, which is an antibody-antigen complex with 577 residues in total. NBIS-AF2-multimer achieves a high monomer TM-score of 0.96, but the complex TM-score is only 0.74, whereas our model achieves 0.97. Antibody/nanobody-antigen complexes are inherently cross-species systems, which lack intra-species binding signals in their respective subunit sequence alignments (MSAs). Consequently, the MSAs generated by the NBIS-AF2-multimer pipeline are likely to be suboptimal, posing significant challenges for accurately predicting interchain interactions. In contrast, our approach may be beneficial in addressing this limitation by leveraging protein-protein structural complementarity to explore potential interaction signals, thereby enhancing the prediction of interchain relationships in such biologically relevant complexes.

In Fig. 2d, we provide a comparative analysis of the prediction results for the eight intertwined complexes, which are characterized by their highly interwoven and tight interfaces, presenting a challenging test for accurate structure prediction. The figure highlights the average values of multimer TM-score, monomer TM-score, and interface DockQ for both DeepSCFold and NBIS-AF2-multimer. Our method demonstrates superior performance across all metrics: the multimer TM-score improves from 0.68 (NBIS-AF2-multimer) to 0.82 (DeepSCFold), the monomer TM-score increases from 0.67 to 0.77, and the interface DockQ rises from 0.45 to 0.60. To further elucidate these findings, Fig. 2g showcases a representative example from T1161, an intertwined complex comprising 96 residues. In this case, the structure model predicted by NBIS-AF2-multimer achieves a monomer TM-score of 0.44, which significantly compromises the overall complex modeling accuracy, yielding a complex TM-score of 0.45. In contrast, DeepSCFold achieves a monomer TM-score of 0.92, resulting in a substantially higher complex TM-score of 0.96. These results underscore the importance of structure-level information inferred from sequences, which not only enhances the determination of inter-chain distances and orientations but also complements the co-evolutionary signals derived from monomeric MSAs constructed through sequence similarity. This dual approach enables more accurate and reliable predictions of complex structures, particularly in challenging cases such as intertwined complexes. A more detailed analysis of the CASP15 results is provided in Supplementary Note 4, along with Supplementary Figs. 2-3 and Supplementary Tables 2-3.



**Fig. 2 | Performance of DeepSCFold in protein complex structure prediction on the CASP15 dataset ( $n = 25$  protein targets).** **a** Average TM-scores and DockQ scores of the complex models generated by DeepSCFold, Yang-Multimer, MULTICOM, NBIS-AF2-multimer, and AlphaFold3. The data are presented as the mean  $\pm$  95% CI, with  $n = 25$  proteins from the benchmark set. **b** Head-to-head comparison of the TM-scores and DockQ scores on each test case between DeepSCFold and NBIS-AF2-multimer, with  $n = 25$  proteins from the benchmark set. The circles indicate heteromers, and crosses indicate homomers. Three representative cases are labeled in the figure and detailed analysis provided in (e, f and g). **c, d** Performance

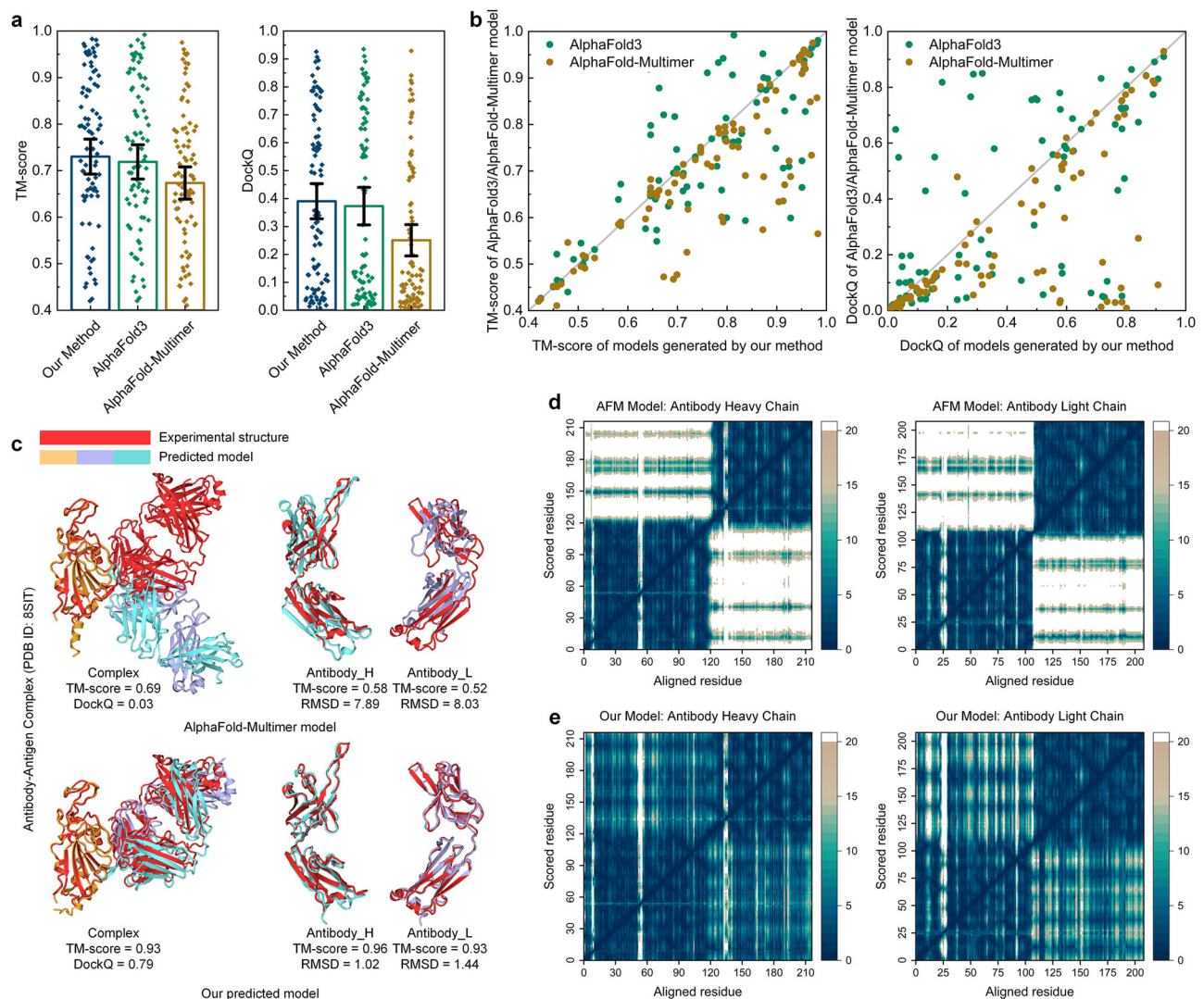
comparison for antibody/nanobody-antigen complexes ( $n = 8$  proteins) and intertwined complexes ( $n = 8$  proteins) from benchmark set, showing average TM-scores of multimers and monomers as well as interface DockQ scores. **e–g** Examples of the multimer structures built by DeepSCFold and NBIS-AF2-multimer for nanobody-antigen complex H1144 (e), antibody-antigen complex H1166 (f) and intertwined complex T1161 (g), with experimental structures shown in red and the monomers of the predicted models colored differently. The H1144 and H1166 complexes are aligned based on the antibody chains. Source data are provided as a Source Data file.

CASP16 featured the emergence of several advanced methods<sup>57,58</sup>, including AF3-server, Zheng-Multimer, Yang-Multimer<sup>59,60</sup>, and MULTICOM\_GATE<sup>61</sup>, which represent the current state of the art in protein complex structure prediction. Our method also participated in CASP16 under the group names Guijunlab-Complex (Group #148) and Guijunlab-Assembly (Group #312), demonstrating competitive performance against these top approaches. A detailed performance evaluations and analyses of our method in the CASP16 benchmark are

provided in Supplementary Note 3, along with Supplementary Fig. 1 and Supplementary Table 1.

### Structural modeling for antibody-antigen complexes with DeepSCFold

Given the inherently weak or absent co-evolutionary signals between antibody and antigen sequences, we conduct a focused investigation into the performance of antibody-antigen complex modeling, with



**Fig. 3 | Results of antibody-antigen complex structure prediction ( $n = 89$  protein targets).** **a** Average TM-scores of complex structures and DockQ scores of antibody-antigen interface for the predicted models by DeepSCFold, AlphaFold-Multimer and AlphaFold3. The data are presented as the mean  $\pm$  95% CI, with  $n = 89$  proteins from the benchmark set. **b** Head-to-head comparison of the TM-scores and DockQ scores on each test case between DeepSCFold, AlphaFold-Multimer and

AlphaFold3, with  $n = 89$  proteins from the benchmark set. **c** Example of an antibody-antigen complex (PDB ID: 8SIT) predicted by DeepSCFold and AlphaFold-Multimer, where experimental structures are colored in red and monomers of the predicted models are colored differently. **d, e** Predicted aligned error (PAE) plots for the antibody heavy and light chains modeled by AlphaFold-Multimer (**d**) and DeepSCFold (**e**). Source data are provided as a Source Data file.

particular emphasis on the impact of complex paired multiple sequence alignments (MSAs). We evaluated DeepSCFold on a curated set of antibody-antigen targets extracted from SAbDab, spanning entries from January 25, 2024, to June 1, 2024. This temporal separation ensures no overlap with the training data of the deep learning model or the protein database utilized for structure modeling, thereby maintaining the integrity of the evaluation. For a fair and objective comparison, both DeepSCFold and AlphaFold-Multimer were provided with identical template and monomeric MSAs. Notably, the complex paired MSA employed by DeepSCFold was constructed by deriving alignments for each subunit from their respective monomeric MSAs. Additionally, we included a comparison with the latest AlphaFold3<sup>26</sup>, with results obtained directly from its online server (<https://alphafoldserver.com/>) to ensure up-to-date benchmarking.

Figure 3a illustrates the complex TM-scores and interface DockQ values of the predicted antibody-antigen complex models, benchmarked against their corresponding experimental structures in the PDB database. The interface DockQ specifically assesses the

interaction between the antibody and antigen chains, excluding the interface between the antibody's heavy and light chains. Detailed evaluation results are provided in Supplementary Table 8. On average, DeepSCFold achieves a complex TM-score of 0.73 and an interface DockQ of 0.39, outperforming both AlphaFold-Multimer (TM-score = 0.67, DockQ = 0.25) and AlphaFold3 (TM-score = 0.72, DockQ = 0.37). Figure 3b provides a comparative analysis of prediction accuracy between DeepSCFold and AlphaFold-Multimer/AlphaFold3 for each target. DeepSCFold achieves higher TM-scores than AlphaFold3 in 59.6% of the test cases and higher DockQ scores in 60.7% of the cases. Additionally, the prediction success rate for the antibody-antigen binding interface (DockQ > 0.23) is improved by 12.4% with DeepSCFold compared to AlphaFold3. These results underscore the complementary nature of the two methods, suggesting that DeepSCFold and AlphaFold3 can synergistically enhance overall prediction performance. Compared to AlphaFold-Multimer, DeepSCFold demonstrates superior performance in 77.5% of test cases based on TM-score and 92.1% based on DockQ, highlighting its ability to effectively

augment the performance of its structural prediction component, AlphaFold-Multimer, through the generation of paired MSAs. Notably, while AlphaFold-Multimer produced models with DockQ < 0.23 for 56 targets, DeepSCFold successfully built antibody-antigen complex models with DockQ > 0.23 for 39.3% of these challenging cases. This improvement suggests that the complex paired MSAs constructed based on structural complementarity provide valuable inter-chain interaction signals, thereby enhancing the accuracy of complex modeling.

A notable example is the complex structure of the SARS-CoV-2 spike receptor-binding domain with the broadly neutralizing antibody CC84.24 Fab (PDB ID: 8SIT). AlphaFold-Multimer's prediction for this protein complex exhibits significant errors in inter-chain orientation, achieving an overall TM-score of 0.63 and an interface DockQ of 0.03 (Fig. 3c). Furthermore, AlphaFold-Multimer generated low-quality monomer models for the antibody, with the heavy chain (TM-score = 0.58, RMSD = 7.89 Å) and the light chain (TM-score = 0.52, RMSD = 8.03 Å) both showing substantial deviations from the experimental structure. We further assess the inter-domain accuracy of single-chain structure model by the predicted aligned error (PAE) plot, which captures the model's global inter-domain structural errors<sup>62</sup>. The PAE plots for the antibody heavy chain and light chain structures predicted by AlphaFold-Multimer are shown in Fig. 3d. These plots reveal two distinct low-error squares corresponding to the two domains of each single-chain structure. While the PAE values within individual domains are low, with each domain achieving a TM-score above 0.9, the inter-domain orientations exhibit high PAE values, indicating that AlphaFold-Multimer fails to accurately predict the relative domain orientations in this case. In contrast, DeepSCFold not only accurately models each domain but also captures the correct inter-domain and inter-chain orientations, as evidenced by the low PAE values across the entire region for both the heavy and light chains (Fig. 3e). These results highlight that, for flexible proteins such as antibodies, while monomer MSAs can effectively infer domain-level topology, accurate structure modeling often requires complex-wide MSAs to capture the interaction signals necessary for resolving their orientations. This emphasizes the importance of incorporating structural complementarity and inter-chain interaction signals in predicting biologically relevant conformations of antibody-antigen complexes.

In addition, we constructed a benchmark set comprising 34 experimentally resolved virus-human complexes and compared DeepSCFold with AlphaFold3 and AlphaFold-Multimer. The results indicate that DeepSCFold more effectively captures the structural patterns of cross-species interactions. Detailed experimental results are provided in Supplementary Note 5, along with Supplementary Fig. 4 and Supplementary Tables 4-5. Furthermore, Supplementary Note 6, together with Supplementary Figs. 5-7 and Supplementary Table 6, provides a detailed analysis of the components involved in structure prediction.

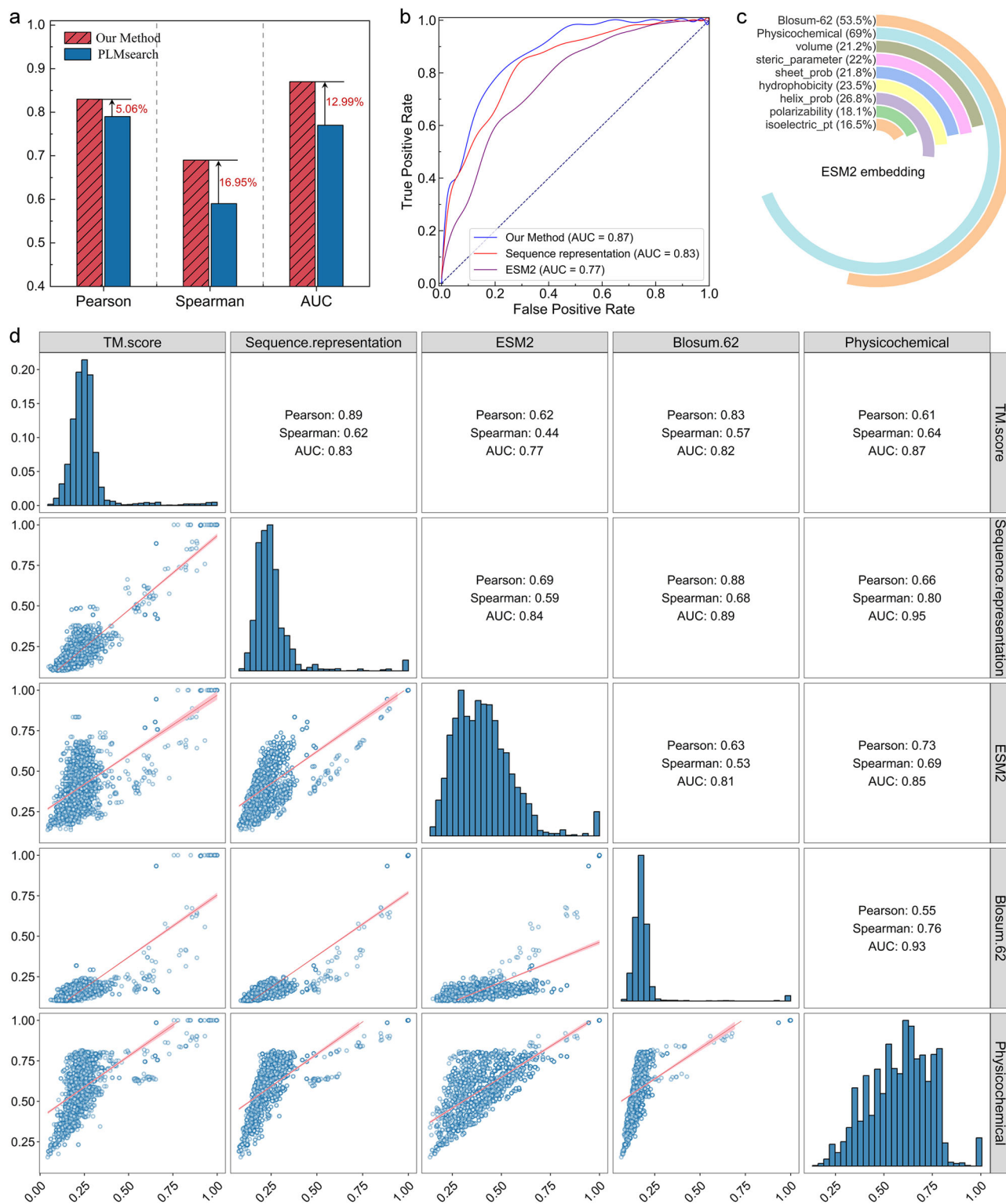
### The performance of sequence-based structural similarity prediction

One core component of DeepSCFold is the prediction of protein-protein structural similarity directly from sequence data. The co-evolutionary signals derived from homologous protein searches are essential for protein structure prediction, as they shed light on residue-residue interactions and evolutionary constraints that shape protein folding and complex formation. Due to the low cost and large scale of sequence data, sequence-based search methods are generally more convenient and faster than structure-based methods, which is especially evident in scenarios involving a large number of sequences<sup>45</sup>, such as metagenomic sequences<sup>53</sup> and antibody variant sequences<sup>63</sup>. Nevertheless, in cases of highly distant evolutionary relationships, sequences may have diverged so extensively that identifying their relatedness becomes challenging<sup>64</sup>. This challenge highlights the need

for developing computational approaches that can directly infer protein-protein structural similarity from sequence information, thereby complementing existing structure prediction methods.

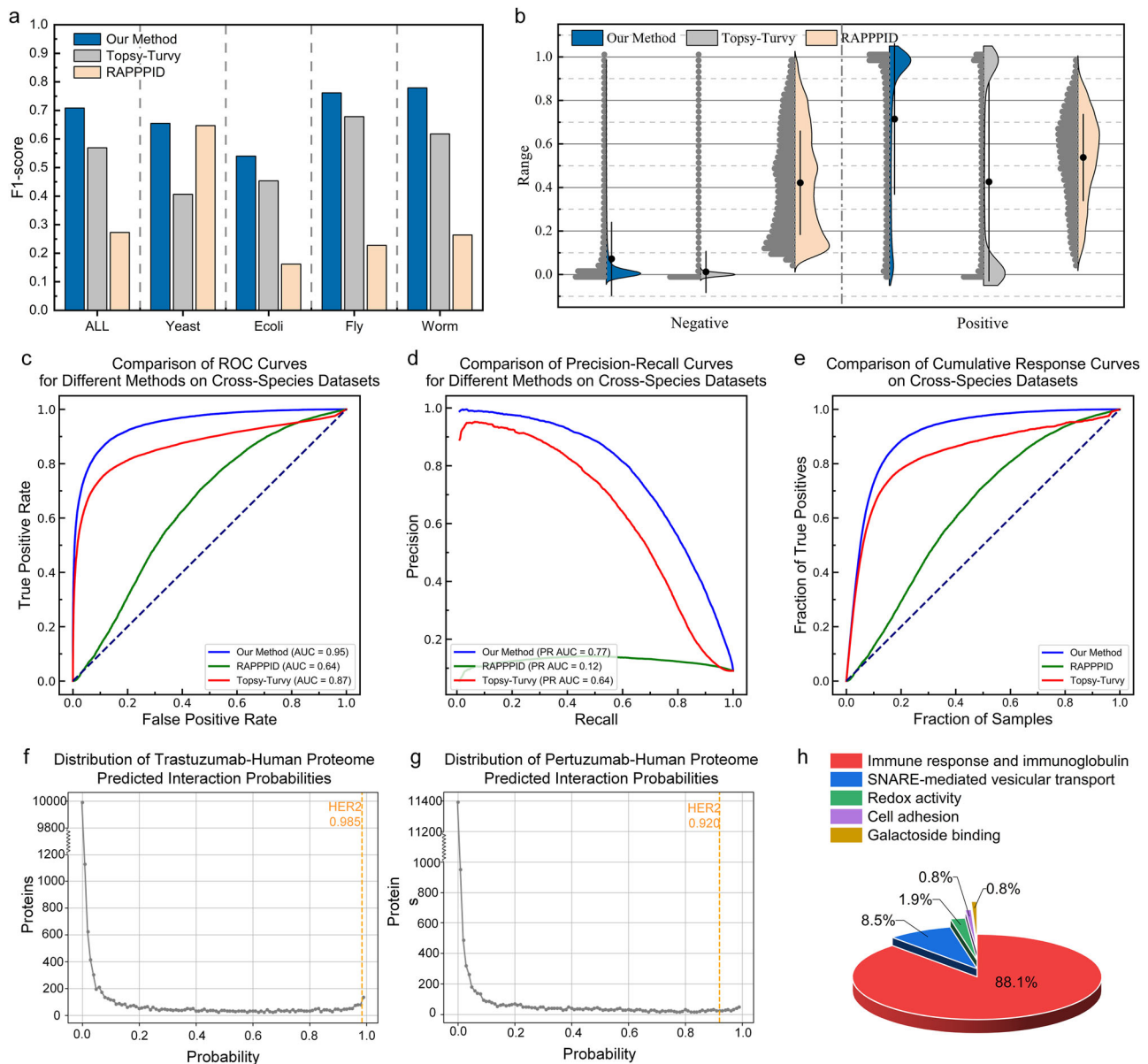
To enhance detection performance while ensuring the versatility and efficiency of sequence search, we developed a deep learning model that directly predicts protein structural similarity from sequences, serving as one core component of DeepSCFold (see details in Methods). We hypothesize that structurally similar protein sequences, despite exhibiting substantial sequence divergence, can effectively address the limitations of conventional sequence alignment methods by providing crucial structural topology information, thereby establishing a robust foundation for constructing high-quality multiple sequence alignments (MSAs) of monomeric proteins, which is essential for achieving high-accuracy structural modeling. This conceptual framework holds particular significance for the alignment analysis of distantly related homologous proteins, offering insights into the evolutionary relationships among protein families with low sequence similarity. Our predicted protein-protein structural similarity (pSS-score) is represented by the template modeling score (TM-score), which is a widely used metric for assessing the global similarity between two protein structures. For a comprehensive and objective analysis, we established a benchmark dataset comprising CASP15 target proteins, implementing a temporal cutoff to guarantee complete independence from the training data, and compared the performance of our method with the state-of-the-art method PLMsearch<sup>45</sup> using Pearson<sup>65</sup>, Spearman<sup>66</sup>, and ROC (AUC) statistical metrics. As shown in Fig. 4a, the correlation between the pSS-score and the true TM-score (calculated by the US-align<sup>67</sup> tool) for protein pairs in the dataset was measured using the Pearson and Spearman metrics. Our method achieved a Pearson correlation of 0.83 compared to 0.79 for PLMsearch, indicating a 5.06% improvement. Notably, the Spearman correlation, which evaluates the rank-order relationship, was 0.69 for our method and 16.95% higher than the 0.59 achieved by PLMsearch. This significant improvement in Spearman correlation demonstrates that our model more effectively preserves the relative ordering of structural similarities, which is crucial for applications involving the ranking of protein similarity. Regarding the receiver operating characteristic (ROC) curve analysis, which measures the model's discriminative capability between positive and negative samples, our method demonstrated superior performance with an area under the curve (AUC) of 0.87, representing a 12.99% improvement over PLMsearch's AUC of 0.77. For this evaluation, positive cases were rigorously defined as protein pairs exhibiting true TM-scores within the highest quartile of the dataset distribution. This further demonstrates the improved classification ability of our method in identifying high-confidence protein pairs, demonstrating robust performance in structural similarity prediction. The details of these results are provided in Supplementary Table 9.

The comparative analysis presented in Fig. 4b systematically investigates the potential factors contributing to the enhanced performance of our method through a comprehensive evaluation against ESM2 embedding and the sequence representation generated from our method. In the ESM2 method, we used the sequence embeddings generated by the ESM2 model and calculated similarity scores between these embeddings, which resulted in an AUC of 0.77. This result shows that the protein language model effectively captures the biological information encoded in the sequence, and both our method and PLMsearch use ESM-based embeddings as a foundation. Importantly, our method further combines additional sequence-derived features to generate the sequence representation, such as physicochemical properties and amino acid substitution probabilities<sup>68</sup>. To evaluate their impact, we extract sequence representations through the multi-scale retention module with the concatenated features, and then calculated the similarity scores using the same procedure as the ESM2 baseline. The results show that



**Fig. 4 | Comparative performance and feature analysis for protein-protein structural similarity prediction ( $n = 2701$  pairs).** **a** Performance comparison between our method and PLMsearch based on Pearson correlation, Spearman correlation, and ROC AUC metrics. The data are presented as the mean, with  $n = 2701$  pairs from the benchmark set. **b** ROC curves showing the classification performance of our method, ESM2, and Sequence representation, with  $n = 2701$

pairs from the benchmark set. **c** Correlation analysis between ESM2 embeddings and additional sequence-derived features, with  $n = 2701$  pairs from the benchmark set. **d** Feature distributions and pairwise correlations, with  $n = 2701$  pairs from the benchmark set. Histograms represent the distribution of each feature, scatter plots show pairwise relationships, and upper triangle statistics include Pearson, Spearman correlations, and AUC values. Source data are provided as a Source Data file.



**Fig. 5 | Comparative analysis and practical applications of protein-protein interaction (PPI) prediction, with  $n = 17,000$  positive samples and 170,000 negative samples.** **a** F1-scores of our method with the state-of-the-art methods Topsy-Turvy and RAPPPID across different species, *Yeast* ( $n = 5000$  positive / 50,000 negative), *E. coli* ( $n = 2000$  positive / 20,000 negative), *Fly* ( $n = 5000$  positive / 50,000 negative), *Worm* ( $n = 5000$  positive / 50,000 negative), and the aggregated dataset (ALL,  $n = 5000$  positive / 50,000 negative). **b** Violin plots showing the distribution of interaction probability scores for positive and negative protein pairs across methods. The data are presented as mean  $\pm$  s.d., with  $n = 17,000$  positive samples and 170,000 negative samples from the benchmark

set. **c–e** Receiver operating characteristic (ROC) curve, Precision-recall (PR) curve, and cumulative response curve (CRC) analysis of our method compared to Topsy-Turvy and RAPPPID, with  $n = 17,000$  positive samples and 170,000 negative samples from the benchmark set. **f, g** Distribution of interaction probabilities of Trastuzumab (**f**) and Pertuzumab (**g**) with the human proteome ( $n = 17,286$  proteins from the STRING database), highlighting high interaction scores with HER2 (Trastuzumab: 0.985, Pertuzumab: 0.920). **h** Clustering result of proteins with high predicted interactions to Trastuzumab and Pertuzumab, with  $n = 270$  human proteins. Source data are provided as a Source Data file.

the AUC of the sequence representation reaches 0.83, indicating that the additional features, which contain biological constraints, positively contribute to the performance compared to the ESM2 embedding alone. In addition, the performance gap between sequence representation (AUC = 0.83) and our full method (AUC = 0.87) demonstrates the importance of our classification module, which effectively utilizes the combined features to achieve more reliable classification and ranking performance.

In Fig. 4c, we conducted a comprehensive analysis of the correlation between ESM2 embeddings and various sequence-derived features to evaluate their potential complementarity. The circular

diagram illustrates these relationships, where each arc corresponds to a distinct feature, with the accompanying percentage quantitatively representing its correlation with the ESM2 embedding. The overall physicochemical feature set exhibited the highest correlation with ESM2 embeddings (69.0%), indicating some overlap in the biological information captured by both. However, individual properties such as hydrophobicity (23.5%), helix probability (26.8%), and polarizability (18.1%) showed relatively low correlations, suggesting that each property contributes unique information not fully captured by the ESM2 model. Notably, the amino acid substitution matrix (Blosom-62) shows only a 53.5% correlation with ESM2 embeddings, demonstrating

that evolutionary information encoded by substitution probabilities provides complementary insights.

Specifically, we present a comprehensive overview of the distribution and correlation among various features and their relationship with true TM-score (Fig. 4d). This experimental design enables the identification of key components driving the performance improvement while providing insights into the relative contributions of different sequence features. The figure consists of histograms along the diagonal, illustrating the distribution of values for each individual item: TM-score, sequence representation, ESM2 embeddings, Blosum-62, and physicochemical properties. The sequence representation exhibited the highest correlation with TM-score (Pearson's  $r = 0.89$ , Spearman's  $\rho = 0.62$ ), demonstrating the effectiveness of integrating diverse biological information. Although ESM2 embeddings alone showed relatively low correlations (Pearson's  $r = 0.62$ , Spearman's  $\rho = 0.44$ ), they still contribute valuable biological information to structural similarity prediction, as detailed result of ablation study in Supplementary Note 6, along with Supplementary Figs. 5–7 and Supplementary Table 6. The Blosum-62 matrix demonstrated a Pearson correlation of 0.83 and a Spearman correlation of 0.57 with TM-score, indicating the contribution of evolutionary information to structural similarity. Physicochemical features also showed a strong relationship with structural similarity, with a Spearman correlation of 0.64 and an AUC of 0.87. Furthermore, the correlations between different features varied, indicating that while there was some overlap, each feature set provided unique information. These results demonstrate the necessity of integrating diverse features, as this combination allows the model to capture a more comprehensive representation of biological characteristics, thereby improving its ability to predict sequence-to-structure relationships.

### The performance of protein-protein interaction probability from sequences

DeepSCFold directly predicts protein-protein interaction (PPI) probabilities through sequence data analysis as another key capability. Many existing methods for constructing protein paired MSAs rely on experimental data and manually annotated information, such as species-specific annotations. This reliance may limit their applicability, particularly when dealing with less-studied or unknown species. If the interaction probability between two given sequences could be directly predicted, it would greatly expand the data sources available for MSA construction. This approach would allow the use of large-scale, unannotated protein sequence databases, such as the BFD<sup>51,52</sup> and MGnify<sup>53–55</sup> metagenomic databases, which have been shown to contribute significantly to protein structure prediction.

To evaluate the performance of our method on interaction prediction, we trained our model on a dataset consisting of human protein pairs, and compared the results with those of state-of-the-art PPI prediction methods, Topsy-Turvy<sup>69</sup> and RAPPID<sup>70</sup>, across four distinct species: *Saccharomyces cerevisiae* (Yeast), *Escherichia coli* (*E. coli*), *Drosophila melanogaster* (Fly), and *Caenorhabditis elegans* (Worm). Figure 5a illustrates the F1-scores of our method compared to Topsy-Turvy and RAPPID across all test species, including an aggregated result (“ALL” in Fig. 5a). Our method consistently achieved higher F1-scores, indicating its strong ability to balance precision and recall, as well as its robustness in predicting interactions within diverse biological systems. Notably, the model performed better in *Drosophila* and *C. elegans*, two organisms known for their intricate PPI networks due to complex developmental processes and signaling pathways. These results indicate that our method has more robust cross-species generalization performance and effectively captures potential co-evolutionary signals. The details of results are provided in Supplementary Table 10.

The violin plots in Fig. 5b display the distributions of interaction probability scores for both negative and positive protein pairs across

all test datasets. For each method, the plot indicates the spread, median, and overall distribution of predicted scores. For negative samples, our method shows a narrow distribution close to zero, highlighting its ability to accurately classify non-interacting pairs with a low false negative rate. However, the most significant observation is seen with positive samples. Moreover, our method exhibits a concentrated distribution of probabilities above 0.8 for positive samples, indicating high certainty in correctly identifying the interaction pairs. In contrast, the distribution of Topsy-Turvy is divided into two regions, one with high confidence and the other with low confidence, revealing the limitations of its predictions. The results of RAPPID showed that the distributions were not well-separated, leading to low overall discrimination performance. In summary, our method shows better performance in distinguishing interaction pairs, with potential to identify structural complementarity between proteins.

To further assess the robustness of DeepSCFold in predicting protein-protein interactions (PPIs), we employed three standard evaluation metrics: the receiver operating characteristic (ROC) curve, the precision-recall (PR) curve, and the cumulative response curve (CRC). The ROC curve analysis compares the true positive rate (sensitivity) to the false positive rate across different decision thresholds (Fig. 5c). DeepSCFold achieved an area under the curve (AUC) of 0.95, outperforming RAPPID (AUC = 0.64) and Topsy-Turvy (AUC = 0.87). The higher AUC indicates that DeepSCFold consistently performs well in distinguishing true interactions from non-interactions across a wide range of confidence thresholds. The PR curve analysis (Fig. 5d) provides insight into model performance under imbalanced conditions, where the positive class (interacting protein pairs) is less frequent, as is typical in real-world PPI datasets. The PR AUC for DeepSCFold was 0.77, outperforming RAPPID (0.12) and Topsy-Turvy (0.64). This higher PR AUC reflects DeepSCFold's ability to maintain high precision without sacrificing recall, which is essential for minimizing false positives while retaining a large proportion of true interactions. The cumulative response curve in Fig. 5e illustrates the fraction of true positive PPIs identified as a function of the sample size evaluated. The curve of DeepSCFold rises more steeply than those of Topsy-Turvy and RAPPID, indicating that it can capture a considerable proportion of true interactions with fewer samples. This rapid enrichment of true positives makes DeepSCFold effective for identifying high-confidence PPIs early in exploratory studies, thereby accelerating the discovery of potential interactions. The analysis of our method compared to Topsy-Turvy and RAPPID on each species are provided in Supplementary Fig. 8. In addition, a detailed analysis of the contribution of input features to protein-protein interaction prediction is provided in Supplementary Note 6, along with Supplementary Figs. 5–7 and Supplementary Table 6.

To demonstrate the practical utility of DeepSCFold, we performed an in-depth analysis of the interaction probabilities of two well-studied antibodies, Trastuzumab and Pertuzumab<sup>71</sup>, with the entire human proteome (17,286 proteins from the STRING<sup>72</sup> database). The distribution of predicted interaction scores (Fig. 5f for Trastuzumab and Fig. 5g for Pertuzumab) confirms that the model assigns high interaction probabilities to known interactions with HER2, a critical receptor targeted by both antibodies. The results for Trastuzumab showed an interaction score of 0.985 with HER2, while Pertuzumab's score was 0.920, aligning with biological evidence of their binding specificity. The sharp distribution of interaction probabilities in these analyses highlights DeepSCFold's potential in large-scale interaction screening, paving the way for more efficient drug discovery and functional annotation of proteins.

For Trastuzumab and Pertuzumab, we extracted proteins with predicted interaction probabilities higher than those for their known primary target, HER2. This resulted in 270 unique proteins with potentially strong interactions. To assess whether these predictions might reflect random noise or exhibit biological patterns, we

performed clustering analysis of these 270 proteins using STRING. The proteins were grouped into five clusters, each with distinct functional enrichment profiles. As shown in Fig. 5h, Cluster 1 (red) comprises 229 proteins primarily associated with adaptive immune response and immunoglobulin-like domains, which may share structural features with antibodies and thus receive higher interaction scores. Other smaller clusters exhibit enrichment in functions such as SNARE-mediated vesicular transport, redox activity, cell adhesion, and galactoside binding. These findings indicate that DeepSCFold may tend to assign high interaction probabilities to proteins with structural or functional relevance to the antibody scaffold or binding context. This provides preliminary evidence that the model's predictions, even when not target-specific, may still reflect biologically meaningful patterns rather than arbitrary associations.

## Discussion

Accurately capturing inter-chain orientations and determining the structure of protein complexes remains a significant challenge, as complex structure prediction is less accurate than single-chain modeling. In this work, we have proposed DeepSCFold, a protocol for automated protein complex prediction that employs deep learning networks to capture potential structural complementarity and construct complex MSAs. The results demonstrate that DeepSCFold could provide interaction signals that effectively improve the structural prediction accuracy of AlphaFold-Multimer.

It has been reported that the ways of protein-protein interactions in nature may be limited, with similar binding modes can be identified for almost all known protein-protein interactions. To capture interaction signals indicative of protein-protein binding modes, DeepSCFold employs deep learning models to predict protein structural similarity and interaction probabilities, which are subsequently used to construct complex MSAs. Notably, current sequence concatenation mechanisms based on species annotations are limited to genomic sequences and thus cannot fully utilize the highly informative homologous sequences available in metagenomic databases to guide multi-chain structure assembly. DeepSCFold avoids this limitation and provides a possibly way for the monomeric MSAs concatenation. By integrating complex MSAs with the state-of-the-art AlphaFold-Multimer modeling method, DeepSCFold enhances the accuracy of complex structure prediction, as borne out by experimental results on the CASP15 protein complex dataset.

Antibody-antigen complexes are widely recognized as lacking inter-chain co-evolution, as it is not possible to find orthologs between host and pathogenic proteins as there is no species overlap (e.g., humans and a virus, originating from different kingdoms). However, when applied to antibody-antigen complex modeling, DeepSCFold achieved a 24.7% higher success rate in predicting antibody-antigen interfaces (DockQ > 0.23) compared to AlphaFold-Multimer, using the same monomeric MSAs and templates. This improvement provides further evidence supporting our hypothesis that the enhanced performance stems from DeepSCFold's ability to capture shape and physicochemical complementarity through its structure-aware complex MSA construction, overcoming the inherent limitations of inter-chain co-evolution in antibody-antigen systems.

Despite the improved performance in complex modeling, there are still some challenges for DeepSCFold. Users are required to provide stoichiometric information for the complex (e.g., the copy number of each component chain) prior to executing DeepSCFold, which may limit its practical applicability. Accurate model quality assessment (MQA) is crucial for reliably distinguishing high-accuracy predictions from incorrect ones, especially in the absence of experimental structures for validation. Moreover, MQA provides users with a self-evaluation confidence score for the model, offering valuable insights into the reliability of the predicted models. Extending DeepSCFold beyond protein-protein complexes to simulate structures formed by

interactions between DNA, RNA, and other ligands and proteins is also a topic to explore in our ongoing research.

## Methods

### Training set

For training a sequence-based structural similarity prediction model, we collected protein complex structures from the Protein Data Bank (PDB, <https://www.rcsb.org/>) up until June 2022 and applied the following filtering criteria: (1) resolution better than 3.5 Å; (2) with the length of each chain greater than 50 residues; (3) the number of chains in the complex does not exceed three; and (4) the complex structure does not contain DNA or RNA. After filtering, a total of 36,955 protein complexes were retained. Next, we employed DomainParser<sup>73</sup> to split the protein chains in these complexes into domains, and then clustered the resulting candidate domains using MMseqs2<sup>36</sup> with a 30% sequence identity threshold. After this step, we obtained 15,636 non-redundant domains. To assess the structural similarity between these domains, we performed pairwise comparisons of the non-redundant domains using USalign<sup>67</sup> and calculated the TM-score, which ranges from 0 to 1, with higher scores indicating greater structural similarity. To address the potential bias in the data distribution, we divided the TM-scores into 10 bins, each spanning a 0.1 interval, ensuring that each bin contained an equal number of domain pairs. This process resulted in ~100,000 domain pairs for subsequent training.

For training a sequence-based interaction probability prediction model, we extracted positive examples from the STRING<sup>72,74</sup> database (v11.5) based on physical binding interactions associated with high experimental evidence scores. From this set, we removed protein pairs involving very short proteins (shorter than 50 amino acids) and, due to GPU memory constraints, also excluded proteins longer than 1000 amino acids. Next, we removed protein-protein interactions with high sequence redundancy from the dataset to prevent the model from memorizing interactions based on sequence similarity alone. Specifically, we clustered the proteins at a 40% sequence identity threshold using MMseqs2<sup>36</sup>. A PPI (A-B) was considered redundant and excluded if another PPI (C-D) had already been selected, where the protein pairs (A, C) and (B, D) were both found within the same clusters. After this step, we obtained 39,183 positive samples. Non-interactions ('negatives') were sampled randomly from the respective proteins in the positive set, with a positive-to-negative sample ratio of 1:10.

### Test set

To assess the performance of protein complex prediction, a total of 25 multimer targets were selected from the CASP15 dataset, as listed under the MULTIMERS category on the official CASP website (<https://predictioncenter.org/casp15/>). These targets were chosen based on their suitability for structural modeling using AlphaFold-Multimer on a single NVIDIA A100 GPU, which is widely used for large-scale protein structure predictions. All selected targets have experimentally determined structures available in the Protein Data Bank (PDB), providing a reliable benchmark for comparison. To evaluate the performance of DeepSCFold on protein complexes lacking inter-chain co-evolution information, we further constructed an independent test set of antibody-antigen protein complexes. We collected antibody-antigen complexes with antigen types classified as protein from the SABDab<sup>48</sup> database, spanning the period from January 25, 2024, to June 1, 2024. The complexes included in the dataset had a sequence length of fewer than 2000 amino acids and a resolution <5 Å. For complex entries sharing the same protein ID, clustering was performed using MMseqs2<sup>36</sup> with a sequence identity threshold of 100%, leading to a final set of 89 Antibody-Antigen protein complexes.

To objectively evaluate the performance of structural similarity prediction model, we constructed a test dataset with a temporal cutoff to ensure its independence from the training data. We split the protein targets from CASP15 into individual chains and then clustered the

candidate chains using MMseqs2 with a 100% sequence identity threshold, resulting in a final set of 74 unique single-chain proteins. Finally, USalign was employed to perform pairwise comparisons and calculate TM-scores, resulting in a total of 2,701 protein pairs. Furthermore, we evaluated the performance of DeepSCFold in predicting interaction probabilities using a publicly available interaction dataset<sup>69,72</sup>. The dataset includes data from four species: *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Escherichia coli*. For each of the four model organisms, there are 5000 positive and 50,000 negative interactions, except for *Escherichia coli* (2000/20,000), resulting in a 1:10 positive-to-negative ratio.

The prediction results on these datasets are included in the Supplementary Data file. The datasets are provided through the general repositories Zenodo.

### Structural similarity and interaction probability prediction

Paired-MSA is crucial for protein multimer structure modeling. However, the construction method that only uses traditional sequence similarity combined with existing complex structure information may still face difficulties in providing high-quality paired MSAs. To efficiently construct paired MSAs, we developed a sequence-based deep learning model to capture the relationships between sequences. The process begins by extracting sequence-derived features and integrating them with embeddings obtained from a protein language model (PLM). Next, the combined features of the two query sequences are independently encoded into sequence representations through a multi-scale retention module. To capture inter-sequence relationships, these representations are then processed by a cross-attention module, which generates a sequence pair representation. Finally, this pair representation is passed through a down-sampling module to predict the structural similarity score (pSS-score) or the interaction probability score (pIA-score).

### Model architecture and input features

We develop a deep learning model to predict the inter-chain interaction and the structure similarity from the input sequences. The input features for the network encompass a diverse set of information, including residue type, positional information, evolutionary relationships, physicochemical properties, and contextualized embeddings. Specifically, these features include one-hot encoding, normalized distance to sequence end, Blosum-62 substitution matrix, physicochemical features<sup>68</sup>, and embeddings from the pre-trained ESM2<sup>40</sup> language model. Together, these features provide a comprehensive representation of each residue, capturing sequence, structural, and functional properties. Detailed descriptions of all input features are provided in Supplementary Table 11.

The network architecture consists of a multi-scale retention module, a criss-cross attention module, and a down-sample module. Initially, the input query protein sequence features are processed and fused to generate task-specific sequence representations. The ESM2 embeddings are first passed through a linear layer for dimensionality reduction, followed by ReLU activation and LayerNorm to introduce non-linearity and normalize the feature distributions. This ensures that the embedding dimensions are appropriately aligned with the 1D features, preventing a large dimensional gap that could otherwise obscure the contribution of the 1D features. The reduced ESM2 embeddings are then concatenated with the 1D input features to form a combined feature embedding. Subsequently, the output is fed into a multi-scale retention module to generate a sequence representation by capturing long-range dependencies. This module consists of 8 layers, with each layer combining a multi-scale retention mechanism and a feed-forward network to refine the information. Next, the relationships between the proteins are considered. The sequence representations of the two queries are combined through matrix multiplication, followed by Conv2D and ELU activation to generate a

combined sequence representation. This representation is then processed through the criss-cross attention module that captures interactions along both the horizontal and vertical axes to generate a sequence pair representation. The module computes separate attention maps for each direction using the query, key, and value projections, followed by softmax normalization to generate attention weights. These attention weights are applied to the value tensors, producing an updated sequence pair representation that captures the relationships between the query proteins. Finally, the sequence pair representation is fed into the down-sample module, which consists of four down-sample blocks and four residual blocks. Each down-sample block is composed of a LayerNorm followed by a convolutional layer, with the first block using a  $4 \times 4$  convolution and the other three blocks using  $2 \times 2$  convolutions. The residual blocks are structured with a series of operations: a  $7 \times 7$  convolution, LayerNorm, a  $1 \times 1$  convolution, GELU activation, and a second  $1 \times 1$  convolution. The output of the down-sample module is passed through a fully connected layer, followed by a sigmoid function to generate the final prediction scores.

### Training

During the training, 90% of the protein pairs are randomly selected from the development set as the training set, and the remaining 10% as the validation set. The network is implemented in Python using PyTorch and trained for up to 30 epochs. Due to GPU memory constraints, the training batch size starts at 4 and is progressively reduced as training progresses: it is set to 4 for the first 10 epochs, reduced to 2 for the next 10 epochs, and finally lowered to 1 for the last 10 epochs. The Adam optimizer is employed to minimize the loss, with the total loss calculated using the Log-Cosh Loss function for the structural similarity prediction model and the Focal Loss function for the protein-protein interaction prediction model. The learning rate is initially set to  $1e-5$  and decreases progressively during training, following an exponential decay strategy. The network model with the least validation loss is selected for downstream tasks. It takes about 14 days to train the network using one NVIDIA Tesla V100 GPU and 80 CPU cores.

### MSA sampling and paired-MSA construction

To obtain more diverse MSAs, we used a combination of sequence alignment tools including HHblits, JackHMMER, and MMseqs2 to search a range of sequence databases, such as UniRef30<sup>49</sup>, UniRef90<sup>50</sup>, UniProt<sup>50</sup>, Metaclust<sup>51</sup>, BFD<sup>51,52</sup>, MGnify<sup>53-55</sup>, and the ColabFold DB<sup>56</sup>. MSAs were generated individually for each monomer. Detailed descriptions of these databases are provided in Supplementary Note 2. Then, we use the predicted pSS-score as a supplement to the traditional sequence similarity in the ranking and selection of the searched MSAs. When the number of candidate MSAs exceeds a predefined threshold  $T_{max}$  (which we set to 10,000 in our experiments), we first use sequence similarity to rank all candidates and select the top  $T_{max}$  sequences. We then predict the pSS-score for this subset and perform the final ranking and selection based on these scores.

To enhance structural diversity and avoid local minima during prediction, we construct up to 19 distinct paired MSAs per complex using multiple strategies and sequence resources. For multimers, we use the developed deep learning model to predict the pIA-scores for each pair of sequence alignments from different subunit MSAs. The subunit MSAs are then concatenated based on these interaction probabilities to construct paired MSAs. Additionally, we use information from multiple sources, such as species annotations, UniProt accession number, and protein complexes from the Protein Data Bank (PDB), to further construct a series of paired MSAs. These paired MSAs are directly integrated into the AlphaFold-Multimer framework without architectural modifications. Specifically, we used the official implementation of AlphaFold-Multimer and followed its default template usage strategy. For each input paired MSA, we set the inference-time parameter `num_multimer_predictions_per_model=5`,

resulting in 25 candidate structures per MSA. Consequently, each complex can yield up to 475 predicted structures (19 paired MSAs  $\times$  25 structures) during structure sampling.

Additionally, in cases where the self-assessment scores returned by the structure prediction models (i.e., the “iptm+ptm” score from AlphaFold-Multimer) are below 0.5, we consider the current sampling insufficient. In such scenarios, we fall back to using sequence similarity instead of the pSS-score to further expand the sampling pool.

### Quality assessment of built models

In this study, we employed an in-house developed single model quality assessment method based on multi-scale protein representations using hierarchical networks. To represent the protein, physical and geometric properties of residues are extracted at the 1D scale, interaction relationships at the 2D scale, and structural topology at the 3D scale. These multi-scale representations are input into a multi-layer network module, which includes graph attention, transformer, and convolutional networks, to predict model quality. Building on this scoring model, we further developed a multi-model evaluation method based on structural consensus. First, the method selects the top N predicted models as the high-quality model pool using the single model quality assessment and model self-assessment scores. Next, all predicted models are aligned with the high-quality model pool to calculate their respective quality scores, and the top 5 models are selected for final evaluation.

### Evaluation metrics

We use two types of metrics to evaluate the quality of the protein complex model built by DeepSCFold. The first type of metrics is one that measure the closeness between the built model and the experimental structure in PDB. In this respect, we adopt the TM-score<sup>75</sup> between the predicted complex model and the corresponding experimental structure calculated by TM-score tool downloaded at (<https://zhanggroup.org/TM-score/>). The second type is one that measures the interface quality of predicted protein complex model. In this regard, we report DockQ<sup>76</sup> score calculated by public tool (<https://github.com/bjornwallner/DockQ>). Besides the above metrics, we also reported the Pearson, Spearman and ROC-AUC/PR-AUC to evaluate the performance of protein-protein structural similarity and interaction probability prediction models. The detailed descriptions of above metrics are given in Supplementary Note 1.

### Statistics and reproducibility

All data were carefully collected and analyzed using standard statistical methods. Comprehensive information on the statistical analyses used is included in various places, including the figures, figure legends and results.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The authors declare that the data supporting the results and conclusions of this study are available within the paper and its Supplementary Information. The all datasets generated in this study have been deposited in the GitHub under accession code 616f52d (<https://github.com/iobio-zjut/DeepSCFold>) and the general repositories Zenodo under accession code No.17015295 (<https://zenodo.org/records/17015295>)<sup>77</sup>. The original sources of the data are as follows: The CASP15 data including the experimental structures are available at [<https://predictioncenter.org/casp15/index.cgi>]; The antibody-antigen data from SABDab database is available at [<https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab>]; The protein interaction data from STRING database is available at [<https://string-db.org>]. Source data are

provided with this paper. The prediction results on these datasets are included in the Supplementary Data file. Source data are provided with this paper.

### Code availability

The code used to develop the model/perform the analyses and generate results in this study is publicly available and has been deposited in GitHub at (<https://github.com/iobio-zjut/DeepSCFold>), under MIT license. The specific version of the code associated with this publication is archived in Zenodo and is accessible via (<https://doi.org/10.5281/zenodo.17013108>)<sup>78</sup>. The online server of DeepSCFold is publicly accessible at (<http://zhanglab-bioinf.com/DeepSCFold/>).

### References

1. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA* **106**, 67–72 (2009).
2. Bitbol, A.-F., Dwyer, R. S., Colwell, L. J. & Wingreen, N. S. Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci. USA* **113**, 12180–12185 (2016).
3. Gueudré, T., Baldassi, C., Zamparo, M., Weigt, M. & Pagnani, A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proc. Natl. Acad. Sci. USA* **113**, 12186–12191 (2016).
4. Szurmant, H. & Weigt, M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr. Opin. Struct. Biol.* **50**, 26–32 (2018).
5. Rajagopala, S. V. et al. The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* **32**, 285–290 (2014).
6. Liu, J. et al. Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15. *Commun. Biol.* **6**, 1140 (2023).
7. Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D. & Bates, P. A. Critical assessment of methods for predicting the 3D structure of proteins and protein complexes. *Annu. Rev. Biophys.* **52**, 183–206 (2023).
8. Singh, A., Dauzhenka, T., Kundrotas, P. J., Sternberg, M. J. E. & Vakser, I. A. Application of docking methodologies to modeled proteins. *Proteins: Struct., Funct., Bioinforma.* **88**, 1180–1188 (2020).
9. Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. der Dtsch. Chemischen Ges.* **27**, 2985–2993 (1894).
10. Koshland, D. E. Jr Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci.* **44**, 98–104 (1958).
11. Chen, R., Li, L. & Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins: Struct., Funct., Bioinforma.* **52**, 80–87 (2003).
12. Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS one* **6**, e24657 (2011).
13. Pierce, B. G. et al. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).
14. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
15. De Vries, S. J. et al. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins: Struct., Funct., Bioinforma.* **69**, 726–733 (2007).
16. De Vries, S. J., Van Dijk, M. & Bonvin, A. M. J. J. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protoc.* **5**, 883–897 (2010).
17. Yan, Y., Zhang, D., Zhou, P., Li, B. & Huang, S.-Y. HDock: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res.* **45**, W365–W373 (2017).
18. Yan, Y., Tao, H., He, J. & Huang, S.-Y. The HDock server for integrated protein-protein docking. *Nat. Protoc.* **15**, 1829–1852 (2020).

19. Li, H., Huang, E., Zhang, Y., Huang, S. Y. & Xiao, Y. HDock update for modeling protein-RNA/DNA complex structures. *Protein Sci.* **31**, e4441 (2022).
20. Lensink, M. F. et al. Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins: Struct., Funct., Bioinforma.* **87**, 1200–1221 (2019).
21. Lensink, M. F. et al. Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment. *Proteins: Struct., Funct., Bioinforma.* **89**, 1800–1823 (2021).
22. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
23. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv*, <https://doi.org/10.1101/2021.10.04.463034> (2021).
24. Zheng, W. et al. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nat. Methods* **21**, 279–289 (2024).
25. Zheng, W., Wuyun, Q., Freddolino, P. L. & Zhang, Y. Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. *Proteins: Struct., Funct., Bioinforma.* **91**, 1684–1703 (2023).
26. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
27. Kryzhtafovich, Schwede, A., Topf, T., Fidelis, M. & Moult, K. J. Critical assessment of methods of protein structure prediction (CASP) —Round XV. *Proteins: Struct., Funct., Bioinforma.* **91**, 1539–1549 (2023).
28. Elofsson, A. Progress at protein structure prediction, as seen in CASP15. *Curr. Opin. Struct. Biol.* **80**, 102594 (2023).
29. Olechnovič, K., Valančauskas, L., Dapkūnas, J. & Venclovas, Č. Prediction of protein assemblies by structure sampling followed by interface-focused scoring. *Proteins: Struct., Funct., Bioinforma.* **91**, 1724–1733 (2023).
30. Wallner, B. Improved multimer prediction using massive sampling with AlphaFold in CASP15. *Proteins: Struct., Funct., Bioinforma.* **91**, 1734–1746 (2023).
31. Peng, Z., Wang, W., Wei, H., Li, X. & Yang, J. Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. *Proteins: Struct., Funct., Bioinforma.* **91**, 1704–1711 (2023).
32. Roney, J. P. & Ovchinnikov, S. State-of-the-art estimation of protein model accuracy using AlphaFold. *Phys. Rev. Lett.* **129**, 238101 (2022).
33. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
34. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinforma.* **20**, 1–15 (2019).
35. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).
36. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
37. Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).
38. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Pairing interacting protein sequences using masked language modeling. *Proc. Natl. Acad. Sci. USA*. **121**, e2311887121 (2024).
39. Chen, B. et al. Improved the heterodimer protein complex prediction with protein language models. *Brief. Bioinforma.* **24**, bbad221 (2023).
40. Devlin, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*. **1**, 4171–4186 (2019).
41. Aloy, P. & Russell, R. B. Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* **22**, 1317–1321 (2004).
42. Gao, M. & Skolnick, J. Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. USA*. **107**, 22517–22522 (2010).
43. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7**, 10480 (2017).
44. Kundrotas, P. J., Zhu, Z., Janin, J. & Vakser, I. A. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. USA*. **109**, 9438–9441 (2012).
45. Liu, W. et al. PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nat. Commun.* **15**, 2775 (2024).
46. Katchalski-Katzir, E. et al. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.* **89**, 2195–2199 (1992).
47. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **272**, 106–120 (1997).
48. Schneider, C., Raybould, M. I. J. & Deane, C. M. SABDab in the age of biotherapeutics: updates including SABDab-nano, the nanobody structure tracker. *Nucleic Acids Res.* **50**, D1368–D1372 (2022).
49. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
50. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
51. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).
52. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).
53. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
54. Gurbich, T. A. et al. MGnify genomes: a resource for biome-specific microbial genome catalogues. *J. Mol. Biol.* **435**, 168016 (2023).
55. Richardson, L. et al. MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
56. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
57. Kagaya, Y. et al. Structure Modeling Protocols for Protein Multimer and RNA in CASP16 With Enhanced MSAs, Model Ranking, and Deep Learning. *Proteins: Structure, Function, and Bioinformatics*, <https://doi.org/10.1002/prot.70033> (2025).
58. Elofsson, A. et al. AlphaFold3 at CASP16. *Proteins: Structure, Function, and Bioinformatics*. 1–13, <https://doi.org/10.1002/prot.70044> (2025).
59. Wang, W., Luo, Y., Peng, Z. & Yang, J. Accurate biomolecular structure prediction in CASP16 with optimized inputs to state-of-the-art predictors. *Proteins: Structure, Function, and Bioinformatics*, <https://doi.org/10.1002/prot.70030> (2025).
60. Luo, Y., Wu, H., Wei, H., Peng, Z. & Yang, J. Predicting the Oligomeric State of Proteins Using Multiple Templates Detected by Complementary Alignment Methods. *Proteins: Structure, Function, and Bioinformatics*, <https://doi.org/10.1002/prot.70017> (2025).
61. Liu, J., Neupane, P. & Cheng, J. Improving AlphaFold2- and AlphaFold3-based protein complex structure prediction with MULTICOM4 in CASP16. *Proteins: Structure, Function, and Bioinformatics*. <https://doi.org/10.1002/prot.26850> (2025).
62. Varadi, M. et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).

63. Shan, S. et al. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc. Natl Acad. Sci.* **119**, e2122954119 (2022).
64. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Struct., Funct., Bioinforma.* **77**, 499–508 (2009).
65. De Winter, J. C. F., Gosling, S. D. & Potter, J. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. *Psychol. Methods* **21**, 273 (2016).
66. Ali Abd Al-Hameed, K. Spearman's correlation coefficient in statistical analysis. *Int. J. Nonlinear Anal. Appl.* **13**, 3249–3255 (2022).
67. Zhang, C., Shine, M., Pyle, A. M. & Zhang, Y. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods* **19**, 1109–1115 (2022).
68. Meiler, J., Müller, M., Zeidler, A. & Schmäschke, F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. modeling Annu.* **7**, 360–369 (2001).
69. Singh, R., Devkota, K., Sledzieski, S., Berger, B. & Cowen, L. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics* **38**, i264–i272 (2022).
70. Szymborski, J. & Emad, A. RAPPID: towards generalizable protein interaction prediction with AWD-LSTM twin networks. *Bioinformatics* **38**, 3958–3967 (2022).
71. Song, G. et al. Broadly neutralizing antibodies targeting a conserved silent face of spike RBD resist extreme SARS-CoV-2 antigenic drift. *Cell Reports*. **44**, 115948 (2025).
72. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids Res.* **51**, D638–D646 (2023).
73. Xu, Y., Xu, D. & Gabow, H. N. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* **16**, 1091–1104 (2000).
74. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids Res.* **47**, D607–D613 (2019).
75. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct., Funct., Bioinforma.* **57**, 702–710 (2004).
76. Basu, S. & Wallner, B. DockQ: a quality measure for protein–protein docking models. *PLoS one* **11**, e0161879 (2016).
77. Hou, M. et al. DeepSCFold: High-accuracy protein complex structure modeling based on sequence-derived structure complementarity [Data set]. *Zenodo*. <https://doi.org/10.5281/zenodo.17015295> (2025).
78. Hou, M. et al. DeepSCFold: High-accuracy protein complex structure modeling based on sequence-derived structure complementarity (v1.0). *Zenodo*. <https://doi.org/10.5281/zenodo.17013108> (2025).
- Engineering at Zhejiang University of Technology. This work was supported by the National Key R&D Program of China [2022ZD0115103, G.Z.], the National Nature Science Foundation of China [62173304, G.Z.; 62203389, X.Z.], the “Pioneer” and “Leading Goose” R&D Program of Zhejiang [2025C01190, G.Z.], the Zhejiang Province High-level Talent Special Support Program [2023R5248, G.Z.], and Fundamental Research Funds for the Provincial Universities of Zhejiang [RF-C2024006, X.Z.].

## Author contributions

G.Z. conceived the project. J.Z. helped supervise the research. G.Z., M.H., and Y.X. designed the experiment. M.H., Y.X., P.W., Z.L., and D.H. performed the experiment and collected the data. G.Z., J.Z., X.Z., M.H., and P.W. analyzed the data. M.H., Y.X. G.Z., J.Z., and X.Z. wrote the manuscript, and all authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at (<https://doi.org/10.1038/s41467-025-65090-7>).

**Correspondence** and requests for materials should be addressed to Jianyang Zeng or Guijun Zhang.

**Peer review information** *Nature Communications* thanks Jianlin Cheng, Yogesh Kalakoti and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

## Acknowledgements

We thank members of the Guijun Zhang lab for discussion and feedback. Computational resources were provided by the College of Information