TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 07/09 pp843–854
DOI: 10.26599/TST.2021.9010064
Volume 27, Number 5, October 2022

Protein Residue Contact Prediction Based on Deep Learning and Massive Statistical Features from Multi-Sequence Alignment

Huiling Zhang†, Min Hao†, Hao Wu†, Hing-Fung Ting, Yihong Tang, Wenhui Xi*, and Yanjie Wei*

Abstract: Sequence-based protein tertiary structure prediction is of fundamental importance because the function of a protein ultimately depends on its 3D structure. An accurate residue-residue contact map is one of the essential elements for current *ab initio* prediction protocols of 3D structure prediction. Recently, with the combination of deep learning and direct coupling techniques, the performance of residue contact prediction has achieved significant progress. However, a considerable number of current Deep-Learning (DL)-based prediction methods are usually time-consuming, mainly because they rely on different categories of data types and third-party programs. In this research, we transformed the complex biological problem into a pure computational problem through statistics and artificial intelligence. We have accordingly proposed a feature extraction method to obtain various categories of statistical information from only the multi-sequence alignment, followed by training a DL model for residue-residue contact prediction based on the massive statistical information. The proposed method is robust in terms of different test sets, showed high reliability on model confidence score, could obtain high computational efficiency and achieve comparable prediction precisions with DL methods that relying on multi-source inputs.

Key words: multi-sequence alignment; residue-residue contact prediction; feature extraction; statistical information; Deep Learning (DL); high computational efficiency

1 Introduction

Proteins are the basic substance of all lifeforms[1], with a wide variety of functions in cells: playing a catalytic role; transporting metabolites; participating in the processes of immunity, cell differentiation, apoptosis, etc. The functions of a protein are largely determined by the corresponding three-dimensional (3D) structure; therefore, understanding the 3D structures and the folding mechanism of proteins has been a research issue of great importance in biology for decades. Although the 3D structures of proteins can be determined through wet-lab experimental techniques, these techniques are time-consuming and expensive. Consequently, the development of accurate algorithms to predict protein's tertiary structures from the corresponding amino acid sequences has been considered a "holy grail" for researchers in computational biology and bioinformatics. In recent years, *ab initio* protein folding has become

• Huiling Zhang, Wenhui Xi, and Yanjie Wei are with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: hl.zhang@siat.ac.cn; wh.xi@siat.ac.cn; yj.wei@siat.ac.cn.
• Min Hao is with College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China. E-mail: minphael@swu.edu.cn.
• Hao Wu is with School of Software Engineering, University of Science and Technology of China, Hefei 230051, China. E-mail: wuhaost@mail.ustc.edu.cn.
• Hing-Fung Ting is with Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China. E-mail: hfting@cs.hku.hk.
• Yihong Tang is with School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: tyh@bupt.edu.cn.
† Huiling Zhang, Min Hao, and Hao Wu contribute equally to this work.
* To whom correspondence should be addressed.
Manuscript received: 2021-07-20; revised: 2021-08-17; accepted: 2021-08-20

© The author(s) 2022. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

a more practical approach since the performance of template-based modeling depends on whether reliable template structures can be found in the Protein Data Bank (PDB). However, in the case of a lack of homologous structural templates, the accuracies of template-based modeling techniques will decline sharply. On the contrary, *ab initio* methods can work on the prediction task without relying on global template structures from PDB. The residue contact map is one of the core inputs for accurate *ab initio* protocols of protein 3D structure prediction.

A residue-residue contact (or simply a "contact") is a pair of residues that are non-local in the primary structure but close in the tertiary structure within a specific distance threshold. As a "simplified" two-dimensional (2D) representation of the 3D structure, protein contacts can provide necessary structural restraints toward the reconstruction of a protein's 3D structure. Accurate prediction of protein contact maps can pave an avenue for successful protein tertiary structure modeling. The past two decades have witnessed progress in how residue contacts can be applied successfully to 3D structure prediction modeling protocols[2–7]. The applications of the predicted contacts can also be extended to molecular dynamics simulations[8, 9] and protein-protein interaction prediction[10].

Residue contact prediction has been one of the most challenging issues in structural biology and bioinformatics. Prediction methods from the early stage can be categorized into three main classes: correlated-mutation-analysis/mutual-information-based techniques, mathematical-optimization-based methods, and Machine-Learning (ML)-based algorithms. Correlated-mutation-analysis/mutual-information-based methods, such as those described by Pollock and Taylor[11], MIp[12], and MIc[13], can achieve prediction precisions of approximately 20%–30% (contact definition as $C\beta$-$C\beta < 8 \times 10^{-10}$), and these techniques are local statistical models for a residue pair is treated statistically independent of other residue pairs. Residue contact prediction methods based on integer linear programming techniques proposed by Rajgaria et al.[14, 15] are *de novo* contact prediction techniques that rely on physical constraints instead of evolutionary information. These methods can be suitable solutions for proteins lacking homologous information, but can also significantly reduce the prediction accuracy for proteins with abundant evolutionary information. Traditional ML-based methods, such as SVMcon[16], NNcon[17], and SVMSEQ[18], are developed on small training sets, and they are local models that predict each residue pair without considering the status of other pairs. Therefore, these methods still show low prediction precisions and coverages, especially for long-range contacts. Some methods[19, 20] that merge machine learning techniques with integer linear programming can improve the prediction performance to certain extent.

The application of Direct Coupling Analysis (DCA)[21] to residue contact prediction shows a highlighted mark in the evolution of the prediction methods, which emphasizes the significance of disentangling direct interactions between residues from the indirect ones. Other methods falling in this category include PSICOV[22] based on sparse inverse covariance estimation and some variants of DCA, such as EVfold (mfDCA)[23], gDCA[24], plmDCA[25], GREMLIN[26], and CCMpred[27]. Although DCA-based methods can obtain more useful contacts than the mutual-information-based methods and traditional ML-based methods, their performance is highly dependent on the quality of MSA. Some meta-predictors, such as PconsC[28], MetaPSICOV[29], and NeBcon[30], try to overcome the shortcomings of pure DCA-based methods through combining the output of different DCA-based methods with other sequence-based information to create more robust predictions.

Recently, great success in contact prediction has been achieved through the introduction of deep learning techniques. Deep-learning-based method DeepCOV[31] uses only covariance information from MSA as input for accurate residue-residue contact prediction. Many deep-learning-based methods rely on many different input sources from third-party softwares, such as Position-Specific Scoring Matrix (PSSM), solvent accessibility, predicted secondary structure, and mid-files/prediction-scores from other contact predictors. Representative methods are RaptorX-Contact[32], DeepConPred2[33], DNCON2[34], DEEPCON[35], SPOT[36], and MapPred[37], and these methods usually require more computational resources than pure MSA-based methods.

In this study, we further demonstrate how to mine effective information in MSA as much as possible in order to achieve the best performance of the prediction model. Comparison with seven methods (as described in Table 1) from different categories, such as ML, ECA, and DL, reveals that the use of deep neural network models, and extensively extracted statistics (such as

Table 1 Summary of the evaluated methods in this work.

Method	Category	Main input	Description
NNcon	Traditional machine learning	PSSM; SA; SS	A 2D-recursive neural network based method trained on 424 protein chains.
SVMSEQ	Traditional machine learning	PSSM; SA; SS	A support vector machine-based method trained on the pdbselect25 dataset.
PSICOV	Direct coupling analysis	MSA	A method based on sparse inverse covariance estimation to recognize direct correlations of residue pairs in the MSA.
CCMpred	Direct coupling analysis	MSA	A pseudolikelihood-based method using co-evolution information from MSA as the implementation of GREMLIN for use with GPU/CPU and parallel computing.
MetaPSICOV	Consensus machine learning	MSA; EVFold(FreeContact); PSICOV; CCMpred; PSSM; SS; SA	A neural network based meta-predictor through combing the sequence-derived features and several distinct coevolution approaches; MetaPSICOV is trained on 624 protein chains.
DeepCov	Deep learning	MSA	A fully convolutional neural network based method operating covariance data derived directly taken from MSA; DeepCov is trained on 3456 protein chains.
DNCON2	Deep learning	PSSM; SI; MSA; SS; SA from PSIPRED and SCRATCH; and midfiles from MetaPSICOV, DNCON, PSICOV, CCMpred, and FreeContact	A method adopted a two-level hierarchical system. Five convolutional neural networks in Level one produce preliminary predictions, while the sixth network in Level two combines previous results to predict the final contact map; DNCON2 is trained on 1230 protein chains.

Note: MSA: multi-sequence alignment; PSSM: position-specific scoring matrix profile; SA: solvent accessibility; SS: secondary structure; SI: sequence-based information; HMM: profile hidden Markov model related feature from HHblits.

self-information, partial entropy, covariance information, mutual information, normalized mutual information, and cross-entropy) from the MSA (with sufficient effective sequences) can achieve state-of-the-art precisions.

2 Material and Method

2.1 Definition of residue contact and number of effective sequences

There are several different definitions for residue-residue contact. A residue pair can be in contact if (1) the Euclidean distance between the $C\beta$ ($C\alpha$ for Glycine) atoms interacting residue pair is less than 8×10^{-10}; (2) the distance $C\alpha$ atoms from the interacting residue pair is less than 12×10^{-10}[14, 15]; (3) the distance between any two atoms from the two residues is less than the sum of their van der Waals radii plus a threshold of 0.6×10^{-10}[38]; (4) the minimal distance between backbone heavy atoms or side-chain atoms in the residue pair is less than 5.5×10^{-10}[39]. In this work, we use the first contact definition which is the most widely adopted one and consistent with the CASP competitions. The contacts are classified into four categories according to the sequence separation (seq_sep) of the two interacting residues: all-range, short-range, medium-range, and long-range contacts defined as $seq\_sep > 6$, $6 \leqslant seq\_sep < 12$, $12 \leqslant seq\_sep < 24$, and $seq\_sep > 24$, respectively.

High-quality MSA is the core component for the success of most recently developed contact prediction approaches. In this study, JackHMMER from HHsuite is used to search against the NCBI-nr database for MSA generation with a cut-off E-value of 1.0×10^{-4}. The number of effective sequences (N_{eff}) in the alignment indicate a great influence on the prediction performance[40]. N_{eff} used in this work is defined following the definition and criteria prescribed by Zhang et al.[40].

2.2 Training/testing sets and model architecture

We use the same training set as DeepCov for benchmark comparison. The evaluation of the proposed model is based on two highly independent test sets. The first one is taken from the work by Zhang et al.[40], which consists of 103 proteins with $N_{eff} > 5 \times L$ (L is the sequence length). The second one is a test set containing 61 TransMembrane (TM) protein chains through culling all α-helix TM (αTM) proteins in PDBTM[41] against

the training sets of all methods (the proposed and the peer methods) used for evaluation in this work, with sequence identity < 25%, R-factor < 0.3, and resolution better than 2.0×10^{-10}. The two test sets are denoted as Testset1 (103 proteins with the majority as globular proteins) and Testset2 (61 αTM proteins).

The network architecture involves an input layer, a Maxout layer to reduce the number of input feature channels from 528 to 64, one or more 2D convolutional layers, and a final sigmoid output layer, each with batch normalization applied. The model is trained for 100 epochs using the Adam optimizer at an initial learning rate of 0.001. The learning rate is decreased by a factor of 0.5 if the S-score training loss do not decrease over the last five epochs. The training dataset is shuffled after each epoch. The final model is selected by taking the epoch with minimum S-score loss on the validation dataset.

2.3 Statistical features from multi-sequence alignment

MSA generally represents the process, algorithmic solution, and the result of three/more biological (protein or nucleic acid) sequences, and the problem has been extensively studied[42–44]. MSA has demonstrated its power in research domains, such as phylogenetic reconstruction, functional region illumination, structure prediction biomolecules. MSA also plays a central role in protein contact prediction because the evolutionary information extracted from the alignment is a significant contributor to many prediction methods. In this study, we use six categories of statistical information extracted from the MSA as input features for the DL model. The six categories of features are self-information, partial entropy, mutual information, normalized mutual information, cross-entropy, and covariance. The definitions and descriptions of these features are presented as follows:

(1) Self-information:

$$I_i^a = \log_2(p_i^a/\langle p_a\rangle) \tag{1}$$

(2) Partial entropy:

$$S_i^a = p_i^a \log_2(p_i^a/\langle p_a\rangle) \tag{2}$$

(3) Covariance:

$$C_{ij}^{ab} = p_{ij}^{ab} - p_i^a\, p_j^b \tag{3}$$

(4) Mutual information:

$$MI(i,j) = \sum_{i,j} p(i,j)\log(p(i,j)/p(i)p(j)) \tag{4}$$

(5) Normalized mutual information:

$$NMI(i,j) = MI(i,j)/\sqrt{S(i)S(j)} \tag{5}$$

(6) Cross-entropy:

$$H(i,j) = S(i) + S(j) - MI(i,j) \tag{6}$$

where p_i^a (p_j^b) stands for the frequency of finding amino acid of Type a (b) in the MSA at Column $i(j)$; $\langle p_a\rangle$ is the average frequency of amino acid of Type a in the entire database; p_{ij}^{ab} is the co-occurrence of amino acid of Types a and b in the MSA at Columns i and j. $p(i,j)$, $p(i)$, and $p(j)$ are the joint probability distribution at positions i and j, marginal probability distribution at position i, and marginal distribution at position j, respectively. $S(i)$ and $S(j)$ are the entropies at positions i and j, respectively.

Self-information (with a feature vector of length 42) and partial entropy information (with a feature vector of length 42) are 1-dimensional (1D) features, while mutual information (with a feature vector of length 1), normalized mutual information (with a feature vector of length 1), cross-entropy (with a feature vector of length 1), and covariance (with a feature vector of length 441) are 2D features. We use 528 statistical features in total.

2.4 Evaluation criteria

A predicted contact map is an $L \times L$ matrix with each element as the probability or score of the model's estimate. The prediction precision of the top-L/n (in peer works $n = 1, 2, 5, 10, \ldots$, and in this work n=1, 2, and 5) is the most widely used evaluation criteria by CASP and the peer works. Other useful evaluation criteria include the precision, coverage, and Matthew's Correlation Coefficient (MCC) for the predictions above a specific probability/score threshold[40]. The STandard Deviation (STD) is adopted to assess the dispersion of precisions, coverages, and MCCs on the whole test set.

The Jaccard index (Jaccard similarity coefficient) is used to analyze the prediction similarity between different methods. The Jaccard index is defined as $J(X,Y) = |X\cap Y|/|X\cup Y|$, where $|X\cap Y|$ and $|X\cup Y|$ represent the number of elements in the intersection and the union of X and Y, respectively.

3 Results and Discussion

The performance of the proposed method is evaluated against seven peer methods. The seven peer methods are representatives for traditional-ML-based (e.g., NNcon and SVMSEQ), DCA-based (e.g., PSICOV and CCMpred), ML-based meta-predictor (e.g., MetaPSICOV), and DL-based (e.g., DeepCov and DNCON2) methods. The evaluation is conducted in terms of a wide range of factors, such as method

similarity, prediction precisions/coverages/MCCs, model score reliability, physical-chemical interactions, and running time. The results and analyses in Sections 3.1 – 3.5 are based on Testset1, and the results in Section 3.6 are based on Testset2.

3.1 Prediction similarity between different contact predictors

The proposed method and the peer methods are different from each other in terms of both input features and methodology. Insightful observations of the prediction similarity between the predictors are significant for understanding the methodology relevance and the applicable scenarios of these methods.

Figure 1 illustrates the Jaccard indices and the prediction similarities between the eight methods used for comparison. The Jaccard index between each two of the eight methods is first calculated for every single protein and then averaged on Testset1. Through hierarchical clustering analysis (with Ward's method) of the overall Jaccard index matrix, we obtain the dendrogram heatmap, shown in Fig. 1. As we can see, these eight methods are categorized into three different groups, namely, traditional-ML-based, DCA-based, and DL-based methods. The proposed method is included in the DL category together with Jaccard indices of 0.5 with DeepCov and 0.6 with DNCON2, but the category of traditional-ML-based methods encompasses MetaPSICOV that combines several DCA-based methods. This could be explained by the consensus strategy and the contact prediction strategy adopted by MetaPSICOV: MetaPSICOV combines different DCA-based methods using ML techniques. Both MetaPSICOV and traditional ML methods use local prediction strategies. The results of the traditional-ML-based methods show high-level dissimilarity with those of DCA-based methods, but similar with those of DL-based methods. The main reason for this phenomenon is that the traditional-ML-based methods do not share too many similar inputs and features with the DCA-based methods, while all the DL-based methods are developed with the inspiration of DCA, and many of the DL-based methods, such as DNCON2, use the results of DCA-based methods as input features of the predictors. Another reason is that both the DL-based and DCA-based methods are global prediction methods, while the traditional-ML-based methods are local prediction methods (the definitions of the above-mentioned local prediction methods and global prediction methods are given elsewhere[4, 21]).

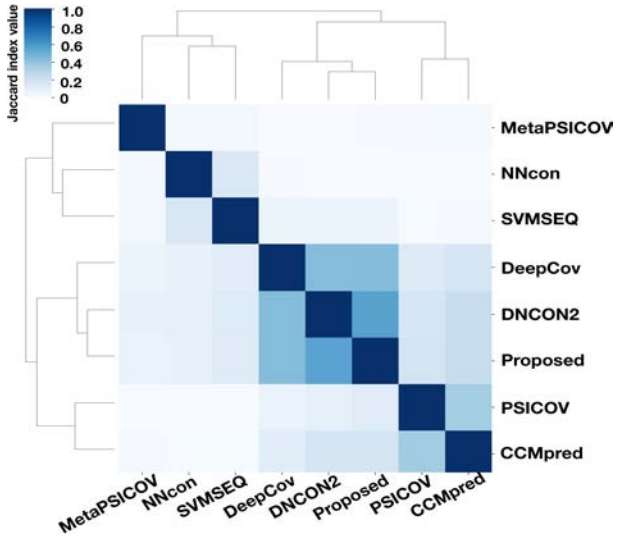


Fig. 1 Dendrogram heatmap of the prediction similarity between eight different methods. The Jaccard index between each two methods is calculated by averaging the Jaccard index value of each protein on the whole test set.

3.2 Performance on model reliability

In the "top-L/n" criteria, the contacts in the top-L/n predictions are selected through ranking the probabilities (scores) of the models, while in the "probability/score threshold" (for example, probability/score > 0.5) criteria, all residue pairs with probability/score larger than a specific probabilities/scores threshold are considered as contacts. Therefore, the credibility of the probability (score) generated by the prediction model ultimately determines the robustness and reliability of the corresponding model. The prediction probabilities/scores for NNcon, SVMSEQ, MetaPSICOV, PSICOV, CCMpred, DeepCOV, DNCON2, and the proposed method span at the ranges of [0.001, 1.000], [0.017, 0.861], [0, 0.989], [0.001, 0.904], [0, 5.270], [0, 1], [0, 1], and [0, 1], respectively.

Figure 2 presents the prediction performance given by different methods according to the prediction precisions, coverages, MCCs, and the corresponding STDs with the increase of probability/score threshold. As the probability (score) thresholds increase, the prediction coverages for all eight methods decrease gradually. With the increase of the probability thresholds, the prediction precision curves of all DL-based methods increase monotonically, while some DCA-based and traditional-ML-based methods are non-monotonic. The

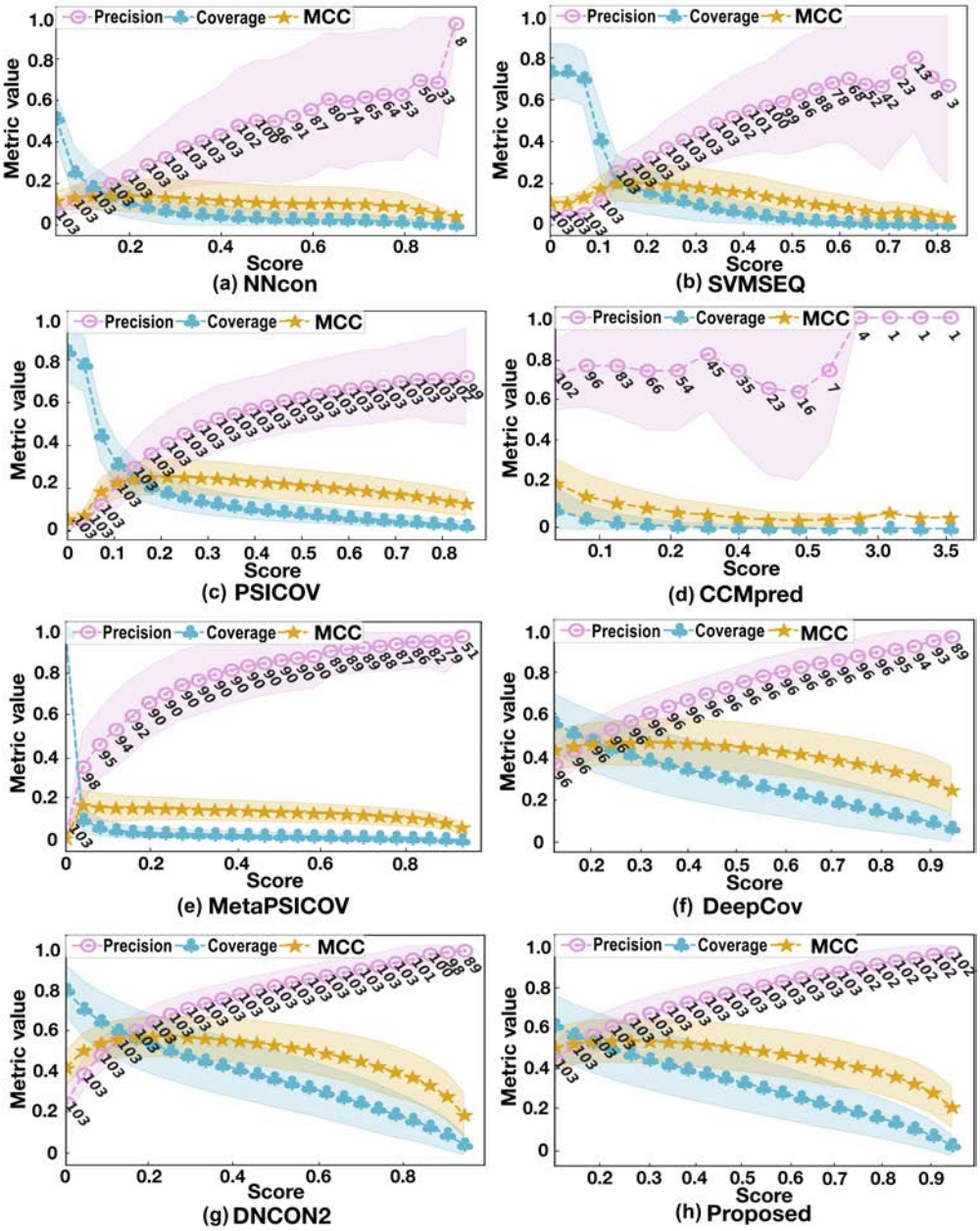


Fig. 2 Trend curves of different evaluation metrics with the increase of the probability (score) on Testset1. The prediction precisions, coverages, and MCCs are shown in pink, yellow, and blue, respectively, and their corresponding STDs are indicated as the colored shadow around the corresponding curves. The numbers nearby the pink precision curve are the number of proteins in Testset1.

prediction precisions of DCA-based method PSICOV, meta-method MetaPSICOV, and the DL-based methods increase with no turning points with the change of the probability threshold. However, the trend of the curves is not monotonic for traditional-ML-based methods NNcon/SVMSEQ and DCA-based method CCMpred. With the increase of the probability (score) threshold, the precision curves of NNcon/SVMSEQ/CCMpred have turning points at some probability (score) thresholds. Compared with DL-based methods, all the traditional-ML-based and DCA-based methods have much lower prediction, coverages, and MCCs, and much higher STDs of prediction precisions. Notably, the proposed method could retain predictions for 102 proteins with an average prediction precision of 96.34% at the probability threshold of 0.95, while the other two DL-based models, DeepCOV and DNCON2, retain only 89 proteins at the same probability threshold.

3.3 Running time comparison between DL methods

Three DL-based methods evaluated in this study share some similar inputs with direct-coupling-based and traditional-/cosensus-ML-based methods (Table 1 for details); therefore, only the runtime of the DL-based-methods is evaluated in this section. The hardware used for the runtime tests is a computing node with the Intel Xeon Gold 6230 CPUs. The runtime is evaluated on proteins with sequence lengths (numbers of residues) of 200 and 500. For the same benchmark protein, we adjust the number of sequences in the MSA to observe how the running time changes with a varying number of sequences. As shown in Fig. 3, for the proteins, the running times of all methods increase with the increase in the number of sequences.

Running time of the proposed method is close to that of DeepCov, but 0.2–0.3 times than that of DNCON2. The only input data for the proposed method and DeepCov are MSA, while DNCON2 also relies on the results from other contact prediction methods, mid-files from MetaPSICOV, SS/SA from PSIPRED and SCRATCH, and PSSM from PSI-BLAST, etc. It is clear that the running time is proportional to the complexity of the input data types. DNCON2 requires several different types of data as input; therefore, it takes much more time than DeepCov and the proposed method relies only on MSA as input.

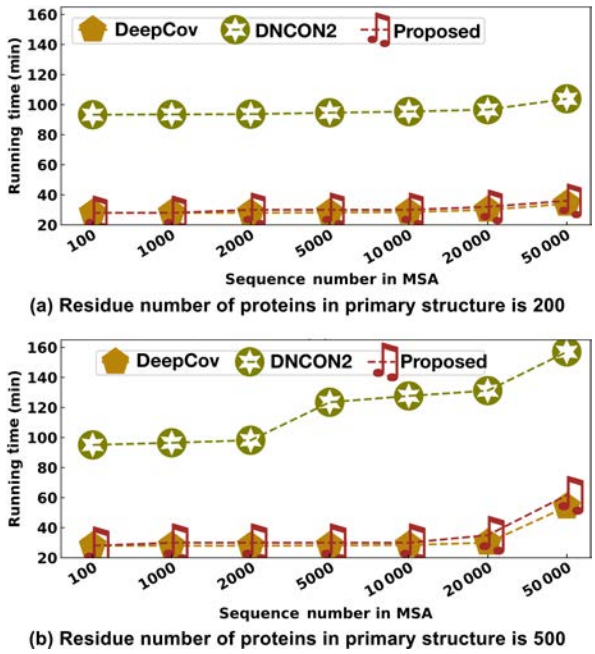


Fig. 3 Runtime comparison among the proposed method, DeepCov, and DNCON2.

3.4 Overall prediction precision on Testset1

To validate the effectiveness of the proposed method, the prediction precisions based on the top-L/n ($n = 1, 2, 5$) criteria at different sequence separations of our method are evaluated against those of other seven peer methods.

As illustrated in Fig. 4, it is obvious that the prediction precisions of our predictor are significantly better than those of NNcon, SVMSEQ, PSICOV, CCMpred, and MetaPSIOV for short-, medium-, long-, and all-range contacts (the definitions are based on different sequence separations as shown in Section 2.2). For example, the proposed method outperforms NNcon, SVMSEQ, PSICOV, CCMpred, and MetaPSICOV by 50.3%, 43.4%, 31.3%, 19.0%, and 30.5% for top-L all-range predicted contacts, respectively. Since there are only limited numbers of short-range and medium-range native contacts in proteins, and the contacts in these two ranges are relatively easy to predict, the proposed method does not show any obvious precision improvements (2% for top-L short-range and 3% for top-L medium-range) on these two ranges compared with another pure-MSA-dependent method DeepCov. For long-range contacts that are more important to 3D structure reconstruction and more difficult to predict than short-range and medium-range

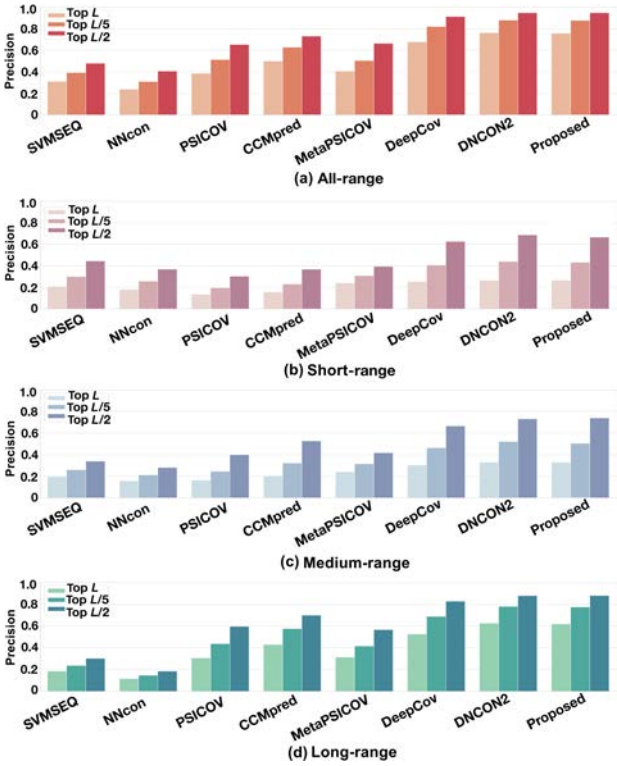


Fig. 4 Prediction precisions of the proposed method in comparison with those of other seven methods on Testset1.

contacts, the proposed method outperforms DeepCOV by 7% and 6% in precisions for top-L long-range and all-range predictions, respectively. The proposed method shows close prediction precisions with DNCON2, which combines different categories of information as input, while our model uses features from MSA only (the results in Section 3.3 indicate that DNCON2 is 3–5 times slower than the proposed method, and the running time gap increases with an increasing number of sequences in MSA).

3.5 Prediction performance of physical-chemical interaction

The tertiary structure of a protein is the spatial arrangement of all the secondary structures and stabilized by the bonding forces between "side chains" through hydrogen bonding, Van der Waals forces, disulfide bonds, salt bridges, and non-polar hydrophobic interactions. These interactions are prone to fracture due to external forces (e.g., heat), resulting in protein deactivation. Previous studies[45, 46] have suggested that evolutionary couplings can detect side-chain interactions.

In this section, we analyze the prediction coverage of physical-chemical interactions related to disulfide bond, salt bridge, and hydrophobic interactions. The coverage of the physical-chemical interactions is defined as the number of true positive contacts related to disulfide bond/salt bridge/hydrophobic interactions in top-L/n ($n = 0.2, 0.5, 1, 2, 5$) contact predictions divided by the total number of disulfide bond/salt bridge/hydrophobic interactions in a protein. The coverages of the physical-chemical interactions by different methods are shown in Table 2, and the precisions of residue-residue contact prediction are tabulated in Table 3). The proposed method shows prediction coverages of 58.33%, 58.33%, 29.76%, 16.66%, and 7.14% for top-5L, top-2L, top-L, top-$L/2$, and top-$L/5$ contacts with disulfide bonds, respectively, which are 4.76%, 11.91%, 3.57%, 2.38%, and 7.14% higher than those of DeepCov, respectively. The proposed method achieves prediction coverages of 65.76%, 45.79%, 32.56%, 19.33%, and 11.63% for top-5L, top-2L, top-L, top-$L/2$, and top-L/5 contacts with salt bridges, respectively, which are 12.70%, 6.11%, 3.27%, 0%, and 3.04% higher than those of DeepCov, respectively. The prediction coverages for top-5L, top-2L, top-L, top-$L/2$, and top-$L/5$ contacts with hydrophobic interactions by the proposed method are 81.49%, 69.59%, 56.58%, 42.65%, and 25.99%, respectively, which are 6.70%, 7.91%, 7.78%, 5.93%, and 3.63% higher than those DeepCov.

Table 2 Coverage of the physical-chemical interactions in top-L/n (n= 0.2, 0.5, 1, 2, 5) predicted contacts by different methods.

(%)

Type	Method	Top-5L	Top-2L	Top-L	Top-$L/2$	Top-$L/5$
Disulfide bond	NNcon	85.71	85.71	46.42	40.47	14.28
	PSICOV	52.38	52.38	38.09	23.80	16.66
	CCMpred	63.09	45.23	48.71	27.38	7.14
	DeepCov	53.57	46.42	26.19	14.28	0
	Proposed	58.33	58.33	29.76	16.66	7.14
Salt bridge	NNcon	14.74	8.97	6.69	4.41	2.50
	PSICOV	49.75	40.20	33.51	26.09	15.44
	CCMpred	61.63	49.93	41.26	34.31	20.70
	DeepCov	53.06	39.68	29.29	19.33	8.59
	Proposed	65.76	45.79	32.56	19.33	11.63
Hydrophobic interaction	NNcon	38.43	23.80	17.47	11.62	6.21
	PSICOV	50.68	38.16	27.97	20.23	11.00
	CCMpred	61.59	49.18	37.90	23.90	11.39
	DeepCov	74.79	61.68	48.80	36.72	22.36
	Proposed	81.49	69.59	56.58	42.65	25.99

Table 3 Prediction precisions for top-L/n (n = 0.2, 0.5, 1, 2, 5) all-range predicted contacts by different methods.

(%)

Method	Top-5L	Top-2L	Top-L	Top-L/2	Top-L/5
NNcon	11.12	17.14	23.75	30.89	40.59
PSICOV	15.42	26.77	38.57	51.33	65.38
CCMpred	19.97	35.54	49.99	62.84	73.14
DeepCov	27.29	50.60	68.82	82.15	91.40
Proposed	31.33	57.27	75.83	87.83	93.92

3.6 Prediction performance on Testset2

α-helical transmembrane (αTM) proteins that are responsible for interactions between cells and their environment represent an important protein category. Due to the insufficient number of resolved 3D structures of αTM proteins, predicting the residue contacts among the transmembrane segments of αTM proteins paves the way for protein folding, which can be further applied to the protein function discovery.

In this section, we test the performance of the proposed model on 61 αTM proteins in Testset2. Figure 5 compares the prediction results of the proposed method to the outcomes of four peer approaches based on the 61 αTM protein in Testset2. The proposed method obtains an average precisions of 33.9%, 43.5%, and 52.5% for top-L, top-$L/2$, and top-$L/5$ all-range predicted contacts, respectively. Among the five methods used for evaluation in Fig. 5, DeepCOV shows the

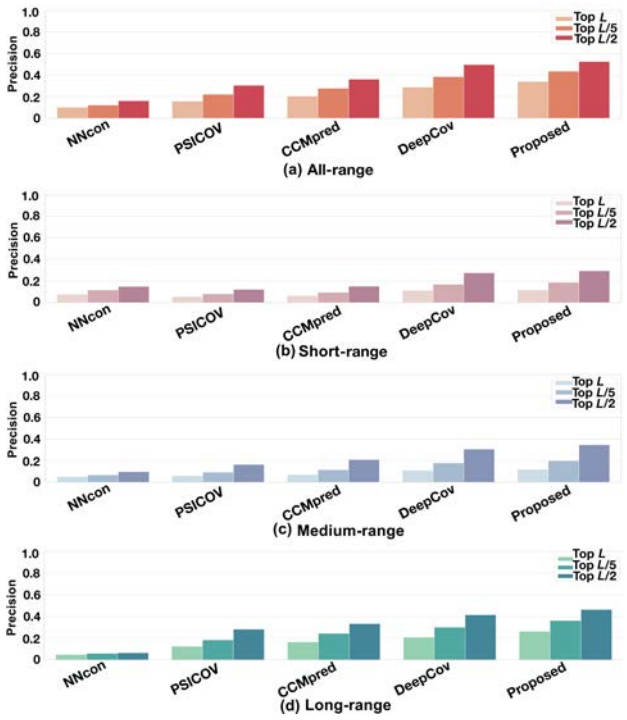


Fig. 5 Prediction precisions of the proposed predictor in comparison with other four peer methods on Testset2.

second-highest precisions that are still 5.2%, 4.9%, and 4.0% lower than those of the proposed method for top-L, top-$L/2$, and top-$L/5$ all all-range predicted contacts, respectively.

Although the proposed method shows higher prediction precisions than other peer methods, significant decreases are observed by comparing the results between Figs. 4 and 5, indicating the results based on Testset2 are less accurate than those of based on Testset1. The main reasons are summarized as below: (1) the evaluated methods are not specifically developed for contact prediction of αTM proteins; (2) there are more targets with $N_{eff} < 5 \times L$ in Testset2 than those in Testset1; and (3) there are 11, 21, and 41 αTM protein targets with native contacts less than $L/5$, $L/2$, and L, respectively. Changing the first two impactors is temporarily beyond the scope of this study, hence we select a subset of 20 targets from Testset1 with the number of native contacts greater than L. Table 4 shows the precisons for short-, medium-, and long-range predictions of the five methods, demonstrating that the results of all five methods outperform the results given in Fig. 5 by a substantial margin. Specifically, the proposed method achieves precisions of 17.97%, 29.43%, and 46.88% for top-L, top-$L/2$, and top-$L/5$ short-range predictions, 17.13%, 28.95%, and 46.52% for top-L, top-$L/2$, and top-$L/5$ medium-range predictions, respectively, and 37.29%, 49.16%, and 65.09% for top-L, top-$L/2$, and top-$L/5$ long-range predictions, respectively. The results demonstrate that the proposed DL model could better utilize the massive MSA-based features and show its effectiveness in residue contact prediction for αTM proteins.

4 Conclusion

Deep learning has demonstrated great power in applications across various fields, such as protein contact prediction. In this study, we developed a feature extraction method based only on the MSA, after which we trained a deep learning model for residue contact prediction based on the massive features (i.e., six types of statistical information) extracted from the MSA. The model was validated on two independent test sets in terms of a wide range of perspectives. It is demonstrated that the proposed method is robust in terms of the protein structural types, and shows high reliability on model confidence score. The results also indicate the effectiveness of the massive statistical features extracted from the MSA for residue contact prediction. Compared with the models that use only the covariance feature from MSA, the proposed method that adopts six types of MSA-based features shows no significant increase in computational time but satisfactory improvement in prediction precision. With enough effective sequences in the MSA, the proposed method using only MSA as

Table 4 Prediction precisions of the proposed predictor in comparison with other four methods for αTM proteins with the number of native contacts larger than $1.0 \times L$.

(%)

Method	Short-rang			Medium-rang			Long-rang		
	Top-L	Top-$L/2$	Top-$L/5$	Top-L	Top-$L/2$	Top-$L/5$	Top-L	Top-$L/2$	Top-$L/5$
NNcon	11.72	16.65	22.93	8.78	12.26	17.39	8.49	11.54	14.43
PSICOV	9.05	13.20	22.31	8.81	13.05	22.81	19.83	28.85	42.87
CCMpred	10.96	16.87	27.35	11.11	18.06	30.99	28.73	41.32	52.81
DeepCov	16.58	26.39	41.47	15.80	25.33	41.11	30.54	43.87	59.46
Proposed	17.97	29.43	46.88	17.13	28.95	46.52	37.29	49.16	65.09

input could achieve comparable prediction precision with the DL methods relying on multi-source inputs, while guarantee higher computational efficiency.

Acknowledgment

This work was partly supported by the Strategic Priority CAS Project (No. XDB38050100), the National Key Research and Development Program of China (No. 2018YFB0204403), the National Natural Science Foundation of China (No. U1813203), the Shenzhen Basic Research Fund (Nos. RCYX2020071411473419, JCYJ20200109114818703, and JSGG20201102163800001), CAS Key Lab (No. 2011DP173015), Hong Kong Research Grant Council (No. GRF-17208019), and the Outstanding Youth Innovation Fund (Doctoral Students) of CAS-SIAT (No. Y9G054).

References

[1] J. S. Zhang, W. K. Li, M. Zeng, X. M. Meng, L. Kurgan, F. X. Wu, and M. Li, NetEPD: A network-based essential protein discovery platform, *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 542–552, 2020.

[2] D. S. Marks, T. A. Hopf, and C. Sander, Protein structure prediction from sequence variation, *Nat. Biotechnol.*, vol. 30, no. 11, pp. 1072–1080, 2012.

[3] B. Adhikari, D. Bhattacharya, R. Z. Cao, and J. L. Cheng, CONFOLD: Residue-residue contact-guided *ab initio* protein folding, *Proteins: Struct., Funct., Bioinformatics*, vol. 83, no. 8, pp. 1436–1449, 2015.

[4] J. B. Xu, Distance-based protein folding powered by deep learning, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 34, pp. 16856–16865, 2019.

[5] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. L. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, et al., Improved protein structure prediction using potentials from deep learning, *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.

[6] J. Y. Yang, I. Anishchenko, H. Park, Z. L. Peng, S. Ovchinnikov, and D. Baker, Improved protein structure prediction using predicted interresidue orientations, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 3, pp. 1496–1503, 2020.

[7] M. Baek, F. Dimaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al., Accurate prediction of protein structures and interactions using a three-track neural network, *Science*, vol. 373, no. 6557, pp. 871–876, 2021.

[8] A. Raval, S. Piana, M. P. Eastwood, and D. E. Shaw, Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations, *Protein Sci.*, vol. 25, no. 1, pp. 19–29, 2016.

[9] E. A. Lubecka and A. Liwo, Introduction of a bounded penalty function in contact-assisted simulations of protein structures to omit false restraints, *J. Comput. Chem.*, vol. 40, no. 25, pp. 2164–2178, 2019.

[10] Q. Cong, I. Anishchenko, S. Ovchinnikov, and D. Baker, Protein interaction networks revealed by proteome coevolution, *Science*, vol. 365, no. 6449, pp. 185–189, 2019.

[11] D. D. Pollock and W. R. Taylor, Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution, *Protein Eng. Des. Sel.*, vol. 10, no. 6, pp. 647–657, 1997.

[12] S. D. Dunn, L. M. Wahl, and G. B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction, *Bioinformatics*, vol. 24, no. 3, pp. 333–340, 2007.

[13] B. C. Lee and D. Kim, A new method for revealing correlated mutations under the structural and functional constraints in proteins, *Bioinformatics*, vol. 25, no. 19, pp. 2506–2513, 2009.

[14] R. Rajgaria, S. R. McAllister, and C. A. Floudas, Towards accurate residue-residue hydrophobic contact prediction for α helical proteins via integer linear optimization, *Proteins: Struct., Funct., Bioinformatics*, vol. 74, no. 4, pp. 929–947, 2009.

[15] R. Rajgaria, Y. Wei, and C. A. Floudas, Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD, *Proteins: Struct., Funct., Bioinformatics*, vol. 78, no. 8, pp. 1825–1846, 2010.

[16] J. L. Cheng and P. Baldi, Improved residue contact prediction using support vector machines and a large feature set, *BMC Bioinformatics*, vol. 8, no. 1, p. 113, 2007.

[17] A. N. Tegge, Z. Wang, J. Eickholt, and J. L. Cheng, NNcon: Improved protein contact map prediction using 2D-recursive neural networks, *Nucl. Acids Res.*, vol. 37, no. S2, pp. W515–W518, 2009.

[18] S. T. Wu and Y. Zhang, A comprehensive assessment of sequence-based and template-based methods for protein contact prediction, *Bioinformatics*, vol. 24, no. 7, pp. 924–931, 2008.

[19] Z. Y. Wang and J. B. Xu, Predicting protein contact map using evolutionary and physical constraints by integer programming, *Bioinformatics*, vol. 29, no. 13, pp. i266–i273, 2013.

[20] H. L. Zhang, Q. S. Huang, Z. D. Bei, Y. J. Wei, and C. A. Floudas, COMSAT: Residue contact prediction of transmembrane proteins based on support vector machines and mixed integer linear programming, *Proteins: Struct., Funct., Bioinformatics*, vol. 84, no. 3, pp. 332–348, 2016.

[21] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 1, pp. 67–72, 2009.

[22] D. T. Jones, D. W. A. Buchan, D. Cozzetto, and M. Pontil, PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments, *Bioinformatics*, vol. 28, no. 2, pp. 184–90, 2012.

[23] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution

captures native contacts across many protein families, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[24] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani, Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners, *PLoS One*, vol. 9, no. 3, p. e92721, 2014.

[25] M. Ekeberg, C. Lövkvist, Y. H. Lan, M. Weigt, and E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models, *Phys. Rev. E*, vol. 87, no. 1, p. 012707, 2013.

[26] H. Kamisetty, S. Ovchinnikov, and D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence-and structure-rich era, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 39, pp. 15674–15679, 2013.

[27] S. Seemayer, M. Gruber, and J. Söding, CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations, *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, 2014.

[28] M. J. Skwark, A. Abdel-Rehim, and A. Elofsson, PconsC: Combination of direct information methods and alignments improves contact prediction, *Bioinformatics*, vol. 29, no. 14, pp. 1815–1816, 2013.

[29] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner., MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins, *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, 2015.

[30] B. He, S. M. Mortuza, Y. T. Wang, H. B. Shen, and Y. Zhang, NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers, *Bioinformatics*, vol. 33, no. 15, pp. 2296–2306, 2017.

[31] D. T. Jones and S. M. Kandathil, High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features, *Bioinformatics*, vol. 34, no. 19, pp. 3308–3315, 2018.

[32] S. Wang, S. Q. Sun, Z. Li, R. Y. Zhang, and J. B. Xu, Accurate de novo prediction of protein contact map by ultra-deep learning model, *PLoS Comput. Biol.*, vol. 13, no. 1, p. e1005324, 2017.

[33] W. Z. Ding, W. Z. Mao, D. Shao, W. X. Zhang, and H. P. Gong, DeepConPred2: An improved method for the prediction of protein residue contacts, *Comput. Struct. Biotechnol. J.*, vol. 16. pp. 503–510, 2018.

[34] B. Adhikari, J. Hou, and J. L. Cheng, DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks, *Bioinformatics*, vol. 34, no. 9, pp. 1466–1472, 2018.

[35] B. Adhikari, DEEPCON: Protein contact prediction using dilated convolutional neural networks with dropout, *Bioinformatics*, vol. 36, no. 2, pp. 470–477, 2020.

[36] J. Hanson, K. Paliwal, T. Litfin, Y. D. Yang, and Y. Q. Zhou, Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks, *Bioinformatics*, vol. 34, no. 23, pp. 4039–4045, 2018.

[37] Q. Wu, Z. L. Peng, I. Anishchenko, Q. Cong, D. Baker, and J. Y. Yang, Protein contact prediction using metagenome sequence data and residual neural networks, *Bioinformatics*, vol. 36, no. 1, pp. 41–48, 2020.

[38] A. Lo, Y. Y. Chiu, E. A. Rødland, P. C. Lyu, T. Y. Sung, and W. L. Hsu, Predicting helix-helix interactions from residue contacts in membrane proteins, *Bioinformatics*, vol. 25, no. 8, pp. 996–1003, 2009.

[39] T. Nugent and D. T. Jones, Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm, *PLoS Comput. Biol.*, vol. 6, no. 3, p. e1000714, 2010.

[40] H. L. Zhang, Z. D. Bei, W. H. Xi, M. Hao, Z. Ju, K. M. Saravanan, H. P. Zhang, N. Guo, and Y. J. Wei, Evaluation of residue-residue contact prediction methods: From retrospective to prospective, *PLoS Comput. Biol.*, vol. 17, no. 5, p. e1009027, 2021.

[41] D. Kozma, I. Simon, and G. E. Tusnády, PDBTM: Protein data bank of transmembrane proteins after 8 years, *Nucl. Acids Res.*, vol. 41, no. D1, pp. D524–D529, 2013.

[42] Y. Zhang, J. W. T. Chan, F. Y. L. Chin, H. F. Ting, D. S. Ye, F. Zhang, and J. Y. Shi, Constrained pairwise and center-star sequences alignment problems, *J. Comb. Optim.*, vol. 32, no. 1, pp. 79–94, 2016.

[43] W. T. Chan, Y. Zhang, S. P. Y. Fung, D. S. Ye, and H. Zhu, Efficient algorithms for finding a longest common increasing subsequence, *J. Comb. Optim.*, vol. 13, no. 3, pp. 277–288, 2007.

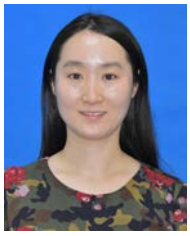
[44] C. X. Zhang, W. Zheng, S. M. Mortuza, Y. Li, and Y. Zhang, DeepMSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins, *Bioinformatics*, vol. 36, no. 7, pp. 2105–2112, 2020.

[45] A. J. Hockenberry and C. O. Wilke, Evolutionary couplings detect side-chain interactions, *PeerJ*, vol. 7, p. e7280, 2019.

[46] M. Chonofsky, S. H. P. De Oliveira, K. Krawczyk, and C. M. Deane, The evolution of contact prediction: Evidence that contact selection in statistical contact prediction is changing, *Bioinformatics*, vol. 36, no. 6, pp. 1750–1756, 2020.

Min Hao received the BEng degree in electronic information engineering in 2006, the MEng degree in signal and information processing in 2009, and the PhD degree in applied mathematics in 2020, all from Southwest University of China, where he is now a lecturer. His research interests include intelligent information processing, pattern recognition, bioinformatics, and affective computing.

Hao Wu received the BEng degree in automation from the University of Science and Technology of China in 2019, where he is currently a master student. His research focuses on deep learning and biomedical image processing.



Huiling Zhang received the BEng degree in communication engineering in 2008 and the MEng degree in signal and information processing in 2011 from Southwest University, China. She is currently a PhD candidate in computer application technology at the University of Chinese Academy of Sciences, China. Her research interests include high perfomance computing, data mining, pattern recognition, bioinformatics, and affective computing.

Hing-Fung Ting received the BS degree from The Chinese University of Hong Kong, China in 1985, and PhD degree from Princeton University, USA in 1993. He is now an associate professor at the Department of Computer Science, The University of Hong Kong. His research interests include bioinformatics, computational compleixty and design, and analysis of algorithms, and has published over 100 papers in these areas. He is also an active member of the ACM (HK).

Yihong Tang is a junior undergraduate at the Department of Computer Science and Technology, School of Computer Science, Beijing University of Posts and Telecommunications (BUPT). He is also a research assistant at GAMMA Lab of Beijing University of Posts and Telecommunications and Mobility AI Lab at The Hong Kong Polytechnic University. His main research interests include deep learning and its applications, graph mining, and intelligence transportations.

Wenhui Xi received the BS and PhD degrees in applied physics from Nanjing University, China in 2007 and 2012, respectively. From 2012 to 2018, he was a postdoctoral researcher at Fudan University, China and University of Oklahoma, USA. He is currently an associate professor at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include bioinformatics and molecular simulation of proteins and peptides.

Yanjie Wei received the BS degree in applied physics from Sichuan University, China in 2004, and the PhD degree in computational biophysics from Michigan Technological University, USA in 2007. From Mar. 2008 to Jun. 2011, he was a postdoctoral researcher at the Department of Biological and Chemical Engineering, Princeton University, USA. He is now a professor at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include computational biology and bioinformatics, focus on protein folding and structure prediction, and gene sequence analysis.