OXFORD

Briefings in Bioinformatics, 2022, **23(4)**, 1–12

https://doi.org/10.1093/bib/bbac169

Advance access publication date: 3 June 2022

Problem Solving Protocol

HDIContact: a novel predictor of residue–residue contacts on hetero-dimer interfaces via sequential information and transfer learning strategy

Wei Zhang, Qiaozhen Meng, Jianxin Wang and Fei Guo

Corresponding authors: Fei Guo, E-mail: guofei@csu.edu.cn; Jianxin Wang, E-mail: jxwang@mail.csu.edu.cn

Abstract

Proteins maintain the functional order of cell in life by interacting with other proteins. Determination of protein complex structural information gives biological insights for the research of diseases and drugs. Recently, a breakthrough has been made in protein monomer structure prediction. However, due to the limited number of the known protein structure and homologous sequences of complexes, the prediction of residue–residue contacts on hetero-dimer interfaces is still a challenge. In this study, we have developed a deep learning framework for inferring inter-protein residue contacts from sequential information, called HDIContact. We utilized transfer learning strategy to produce Multiple Sequence Alignment (MSA) two-dimensional (2D) embedding based on patterns of concatenated MSA, which could reduce the influence of noise on MSA caused by mismatched sequences or less homology. For MSA 2D embedding, HDIContact took advantage of Bi-directional Long Short-Term Memory (BiLSTM) with two-channel to capture 2D context of residue pairs. Our comprehensive assessment on the Escherichia coli (E. coli) test dataset showed that HDIContact outperformed other state-of-the-art methods, with top precision of 65.96%, the Area Under the Receiver Operating Characteristic curve (AUROC) of 83.08% and the Area Under the Precision Recall curve (AUPR) of 25.02%. In addition, we analyzed the potential of HDIContact for human–virus protein–protein complexes, by achieving top five precision of 80% on O75475-P04584 related to Human Immunodeficiency Virus. All experiments indicated that our method was a valuable technical tool for predicting inter-protein residue contacts, which would be helpful for understanding protein–protein interaction mechanisms.

Keywords: hetero-dimer interfaces, inter-protein contact prediction, sequential information, transfer learning, two-channel

Introduction

Proteins maintain the functional order of cell in life by interacting with other proteins [1]. As the structure informs function and diversity, the determination of protein complex structural information can help us correctly understand molecular mechanisms and related biological processes at the structural level [2, 3]. Traditional methods to determine the complex three-dimensional (3D) structural information are experimental methods, such as nuclear magnetic resonance spectroscopy [4], X-ray crystallography [5] or cryo-electron microscopy. However, technical difficulty and high cost cannot be ignored. It prompts the emergence of computational methods, which have gradually been developed to make up for the shortcomings of experimental methods.

A variety of protein docking methods have been developed to determine the 3D structure of two interacting proteins [6–9]. For the target of determining the 3D structure for most interacting proteins in organisms, the number of monomer protein structures with structural resolution in the Protein Data Bank (PDB) is still very limited [10]. However, the sequential information has increased dramatically, which benefits from high-throughput sequencing technologies and large-scale genome projects [11, 12]. For determining the 3D structure of complexes, the prediction of residue–residue contacts based on sequential information provides a valuable solution [13–16]. It is important that the prediction of inter-protein residue contact is useful for modeling 3D structure of complex [17, 18].

Unlike methods such as DNCON2_inter [19], DRCon [20] and DeepHomo [21], we mainly focused on the interactions between hetero-dimers rather than homo-oligomers. Computing methods for the prediction of residue–residue contacts on hetero-dimer interfaces from sequential information are mainly divided into Direct Coupling Analysis (DCA)-based, Machine Learning (ML)-based and Deep Learning (DL)-based methods. DCA is a group of methods to harvest information about

Wei Zhang is a master degree candidate in Tianjin University. Her research interests include Bioinformatics and Machine Learning.
Qiaozhen Meng is a PhD candidate in Tianjin University. Her research interests include Bioinformatics and Machine Learning.
Jianxin Wang is a Professor in Central South University. His main research interests include Bioinformatics and Algorithms.
Fei Guo is a Professor in Central South University. Her research interests include Bioinformatics and Computational Biology.

Received: January 17, 2022. **Revised:** March 7, 2022. **Accepted:** April 16, 2022

co-evolving residues in a protein family by learning a generative model from MSA [22–26]. It tries to distinguish between interacting and non-interacting pairs by extracting information directly from a single MSA, for making an effective prediction of monomer protein structure in the Critical Assessment of Protein Structure Prediction (CASP) experiment [27–30]. Subsequently, this method has been extended from intra-protein residue contact prediction to inter-protein [21, 31–34] based on concatenated MSA, including plmDCA [26, 35], CCMpred [25], Germlin [32] and EVcomplex [18, 31, 36]. In bacteria, two single MSAs can be concatenated according to their genomic distance, because the genes encoding the interacting proteins are usually located in the same operon [31, 32]. In addition to DCA-based methods, some ML methods have also been developed to predict the residue–residue contacts of the heterodimer interfaces, such as PAIRpred [37], PPiPP [38] and BIPSPI [39]. While these methods have more potential in reducing the influence of noise on MSA caused by mismatched sequences or less homology, limitations still need to be overcome, for example, they only focus on local information and cannot directly predict the contact map of hetero-dimer interfaces. The introduction of DL methods has made it possible to directly predict inter-protein residue contact. ComplexContact built two paired MSA, and applied a DL method originally developed for intra-protein contact prediction to predict interfacial contacts from paired MSAs. However, ComplexContact [33, 34] was not trained on the protein complex data. It is a significant need to develop more deep learning methods directly targeting hetero-dimers structure prediction.

Recent advances in protein language model open up many opportunities to improve intra-protein residue contact prediction [40, 41]. In particular, Rajani NF [42] studied the self-attention mechanism and discovered a correlation between self-attention maps and contact patterns. And Xu group developed GLINTER [43], which used coevolution signals generated by MSA Transformer and graph representation of protein structures generated by GCN recently. In this paper, to avoid the challenge of monomer structures, we proposed a deep learning model HDIContact for generating contact map of hetero-dimer interfaces via sequential information. In order to solve the problem of insufficient homologous sequences of protein complexes, HDIContact first concatenated MSAs of monomer based on the genome distance for complex, and transferred the skill of extracting co-evolutionary patterns to inter-protein contact prediction for producing MSA 2D embedding, which was learned by MSA transformer [44] on 26 million monomer MSAs through Masked Language Modeling (MLM). And then it learned the contextual information of all the inter-protein residue pairs on MSA 2D embedding from two-channel (L-BiLSTM and R-BiLSTM). We compared our novel model with other state-of-the-art methods on E. coli test dataset, considering that the concatenated MSA of complex from prokaryotes was more likely to appear more homologous sequences. In order to explore the potential of our method for human–virus complexes, protein complex related to HIV was selected as an independent test case for further comparison.

Methods

Deep learning architecture

In this study, we proposed a novel deep learning method to predict residue–residue contacts on the hetero-dimer interfaces from sequential data, called HDIContact. It combined transformer with BiLSTM, and used transfer learning strategy and two-channel mechanism to extract and integrate context-specific co-evolution information from homologous sequences for detecting inter-protein residue contacts. HDIContact was mainly divided into two steps: (I) generating MSA 2D embedding ($L \times R \times 144$); (II) capturing 2D context of residue pairs. The overall framework was shown in Figure 1.

Generating MSA 2D embedding

Transformer is a powerful sequence model that can transfer information from any locations to other locations [45]. MSA transformer is a protein language model recently proposed by FaceBook, which can extract information on patterns from columns of co-evolution in the MSA by pre-training an model on a large dataset of 26 million monomer MSAs [44]. Faced with the scarcity of homologous data for protein complexes, we applied knowledge and skills learned by MSA transformer on monomer MSAs to produce MSA embedding of complex through transfer learning strategy.

MSA transformer took a set of aligned sequences as input. Then word embedding and positional encoding were performed on the input to obtain $x \in R^{M\times(L+R)\times d}$, where M was the number of homologous sequences in the MSA, L and R were the sequence lengths of the ligand and receptor, respectively, d was the hidden dimension and $x_{mi:}$ was the encoding representation of residue i on sequence m. After that, each MSA transformer layer took a matrix of $M \times (L + R) \times d$ size as input and output. Each layer adapted axial attention approach [46, 47] to alternate attention over rows and columns of the 2D state. The tied row attention was computed as follows:

$$\sum_{m=1}^{M} \frac{Q_m K_m^T}{\sqrt{Md}}, \qquad (1)$$

where Q_m and K_m were obtained by linear transformation of $x_{mi:}$ as the matrix of queries and keys for the m-th row of input.

Here, the pre-trained model was consisted of 12 MSA transformer layers. It adapted a multi-head (12 heads for each layer) attention mechanism to use multiple matrix of queries Q and keys K to focus on different contexts at the same time. First, we used jackhammer [48] to search for the homologous sequences separately, and concatenated them based on the genome distance. Then,

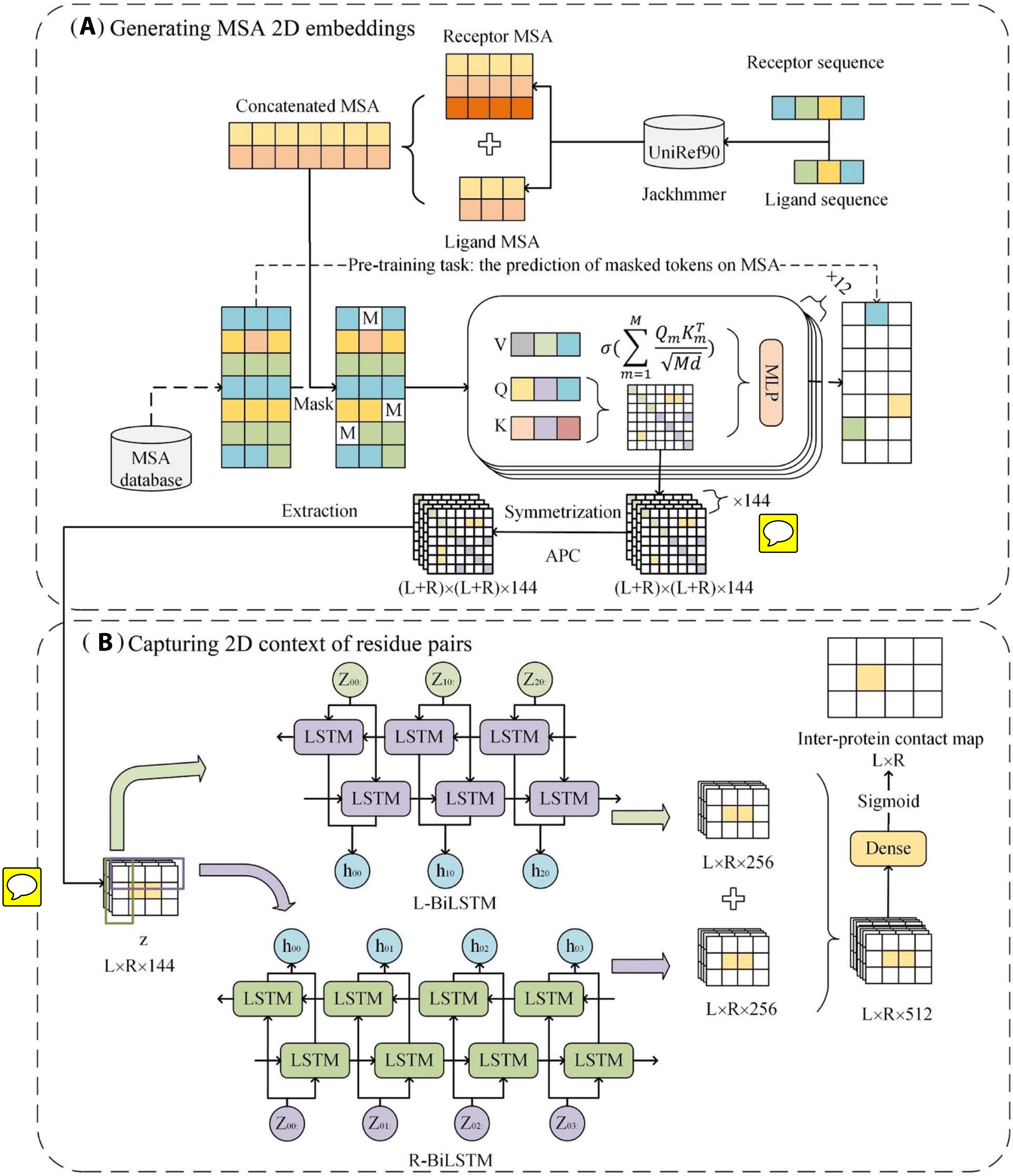

Figure 1. The overall framework of HDIContact. **(A)** We concatenated MSA based on the genome distance and generated MSA 2D embeddings through transfer learning strategy; **(B)** We captured 2D context of residue pairs from MSA 2D embedding on hetero-dimer interfaces based on two-channel (L-BiLSTM and R-BiLSTM).

we passed the concatenated MSA through the pre-train model to obtain the $(L+R) \times (L+R) \times 144$ row attention maps F (one map for each head in each layer). Next, we performed symmetrization for reducing the influence of receptor's and ligand's order when concatenated MSA, as follows:

$$F^{S}_{::k} = F_{::k} + F^{T}_{::k}, k = 1, 2, ..., 144 \tag{2}$$

We again applied Average Product Correction (APC) to each attention map independently, which was commonly used to correct for background effects in protein contact prediction [49].

$$F^{S+APC}_{ijk} = F^{S}_{ijk} - \frac{F^{S}_{i:k} F^{S}_{:jk}}{F^{S}_{::k}},$$

$$i, j = 1, 2, ..., L+R, k = 1, 2, ..., 144 \tag{3}$$

Finally, after preprocessing, we selected the attention maps of inter-protein as MSA 2D embedding.

Capturing 2D context of residue pairs

Long Short-Term Memory (LSTM) was specially designed to solve the problem of gradient disappearance caused by long-term dependence when processing long sequences in general Recurrent Neural Network[50]. LSTM was mainly composed of forget gate *f*, input gate *i* and output gate *o*, whose value at time-step *t* were computed as follows:

$$\begin{aligned} f_t &= \sigma(w_f \cdot [h_{t-1}, z_t] + b_f) \\ i_t &= \sigma(w_i \cdot [h_{t-1}, z_t] + b_i) \\ o_t &= \sigma(w_o \cdot [h_{t-1}, z_t] + b_o), \end{aligned} \qquad (4)$$

where z_t was the current input vector and h_{t-1} was the hidden state at time-step t, σ denoted the sigmoid activation function.

The cell state C_t updated old information (last cell state C_{t-1}) and stored new information (candidate cell state $\tilde{C}_t$) by forget gate f_t and input gate i_t, respectively, where $\tilde{C}_t$ was created by a tanh (hyperbolic tangent) layer.

$$\begin{aligned} C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ \tilde{C}_t &= tanh(w_c \cdot [h_{t-1}, z_t] + b_c) \end{aligned} \qquad (5)$$

After selectively forgetting and memorizing part of the information, hidden state h_t which included important long-term memory information was retained as follows:

$$h_t = O_t * tanh(C_t) \qquad (6)$$

Here, BiLSTM was a combination of forward and backward LSTM. Backward LSTM added a reverse operation on the basis of LSTM [51]. The final output of BiLSTM was the concatenation of the forward and backward LSTM. For MSA 2D embedding $(L \times R \times 144)$ z, R-BiLSTM learned the contextual features from $R \times 144$ embedding information between all residues on receptor and each residue on ligand, and generated *L* receptor hidden embedding ($R \times 256$) with the same length as receptor. While the other L-BiLSTM was the opposite, and learned $L \times 144$ embedding information. The hidden embedding learned from two-channel (receptor and ligand) was concatenated to obtain the contextual information of all residue pairs. Finally, after the dense layer, we got the predicted inter-protein contact map.

Training and test datasets

We selected public datasets with a sufficient number of homologous sequences and known hetero-dimer 3D structures for inter-protein residue contacts prediction. Baker dataset was employed [32] as training dataset, which was a set of 32 protein pairs. They were nonredundant complexes with experimentally solved structures in the PDB. Futhermore, E. coli dataset [31], a set of 59 protein pairs from prokaryotes, was selected as validation dataset and test dataset. We used blastclust to cluster all the concatenated sequences in Baker dataset and E. coli dataset under 80% coverage, and found that the identity of all sequences was below 5%. We took 20% of the E. coli dataset as a verification dataset with a total of 12 protein pairs, according to the distribution of sequence length shown in Table S1. The test dataset contained the remaining 80% of the E. coli dataset with a total of 47 protein pairs.

We generated MSA for each protein from all pairs by running jackhmmer [48] with three iterations and E-value = 1E − 3 to search through the UniRef90 library dated in October 2020. We concatenated pairs based on the genome distance according to the strategy employed by EVcomplex [31] and Baker [32]. We had performed de-redundancy on the concatenated MSA to ensure that the identity between sequences would not exceed 90%, and the coverage rate would be above 75%. The distribution of MSA depth on training dataset, validation dataset and test dataset is shown in Figure 2. Thereamong, MSA depth could be regarded as the number of nonredundant sequence homologs in an MSA under a sequence identity $<$ 90% cutoff [52]. It could be seen that all cases had a deeper MSA, which would provide more reliable co-evolution information for prediction.

Implementation

We used pytorch to implement our deep learning model and selected Adam as gradient descent optimization algorithm to optimize our model [53]. Learning rate was set to 0.001. And we chose sigmoid as activation function of output layer. What's more, due to the extremely unbalance between the number of interacting and non-interacting residue pairs, we employed Focal Loss [54] to train for each residue pair of *i* and *j* as follows:

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \qquad (7)$$

where α_t and p_t were defined as follows:

$$\alpha_t = \begin{cases} \alpha, & y = 1 \\ 1 - \alpha, & y = 0 \end{cases} \quad and \quad p_t = \begin{cases} p_{ij}, & y = 1 \\ 1 - p_{ij}, & y = 0, \end{cases} \qquad (8)$$

where p_{ij} was the probability of contact between residues *i* and *j*, α was the parameter to balance the importance of positive and negative examples, γ (set to 1.5) was the tunable focusing parameter to focus learning on hard examples and down-weight the numerous easy examples.

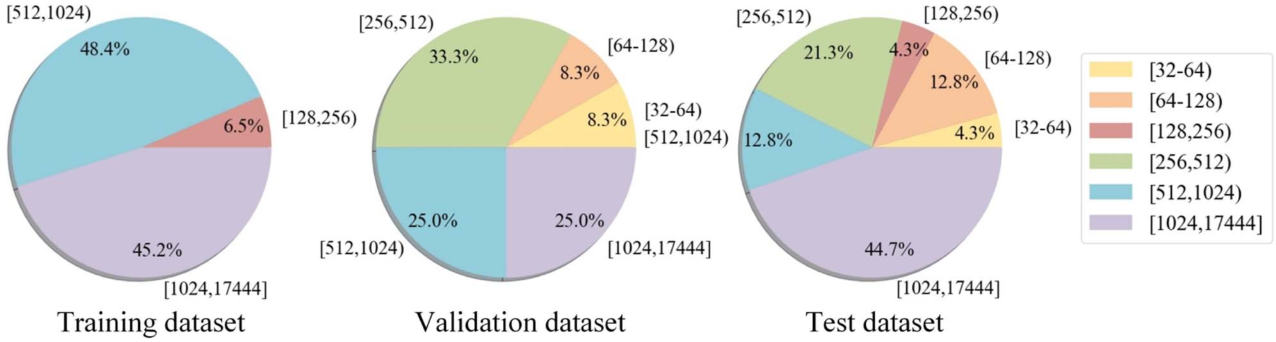

Figure 2. Distribution of MSA depth on training dataset, validation dataset and test dataset. MSA depth was divided into six groups, including $[32, 64), [64, 128), [128, 256), [256, 512), [512, 1024), [1024, 17444]$.

Results

Evaluation criteria

In order to evaluate the performance of inter-protein residue–residue contacts prediction, we used three criteria: precision, AUROC and AUPR. A pair of residues from two chains is considered to be in contact when the minimum distance of their atoms are within 10 Å [55]. The prediction precision is defined as the percentage of correctly predicted contacts among top N $(1, 2, \cdots, 99, 100)$ or L/K $(K = 30, 20, 10, 5, 2)$ predictions, where L is the total length of the two protein chains. It was commonly used in both intra- and inter-protein residue contact predictions [33].

Performance on E. coli test dataset

Comparison with state-of-the-art methods

To evaluate the performance of our proposed method for inter-protein residue–residue contact prediction on E. coli test dataset, we compared HDIContact with a series of state-of-the-art methods involving DCA-based, ML-based and DL-based methods. DCA-based method included plmDCA [26, 35], CCMpred [25], Germlin [32], EVcomplex [31, 56], ML-based method included BIPSPI [39] and DL-based method included ComplexContact [33, 34].

Average precision of all the targets were used to represent the performance of all methods on E. coli test dataset, when top N $(N = 1, 2, \cdots, 99, 100$, or the number of native contacts $P)$ predicted contacts were considered, as shown in Figure 3A. It could be seen that HDIContact achieved a much better average precision than other methods. For top 1, 10 and 100 predicted contacts, HDIContact obtained best precision of 65.96%, 55.96% and 46.06%, as shown in Table 1. HDIContact was the only one with at least 50% precision when up to 50 predicted contacts were considered on E. coil dataset. When the predicted contacts were determined according to the number of native contacts, the improvement of HDIContact was much more significant, and the precision has been improved almost over four times (28.49% versus 7%) than DCA-based methods and ML-based methods. It noticed that HDIContact could more accurately identify the residue pairs that were close in space by analyzing the correlated evolutionary sequence changes across proteins.

We also evaluated top L/K $(K = 30, 20, 10, 5, 2)$ predicted contacts to show the L-dependent precision, where L-dependent precision could weaken the effect of protein complex size. It could be seen from Table 2, HDIContact obtained higher and more reliable precision among seven methods. For top $L/10$, at least half of our predicted contacts were correct, while the other methods were only up to the precision of 46.35%. Besides the precision of top predicted contacts, we also compared the AUROC and AUPR without threshold constraint of all methods on E. coli test dataset. HDIContact obtained the highest AUROC of 83.08%. The AUPR of all methods was relatively low due to extremely unbalanced ratio of interacting and non-interacting pairs, but the AUPR of HDIContact has been improved at least 7% (25.02% versus 17.92%).

Comparison of different model architectures

For the MSA 2D embedding obtained through pre-train MSA transformer, we designed a series of schemes and conducted comparative experiments on E. coli test dataset. Compared with the deep learning architecture of HDIContact, we mainly tried different networks, including linear, BiLSTM, 3 × 3 2D convolution (Conv3) and 5 × 5 2D convolution (Conv5). They all used MSA 2D embedding as the input, and then extracted the contact information of residue pairs, and finally used the dense layer to integrate all the information extracted by the previous layer.

Average precision of all the targets when top N $(N = 1, 2, \cdots, 99, 100$, or the number of native contacts $p)$ or L/K $(K = 30, 20, 10, 5, 2)$ predicted contacts were considered is shown in Figure 4. HDIContact achieved better performance among five different model architectures with the precision of 65.96%, 55.96% and 46.06% for top 1, 10 and 100 predicted contacts, while the second-better method Conv3 gave lower precision of 59.57%, 51.91% and 41.38%, as shown in Table 3. AUROC and AUPR of different model architectures showed that our method yielded an overall much better performance, as shown in Table 4. It was worth noting that the precision of linear decreased much

Table 1. Comparison of precision by HDIContact and other methods on E. coli test dataset considering top *N* predicted contacts. *P* represented the number of native contacts on the target.

Methods	Top *N* precision (%)						
	1	**5**	**10**	**20**	**50**	**100**	*P*
HDIContact	65.96	58.3	55.96	53.83	50.89	46.06	28.49
BIPSPI	17.02	10.64	10.85	9.89	10.38	9.74	7.61
plmDCA	53.19	41.7	32.98	27.45	18.64	13.68	7.09
CCMpred	53.19	42.98	35.32	27.87	19.02	14.06	7.11
Germlin	53.19	42.55	34.68	28.3	19.57	14.32	7.16
EVcomplex	42.55	39.57	32.98	27.98	19.7	15.15	7.32
ComplexContact	29.79	44.26	45.11	46.17	45.06	40.7	22.74

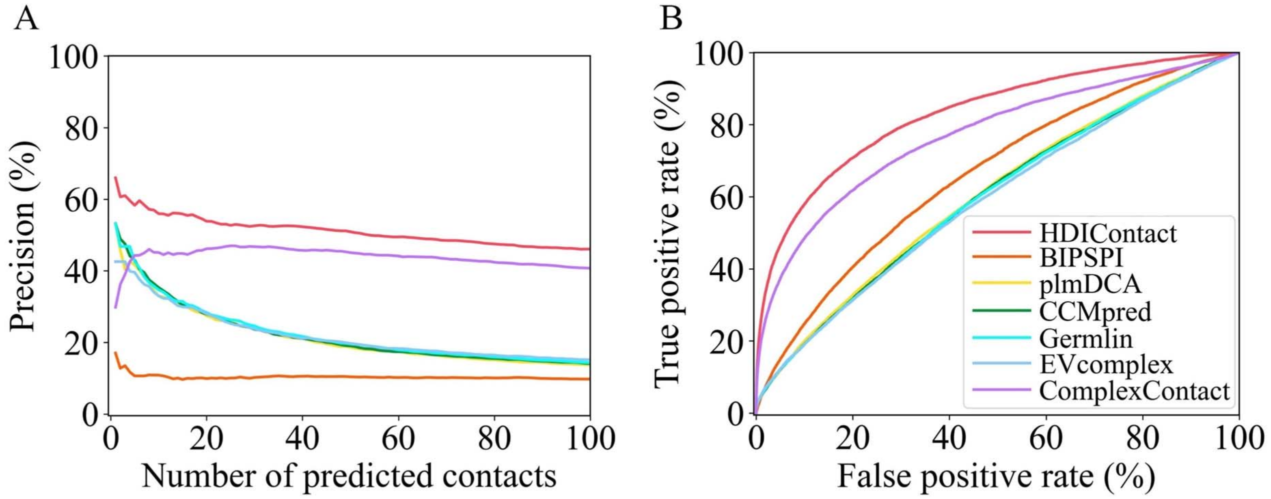

Figure 3. Performance of HDIContact and other state-of-the-art methods on E. coli test dataset. **(A)** Precision by all methods considering top *N* predicted contacts. **(B)** AUROC of all methods.

Table 2. Comparison of AUPR, AUROC and precision by HDIContact and other methods on E. coli test dataset considering top *L*/*K* predicted contacts. *L* was the sum amino acid sequence length of a protein pair.

Methods	AUPR	AUROC	Top *L*/*K* precision (%)				
	(%)	(%)	*L*/30	*L*/20	*L*/10	*L*/5	*L*/2
HDIContact	25.02	83.08	56.83	55.64	51.88	49.23	41.34
BIPSPI	6.11	66.1	10.08	8.91	9.92	10.23	9.68
plmDCA	5.19	60.48	33.48	28.93	21.82	15.6	10.51
CCMpred	5.15	59.89	33.7	29.46	21.85	15.99	10.85
Germlin	5.17	59.74	31.88	29.66	22.42	16.47	10.95
EVcomplex	5.12	59.12	32.87	29.24	22.13	16.46	11.46
ComplexContact	17.92	77.09	47.09	46.04	46.35	43.16	33.79

faster than other methods and got the lowest AUROC value when more top predicted contacts were considered. The Conv5 obtained worse accuracy than the Conv3 (48.94% of Conv5 verse 59.57% of Conv3 in top one predicted contacts). It might be due to a larger convolution kernel, which made some information loss during the integration and extraction process of local characteristics. While HDIContact used two-channel mechanism. Receptor channel R-BiLSTM learned the contextual features from *R* × 144 embedding information between all residues on receptor and each residue on ligand, while ligand channel L-BiLSTM was the opposite, and learned *L* × 144 embedding information. It allowed the model to achieve higher accuracy with fewer parameters (46.06% of our verse 39.91% of BiLSTM in top 100 predicted contacts).

Impact of key components

Impact of MSA

To study the impact of MSA, we used embedding features generated by model pre-trained on protein sequences (depth 0) instead of embedding features generated on MSA. As shown the results of depth 0, top precision values were only about 1% without the MSA of complex.

Table 3. Comparison of precision by different model architectures on E. coli test dataset considering top *N* predicted contacts. *P* represented the number of native contacts on the target.

Methods	**Top** *N* **precision (%)**						
	1	**5**	**10**	**20**	**50**	**100**	*P*
HDIContact	65.96	58.3	55.96	53.83	50.89	46.06	28.49
Conv5	48.94	45.11	44.26	43.62	40.81	38.26	24.99
Conv3	59.57	54.04	51.91	50.21	45.79	41.38	25.54
BiLSTM	63.83	55.32	54.47	50	44.85	39.91	23.45
Linear	61.7	56.6	53.62	47.98	41.02	35.19	19.9

Table 4. Comparison of AUPR, AUROC and precision by different model architectures on E. coli test dataset considering top *L*/*K* predicted contacts. *L* was the sum amino acid sequence length of a protein pair.

Methods	**AUPR**	**AUROC**	**Top** *L*/*K* **precision (%)**				
	(%)	**(%)**	*L*/30	*L*/20	*L*/10	*L*/5	*L*/2
HDIContact	25.02	83.08	56.83	55.64	51.88	49.23	41.34
Conv5	20.86	81.91	44.47	44.78	41.72	40.12	35.34
Conv3	21.88	81.57	53.33	51.9	47.87	44.41	37.33
BiLSTM	19.84	80.36	53.17	52.49	48.32	42.94	35.47
Linear	15.9	77.46	51.22	50.48	45.55	39.14	30.33

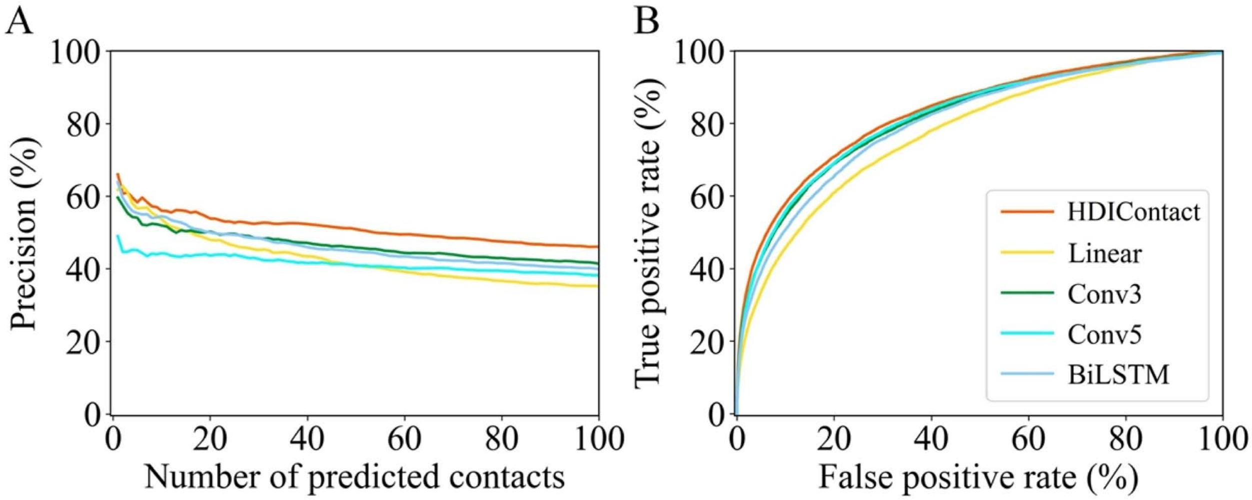

Figure 4. Performance of different model architectures on E. coli test dataset. **(A)** Precision by different model architectures considering top *N* predicted contacts. **(B)** AUROC of different model architectures.

BPISPI, which only used MSA of monomer, obtained the worst precision. All these proved that MSA was crucial as the input of HDIContact and other DCA-based methods.

Considering the importance of MSA, we have examined the effect of the MSA depth on the performance of all methods, except for ComplexContact and BIPSPI whose web servers do not provide MSA mode for input. MSA depth could be regarded as the number of nonredundant sequence homologs in an MSA under a sequence identity < 90% cutoff [52]. On E.coil test dataset, we explored the impact on the performance by randomly down-sampling the non-redundant MSA to a specified depth. Among them, down-sampling was not performed when the original MSA depth did not exceed the specified depth. Figure 5 showed that the precision of top 5, top 50 and top *L*/30 predicted contacts with respect to MSA depth on E. coil test dataset. For top predicted contacts, all methods generally obtained a better performance with increased MSA depth. A deeper MSA provided more reliable evolutionary information, which led to better performance of all test methods. In addition, HDIContact, only using MSA with 64 depth, could achieve the same top precision as other methods that used MSA with 1024 depth. And HDIContact gave the precision of 40.43% for top five predicted contacts, 31.96% for top 50 predicted contacts and top 36.34% for top 50 predicted contacts with 64 depth MSA, respectively. It proved that our method was able to learn co-evolution information from shallower MSA.

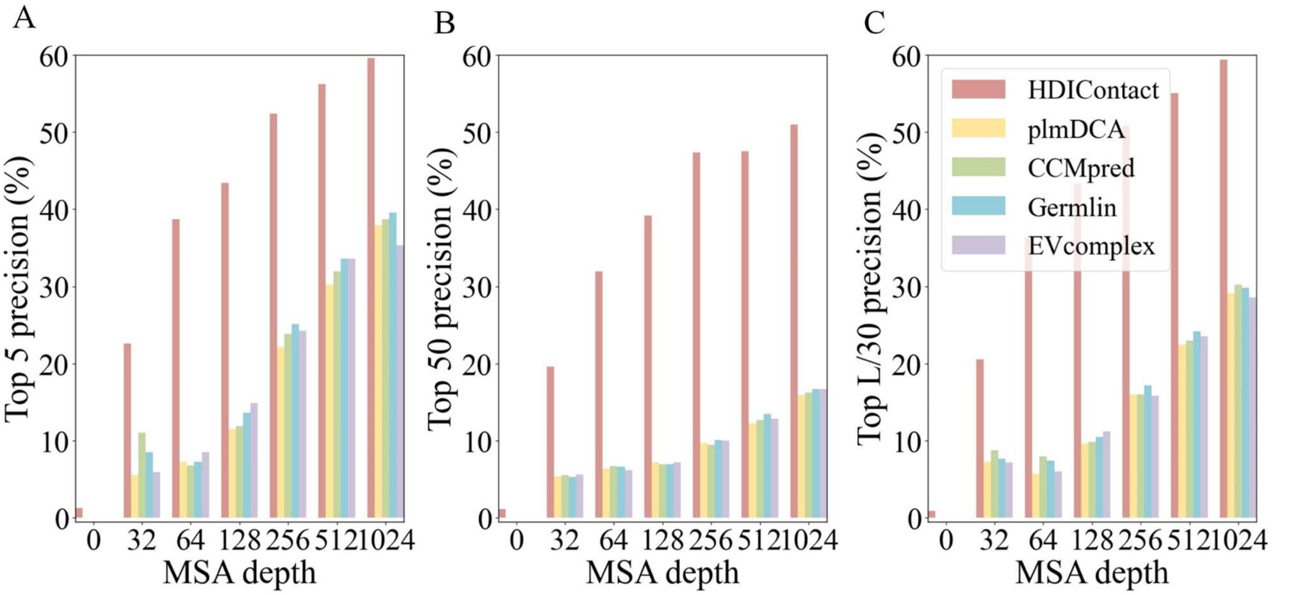

Figure 5. The impact of MSA. Precision of top 5 **(A)**, top 50 **(B)** and top *L*/30**(C)** predicted contacts by HDIContact, plmDCA, CCMpred, Germlin and EVcomplex with respect to MSA depth on E. coil test dataset.

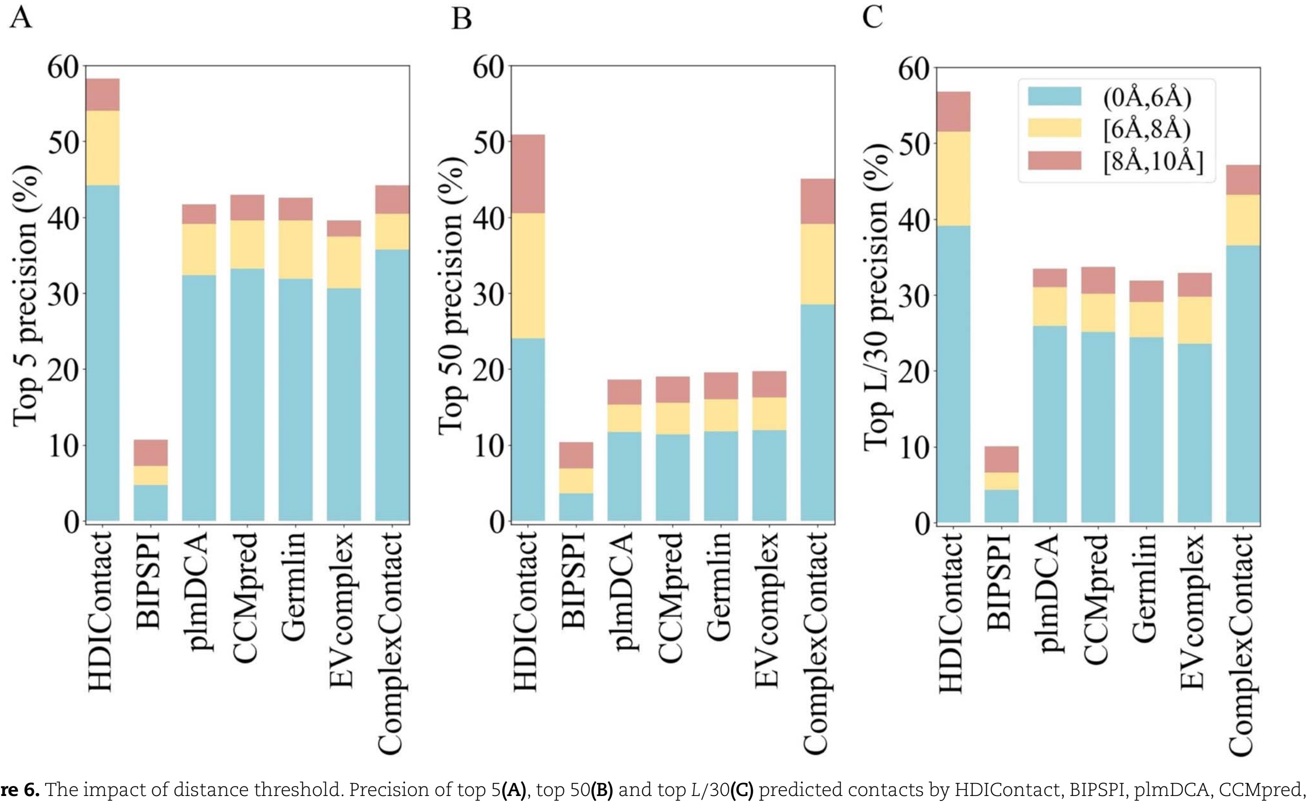

Figure 6. The impact of distance threshold. Precision of top 5**(A)**, top 50**(B)** and top *L*/30**(C)** predicted contacts by HDIContact, BIPSPI, plmDCA, CCMpred, Germlin, EVcomplex and ComplexContact in different bins of distance threshold on E. coil test dataset.

Impact of distance threshold

We determined whether a residues pair from two chains was considered to be in contact based on the minimum distance of their heavy atoms. The value of distance threshold would change the number of native inter-protein residue contacts, which was related to the difficulty of target prediction. For inter-protein contact residue pairs, according to the range of minimum distance between their atoms, they could be divided into three bins: (0Å, 6Å), [6Å, 8Å), [8Å, 10Å]. We calculated top precision of all methods for predicting inter-protein residue–residue contacts in different bins, and then evaluated the impact of the distance threshold on the performance of the method. Figure 6 gave the average precision

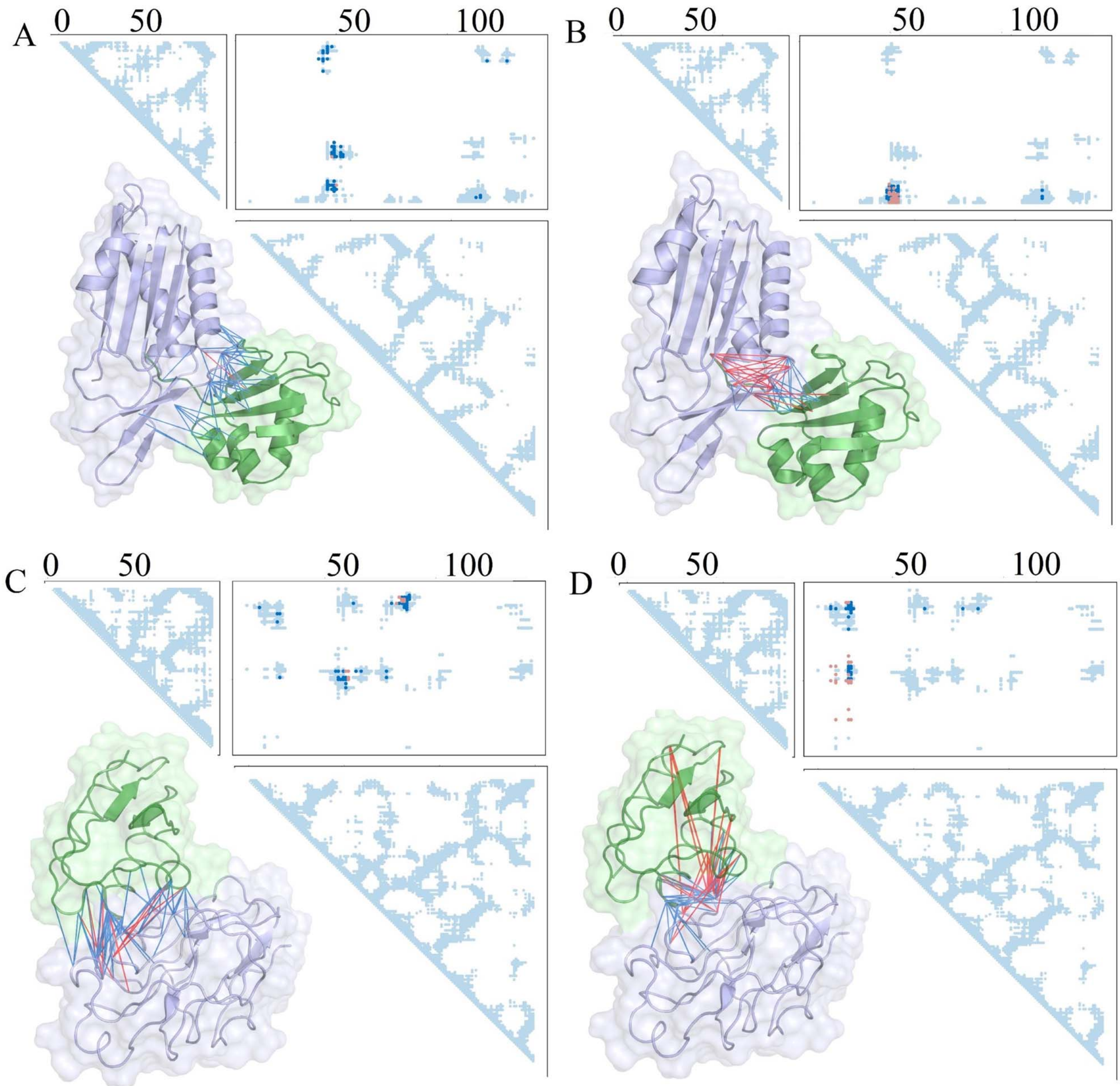

Figure 7. Top 50 predicted contacts by HDIContact (**A** and **C**) and BIPSPI (**B** and **D**) on D, E chains of 1FM0 **(**A and B) and B, A chains of 1GGR (C and D). Two monomers of native structures were shown in cartoons and colored in green and purple, respectively. The correct and incorrect predictions were shown in blue and red. The contact map showed native (skyblue), correctly (darkblue) and incorrectly (red) predicted contacts.

of our method for top 5, top 50 and top $L/30$ predicted contacts in the different bins of distance threshold. A general trend could been seen that the precision of all methods were improved with increased distance threshold. Target with larger distance threshold might have more inter-protein residue contacts, which might make it easier to predict the contacts. HDIContact still achieved better top precision with a more restricted threshold (6 Å), whose top precision was higher than the precision calculated by most methods based on the 10 Å distance threshold. It proved that our method could more accurately extract and integrate the co-evolution information contained in MSA.

Case study

Analysis on E. coil protein–protein complex

We selected D, E chains of 1FM0 and B, A chains of 1GGR as two examples of E. coil protein–protein complex. Figure 7 showed top 50 predicted contacts by HDIContact and BPISPI with the native contact map for these two targets. Comparisons of top precision, AUROC and AUPR between HDIContact and the other methods were shown in Supplementary Tables S2–5. Our method achieved higher precision of 96% and 80% on two targets of 1FM0 and 1GGR, while the precision of BIPSPI were only 50% and 58%, with 40% and 8% for plmDCA, 40% and 8% for CCMpred, 36% and 10% for Germlin, 40% and

10% for EVcomplex, 76% and 14% for ComplexContact, respectively. It could be seen that the contacts predicted by HDIContact were evenly distributed on the interface (Figure 7A/C), while most of the contacts predicted by BPISPI were concentrated in one area of the interface (Figure 7B/D). Specifically, the predicted contacts by HDIContact were mostly close to the native contacts on the contact map, even though they might not overlap. It meant that such near-native contacts by ours were still roughly correct, even though they might be classified to be incorrect contacts according the cutoff of 10 Å. HDIContact successfully predicted the contact between residues 78, 79 and residues 122, 124.

Analysis on human–virus protein–protein complex

To explore the potential of HDIContact on the human–virus protein–protein complex, a protein complex O75475-P04584 (PDB id: 3F9K) related to HIV was selected as the independent test case. O75475-P04584 was the complex formed by HIV-2 integrase and Lens epithelium derived growth factor (LEDGF). Human protein was LEDGF (chain R of 3F9K, O75475, green) from homo sapiens, as the binding partner of lentiviral integrase (IN) proteins. It is required for efficient viral replication [57]. Virus protein was integrase (chain E of 3F9K, P04584, purple) from Human immunodeficiency virus type 2. Unlike E. coil case, whose MSA could be concatenated according to their genome distance [31, 32], we concatenated MSA from monomers based on species for human–virus protein–protein complex. A single representative for each species was chosen based on the highest identity to the query sequence. By filtering, a nonredundant MSA with 334 depth, $<$ 90% identity, $>$ 75% coverage was obtained. As shown in Tables S6–7, all methods had a lower top precision, which showed the difficulty of prediction for O75475-P04584. It might be due to the low quality of MSA, which contained more mismatched sequences. HDIContact and ComplexContact were the only methods that had correctly predicted contacts in the specified number of predictions, using low-quality MSA. HDIContact achieved the precision of 80% for top five, 36% for top 50 and 55.56% for top $L/30$, while ComplexContact only achieved the precision of 40% for top five, 6% for top 50 and 33.33% for top $L/30$. Meanwhile, the AUPR and AUROC of HDIContact reached 15.28% and 88.35%, respectively. Figure 8 showed top five inter-protein residue–residue contacts predicted by HDIContact for O75475-P04584. It could be seen that although residue 3 GLU and residue 438 VAL were classified as incorrect contacts according the cutoff of 10 Å, it was still relatively close to native contacts (red). It proved that our method could give some insights for the study of human–virus interaction.

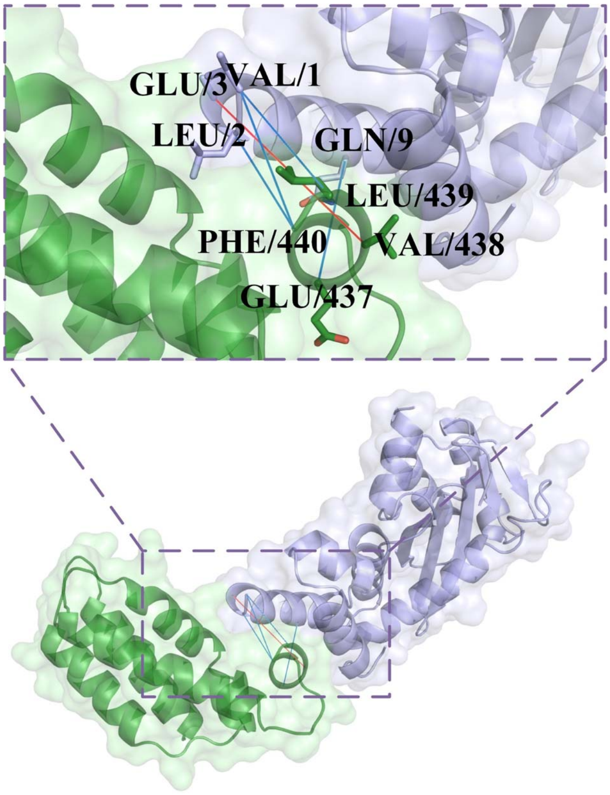

Figure 8. Top five predicted contacts by HDIContact on O75475-P04584. Two monomers of native structures were shown in cartoons and colored in green (human protein) and purple (virus protein), respectively. The correct and incorrect predictions were shown in blue and red.

Conclusion

In this study, we proposed HDIContact, a predictor of residue–residue contacts on hetero-dimer interfaces. On the one hand, it transferred the skill of extracting co-evolutionary patterns learned by pre-train protein language model to inter-protein contact prediction for producing MSA 2D embedding, which could reduce the influence of noise on concatenated MSA caused by mismatched sequences or less homology. On the other hand, it used two-channel mechanism to capture 2D context of residue pair on MSA 2D embedding from two different directions of receptor and ligand, and proved the effectiveness by comparing with other deep learning model architectures. We performed comparative assessment on E. coli test dataset with sufficient homologous sequences, and experiments show that HDIContact achieved the best performance for inter-protein residue contact prediction. In addition, we found that our method was more robust to different depth of MSA and distance thresholds. Also, we proved the potential of HDIContact for human–virus complexes, by achieving top precision on protein complex related to HIV. In the future work, we will integrate predicted contacts into protein–protein docking, in order to improve the accuracy of protein complexes 3D structure prediction.

Key Points

- We introduced a deep learning framework, HDIContact, which combined transfer learning strategy and two-channel mechanism to predict residue–residue contacts on hetero-dimer interfaces.

- HDIContact predicted inter-protein residue contacts from sequential information, which was a valuable solution faced with a sharp increase in sequential information relative to structural information.
- We used precision, AUROC and AUPR three criteria to compare with other state-of-the-art methods on E. coli test dataset, and analyzed the impact of MSA and distance threshold.
- HDIContact outperformed other state-of-the-art methods. And we explored the potential of HDIContact for human–virus protein–protein complexes, by selecting HIV-related protein complex as independent test case.

Supplementary information

Supplementary data are available at *Briefings in Bioinformatics* online.

Availability

The source code, trained model, the processed files of the datasets and corresponding results are available at https://github.com/guofei-tju/zw-tju-HDIContact.

Acknowledgements

Authors thank professor David Baker for kindly providing Baker dataset and professor Debora S. Marks for kindly providing E. coil dataset.

Funding

This work is supported by a grant from National Key R&D Program of China (2020YFA0908400) and National Natural Science Foundation of China (NSFC 62172296, 61972280).

References

1. Buxbaum E. *Fundamentals of protein structure and function*, Vol. **31**. New York: Springer, 2007.
2. Altman RB, Dugan JM. Defining bioinformatics and structural bioinformatics. *Structural Bioinformatics* 2003;**44**:3–14.
3. Fauman EB, Hopkins AL, Groom CR. Structural bioinformatics in drug discovery. *Methods Biochem Anal* 2003;**44**:477–97.
4. O'Connell MR, Gamsjaeger R, Mackay JP. The structural analysis of protein–protein interactions by NMR spectroscopy. *Proteomics* 2009;**9**(23):5224–32.
5. Shi Y. A glimpse of structural biology through X-ray crystallography. *Cell* 2014;**159**(5):995–1014.
6. Zhang W, Meng Q, Tang J, *et al.* Exploring effectiveness of ab-initio protein–protein docking methods on a novel antibacterial protein complex dataset. *Brief Bioinform* 2021;**22**(5):bbab038.
7. Moult J, Fidelis K, Kryshtafovych A, *et al.* Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins: Structure, Function, and Bioinformatics* 2018; **86**:7–15.
8. Huang SY. Exploring the potential of global protein–protein docking: an overview and critical assessment of current programs for automatic ab initio docking. *Drug Discov Today* 2015;**20**(8):969–77.
9. Huang SY. Search strategies and evaluation in protein–protein docking: principles, advances and challenges. *Drug Discov Today* 2014;**19**(8):1081–96.
10. Burley SK, Berman HM, Kleywegt GJ, *et al.* Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein Crystallography* 2017;**1607**:627–641.
11. Consortium EP, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**(7414):57.
12. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell* 2015;**58**(4):586–97.
13. Wuyun Q, Zheng W, Peng Z, *et al.* A large-scale comparative assessment of methods for residue–residue contact prediction. *Brief Bioinform* 2018;**19**(2):219–30.
14. Söding J. Big-data approaches to protein structure prediction. *Science* 2017;**355**(6322):248–9.
15. He B, Mortuza S, Wang Y, *et al.* NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* 2017;**33**(15):2296–306.
16. Yang J, Shen HB. MemBrain-contact 2.0: a new two-stage machine learning model for the prediction enhancement of transmembrane protein residue contacts in the full chain. *Bioinformatics* 2018;**34**(2):230–8.
17. Yu J, Andreani J, Ochsenbein F, *et al.* Lessons from (co-) evolution in the docking of proteins and peptides for CAPRI Rounds 28–35. *Proteins: Structure, Function, and Bioinformatics*. 2017;**85**(3):378–90.
18. Green AG, Elhabashy H, Brock KP, *et al.* Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat Commun* 2021;**12**(1):1–12.
19. Quadir F, Roy RS, Halfmann R, Cheng J. 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.
20. 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):1904–1910.
21. Yan Y, Huang SY. Accurate prediction of inter-protein residue–residue contacts for homo-oligomeric protein complexes. *Brief Bioinform* 2021;**22**(5):bbab038.
22. Baldassi C, Zamparo M, Feinauer C, *et al.* Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PloS one* 2014;**9**(3):e92721.
23. Weigt M, White RA, Szurmant H, *et al.* Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci* 2009;**106**(1):67–72.
24. 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.
25. 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.
26. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J Comput Phys* 2014;**276**:341–56.
27. Pereira J, Simpkin AJ, Hartmann MD, *et al.* High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics* 2021;**89**(12):1687–1699.

28. 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.
29. Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. Nature. *Machine Intelligence* 2021;**3**(7):1–9.
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. Hopf TA, Schärfe CP, Rodrigues JP, *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *elife* 2014;**3**:e03430.
32. Ovchinnikov S, Kamisetty H. Baker D. *Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. elife* 2014;**3**:e02030.
33. 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.
34. Tm Z, Wang S, Xu J. Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis. bioRxiv. 2018;240754.
35. Ekeberg M, Lövkvist C, Lan Y, *et al.* Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E* 2013;**87**(1):012707.
36. Szurmant H, Weigt M. Inter-residue, inter-protein and inter-family coevolution: bridging the scales. *Curr Opin Struct Biol* 2018;**50**:26–32.
37. Afsar Minhas FA, Geiss BJ, Ben-Hur A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins: Structure, Function, and Bioinformatics.* 2014;**82**(7): 1142–55.
38. Ahmad S, Mizuguchi K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PloS one* 2011;**6**(12):e29104.
39. 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.
40. 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.
41. Rao R, Meier J, Sercu T, *et al.* Transformer protein language models are unsupervised structure learners. In: *International Conference on Learning Representations*, 2020.
42. Vig J, Madani A, Varshney LR, *et al.* Bertology meets biology: Interpreting attention in protein language models. arXiv preprint arXiv:200615222. 2020.
43. Xie Z, Xu J. Deep graph learning of inter-protein contacts. *Bioinformatics* 2022;**38**(4):947–53.
44. Rao R, Liu J, Verkuil R, *et al.* Msa transformer. In: *International Conference on Machine Learning*. 2021; 8844–8856.
45. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: *Advances in neural information processing systems*, 2017, 5998–6008.
46. Child R, Gray S, Radford A, *et al.* Generating long sequences with sparse transformers. arXiv preprint arXiv:190410509. 2019.
47. Ho J, Kalchbrenner N, Weissenborn D, *et al.* Axial attention in multidimensional transformers. arXiv preprint arXiv:191212180. 2019.
48. Yoon BJ. Hidden Markov models and their applications in biological sequence analysis. *Curr Genomics* 2009;**10**(6):402–15.
49. 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.
50. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**(8):1735–80.
51. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005;**18**(5-6):602–10.
52. Haldane A, Levy RM. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Physical Review E* 2019;**99**(3):032405.
53. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:14126980* 2014.
54. Lin TY, Goyal P, Girshick R, *et al.* Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*, 2017, 2980–8.
55. Méndez R, Leplae R, De Maria L, *et al.* Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins: Structure, Function, and Bioinformatics.* 2003;**52**(1):51–67.
56. Hopf TA, Green AG, Schubert B, *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 2019;**35**(9):1582–4.
57. Hare S, Shun MC, Gupta SS, *et al.* A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLoS Pathog* 2009;**5**(1):e1000259.