OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–14
https://doi.org/10.1093/bib/bbac499
Problem Solving Protocol

DeepHomo2.0: improved protein–protein contact prediction of homodimers by transformer-enhanced deep learning

Peicong Lin†, Yumeng Yan† and Sheng-You Huang

Corresponding author. S.-Y. Huang, School of Physics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, P. R. China. Tel.: +86-27-87543881; Fax: +86-027-87556576; E-mail: huangsy@hust.edu.cn

†Peicong Lin and Yumeng Yan contributed equally to this work.

Abstract

Protein–protein interactions play an important role in many biological processes. However, although structure prediction for monomer proteins has achieved great progress with the advent of advanced deep learning algorithms like AlphaFold, the structure prediction for protein–protein complexes remains an open question. Taking advantage of the Transformer model of ESM-MSA, we have developed a deep learning-based model, named DeepHomo2.0, to predict protein–protein interactions of homodimeric complexes by leveraging the direct-coupling analysis (DCA) and Transformer features of sequences and the structure features of monomers. DeepHomo2.0 was extensively evaluated on diverse test sets and compared with eight state-of-the-art methods including protein language model-based, DCA-based and machine learning-based methods. It was shown that DeepHomo2.0 achieved a high precision of >70% with experimental monomer structures and >60% with predicted monomer structures for the top 10 predicted contacts on the test sets and outperformed the other eight methods. Moreover, even the version without using structure information, named DeepHomoSeq, still achieved a good precision of >55% for the top 10 predicted contacts. Integrating the predicted contacts into protein docking significantly improved the structure prediction of realistic Critical Assessment of Protein Structure Prediction homodimeric complexes. DeepHomo2.0 and DeepHomoSeq are available at http://huanglab.phys.hust.edu.cn/DeepHomo2/.

Keywords: protein–protein interaction, homo-oligomers, residue–residue contact prediction, deep learning, transformer features

Introduction

Protein–protein interactions are involved in most essential biological processes, including signal transduction, immune response, cell adhesion and cell-cycle regulation [1–6]. Many proteins assemble to form symmetric homo-oligomers to perform their functions by interacting with themselves [7, 8]. Therefore, three-dimensional (3D) structures of homo-oligomeric proteins are essential for understanding their biological function [9]. With the development of experimental methods such as X-ray crystallography [10], nuclear magnetic resonance [11] and cryogenic electron microscopy (cryo-EM) [12–15], a number of homo-oligomeric structures have been determined and deposited in the Protein Data Bank (PDB) [16]. However, due to the high cost of experiments and the rapid development of high-throughput sequencing technology [17], the gap between the number of known sequences [18] and that of experimentally determined structures [16] is becoming larger and larger. Hence, it is highly desirable to develop computational methods for predicting the homo-oligomeric structures based on the huge amounts of sequence data.

Many studies have used sophisticated statistical methods like direct-coupling analysis (DCA) [19] to predict intra-protein residue–residue contacts as an intermediate stage of protein structure prediction [20–24]. Recently, deep learning-based methods have also achieved enormous progresses in intra-protein residue–residue contact predictions [25–33]. The statistical methods usually use the maximum-entropy model to detect the statistical coupling between amino acids, and the deep learning-based methods utilize the power of deep learning network to predict the residue-residue contacts by extracting the correlation between high-dimension features. Both statistical and deep learning-based methods normally use the multiple sequence alignment (MSA) to explore the evolutionary information including conservation of single residues and/or interacting residue pairs [34, 35]. The prediction of residue–residue contacts has led to a significantly improved accuracy of predicted structures, as observed in the Critical Assessment of protein Structure Prediction (CASP) experiments [36–41].

Motivated by the success of intra-protein contact predictions in monomer structure prediction, some similar methods have been developed to predict the inter-protein contacts for complex structure prediction [42–47]. However, these methods have limited accuracies due to the challenge of constructing a joint MSA, in which each line contains a pair of interacting proteins (i.e. interlogs [48]). Fortunately, this limitation would become trivial

Peicong Lin is a PhD student in the School of Physics at Huazhong University of Science and Technology. His research interests include protein structure prediction and protein–protein contact prediction.

Yumeng Yan is a PhD student in the School of Physics at Huazhong University of Science and Technology. His research interests include molecular docking and protein–protein contact prediction.

Sheng-You Huang is a full professor in the School of Physics at the Huazhong University of Science and Technology. His research interests are molecular docking and scoring functions for protein–ligand interactions, protein–protein interactions and protein–peptide interactions. More information about his research group can be found at http://huanglab.phys.hust.edu.cn.

Received: June 30, 2022. **Revised:** October 8, 2022. **Accepted:** October 21, 2022

for homo-oligomers as their interacting partners have identical sequences. However, another critical problem arises. That is, how to distinguish the intra-protein from inter-protein contacts. To address this problem, various methods including statistical methods like DCA [49] and deep learning-based methods [42, 50–53], have been proposed for the prediction of inter-protein contacts in homo-oligomers. For example, Xie and Xu proposed a deep learning-based method GLINTER, leveraging a rotational invariant representation of protein 3D structures for interface contact prediction of dimers [51]. Roy *et al.* developed a deep dilated convolutional residual network method DRCon to predict inter-chain residue–residue contacts in homodimers [52]. Quadir *et al.* developed a server DeepComplex to predict the quaternary structure of complexes based on the predicted inter-chain contacts [53]. In addition, we have also developed a deep learning-based method, named DeepHomo, to predict the inter-protein contacts for homodimers by using both sequence and structure features [42].

Despite the significant successes of current methods, their performances depend on the depth of MSA, which directly determines the accuracy of DCA methods in capturing the co-evolution information. Therefore, it calls for more advanced methods, which utilize a different method to extract global co-evolution information from the MSA more efficiently, to improve the ability to predict inter-protein interactions. In addition, current methods normally rely on the 3D structure of monomer proteins. However, the structure of monomer proteins may not be always available in real applications, especially for those challenging proteins that have few homologous sequences. Therefore, it is also needed to develop a specialized model for predicting inter-protein contacts of homo-oligomers without using structure information. Meeting the need, we have proposed a deep learning model, named DeepHomo2.0, to predict inter-protein residue–residue contacts for homodimers by utilizing the advantage of Transformer in protein language models [54–57]. Our model efficiently leverages co-evolution information, Transformer features and structure features. With the inclusion of Transformer features and fine-tuning of the network architecture, DeepHomo2.0 achieved a significant improvement over DeepHomo1.0 in predicting inter-protein residue–residue contacts on diverse test sets. Moreover, even the sequence version without using structural features, named as DeepHomoSeq, can also achieve a comparable performance to state-of-the-art methods that utilize structure information.

Materials and Methods

Deep learning architecture

Figure 1 shows the overall architecture of DeepHomo2.0, which is modified based on that of DeepHomo1.0 – the first version of DeepHomo [42]. The architecture of the sequence version (DeepHomoSeq) is the same as that of DeepHomo2.0, excluding the input structure features. It consists of two main modules, i.e. 1D convolutional neural network and 2D convolutional neural network, which are mainly composed of ResNet-v2 blocks [58]. Specifically, starting from a structure and the corresponding MSA, the 1D convolutional neural network processes the 1D features and obtains the high-dimensional information through three same blocks, in which the convolution kernels are set to 1 × 17 and their filters are set to 32. The output of the 1D convolutional neural network is converted to a 2D matrix by outer concatenation like RaptorX-Contact [32]. Then, the 2D matrix is concatenated with the pairwise features as the 2D input, which is fed into the 2D convolutional network. The 2D convolution network has nine blocks and the numbers of filters are set to 32, 64, 64, 64, 64, 64, 64, 64 and 64, respectively. Finally, the 2D output is converted into the matrix of contact probability for all residue pairs after passing through a batch-normalized layer, an exponential linear unit activation layer, a 1×1 convolution layer and a sigmoid activation layer. All the 2D convolutional layers take the 3×3 convolution kernel and the zero padding that keeps the consistent dimensions of the output tensor at each layer.

Data sets

We used the same training, validation and test sets as those for DeepHomo1.0 [42]. The whole data set consists of 4132 non-redundant homodimeric protein structures with C2 symmetry, which is composed of two identical subunits and formed by a rotation around a single rotation axis of one subunit. The data set was collected from the PDB and clustered with a sequence identity of 30% using MMseqs2, where the default coverage of 0.8 between sequences was used [59, 60]. The 4132 homodimers were then randomly divided into training set (3532 cases), validation set (300 cases) and test set (300 cases). This test set including 300 complexes is referred to as the PDB test set. In addition, we also used the CASP–Critical Assessment of Predicted Interactions (CAPRI) test set of 28 targets from CASP–CAPRI processed by DeepHomo1.0. For each test case, both the experimental and predicted monomer structures were evaluated, where AlphaFold2 was used to predict the monomer structures after excluding the corresponding templates by setting the "*max_template_date*" to the day before the release date of the target [61]

Input features

Since DeepHomo2.0 and DeepHomoSeq share the similar architecture, we take DeepHomo2.0 as an example to describe the input features. The input features can be grouped into two categories according to their dimensions. The first category is 1D sequential features, including Transformer embedding vector, protein sequence profile like position-specific scoring matrix (PSSM), 8-state secondary structure types and the solvent-accessible surface area of each amino acid. The second category is the 2D pairwise features, including Transformer row attention, direct co-evolution information calculated by CCMpred [62], intra-protein residue–residue distance map and docking map created by our fast Fourier transform (FFT)-based docking program, HSYMDOCK [63, 64]. For the DeepHomoSeq model, it only uses the sequence features including PSSM, Transformer embedding vector, Transformer row attention and direct co-evolution information.

To derive the above sequence features, we needed to generate an MSA. For a query sequence with L amino acids, we first ran HHblits [65] using a minimum coverage of 40% and a maximum pairwise sequence identity of 99% [66] against the UniRef30_2020_03 database [67] for three iterations, where the threshold of e-value was set to its default value. Next, according to the generated MSA, CCMpred was used to calculate the raw scores and average product correction of direct coupling analysis (DCA-DI and DCA-APC). The features of DCA-DI and DCA-APC were represented as two matrices, and each of them has a dimension of $L \times L \times 1$. Then, the protein sequence profile PSSM was produced by the LoadHMM script in the RaptorX-Contact [32] package and represented as a matrix with dimension $L \times 20$.

For the structure features, we processed each monomer structure in the same way as that for DeepHomo1.0. For the 1D structure features, we used the Dictionary of Protein Secondary Structure (DSSP) program to assign the secondary structure type for

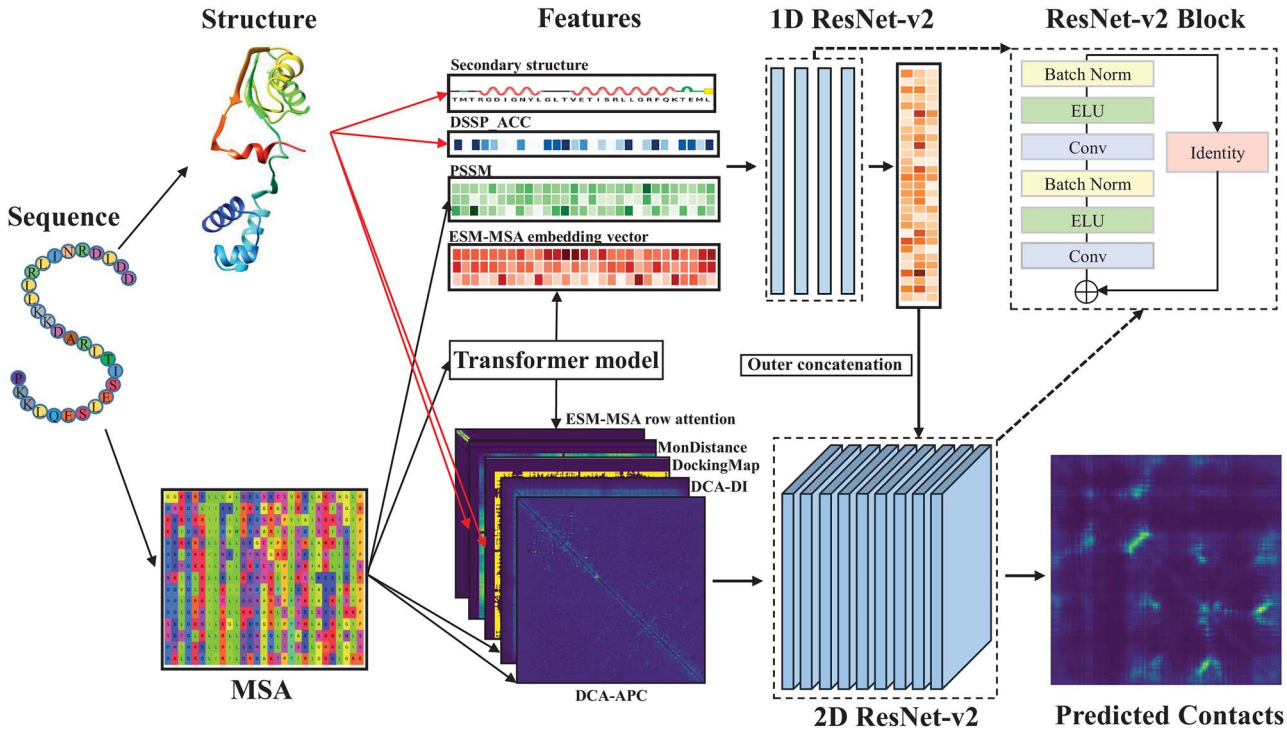

Figure 1. The framework of the DeepHomo2.0. The network of DeepHomo2.0 consists of two main components: 1D convolutional neural network, including PSSM, ESM-MSA embedding vector, secondary structure type and solvent-accessible surface area features (upper layer); 2D convolutional neural network, including intra-protein residue–residue distance map, docking map, DCA-DI, DCA-APC and ESM-MSA row attention features (bottom layer). Two components consists of several ResNet-v2 blocks (upper right).

each amino acid [68]. Then, the assigned secondary structure types were converted into a matrix with dimension $L \times 8$ according to one-hot encoding. In addition, we also used solvent-accessible surface area of each residue produced by DSSP (DSSP_ACC) as the hydrophobicity feature. For the other 2D structure features, the intra-protein residue–residue distance map consisted of the minimal heavy atom distances with a size of $L \times L \times 1$. The docking map was generated from a modified version of our in-house FFT-based docking program, HSYMDOCK, and was represented as a dimension of $L \times L \times 1$.

The most important part of our model is that we used the Transformer features produced by the pre-trained unsupervised protein language model, evolutionary scale modeling-multiple sequence alignment (ESM-MSA). ESM–MSA interleaves row and column attention across the MSA and has been trained with a variant of the masked language modeling objective across various protein families [55]. By inferring through this model, we obtained the ESM–MSA embedding vector and ESM–MSA row attention. Due to the limited memory, we did not use the full sequence of the MSA. Instead, we used the hhfilter [69] with the *-diff* 512 option to filter the MSAs by selecting the most diverse sequences. If it returned more than 512 sequences, we selected the top 512 sequences from the filtered MSA. The ESM–MSA embedding vector was a matrix with dimension $L \times 768$ and the ESM–MSA row attention was a matrix with dimension of $L \times L \times 144$. Finally, we concatenated all the 1D features and 2D features, respectively, resulting in the 1D input of dimension of $L \times 797$ and 2D input with dimension $L \times L \times 148$.

Similar to DeepHomo1.0, we defined a pair of residues as being contacted if any two heavy atoms of the two residues are within 8 Å and the two residues are from different chains. An ideal contact map for a homodimeric interface should be diagonally symmetric, which is formed by two identical subunits and rotation around a rotation axis of one subunit. However, many homodimeric chains may be slightly different from each other, leading to slightly different contact maps. Hence, we took several operators to ensure the symmetry of the native contacts. First, we applied an "OR" operation to all the interfaces. Namely, if a pair (i,j) is in contact in any interface, the pair (i,j) of the native contacts was set to 1. Second, we applied a transpose operator to the native contacts to obtain the transpose matrix, which was added to the matrix of native contacts to ensure the diagonal symmetry. Any value that exceeded 1 in the matrix was set to 1. In other words, if a pair (i,j) was contacted in the native structure, the residue pair (j,i) was set to 1 too.

Implementation and training

In the training process, due to the extreme unbalance between non-contact residue pairs and contact pairs, we utilized the Focal Loss [70] as the loss function, and optimized the model by Adam optimizer [71]. The Focal Loss function for each pair of residues i and j is described as follows

$$\text{Focal\_Loss}(\mathbf{p_t}) = \alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

Here, the parameters α_t and p_t are defined as

$$\alpha_t = \begin{cases} \alpha, & y=1 \\ 1-\alpha, & y=0 \end{cases} \tag{2}$$

and

$$p_t = \begin{cases} p_{ij}, & y=1 \\ 1-p_{ij}, & y=0 \end{cases} \tag{3}$$

where α and γ are set to 0.25 and 1.50 [42], and p_{ij} is the predicted contact score of residue i and residue j. Since the number of negative labels for the ground truth is far more than the number of positive ones, we set a low value of α to reduce the contribution of negative labels for the loss function. We also set a high value of γ to make the loss function focus on the hard labels. In addition, we also used a mini-batch size of 1 and a learning rate of 0.001 and without any regularization. Due to the GPU memory limitation, the length of protein sequence is restricted to 400 amino acids. Specifically, for the target protein with a sequence length of more than 400 residues, a continuous subsequence with a length of 400 is randomly sampled to represent the protein. We trained our deep learning-based model (DeepHomo2.0 and DeepHomoSeq) on a PDB training set of homodimers. The models were implemented using pytorch (v.1.8.1) and trained on one A100 GPU for 5 h (4.5 h). During the evaluations, without considering the consumption of generating input features, DeepHomo2.0 (DeepHomoSeq) took an average of 3.803 (3.089) s on the PDB test set of 300 homodimeric complexes.

Results and Discussion

Protein–protein contact predictions

Evaluation on the complexes from the PDB

We firstly evaluated the performance of DeepHomo2.0 and DeepHomoSeq on the PDB test set with the experimental monomer structures from the PDB. In addition, given that the monomer structure may not be available in real situations, we also evaluated our approaches on the monomer structures predicted by AlphaFold2 for more realistic applications. For comparison, we also calculated the results of seven other approaches, including two coevolution-based methods (DCA-DI-based and DCA-APC-based) and four machine learning-based methods (DeepHomo1.0, GLINTER, BIPSPI_seq, BIPSPI_struc and ESM-MSA-based).

In this study, we adopted a similar strategy used by Uguzzoni *et al* [49] for coevolution-based methods. This strategy was performed as follows: (i) Generated the MSA and calculated the corresponding DCA-DI and DCA-APC scores; (ii) According to the input monomer structure, the residue pairs with intra-monomer distances of more than 12 Å were regarded as the candidates of inter-protein contacts. (iii) The candidates were ranked according to the coevolutionary scores. In addition, since ESM–MSA is a deep learning-based method which uses attention heads to predict the intra-protein contacts by a regression model, ESM–MSA adopted the same processing strategy as the coevolution-based methods. BIPSPI [72] is a machine learning-based method, which employs the extreme gradient boosting algorithm [73] with a set of structure-based and/or sequence-based features to predict inter-protein contacts. BIPSPI has two versions that accept the sequence as input (BIPSPI_seq) or the structure as input (BIPSPI_struc). We submitted the structures and sequences of monomers to the BIPSPI web server. GLINTER is a deep learning-based method to predict inter-protein residue–residue contacts [51]. We downloaded the source code of GLINTER from the GitHub website and ran the homodimer script locally to predict the inter-chain contacts.

Table 1 lists the results of DeepHomo2.0, DeepHomoSeq and the other seven methods for the top 1, 10, 25 and 50 predicted inter-protein contacts on the PDB test set with experimental structures. It can be seen from the table that DeepHomo2.0 achieved the overall best performance among the nine methods. Specifically, DeepHomo2.0 obtained a precision of 70.9% when the top10 predictions were considered, compared to 55.1% for DeepHomoSeq, 55.5% for DeepHomo1.0, 63.3% for GLINTER, 41.1% for ESM-MSA, 24.0% for the DCA-APC, 16.3% for DCA-DI, 23.8% for BIPSPI_struc and 6.4% for BIPSPI_seq. In addition, we also listed the top L/10, L/5 and L precision in Table 1, where L is the length of protein. Similarly, DeepHomo2.0 achieved the best performance among the nine methods and obtained an average precision of 46.4% for the top L predicted contacts, compared to 36.8% for GLINTER, 33.4% for DeepHomoSeq and 31.6% for DeepHomo1.0. With the inclusion of ESM-MSA features, DeepHomo2.0 obtained an improvement of nearly 15% in all evaluation criteria, compared with our previous DeepHomo1.0. This indicates that the ESM-MSA features can allow our deep learning-based network to extract the inter-protein residue–residue interactions more effectively. Compared with the traditional DCA method that extracts co-evolution information directly from the covariance of the MSA, the Transformer method seems to have much better power to capture more important pairwise interactions.

Notably, DeepHomoSeq obtained a close precision for the top 1 and 10 predicted contacts and achieved a better performance in the top 25, 50, L/10, L/5 and L predicted contacts than the DeepHomo1.0. Considering that the other methods like DeepHomo1.0 and BIPSPI_struc used the structure information of monomers, the close and better performance of DeepHomoSeq in different evaluation criteria suggests that only sequence information may be enough to accurately predict protein-protein interactions. In order to further investigate the robustness of our methods, we also tested DeepHomo2.0, DeepHomoSeq and the other methods with the predicted structures by AlphaFold2 on the same test set. Table 1 shows the corresponding results of all methods in parentheses. On the whole, the methods using structure information all drop nearly 5–10% precision in the top1, 10, 25, 50, L/10, L/5 and L predicted contacts. Although the overall precision decreased, DeepHomo2.0 still achieved the best performance for all evaluation criteria, and DeepHomoSeq achieved the third best performance among all methods.

We also tested DeepHomo2.0 and DeepHomoSeq on a PDB subset of 218 complexes used by DRCon, named as T218, which were constructed by excluding the same homodimers included in the training set of DRCon. Table 2 lists the precisions of DeepHomo2.0, DeepHomoSeq, DeepHomo1.0, GLINTER and DRCon for the top 1, 10, 25, 50, L/10, L/5 and L predicted contacts on the T218 test set with experimental and predicted structures. It is worth mentioning that the results of DRcon were directly taken from the literature [52], where the precisions for some criteria are not available. It can be seen from the table that DeepHomo2.0 also achieved a better performance than the other methods on the test set with experimental structures, and obtained the precisions of 75.2%, 72.2%, 69.1%, 65.1%, 69.5%, 65.5% and 46.8% when the top 1, 10, 25, 50, L/10, L/5 and L predicted contacts were considered. Similar trends can be observed in the results with AlphaFold2-predicted structures (Table 2).

For a more detailed comparison, Figure 2a shows the average predictions of DeepHomo2.0, DeepHomoSeq and other methods as a function of the number of top predicted contacts on the test set with experimental structures. In addition, Figure 2c also shows the corresponding results using the predicted structures generated by AlphaFold2. The precision mentioned above is defined as the percentage of the true contacts among the top n predicted contacts. It can be seen from the figure that DeepHomo2.0 achieved the best precision for different numbers of top predicted contacts for the cases of both experimental and predicted structures. DeepHomoSeq obtained the overall third best performance with experimental structures, which is comparable to DeepHomo1.0. In addition, DeepHomoSeq obtained better performance than GLINTER

Table 1. Comparison of the precisions (%) by DeepHomo2.0, DeepHomoSeq and seven other methods on the PDB test set of 300 experimental structures (predicted structures by AlphaFold2) when the top 1, 10, 25, 50, L/10, L/5 and L predicted contacts are considered, where the results of our DeepHomo2.0 and DeepHomoSeq are shown in bold.

Method	Top 1	Top 10	Top 25	Top 50	Top L/10	Top L/5	Top L
DeepHomo2.0	**73.0 (64.0)**	**70.9 (62.2)**	**67.7 (59.2)**	**63.9 (55.4)**	**68.1 (59.6)**	**64.4 (56.4)**	**46.4 (39.8)**
DeepHomoSeq	**58.3**	**55.1**	**52.0**	**47.9**	**52.3**	**48.6**	**33.4**
DeepHomo1.0	60.7 (55.0)	55.5 (47.7)	50.5 (44.5)	46.3 (40.2)	51.5 (45.5)	47.7 (41.1)	31.6 (26.6)
GLINTER	68.0 (60.0)	63.3 (56.0)	58.4 (52.3)	52.5 (46.7)	59.3 (52.7)	54.5 (48.1)	36.8 (34.6)
ESM_MSA	45.3 (44.0)	41.1 (38.9)	36.7 (33.1)	31.7 (27.4)	37.9 (34.6)	33.1 (28.5)	19.0 (15.0)
DCA_APC	32.0 (27.0)	24.0 (19.9)	16.7 (13.3)	11.7 (9.0)	18.9 (14.5)	13.5 (10.0)	6.1 (4.5)
DCA_DI	29.7 (24.7)	16.3 (12.5)	10.1 (7.8)	6.7 (5.1)	11.9 (8.5)	7.8 (5.7)	3.6 (2.7)
BIPSPI_Struc	24.0 (17.7)	23.8 (14.6)	21.6 (13.6)	20.4 (12.9)	22.5 (13.6)	20.9 (13.3)	16.2 (10.8)
BIPSPI_Seq	7.3	6.4	5.0	4.6	5.6	4.7	3.8

Table 2. Comparison of the precisions (%) by DeepHomo2.0, DeepHomoSeq and three other methods on the PDB test set of 218 experimental structures (predicted structures by AlphaFold2) when the top 1, 10, 25, 50, L/10, L/5 and L predicted contacts are considered, where the results of our DeepHomo2.0 and DeepHomoSeq are shown in bold.

Method	Top 1	Top 10	Top 25	Top 50	Top L/10	Top L/5	Top L
DeepHomo2.0	**75.2 (64.7)**	**72.2 (62.2)**	**69.1 (59.3)**	**65.1 (55.3)**	**69.5 (59.7)**	**65.8 (56.6)**	**46.8 (39.3)**
DeepHomoSeq	**60.1**	**57.2**	**53.6**	**49.4**	**54.0**	**50.1**	**34.0**
DeepHomo1.0	62.4 (56.4)	55.3 (49.9)	51.4 (44.9)	46.7 (40.0)	52.0 (46.0)	48.2 (41.4)	31.8 (26.3)
GLINTER	69.3 (61.5)	64.0 (56.0)	59.0 (52.4)	53.2 (47.0)	60.2 (52.5)	55.5 (48.5)	37.2 (34.0)
DRCon[a]	–	58.0	–	–	56.5	53.8	41.9

[a]The data of DRCon are directly taken from the literature [52], where "–" means that the data are not available for the situations.

with predicted structures when the considered top numbers are more than 29. Furthermore, to compare the performance of different methods in each protein, Figure 2b shows a comparison of the precisions between DeepHomo2.0 and other approaches with experimental structures for the 300 proteins in the test set when the top 100 predicted contacts were considered. Figure 2d gives a similar comparison with predicted structures by AlphaFold2. It can be seen from Figure 2b and d that DeepHomo2.0 obtained a better precision than other methods for most of the proteins. Especially, most of the points are distributed in the lower right corner, which indicates that DeepHomo2.0 obtained high precision when the other methods obtained low precision.

Although DeepHomoSeq only used the sequence features, it still achieved excellent performance in inter-protein contact prediction. Especially, DeepHomoSeq obtained a significantly higher precision compared with BIPSI_struc and DeepHomo1.0 which used both sequence and structure features. This means that we can only use sequence features to predict inter-protein contacts in blind tests. With the structures predicted by AlphaFold2, our DeepHomo2.0 also achieved the best performance. Therefore, DeepHomo2.0 would be the first choice to predict inter-protein residue–residue interactions if the quality of predicted structures is of good quality. In addition, accurate prediction of inter-protein contacts will greatly help the structure prediction of corresponding protein–protein complexes [42]. Hence, we can use the predicted contacts as the restraints in protein docking. There is still a large number of protein complex structures that have not been determined. Therefore, an accurate inter-protein contact predictor would be an excellent intermediate stage for predicting the complex structures.

Application to realistic CASP–CAPRI targets

CASP–CAPRI challenge is a double blind experiment, which aims to assess the computational methods of modeling protein assemblies in the community [74–76]. We evaluated the performance of DeepHomo2.0 and DeepHomoSeq on the CASP–CAPRI test set. Table 3 lists the precisions of DeepHomo2.0, DeepHomoSeq and seven other methods in inter-protein contact predictions for top 1, 10, 25, 50, L/10, L/5 and L predicted contacts with experimental and predicted structures. It can be seen from the table that DeepHomo2.0 outperformed the other methods no matter whether experimental or predicted structures were used. Specifically, DeepHomo2.0 achieved the precisions of 64.3%, 63.2%, 61.0%, 57.3%, 60.3%, 55.1% and 37.4% for top 1, 10, 25, 50, L/10, L/5 and L predicted contacts with the experimental structures, respectively. Similar performances are also observed for the case with the input of predicted structures. In addition, DeepHomoSeq obtained the third best performance among all methods, and significantly outperformed the ESM-MSA-based, BIPSPI and DCA-based methods. It is worth emphasizing that the ESM-MSA-based method outperformed the BIPSPI and coevolution-based methods and also obtained a high precision. This indicates the importance of ESM-MSA features in predicting inter-protein residue-residue contacts.

DeepHomo2.0 and DeepHomoSeq achieved high precisions in all evaluation criteria on the PDB and CASP-CAPRI test sets. This indicates that our methods with Transformer features have better robustness than the other approaches. DeepHomoSeq can achieve similar or better performance in protein–protein contact predictions, compared with machine learning-based methods (BIPSPI_struc and DeepHomo1.0) that use structure information. Even when high-quality monomer structures were used by BIPSPI_struc and DeepHomo1.0, DeepHomoSeq also obtained a similar or better precision. This is valuable because we may use a sequence-based model to predict inter-protein contacts of homo-oligomers without monomer structures in blind protein–protein interaction predictions like CAPRI.

Performance with more rigorous homology criterion

In protein structure prediction, a sequence identity threshold of 30% is usually used to remove the redundancy between the training and test sets. Similarly, we also used a sequence identity

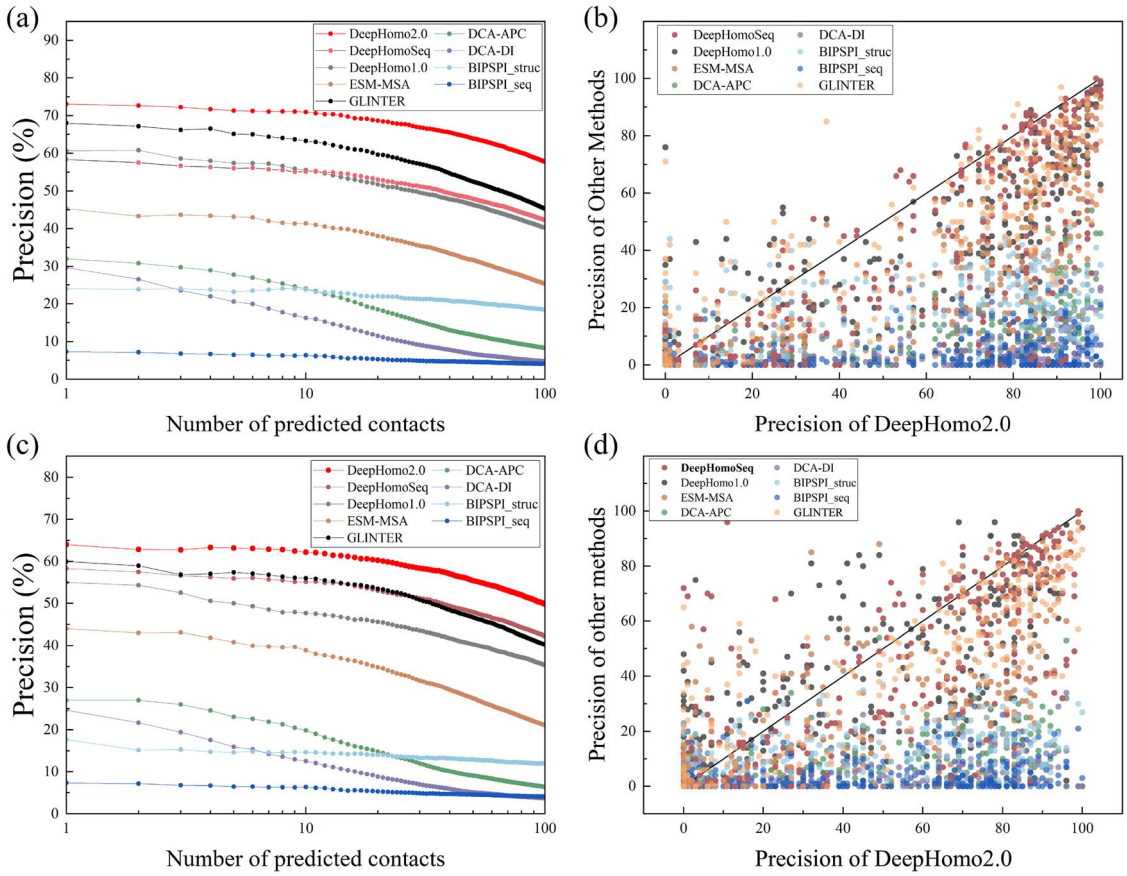

Figure 2. The performance of DeepHomo2.0, DeepHomoSeq and seven other methods on the PDB test set of 300 complexes. Here, DeepHomoSeq and BIPSPI_Seq only use the sequence information. The results in upper row were calculated with the experimental structures from the PDB. The results in lower row were calculated with the predicted structures by AlphaFold2. (a,c) The precision is a function of the number of predicted contacts. (b,d) Comparison of the precision for top 100 predicted contacts between DeepHomo2.0 and the other methods.

Table 3. Comparison of the precisions (%) by DeepHomo2.0, DeepHomoSeq and seven other methods on the CASP-CAPRI set of 28 realistic targets with experimental structures (predicted structures by AlphaFold2) when the top 1, 10, 25, 50, L/10, L/5 and L predicted contacts are considered, where the results of our DeepHomo2.0 and DeepHomoSeq are shown in bold.

Method	Top 1	Top 10	Top 25	Top 50	Top L/10	Top L/5	Top L
DeepHomo2.0	**64.3 (67.9)**	**63.2 (60.7)**	**61.3 (57.3)**	**57.5 (54.6)**	**60.3 (56.9)**	**55.1 (52.3)**	**37.4 (34.6)**
DeepHomoSeq	**60.7**	**55.7**	**52.1**	**47.6**	**49.7**	**45.4**	**29.4**
DeepHomo1.0	57.1 (57.1)	48.2 (47.9)	42.7 (44.3)	39.6 (40.4)	40.2 (38.6)	34.5 (34.0)	23.7 (22.4)
GLINTER	55.6 (51.9)	54.4 (53.3)	50.7 (49.5)	47.9 (44.9)	51.6 (49.2)	46.0 (44.1)	28.5 (25.7)
ESM_MSA	39.3 (42.9)	33.9 (36.4)	31.1 (32.0)	25.6 (26.9)	27.3 (28.7)	22.2 (23.7)	12.6 (12.9)
DCA_APC	14.3 (17.9)	17.5 (18.9)	13.6 (14.0)	10.3 (10.9)	12.4 (13.0)	9.4 (9.7)	3.9 (4.0)
DCA_DI	25.0 (25.0)	12.1 (12.5)	7.7 (8.3)	5.4 (5.6)	7.1 (7.7)	5.0 (5.4)	2.3 (2.3)
BIPSPI_Struc	10.7 (17.9)	17.5 (12.1)	17.9 (11.9)	15.4 (11.6)	17.2 (13.1)	15.6 (12.3)	12.0 (9.0)
BIPSPI_Seq	7.1	4.6	4.4	4.6	4.3	4.3	3.8

of 30% to cluster the homodimers [42]. However, a lower sequence identity of below 30% between two sequences may still share some homology. Therefore, to exclude the homologous sequences of target proteins in the test set from the training set, we further used more rigorous redundancy criteria to verify whether our method could still accurately predict the inter-protein residue–residue contacts. Specifically, we have removed the redundancy between the training set and our test sets with an E-value cutoff of 0.1 by MMseqs2, yielding a reduced training set from 3504 to 2833 complexes.

With the 2833 complexes, we retrained our DeepHomo2.0 and DeepHomoSeq by the same parameters. The two retrained models, named DeepHomo2.0$_{E0.1}$ and DeepHomoSeq$_{E0.1}$, were then compared with the original models on the PDB set of 300 complexes, the PDB subset of 218 complexes, and the CASP_CAPRI set of 28 complexes, respectively. The corresponding results are listed in Supplementary Table S1. It can be seen from the table that DeepHomo2.0$_{E0.1}$ obtained a similar precision to DeepHomo2.0 on the three test sets. Specifically, DeepHomo2.0$_{E0.1}$ obtained a slightly higher precision than DeepHomo2.0 on the PDB test set of 300 complexes and the PDB reduced set of 218 complexes, while DeepHomo2.0 achieved a slightly better performance than DeepHomo2.0 $_{E0.1}$ on the CASP_CAPRI set of 28 complexes. Compared with the results of DeepHomo2.0$_{E0.1}$, the

performance of DeepHomoSeq$_{E0.1}$ seems to be more dependent on the training set. As such, DeepHomoSeq$_{E0.1}$ had a lower performance than DeepHomoSeq on all three test sets (Supplementary Table S1). The reduced performance of DeepHomoSeq$_{E0.1}$ can be understood because DeepHomoSeq and DeepHomoSeq$_{E0.1}$ only utilize the sequence features. Therefore, their precisions tend to be more affected by the sequence homology, although DeepHomoSeq$_{E0.1}$ still performed satisfactorily compared with the other methods (Tables 1–3 and Supplementary Table S1).

Performance with different probability thresholds

For classical binary prediction, a cutoff of probability is usually used to classify positives and negatives. For the above evaluations, we have selected the top n predicted contacts as the positives to obtain the overall performance. However, this may result in a low precision for a single protein when the overall probabilities in the predicted inter-protein contact map are relatively low for the protein. This may be due to the few sequences in MSA generated by the target complex, where hidden information about the contacted pairs is hardly captured by deep learning network [77], or weak co-evolutionary coupling between interacting pairs. For integrating contacts into protein-protein docking, it is critical to exclude those false-positive contacts. On one hand, we want to use as many contacts as possible to filter out wrong binding modes. On the other hand, we need a high precision of contacts to avoid those false-positive solutions. A balance between the number and precision of contacts should be considered when integrating predicted contacts into protein–protein docking.

To address this problem, we have chosen a contact probability cutoff, i.e. P_{struc} for DeepHomo2.0 and P_{seq} for DeepHomoSeq, as a threshold to calculate the precision of DeepHomo2.0 or DeepHomoSeq, respectively. Namely, we defined the precision for the top n predicted contacts using the following formula:

$$\text{Precision}(n) = \frac{\sum_{i=1}^{N} \text{TP}_i(n)}{\sum_{i=1}^{N} \text{P}_i(n)} \quad (4)$$

where N is the total number of the target homodimers in the PDB test set, $\mathbf{TP}_i(n)$ is the number of true positive contacts with a contact probability above $P_{struc\,(seq)}$ among the top n predicted contacts for the ith target, and $\mathbf{P}_i(n)$ is the number of contacts with a probability above $P_{struc\,(seq)}$ among the top n predicted contacts for the ith target. Figure 3b shows the corresponding precisions for all, top 1, top 10, top 25, top 50 and top 100 predicted contacts as a function of P_{struc} or P_{seq} on the PDB test set of 300 homodimers with experimental structures.

It can be seen from the figure that all precisions increase as the contact probability cutoff increases, and eventually reached a maximum precision of 100%. In addition, there exist two characteristic probability points. One is called the "crossover" probability, at which the precision for all contacts turns higher than the precisions of all other top n contacts. The other is the "convergence" probability, at which all the precisions reach the maximum value of 100%. In other words, if one wants to obtain a balance between the number and precision of contacts, a probability cutoff at about the crossover point should be used; If one wants a maximum precision, a probability cutoff at the convergence point should be considered. According to Figure 3, the crossover and convergence probabilities are 0.70 and 0.85 for the structure-based DeepHomo2.0 and 0.55 and 0.81 for the sequence-based DeepHomoSeq, respectively. From Figure 3, one can also see that starting from the case of top25 contacts, their precisions almost overlap after the crossover probability point, indicating that the top 25 contacts may be enough in real applications.

Impact of features, MSA and contact density

Ablation studies of sequence features

To assess the contribution of different factors to DeepHomo2.0 and DeepHomoSeq, and investigate whether there is complementarity between the ESM-MSA features and coevolutionary features, we have conducted an ablation study with DeepHomo2.0 and DeephomoSeq by removing different features. For DeepHomo2.0 (baseline), we designed six models for comparison, where three models with one feature left out including No_PSSM, No_MonDistance and No_DockingMap models. The other three models deleted secondary structure type and solvent accessible surface (No_DSSP), DCA-DI and DCA-APC (No_DCA), and ESM-MSA embedding vector and ESM-MSA row attention (No_ESM), respectively. For DeepHomoSeq (baseline), we also designed five models for comparison including No_PSSM, No_DCA, No_ESM, No_ESM-1D and No_ESM-RA, where No_ESM-1D excludes the ESM-MSA embedding vector feature and ESM-RA exclude the ESM-MSA row attention feature. All six (five) models were retrained with the same hyper-parameters and training set used for the DeepHomo2.0 (DeepHomoSeq) model.

Figure 4 shows the precisions of different models in the ablation experiments. Overall, the ESM-MSA embedding vector and ESM-MSA row attention features are the most important factors, and the PSSM and DCA features also have some contributions to DeepHomoSeq. DeepHomoSeq obtained a precision of 55.1% for the top 10 predicted contacts on the PDB test set, compared with 54.4% for No_PSSM, and 53.9% for No_DCA. The two retrained models have similar performances with DeepHomoSeq. This may be due to that DeepHomoSeq can excellently capture the information of PSSM features from ESM-MSA embedding vector features and the information of DCA feature from ESM-MSA contact prediction feature. In addition, the No_ESM model obtained a precision of 34.3% for the top 10 predicted contacts, compared with 45.3% for No_ESM-RA, and 52.2% for No_ESM-1D.

The comparison indicates that DeepHomoSeq achieved a significantly improved performance by adding the Transformer features. Furthermore, considering that ESM-MSA row attention consists of multiple attention heads that may contain the information of inter-protein interaction, it is not surprising to see that the No_ESM-1D model obtained a better performance than the No_ESM-RA model. In addition, No_ESM-RA model obtained a better performance than No_ESM model, which suggested that the 1D sequence feature may imply the useful information ignored by the 2D pairwise feature. Recently, Alphafold2 directly used the original information of MSA and achieves a great progress in protein structure prediction. As such, the precision could be further improved by using MSA directly to capture hidden information about the inter-protein interactions. Since DeepHomo2.0 utilizes both sequence and structure features, we analyzed the ablation experiment with sequence and structure information. The retrained models that exclude the sequence features of DeepHomo2.0 have the same trend as that of DeepHomoSeq. The No_MonDistance model has little influence on the precision. No_DSSP and No_DockingMap models contribute to DeepHomo2.0 to a certain extent. In summary, the structure features can further improve the capability of predicting the inter-protein residue–residue interactions in our model.

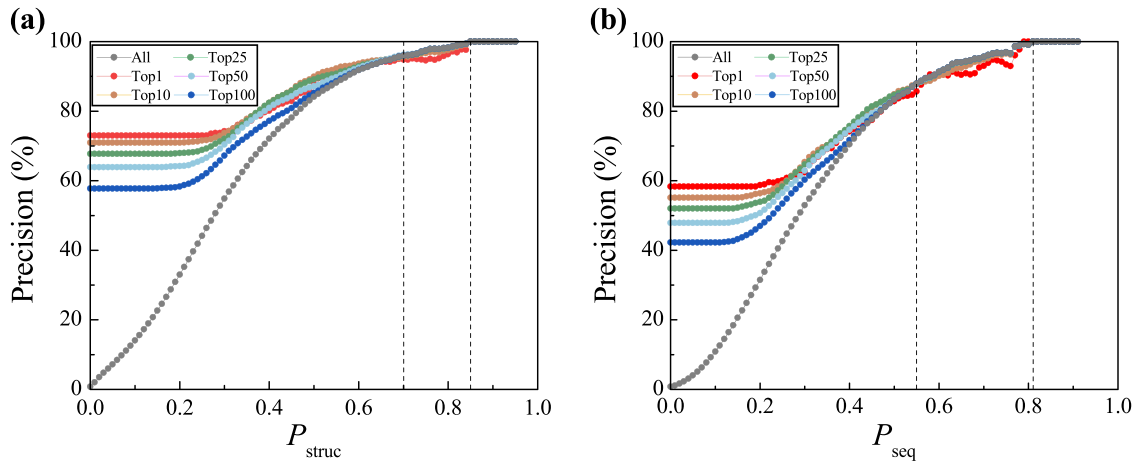

Figure 3. The precision as a function of contact probability cutoff on the PDB test set of 300 complexes. The overall precisions of DeepHomo2.0 (a) and DeepHomoSeq (b) when the top 1, top 10, top 25, top 50, top 100, and all predicted contacts were considered. The dashed lines indicate the crossover and convergence points, respectively.

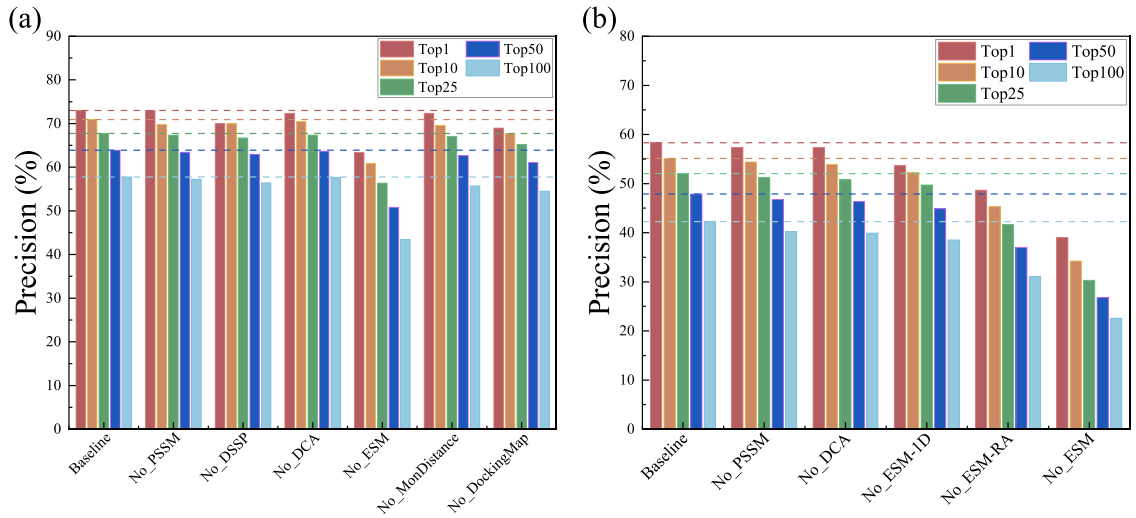

Figure 4. Ablation experiments of DeepHomo2.0 and DeepHomoSeq on the PDB test set of 300 complexes. (a) Six different models for DeepHomo2.0: No_PSSM for excluding PSSM feature, No_DSSP for excluding secondary structure type and solvent-accessible surface features, No_DCA for excluding DCA-DI and DCA-APC features, No_ESM for excluding ESM-MSA embedding vector and ESM-MSA row attention features, No_MonDistance for excluding intra-protein residue–residue distance map feature and No_DockingMap for excluding docking map. (b) Five different models for DeepHomoSeq: No_PSSM, No_DCA, No_ESM, No_ESM-1D for excluding ESM-MSA embedding vector feature and No_ESM-RA for excluding ESM-MSA row attention feature.

Impacts of the MSA depth and contact density

While we have analyzed the contributions of different input features of DeepHomo2.0 and DeepHomoSeq, there are still some factors that may affect the precision or limit the applicability of our model. The depth of multiple sequence alignment is a critical factor in intra-protein contact prediction. The number of native contacts indirectly reflects the difficulty of complex contact prediction [48, 50]. Therefore, we analyzed the impacts of MSA and contact density on our deep learning model. Specifically, for each target, the number of non-redundant sequences was measured as the effective sequence number (M_{eff}) after reweighting [23]. In this study, we used a sequence identity cutoff of 70% to remove the redundant sequences. The M_{eff} was defined as follows

$$\mathrm{M_{eff}} = \sum_{k=1}^{M} 1/m_k \tag{5}$$

where M is the number of sequences in the MSA and m_k is the number of sequences which have a sequence identity of more than 70% with the k-th sequence. The contact density is defined as the number of inter-protein contacts per length of protein sequence, and can be calculated as follows

$$\textit{Contact Density} = \mathrm{Sum}\ (\textit{contacts})/\mathrm{L} \tag{6}$$

where L is the length of the sequence.

Figure 5a and b shows the precisions of DeepHomo2.0 and DeepHomoSeq for the top 10 predicted contacts versus $\ln(M_{eff})$ on the PDB test set with experimental structures. For comparison, the results of DCA-APC, DCA-DI and ESM-MSA-based methods are also shown in Figure 5. It can be seen from the figures that DeepHomo2.0 and DeepHomoSeq obtained a better performance than other methods in low and high $\ln(M_{eff})$. Especially, DeepHomo2.0 (DeepHomoSeq) achieved a precision of 65% (45%) in the top 10 predicted contacts for the targets with $\ln(M_{eff})$ ranging from 0 to 3, while the other methods only obtained a precision of less than 20%. This indicates that DeepHomo2.0 and DeepHomoSeq can pick up the correct hidden information in a large

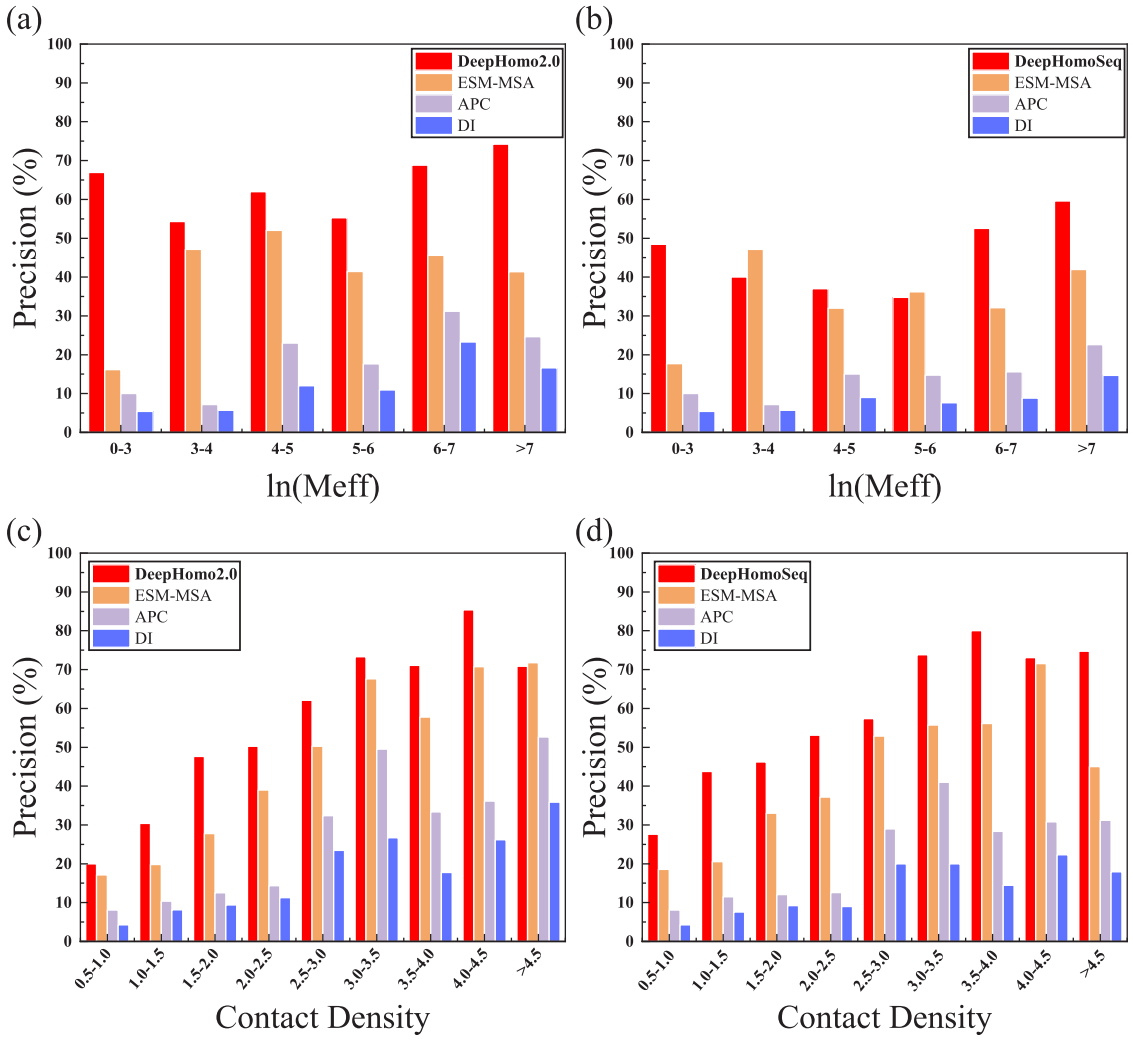

Figure 5. The impact of MSA depth and contact density. The performance of top 10 predicted contacts for DeepHomo2.0, DeepHomoSeq, DCA-DI, DCA-APC and ESM-MSA methods on the PDB test set of 300 complexes. The precision of DeepHomo2.0 with different ranges of ln(Meff) on experimental structures (a). The precision of DeepHomoSeq with different ranges of ln(Meff) (b). The precision of DeepHomo2.0 with different ranges of contact density on experimental structures (c). The precision of DeepHomoSeq with different ranges of contact density (d).

amount of noise. Besides, ESM-MSA and coevolutionary-based methods obtained an unexpected high precision in the range of 4–5 $\ln(M_{eff})$. Further examinations revealed that proteins with a $\ln(M_{eff})$ ranging from 4 to 5 obtained a high contact density (4.29), which would make it easier to predict the contacts. In addition, as a target protein has more non-redundant homologous sequences, the hidden residue–residue interaction information would be more easily captured. Therefore, all methods achieved a high precision with the increased number of homology, and DeepHomo2.0 (DeepHomoSeq) obtained the best performance on the target proteins with a $\ln(M_{eff})$ more than 6.

Figure 5c and d shows the precisions of DeepHomo2.0 and DeepHomoSeq for the top 10 predicted contacts versus contact density on the PDB test set with experimental structures. It can be seen from the figures that DeepHomo2.0 and DeepHomoSeq obtained better performance with increased contact density, and achieved the best performance in almost all contact density ranges. Moreover, we can observe that the complexes with a contact density ranging from 4.0 to 4.5 obtained the highest precision, which is because it obtained a larger depth of MSA. Especially, even the targets with a low contact density, our deep learning model also achieved a higher precision than the other methods.

Through the analysis of the effects of M_{eff} and contact density, we may conclude that both of them are the main factors of deep learning-based and coevolution-based methods. The higher contact density (larger interaction interface) and more homologous sequences will greatly improve the performance of different predicted methods. Moreover, our DeepHomo2.0 and DeepHomoSeq can achieve much better performance even if the M_{eff} and contact density are very low, which shows the robustness of our models.

Examples of contact prediction

Figure 6 shows the top 100 predicted contacts of four different methods in four triangular matrix regions for target 2IA0 from the PDB test set. In each region, native contacts, correct predictions, and incorrect predictions are colored in gray, green and red, respectively. It can be seen from Figure 6 that the top 100 predicted contacts predicted by DeepHomo2.0, DeepHomoSeq and DeepHomo1.0 were all concentrated in local areas, which indicates that our deep learning-based architecture can grab the important interaction between secondary structures, which may be a key interaction for biological functions. However, the top 100 predicted contacts predicted by coevolution-based methods and

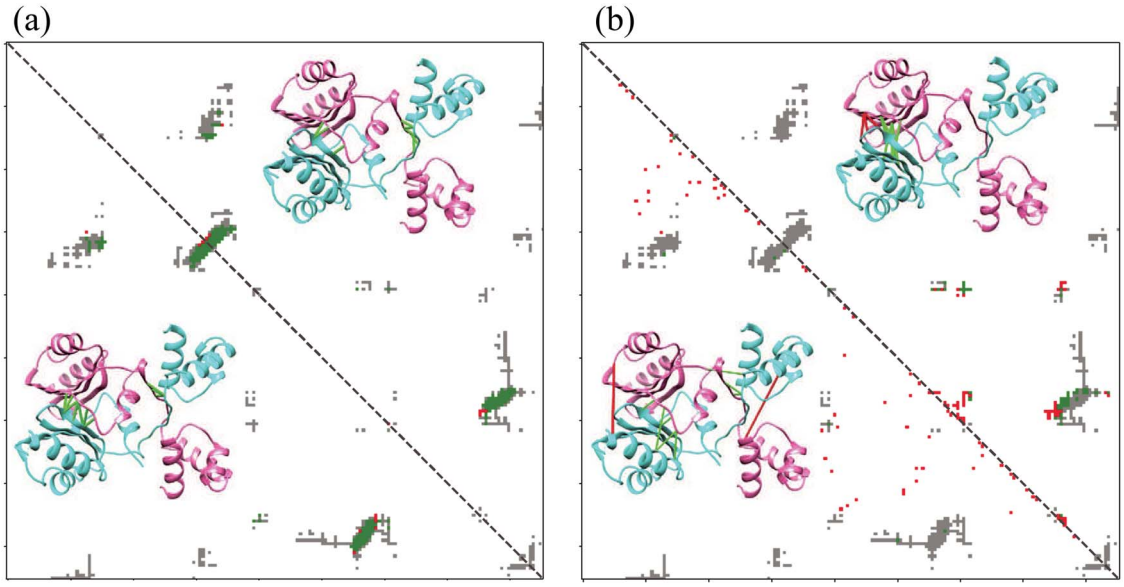

Figure 6. Comparison between the native contacts and the top 100 contacts predicted by four methods on 2IA0 complex. The native contacts, correct predictions and incorrect predictions are colored in gray, green and red, respectively. The diagonal line is indicated by dotted lines. The 3D structures of 2IA0 with the top 10 predicted contacts by four methods are also shown in the map. The correct (incorrect) predicted contacts are colored as green (red). (a) DeepHomoSeq (lower left triangle) and DeepHomo2.0 (upper right triangle). (b) ESM-MSA (lower left triangle) and DeepHomo1.0 (upper right triangle).

Transformer method were distributed throughout the whole inter-protein contact map. In addition, for the 2IA0 complex, DeepHomo2.0 and DeepHomoSeq achieved high precisions of 92.0% and 91.0% for top 100 predicted contacts, respectively, compared with 47.0% for DeepHomo1.0, and 42.0% for ESM-MSA and 8% for DCA-APC.

In order to more intuitively show that correctly predicted contacts were mainly distributed in some secondary structure interaction regions, the three-dimensional structure of homodimer displayed by Chimera [78] is shown in Figure 6. From Figure 6, the top 10 correct contacts predicted by DeepHomo2.0 and DeepHomoSeq were mainly concentrated on the spatial adjacent helix secondary interaction region for 2IA0. Although the contacts predicted by DeepHomo1.0 also included the interaction of two secondary structures, they were far apart in space which leads to false positive predictions. The predicted inter-protein contact maps by ESM-MSA were relatively discrete and distributed in different areas.

Since the actual number of native contacts may be more than 100, we further analyzed the inter-protein contact map to fully analyze the performance of different methods in contact predictions. The full predicted contact maps of five different methods are shown in Figure 7. It can be seen from the figure that the coevolution-based methods have a lot of extra noise, which leads to low precision. DeepHomo1.0 and ESM-MSA methods predicted the correct contacts in most areas, but there were still lots of false positive predictions. The contours of inter-protein contact map predicted by DeepHomo2.0 and DeepHomoSeq were similar to the native contacts. This suggests that DeepHomo2.0 and DeepHomoSeq not only achieved high precisions for top predicted 100 contacts, but also performed well on the overall contact map.

Application to protein–protein docking

Finally, we have integrated predicted residue-residue contacts into our *ab initio* symmetric protein-protein docking program HSYMDOCK to model the structure of protein homodimeric complexes. Specifically, putative binding modes with a basic shape complementarity are first sampled through a FFT-based search algorithm. There, the sampled putative binding modes are evaluated by our iterative knowledge-based scoring function ITScorePP plus a contact-based energy score as follows

$$E_{\text{cont}} = \begin{cases} E_0 & \text{if } r_{ij} \le 8.0\ \text{Å} \\ E_0 * \left(1 - \frac{r_{ij}-8.0}{4.0}\right) & \text{if } 8.0\ \text{Å} < r_{ij} \le 12.0\ \text{Å} \\ 0.0 & \text{if } r_{ij} > 12.0\ \text{Å} \end{cases} \quad (7)$$

where r_{ij} is the distance between two residues that are predicted as a contact. The E_0 represents an empirical constraint energy between two residues for a predicted contact and was set to be −100.0 kcal/mol in the present study. As protein docking is sensitive to false-positive contacts, only those reliable contacts that are above a contact probability threshold were considered during the docking calculations. Here, the contact threshold is empirically set to 0.65, which is around the middle between the crossover and convergence points in the precision-probability cutoff relationship (Figure 3). With our contact-guided protein docking algorithm, we have modeled the complex structures of 28 realistic homodimeric targets from previous CASP–CAPRI experiments, where the monomer structure was taken from the top prediction by the Zhang server in the CASP experiments [79, 80].

Here, the accuracy of a predicted complex structure is divided into four categories, high, medium, acceptable and incorrect, using the CAPRI criteria depending on the combination of three parameters, ligand root mean square deviation (RMSD), interface RMSD and percentage of native contacts. A complex structure with an at least acceptable accuracy is defined as a successful prediction. Figure 8 shows the success rates of DeepHomo2.0 and DeepHomoSeq in complex structure prediction on the data set of 28 realistic CASP–CAPRI homodimeric targets when different numbers of top predictions are considered. As a reference, the corresponding results for the *ab initio* HSYMDOCK program have also been given in the figure. It can be seen from Figure 8 that

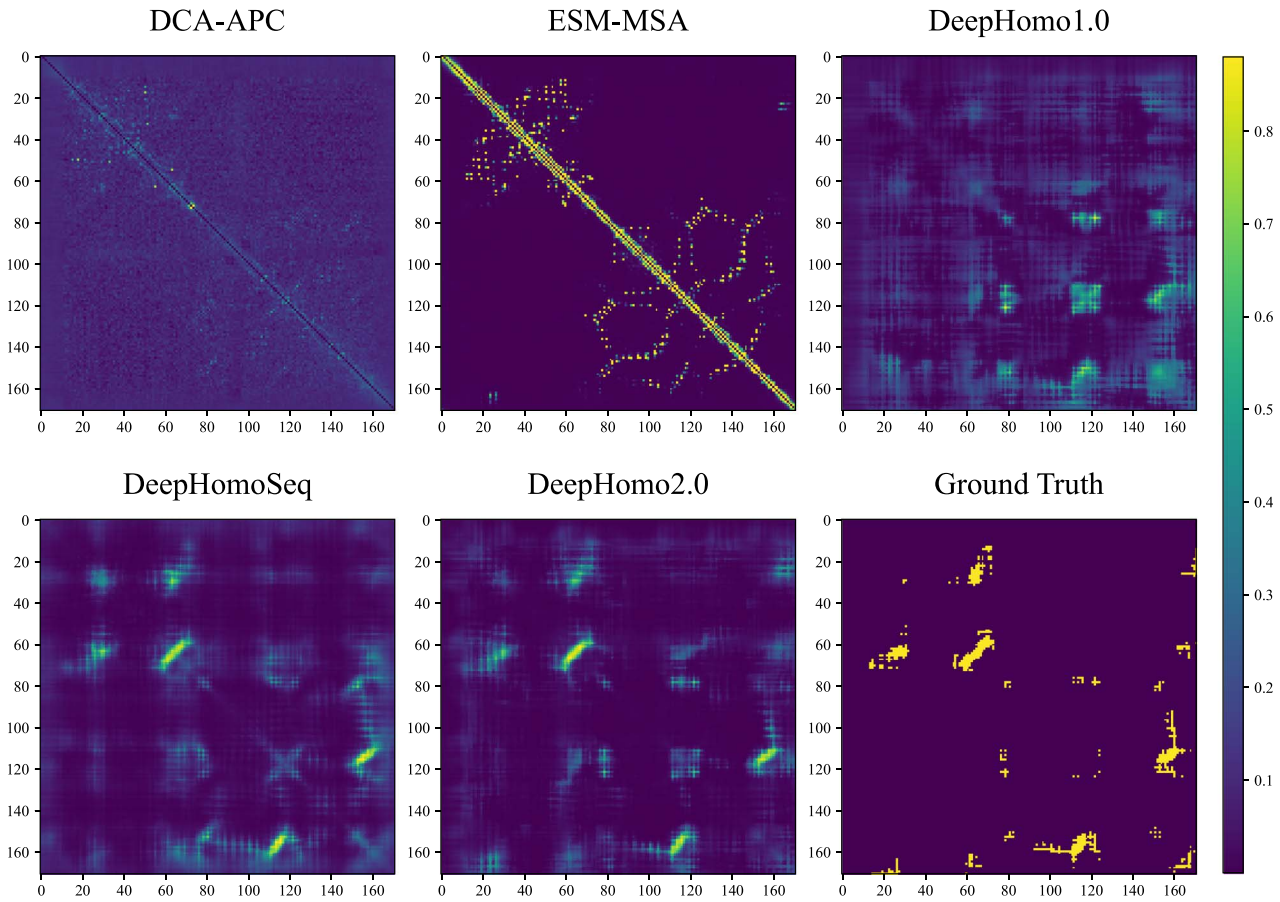

Figure 7. The whole predicted inter-protein contact map of five methods for 2AI0 complex, including DCA-APC, ESM-MSA, DeepHomo1.0, DeepHomoSeq and DeepHomo2.0. The native contact map is shown as a reference.

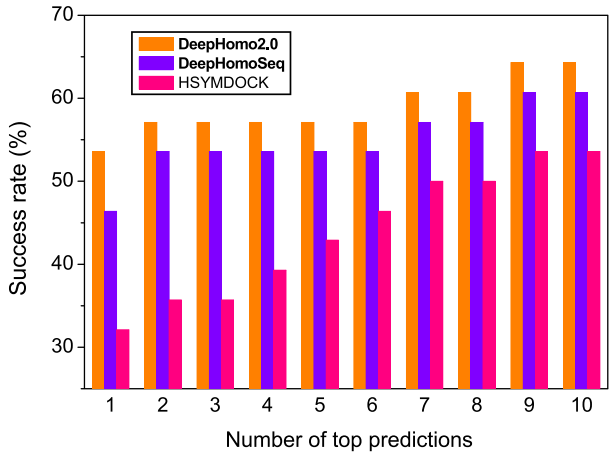

Figure 8. The success rates of DeepHomo2.0, DeepHomoSeq and HSYMDOCK in structure prediction of homo-oligomer complexes on the benchmark of 28 realistic CASP–CAPRI homodimeric targets when different numbers of top predictions were considered.

DeepHomo2.0 and DeepHomoSeq considerably improved the performance of structure prediction for the realistic homodimeric targets over ab initio HSYMDOCK protein docking. When the top 1, 5 and 10 predictions were considered, DeepHomo2.0 achieved the high success rates of 53.6%, 57.1% and 64.3%, respectively, compared with 46.4%, 53.6% and 60.7% for DeepHomoSeq and 32.1%, 42.9% and 53.6% for *ab initio* HSYMDOCK.

In addition, we have also compared our models with AlphaFold-Linker [81] and AlphaFold-Multimer [82]. To mimic real experiments, the monomer structures predicted by AlphaFold2 [61] were used in our docking calculations. The corresponding results were shown in Supplementary Figure S1. It can be seen from the figure that although AlphaFold-Multimer and AlphaFold-Linker obtained a higher success rates when the top 1 prediction was considered, DeepHomo2.0 achieved higher performance than AlphaFold-Linker starting from top 2 predictions and the same performance as AlphaFold-Multimer starting from top 4 predictions. The success rate of DeepHomo2.0 can increase further when more predictions were considered, and reach as high as 96.4% for top 100 predictions. Considering that the training set of AlphaFold-Multimer may share some similarities with the CASP–CAPRI test set, the present results show the robustness of docking-based approaches with DeepHomo2.0.

Computational efficiency

In addition, we also examined the running time of DeepHomo2.0, DeepHomoSeq and the other six methods for contact prediction. For the GLINTER method, we downloaded it from its web site and then ran locally to predict the inter-protein contacts. Since the database used by the authors of GLINTER is smaller than that for DeepHomo2.0, GLINTER would take less time to search the MSA and calculate the input features in its original study. For BIPSPI_seq and BIPSPI_struc, they only have the web service available. Therefore, we tried to estimate their running time by

comparing the times between the commit and the end after submitting a job to the web server. Since our methods need various input features, it would take more time to generate the features. Normally, a prediction can be finished within 30 min for most of the cases. For a typical homodimer like 3KH1 with 200 amino acids, DeepHomo2.0 and DeepHomoSeq consumes a total of 644 and 392 s to predict their inter-protein contacts, compared with 300 s for DCA-DI, 285 s for DCA-APC, 300 s for ESM-MSA, 130 s for GLINTER, 171 s for BISPI_seq and 667 s for BIPSPI_struc, respectively.

Conclusion

We have described two deep learning models (DeepHomo2.0 and DeepHomoSeq) for inter-protein residue–residue contact prediction across homodimeric interfaces. DeepHomo2.0 integrates sequence features of DCA and protein language models, both of which are generated from the MSA and structure features. The sequence-based version, DeepHomoSeq, only uses the sequence features.

The results on the diverse test sets showed that our DeepHomo2.0 outperformed the other eight tested approaches, including two coevolution-based methods, one protein language model and five machine learning-based approaches. Although no structure information is used by DeepHomoSeq, it can obtain a better performance than DeepHomo1.0 and BIPSPI method, which use the structure information. In addition, DeepHomo2.0 achieved the best performance among all methods and obtained a great improvement over our previous method (DeepHomo1.0). Those improvements are likely attributed to the use of Transformer features, which make our models more easily capture the inter-protein interactions in homodimeric proteins. Our further analysis shows that the precision of all methods is related to the depth of MSA and the contact density. With the increase of the two key factors, the precision of top predicted contacts also has an increasing trend. For the targets of low M_{eff} or contact density, the precision of DCA-based methods and protein language model based methods is very low, often less than 20%. On the contrary, our DeepHomo2.0 and DeepHomoSeq can still obtain a good performance for those targets. With inclusion of predicted inter-protein contacts, our integrated docking program significantly improved the structure prediction of homodimeric complexes, and obtain a success rate of 53.6% for DeepHomo2.0 and 46.4% for DeepHomoSeq on the benchmark of realistic CASP–CAPRI targets. These results demonstrate the accuracy of our models and their values in modeling the structure of homodimeric complexes.

Key Points

- Deep learning models were developed for inter-protein contact prediction of homodimers by including transformer features.
- Our models can predict the inter-protein contacts irrespective of structure information.
- Our models much outperformed existing approaches on both experimental and realistic complexes.
- Utilizing predicted contacts significantly improved the docking accuracy in protein-protein docking.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Data availability

DeepHomo2.0 and DeepHomoSeq are available at http://huanglab.phys.hust.edu.cn/DeepHomo2/.

Authors' Contributions

S.H. conceived and supervised the project. P.L., Y.Y. and S.H. designed and performed the experiments. P.L. and S.H. wrote the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grants nos 62072199 and 32161133002) and the startup grant of Huazhong University of Science and Technology.

References

1. Tuller T, Atar S, Ruppin E, *et al.* Common and specific signatures of gene expression and protein-protein interactions in autoimmune diseases. *Genes Immun* 2013;**14**(2):67–82.
2. Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci* 1996;**93**(1):13–20.
3. Huang SY. Search strategies and evaluation in protein-protein docking: principles, advances and challenges. *Drug Discov Today* 2014;**19**(8):1081–96.
4. Davis D, Yaveroğlu ÖN, Malod-Dognin N, *et al.* Topology-function conservation in protein-protein interaction networks. *Bioinformatics* 2015;**31**(10):1632–9.
5. Keskin O, Tuncbag N, Gursoy A. Predicting protein-protein interactions from the molecular to the proteome level. *Chem Rev* 2016;**116**(8):4884–909.
6. Farooq QUA, Shaukat Z, Aiman S, *et al.* Protein-protein interactions: methods, databases, and applications in virus-host study. *World J Virol* 2021;**10**(6):288–300.
7. Lesieur C. The assembly of protein oligomers: old stories and new perspectives with graph theory. *Oligomeriz Chem Biol Compound* 2014:327.
8. André I, Strauss CEM, Kaplan DB, *et al.* Emergence of symmetry in homo-oligomeric biological assemblies. *Proc Natl Acad Sci* 2008;**105**(42):16148–52.
9. Goodsell DS, Olson AJ. Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* 2000;**29**(1):105–53.
10. Dessau MA, Modis Y. Protein crystallization for X-ray crystallography. *J Vis Exp* 2011;**47**:2285.
11. Mani R, Cady SD, Tang M, *et al.* Membrane-dependent oligomeric structure and pore formation of a beta-hairpin antimicrobial peptide in lipid bilayers from solid-state NMR. *Proc Natl Acad Sci U S A* 2006;**103**(44):16242–7.
12. Bai XC, McMullan G, Scheres SHW. How cryo-EM is revolutionizing structural biology. *Trends Biochem Sci* 2015;**40**(1):49–57.
13. Scheres SHW. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* 2012;**180**(3):519–30.
14. Cheng Y. Single-particle cryo-EM-How did it get here and where will it go. *Science* 2018;**361**(6405):876–80.

15. Rohou A. Improving cryo-EM structure validation. *Nat Methods* 2021;**18**(2):130–1.
16. Berman HM, Westbrook J, Feng Z, *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**(1):235–42.
17. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell* 2015;**58**(4):586–97.
18. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**(D1):D480–9.
19. Weigt M, White RA, Szurmant H, *et al.* Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;**106**(1):67–72.
20. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008;**24**(3):333–40.
21. Jones DT, Buchan DWA, Cozzetto D, *et al.* PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;**28**(2):184–90.
22. Ekeberg M, Lövkvist C, Lan Y, *et al.* Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E* 2013;**87**(1):012707.
23. Morcos F, Pagnani A, Lunt B, *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci* 2011;**108**(49):E1293–301.
24. Ovchinnikov S, Park H, Varghese N, *et al.* Protein structure determination using metagenome sequence data. *Science* 2017;**355**(6322):294–8.
25. Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat Mach Intell* 2021;**3**:601–9.
26. Laine E, Eismann S, Elofsson A, *et al.* Protein sequence-to-structure learning: is this the end (-to-end revolution)? *Proteins* 2021;**89**(12):1770–86.
27. Du Z, Su H, Wang W, *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc* 2021;**16**(12):5634–51.
28. Wang S, Li Z, Yu Y, *et al.* Folding membrane proteins by deep transfer learning. *Cell Syst* 2017;**5**(3):202–211.e3.
29. Wu T, Guo Z, Hou J, *et al.* DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics* 2021;**22**(1):1–17.
30. Yang J, Anishchenko I, Park H, *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci* 2020;**117**(3):1496–503.
31. Li Y, Zhang C, Bell EW, *et al.* Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS Comput Biol* 2021;**17**(3):e1008865.
32. Wang S, Sun S, Li Z, *et al.* Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**(1):e1005324.
33. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* 2018;**34**(9):1466–72.
34. Schmitz S, Ertelt M, Merkl R, *et al.* Rosetta design with co-evolutionary information retains protein function. *PLoS Comput Biol* 2021;**17**(1):e1008568.
35. Figliuzzi M, Barrat-Charlaix P, Weigt M. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol Biol Evol* 2018;**35**(4):1018–27.
36. Cheng J, Choe MH, Elofsson A, *et al.* Estimation of model accuracy in CASP13. *Prot Struct Funct Bioinformatics* 2019;**87**(12):1361–77.
37. Ovchinnikov S, Kim DE, Wang RYR, *et al.* Improved de novo structure prediction in CASP 11 by incorporating coevolution information into Rosetta. *Prot Struct Funct Bioinformatics* 2016;**84**:67–75.
38. AlQuraishi M. AlphaFold at CASP13. *Bioinformatics* 2019;**35**(22):4862–5.
39. Monastyrskyy B, D'Andrea D, Fidelis K, *et al.* New encouraging developments in contact prediction: assessment of the CASP 11 results. *Prot Struct Funct Bioinformatics* 2016;**84**:131–44.
40. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, *et al.* Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Prot Struct Funct Bioinformatics* 2018;**86**:51–66.
41. Shrestha R, Fajardo E, Gil N, *et al.* Assessing the accuracy of contact predictions in CASP13. *Prot Struct Funct Bioinformatics* 2019;**87**(12):1058–68.
42. Yan Y, Huang SY. Accurate prediction of inter-protein residue-residue contacts for homo-oligomeric protein complexes. *Brief Bioinform* 2021;**22**(5):bbab038.
43. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* 2014;**3**:e02030.
44. Szurmant H, Weigt M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr Opin Struct Biol* 2018;**50**:26–32.
45. Hopf TA, Schärfe CPI, Rodrigues JP, *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* 2014;**3**:e03430.
46. Zhao Z, Gong X. Protein-protein interaction interface residue pair prediction based on deep learning architecture. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**(5):1753–9.
47. Liu J, Gong X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC Bioinformatics* 2019;**20**(1):609.
48. Zeng H, Wang S, Zhou T, *et al.* ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res* 2018;**46**(W1):W432–7.
49. Uguzzoni G, Lovis SJ, Oteri F, *et al.* Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci* 2017;**114**(13):E2662–71.
50. Quadir F, Roy RS, Halfmann R, *et al.* DNCON2_Inter: predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning. *Sci Rep* 2021;**11**(1):1–10.
51. Xie Z, Xu J. Deep graph learning of inter-protein contacts. *Bioinformatics* 2022;**38**(4):947–53.
52. Roy RS, Quadir F, Soltanikazemi E, *et al.* A deep dilated convolutional residual network for predicting interchain contacts of protein homodimers. *Bioinformatics* 2022;**38**(7):1904–10.
53. Quadir F, Roy RS, Soltanikazemi E, *et al.* Deepcomplex: a web server of predicting protein complex structures by deep learning inter-chain contact prediction and distance-based modelling. *Front Mol Biosci* 2021;**8**:716973.
54. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**(15):e2016239118.
55. Rao R, Liu J, Verkuil R, *et al.* Msa transformer. *International Conference on Machine Learning*. PMLR, 2021, 8844–56.
56. Brandes N, Ofer D, Peleg Y, *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**(8):2102–10.

57. Heinzinger M, Elnaggar A, Wang Y, *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 2019;**20**(1):1–17.
58. He K, Zhang X, Ren S, *et al.* Identity mappings in deep residual networks. In: *European Conference on Computer Vision*. Springer, Cham, 2016, 630–45.
59. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**(11):1026–8.
60. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;**9**(1):2542.
61. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**(7873):583–9.
62. Seemayer S, Gruber M, Söding J. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* 2014;**30**(21):3128–30.
63. Yan Y, Huang SY. CHDOCK: a hierarchical docking approach for modeling Cn symmetric homo-oligomeric complexes. *Biophys Rep* 2019;**5**(2):65–72.
64. Yan Y, Tao H, Huang SY. HSYMDOCK: a docking web server for predicting the structure of protein homo-oligomers with Cn or Dn symmetry. *Nucleic Acids Res* 2018;**46**(W1):W423–31.
65. Remmert M, Biegert A, Hauser A, *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**(2):173–5.
66. Li Y, Hu J, Zhang C, *et al.* ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* 2019;**35**(22):4647–55.
67. Mirdita M, von den Driesch L, Galiez C, *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* 2017;**45**(D1):D170–6.
68. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**(12):2577–637.
69. Steinegger M, Meier M, Mirdita M, *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;**20**(1):1–15.
70. Lin TY, Goyal P, Girshick R, *et al.* Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**(2):318–27.
71. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.
72. Sanchez-Garcia R, Sorzano COS, Carazo JM, *et al.* BIPSPI: a method for the prediction of partner-specific protein-protein interfaces. *Bioinformatics* 2019;**35**(3):470–7.
73. Chen T, He T, Benesty M, *et al.* Xgboost: extreme gradient boosting. R package version 0.4-2. 2015;**1**(4):1–4.
74. Lensink MF, Velankar S, Kryshtafovych A, *et al.* Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment. *Proteins* 2016;**84**:323–48.
75. Lensink MF, Velankar S, Baek M, *et al.* The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins* 2018;**86**:257–73.
76. Lensink MF, Brysbaert G, Nadzirin N, *et al.* Blind prediction of homo- and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins* 2019;**87**(12):1200–21.
77. Neuwald AF, Lanczycki CJ, Hodges TK, *et al.* Obtaining extremely large and accurate protein multiple sequence alignments from curated hierarchical alignments. *Database* 2020;**2020**:baaa042.
78. Pettersen EF, Goddard TD, Huang CC, *et al.* UCSF Chimera – a visualization system for exploratory research and analysis. *J Comput Chem* 2004;**25**(13):1605–12.
79. Yang J, Yan R, Roy A, *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat Methods* 2015;**12**(1):7–8.
80. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 2010;**5**(4):725–38.
81. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* 2022;**13**(1):1265.
82. Evans R, O'Neill M, Pritzel A, *et al.* Protein complex prediction with AlphaFold-multimer. BioRxiv. 2021; 2021.10.04.463034.