JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

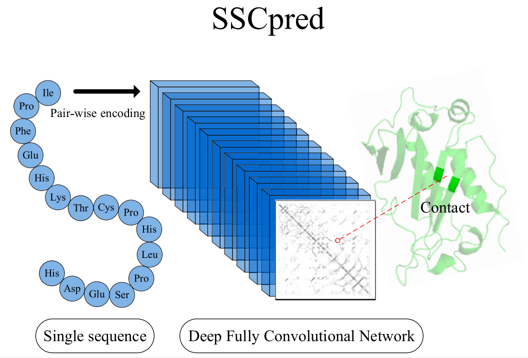
SSCpred: Single-Sequence-Based Protein Contact Prediction Using Deep Fully Convolutional Network

Ming-Cai Chen, Yang Li, Yi-Heng Zhu, Fang Ge, and Dong-Jun Yu*

Cite This: https://dx.doi.org/10.1021/acs.jcim.9b01207 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: There has been a significant improvement in protein residue contact prediction in recent years. Nevertheless, state-of-the-art methods still show deficiencies in the contact prediction of proteins with low-homology information. These top methods depend largely on statistical features that derived from homologous sequences, but previous studies, along with our analyses, show that they are insufficient for inferencing an accurate contact map for nonhomology protein targets. To compensate, we proposed a brand new single-sequence-based contact predictor (SSCpred) that performs prediction through the deep fully convolutional network (Deep FCN) with only the target sequence itself, i.e., without additional homology information. The proposed pipeline makes good use of the target sequence by utilizing the pair-wise encoding technique and Deep FCN. Experimental results demonstrated that SSCpred can produce accurate predictions based on the efficient pipeline. Compared with several most recent methods, SSCpred achieves completive performance on nonhomology targets. Overall, we explored the possibilities of single-sequence-based contact prediction and designed a novel pipeline without using a complex and redundant feature set. The proposed SSCpred can compensate for current methods' disadvantages and achieves better performance on the nonhomology targets. The web server of SSCpred is freely available at http://csbio.njust.edu.cn/bioinf/sscpred/
.



INTRODUCTION

Knowledge of protein structure is essential for various researches and applications such as biotechnology and medical sciences. It can be obtained by wet-lab techniques like X-ray crystallography,[1] nuclear magnetic resonance (NMR) spectroscopy,[2] and cryo-electron microscopy (cryoEM).[3] However, it is impracticable to solve a large number of protein structures experimentally due to the reason that these techniques are very expensive and time-consuming.[4] Up to now, according to the Universal Protein Resource (UniProt, http://www.uniprot.org/),[5,6] there are over 120 annotated million protein sequences but only about 150,000 structures available at the Protein Data Bank (PDB, http://www.rcsb.org/).[7] Moreover, the quantitative gap between sequences and structures still presents an expanding trend due to the emergence of the next-generation sequencing technology. Therefore, there is an urgent need for an efficient *de novo* structure prediction algorithm[8] (predict 3D structures from "scratch" using given protein sequences). During the past decades, significant progress has been made. Nevertheless, state-of-the-art methods are still of widely varying accuracy because of protein 3D structures' complexities.[4,9−13]

Residue contact information provides the global topology constraints of protein structural folds and so could act as a bridge between primary sequences and the 3D structures. A lot of previous research[14−16] has shown that this constraint can provide valuable information to assist structure prediction. According to the assessment of contact predictions in the 12th Critical Assessment of Protein Structure Prediction (CASP12),[16] adding contact prediction results to structure prediction brings up to 26% improvement on the global distance test-total score.

Early studies on protein contact prediction are based on the fact that spatially close residues tend to present a correlated mutation pattern in evolution.[17−19] Such evolutionary information can be exploited from homologous sequences using statistical techniques. There are several types of statistical algorithms to further analyze residue coevolution: statistical coupling analysis,[20] inverse covariance estimation based on direct coupling analysis (DCA),[21,22] and pseudo-likelihood maximization based on DCA.[23] These algorithms mainly try to disentangle direct from indirect residue coevolution, known as the indirect coupling problem, in Multiple Sequence Alignment (MSA). Classic statistical methods including PSICOV,[21]

Received: December 31, 2019
Published: April 27, 2020

CCMpred,[23] and Gremlin[24] solved this problem to some extent, and they gained good results with the use of advanced statistical algorithms and mounting databases.

Machine learning algorithms enormously improve the accuracy of protein contact prediction.[25,26] Unlike statistical methods, most of which show poor performance when targets do not have enough homologous sequences in the database, machine learning algorithms make up for this weakness, such as some early stage machine learning methods: SVMSEQ[25] operates on local window features with a support vector machine (SVM). MetaPSICOV is a meta-predictor that integrates several machine learning-based contact prediction methods with a shallow neural network. This kind of method is inefficient because they both use a local sliding window and predict contacts one by one. Furthermore, SVM and the shallow neural network are obviously not capable of modeling complex interactions between residues.

Recently, explosive data growth and the development of deep learning over the last decades inspire us to rethink the use of deep architectures. As a representative method, RaptorX-Contact[27] integrates kinds of evolutionary information and sequence information utilizing Deep FCN. In similar ways, some exciting progress had been achieved: the average accuracy of CASP12 was increased to almost two times against CASP11.[16] Most recently, the raw covariance matrix[28] calculated from alignments was found to be the most effective feature: DeepCov[29] achieves competitive results only with it, arguing that to accurately predict contact may not need to rely on additional information. ResPRE,[30] alternatively, uses the maximum likelihood approach to estimate the ridge regularized inverse covariance matrix which can address the issue of transitional noise. TripletRes[31] combines three profile matrices, which build on covariance, precision, and pseudo-likelihood maximization, to make good use of MSA. According to the most recent assessment of CASP13,[32] even the best method at CASP12 would have only rated as average in the CASP13 groups. Presently, we can obtain fairly accurate prediction results to use these methods with high-quality alignments.

However, about one-third of the 15,000 protein families lack homologues with known structures,[33] and more than a 90% sequence have few known homologous sequences.[33,34] If there is only a shallow alignment, the accuracy of these state-of-the-art methods drops to a rather low level. We analyzed the overall performance of all predictors in CASP12[16] and found that their predictions of low-homology targets are very inaccurate. A similar trend was found when we tested three state-of-the-art methods. Results are shown in Figure S1 in the Supporting Information.

To improve the accuracy of proteins with low-homology information, we developed a Single-Sequence-based Contact Predictor (SSCpred) which is solely based on single-sequence information. Experiments showed that SSCpred can provide more accurate contact-maps compared to the controlled methods when only shallow alignments are available. The proposed method is robust to the quality of the alignment, and it gets rid of a complex and redundant feature set. To the best of our knowledge, none of the existing methods are specifically optimized for targets with few sequence homologues. Furthermore, the fact that competitive results can be achieved by SSCpred should motivate further researches to continue to explore the possibilities for single-sequence-based *de novo* protein structure prediction.

MATERIALS AND METHODS

Data Sets. To strictly evaluate the proposed method, we constructed three different types of test data sets. The first test data set, denoted as PDB1000, consists of 1000 proteins which are randomly selected from PDB. The sequence identity between any two proteins in PDB1000 is less than 25% according to CD-HIT.[35] The second test data set, denoted as HARD90, contains 90 selected hard targets. A sequence with a value of N_{eff} (number of effective sequences in MSA) less than 1 is considered as a hard target, where N_{eff} is defined in eq 1. The third test data set, denoted as CASP67, is made up of 67 CASP free-modeling (FM) targets from CASP11 and CASP12.

The construction of our training data sets is based on two primary principles: (1) keep the sequence identity between test data sets and corresponding training data sets below 25% and (2) make the training data sets as big as possible to have ample samples for training a robust prediction model. In view of this, we first downloaded all structures and sequences deposited in PDB before April 2019; then, those proteins with lengths longer than 400 or shorter than 40 were removed. The sequence identity of the remaining proteins were culled less than 70% using CD-HIT,[35] and the obtained proteins constitute the initial training data set. Next, to reduce the homology bias, we removed redundant sequences from the initial training data set for the three test data sets separately, resulting in three corresponding training data sets. More specifically, the sequence identity between a test data set and the corresponding training set was culled less than 25%, i.e., any sequence in the initial training data set will be removed if it has a sequence identity with any sequence in the corresponding test data set higher than 25%. Finally, for PDB1000, HARD90, and CASP67, three corresponding training data sets with 15581, 24862, and 24673 sequences, respectively, were obtained. Note that for each training data set, we used the first 1000 sequences as a validation subset during the model training stage. A detailed PDB list of three test data sets and three corresponding data sets are available at http://csbio.njust.edu.cn/bioinf/sscpred/.

Evaluation Indices. In this study, two residues are considered as in contact if the spatial distance between their C_β atoms (C_α in case of glycine) is less than 8 Å. A contact between two residues can be classified into long range, medium range, or short range if their sequence separation is at least 24 residues, between 12 and 23, and between 7 and 11, respectively. Considering that long-range contact is more critical for the subsequent contact-assisted protein structure prediction,[27,36] we mainly focus on the accuracy of long-range predictions, i.e., the percentage of true positives in Top-L/K (K can be 5, 2, 1, 1/2) long-range predictions. MSA is an aligned sequence set consisting of the homologous sequences found from databases by search tools like PSI-BLAST[37] or HHblits.[38] The number of effective sequences measures the effective information that contains in MSA

$$N_{eff} = \frac{1}{\sqrt{L}} \cdot \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m \neq n}^{N} \mathbb{I}[S_{m,n} \geq 0.8]} \tag{1}$$

where L is the length of the target protein. $S_{m,n}$ is the sequence identity between the m-th aligned sequence and n-th aligned sequence. $\mathbb{I}$ is the indicator function, i.e., equals to 1 if $S_{m,n} \geq 0.8$; otherwise 0. Because N_{eff} has a large range of quantities (as low as 0.1 to as high as 1000), we mainly use its logarithmic scale in this article.

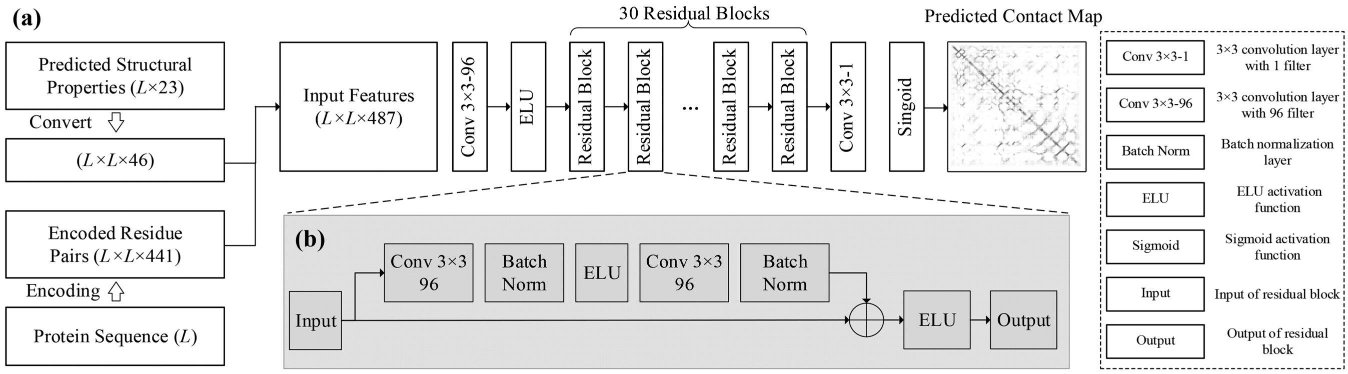


Figure 1. Illustration of SSCpred: (a) network structure and (b) residual block.

In this study, we also measure the diversity of contact maps by Entropy Score (ES):

$$ES = \frac{E|0 - E|C}{E|0} \times 100\% \tag{2}$$

where E represents the entropy in a certain state, and $E|0$ and $E|C$ is the entropy with and without contact constraints respectively:

$$E|X = \frac{1}{\#\,\text{all pairs}} \sum_{i,i \neq j} \ln(U_{ij} - L_{ij}) \tag{3}$$

where U_{ij} and L_{ij} are the lower and upper bounds of the distances between residues. In the calculation, $U_{ij} = 3.8 \times |i - j|$, if residue i and j are not in contact; $U_{ij} = 8$, if they are in contact. $L_{ij} = 3.2$ for all pairs (Only correctly predicted contacts in Top-L/K (K can be 5, 2, 1, 1/2) are taken into consideration).

Feature Representation. SSCpred aims at predicting residue contact without additional evolutionary features. On the basis of this principle, the input feature set of SSCpred only contains **information derived from the single-sequence itself**. Each residue pair is encoded to a one-hot feature vector. There are 20 types of amino acids and unknown types considered as an additional category, so each residue pair is encoded into a 441-dimensional feature vector. The element that corresponds to the specific residue pair type (totally 21 × 21 = 441 types) in the feature vector is set to 1, and other elements are set to 0, e.g., residue pair AA encoded as [1,0, ...,0].

Besides the sequence information, we also use the output of a most recent structure properties prediction method SPIDER3-single,[34] which also only takes a single sequence as input features. Its prediction includes a three-state secondary structure ($L \times 3$), eight-state secondary structure ($L \times 8$), accessible surface area ($L \times 1$), main-chain angles ($L \times 8$), half-sphere exposure ($L \times 2$), and contact number ($L \times 1$). Accordingly, each sequence of length L can be encoded into an L by 23 vector. We then construct a 46-dimensional feature vector of a residue pair by concatenating the corresponding 23-dimensional feature vectors of the two residues.

Finally, we construct a 487-dimensional feature vector of each residue pair by further combining its 441-dimensional one-hot feature vector and 46-dimensional SPIDER3-single predicted feature vector. Accordingly, for a protein sequence with length L, we obtain a corresponding symmetric feature matrix of size of $L \times L$, among which each element at (i, j) is the 487-dimensional feature vector of the reside pair (i, j). The feature matrix will be used as the input of the subsequent neural network.

Pipeline and Network Architecture. The deep fully convolutional network (Deep FCN) is a special convolutional neural network which is very suitable for contact prediction. It can integrate different kinds of input information and directly output the likelihood of two residues being in contact in an end-to-end fashion. But due to the gradient vanishing problem, training of the deep network used to be difficult. The residual connection[39] provides a solution to the problem by skipping over layers and amplifying the output of previous layers.

On the basis of Deep FCN with residual connections, we designed our pipeline, as shown in Figure 1(a): an encoded one-hot feature vector of a residue pair ($L \times L \times 441$) and pair-wise predicted structure properties ($L \times L \times 46$) are combined as the input feature ($L \times L \times 487$). The network gives prediction results as a $L \times L$ matrix, among which each element represents the likelihood of a pair being in contact in the corresponding position.

We tuned the network's hyper-parameters, mainly including filter sizes in each layer, depth of the network, and scale factor of L2 regularization according to its performance on the validation data set (the first 1000 sequences in each training data set). In this study, we tuned hyper-parameters one by one, i.e., other hyper-parameters are fixed when one hyper-parameter was gradually changed and tested. By doing so, we found a group of parameters that would achieve relatively high accuracy and keep the model memory efficient, as shown in Figure 1(a): all the intermediate convolution layers have 96 filters with a 3 × 3 kernel size and use the exponential linear unit (ELU) activation function. The final output layer has one filter, followed by the sigmoid activation function. All the convolution layers use the "same padding".

During the training process, we calculated the loss by the cross-entropy function and optimized the model by the Adam optimizer. Furthermore, we applied L2 regularization to prevent overfitting, i.e., added the scaled sum of the squared values of all weights to the loss function. Besides regularization, we found that dropout as another technique to prevent overfitting is only of marginal help. It might be caused by the disharmony between batch normalization and dropout.[40] Furthermore, the ReLU (rectified linear unit) activation function was replaced with the ELU activation function, because the latter one tends to be more effective in the deep architecture.[41]

The model, which was implemented by TensorFlow for Python, does not take too much time or computing resources for training and prediction on a single Nvidia GTX TITAN X

https://dx.doi.org/10.1021/acs.jcim.9b01207

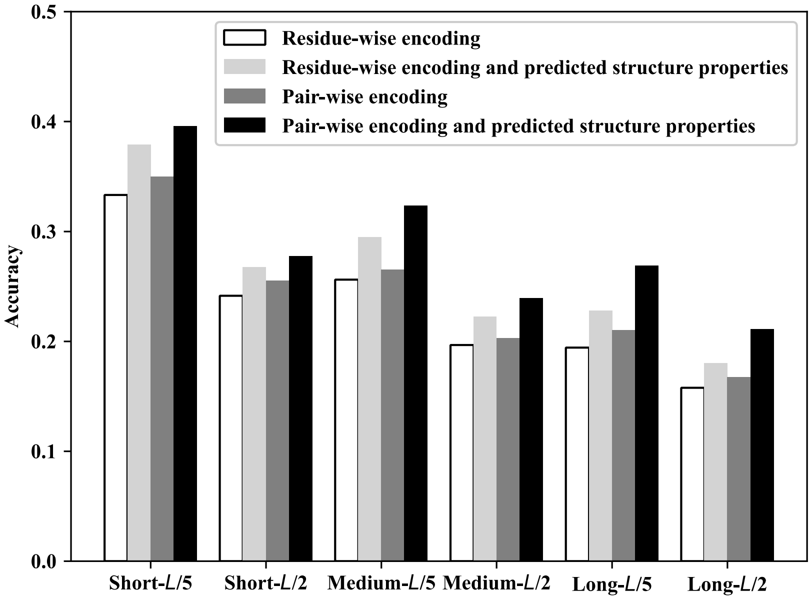


Figure 2. Accuracy of SSCpred trained with different encoding methods and with or without predicted structure properties.

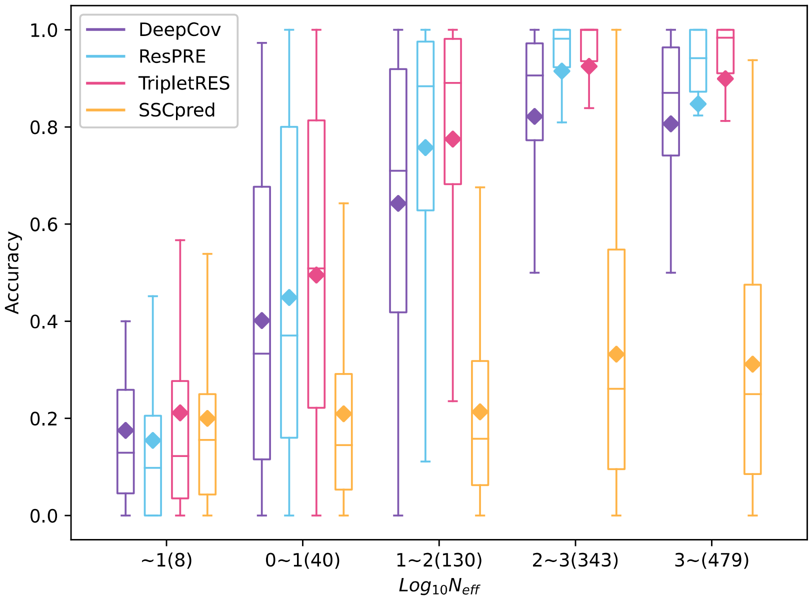


Figure 3. Box plot of Top-$L/2$ Accuracy of four predictors on PDB1000 data set with X-axis (logarithmic scale N_{eff}) binned by 1. Minimum, maximum, first quartile, third quartile, and mean (diamond) accuracy are shown for each bin. The number in parentheses refers to the number of targets in each bin.

graphics card: the network took roughly 18 hours, 15 epochs over the training data set to converge. For prediction, it only needs an average of 2 minutes per target including time for generating features.

■ RESULTS AND DISCUSSION

Comparison of Feature Sets. *Difference Brought by Encoding Methods.* The most common method to encode a protein sequence is residue-wise one-hot encoding, which will produce a 21-dimensional vector for each residue (20 amino acids and unknown types considered as an additional category). It can then be converted to pair-wise which results in 21 + 21 = 42 elements for each pair, or we can use pair-wise one-hot encoding, which results in 21 × 21 = 441 elements for each pair.

To study the contribution of the encoding methods toward the performance of SSCpred, we first compared the accuracy achieved by our model with two encoding methods, respectively. As the gray and black bars in Figure 2 show, the latter slightly performs better on the validation data set, with Top-$L/2$ long-range accuracy of 0.6% higher and Top-$L/5$ long-range accuracy 1.0% higher.

Why have different encoding methods resulted in different performances? We speculated that it is the problem called "multicollinearity", which happens when two or more elements in the feature vector are correlated with each other. For example, the feature vectors of residual pairs AC and AF have the same first half. The problem is known to affect prediction accuracy adversely. So, we should not treat the A (alanine) in AC and AF in the same way here. In other words, considering them as totally different pairs, instead of independent residues, may help the model gain better performance.

In addition to one-hot encoding, there are other encoding methods like the AAINDEX descriptor[42] or BLOSUM descriptor,[43] but we found they make no obvious difference with residue-wise one-hot encoding. One explanation is that such a linear relationship can be easily learned by a neural network.

Predicted Structure Properties as Additional Features. Apart from the sequence information, we also used structure properties prediction results from SPIDER3-Single as additional input features.

In further comparison experiments, after combining sequence information and SPIDER3-Single's result together, better results were yielded. The average accuracy of Top $L/5$ contact predictions reached 22.7% and 26.8%, and the average accuracy of Top $L/2$ contact predictions reached 18.0% and 21.1% with two encoding methods, respectively. The experimental results show that the improvement came from the predicted structure properties and the pair-wise encoding.

Comparison of SSCpred with Other Predictors. Existing contact prediction pipelines usually use PSI-BLAST,[37] HHblits,[38] or other sequence database search methods to generate MSA. Some of them produce inaccurate or even abnormal results on shallow alignments, and some use different homology search tools to operate on different databases, which is particularly common among meta-predictors. For these reasons, it was hard for us to consistently quantify the homology information used in their pipeline. So we chose to compare SSCpred against three methods: TripletRes,[31] DeepCov,[29] and ResPRE.[30] They all directly take MSA as the only input feature and provide standalone packages, so we could get stable large-scale prediction results and make a relatively fair comparison. In this study, they all use identical alignment generated from HHblits version 3.0.3 with default parameters to search on the Uniprot20_2016_02 database.

Comparison on PDB1000 Data Set. Considering the large size of our training data sets, it is necessary to evaluate SSCpred's performance on the same scale. So we first performed experiments on 1000 PDB targets. Figure 3 shows the performance of SSCpred on the data set and comparison with DeepCov, ResPRE, and TripletRes. The average Top-$L/5$ long-range and Top-$L/2$ long-range accuracy of SSCpred on this data set are 29.2% and 23.4%, respectively.

In general, the accuracy of all the other three methods varies greatly with the change of N_{eff}, indicating that their performance highly relies on the quality of alignments. For SSCpred, ideally, prediction accuracy should be irrelevant to N_{eff}, but it seems that there is still a correlation between the two. One explanation is that sequences with higher N_{eff} tend to share higher identity with sequences in our training data set (even though no two sequences share more than 25% sequence identity between the training data set and test data set).

In this PDB1000 data set, only 5% of targets can be classified as hard targets according to our standard, which results in a big gap between SCCpred and the other three methods. One might argue that our method is of limited practical value. But it should be noticed that *de novo* protein structure prediction mainly focuses on FM targets, which usually lack homologous sequences. The distribution of N_{eff} in CASP FM targets, instead of a randomly selected PDB subset, is closer to the practical situation: 14 out of 67 targets can be defined as hard targets in our CASP67 data set, i.e., a significant part of sequences that users will input to SSCpred can be regarded as hard targets. The prediction of low homology targets is the weakness for most contact prediction methods, and inaccurate contact prediction results may even have negative impacts on further structure prediction instead of assisting it.[44] Therefore, reliable contact maps for hard targets are essential to improve the overall performance of *de novo* protein structure prediction.

Comparison of HARD90 Data Set. Table 1 shows the performance of SSCpred on the HARD90 data set in control with three methods. For the 90 sequences, SSCpred's average accuracy of Top-$L/5$ contact predictions is 20.2%, exceeding that of DeepCov, ResPRE, and TripletRes by 5.3%, 3.8%, and 0.3%, respectively. The average Top-$L/2$ long-range accuracy is 16.2%, exceeding that of DeepCov, ResPRE, and TripletRes by 4.2%, 3.6%, and 1.0%, respectively. SSCpred achieves competitive results on the HARD90 data set with the minimum amount of sequence information.

Table 1. Average Accuracy by Four Predictors on HARD90 Data Set

	Short		Medium		Long	
Methods	$L/5$	$L/2$	$L/5$	$L/2$	$L/5$	$L/2$
DeepCov	0.201	0.160	0.182	0.147	0.149	0.119
ResPRE	0.329	0.230	0.270	0.200	0.164	0.126
TripletRes	0.322	0.234	0.260	0.192	0.199	0.152
SSCpred	0.369	0.261	0.271	0.212	0.202	0.162

In Figure 4, we show a head-to-head comparison between three predictors on each target in the HARD90 data set. The sizes of dots are related to N_{eff}. The higher the N_{eff}, the bigger the size. Smaller dots are scattered more at the bottom right part in Figure 4(a) and (b), which means that SSCpred has the unique advantage of predicting targets that are hard to predict for other homology-based methods. Especially, SSCpred achieves high accuracy on several targets represented as dots near the right border in Figure 4(a) and (b), indicating that even without homology information accurate predictions can be produced by the proposed method.

Only accuracy is not sufficient evidence that the prediction results of SSCpred can effectively assist protein 3D structure prediction, so we further calculated their Entropy Score (ES). As shown in Table 2, the prediction results of SSCpred present a higher diversity. It means that the proposed method has applicative value in assisting structure prediction.

It can be noticed that the dots in Figure 4(a) and (b) are more dispersed, suggesting that SSCpred can generate diversified results. So here we further calculate the average difference of long-range contact prediction probabilities between all predictors (to eliminate the bias, the probabilities are rescaled using min−max normalization). The difference between any two out of the three predictors is about 0.07. However, the difference between SSCpred and other predictors is larger than 0.15. Besides, considering that SSCpred makes predictions using a different pipeline, we believe SSCpred is suitable to be combined with other complementary methods to further improve the overall performance of structure prediction.

Comparison of CASP67 Data Set and a Case Study. We evaluated the performance of SSCpred on the publicly available CASP11 and CASP12 FM targets. In Table 3, the

https://dx.doi.org/10.1021/acs.jcim.9b01207

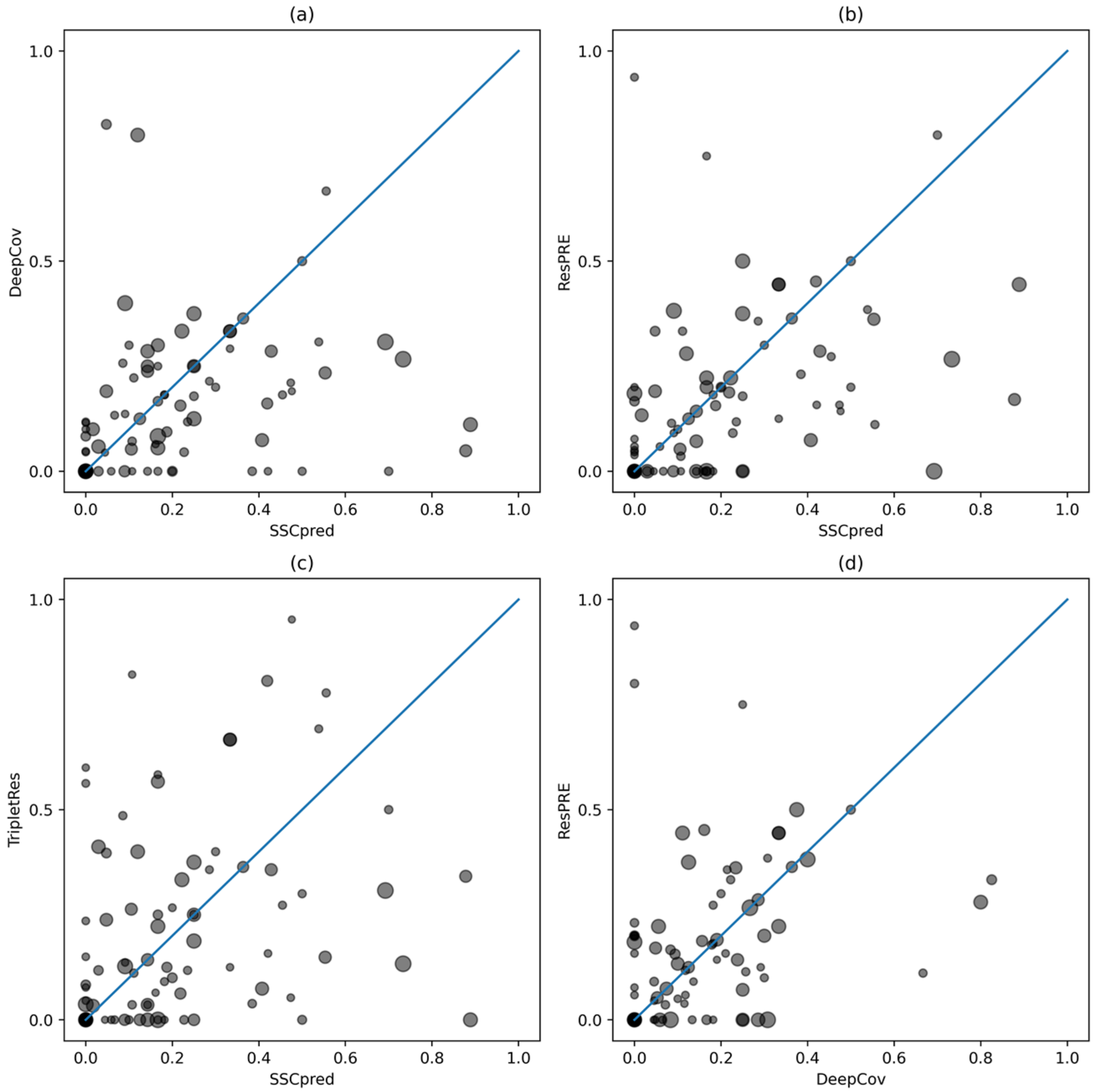


Figure 4. Comparison of Top-$L/2$ long-range precision by different predictors on the HARD90 data set: (a) SSCpred vs DeepCov, (b) SSCpred vs ResPRE, (c) SSCpred vs TripletRes, and (d) DeepCov vs ResPRE. The size of each dot is related to N_{eff}. The higher the N_{eff} value is, the larger the dot is.

Table 2. ES by Four Predictors on HARD90 Data Set

	Short		Medium		Long	
Methods	$L/5$	$L/2$	$L/5$	$L/2$	$L/5$	$L/2$
DeepCov	0.0068	0.0118	0.0043	0.0078	0.0021	0.0037
ResPRE	0.0042	0.0083	0.0029	0.0058	0.0021	0.0037
TripletRes	0.0065	0.0118	0.0040	0.0073	0.0024	0.0042
SSCpred	0.0070	0.0128	0.0042	0.0079	0.0027	0.0049

Table 3. Performance of SSCpred on CASP67 Data Set

	$\text{Log}_{10}N_{eff}$	Long-range $L/5$	Long-range $L/2$
Average	1.059	0.282	0.214
Median	1.116	0.217	0.180
Maximum	3.243	0.833	0.784
Minimum	−1.188	0.000	0.000

average accuracy of top $L/5$ and $L/2$ long-range contact predictions are 28.2% and 21.4%, respectively. Among these predictions, we selected part of relatively accurate predictions of SSCpred as shown in Table 4. All detailed results are listed in Table S1 in the Supporting Information.

To obtain an estimation of the average $\log_{10}N_{eff}$ values of targets in the PDB database, we calculate it for over 5000 randomly selected targets in the PDB database. The search

Table 4. Part of Predictions by Four Predictors on CASP11 and CASP12 Targets

	DeepCov		ResPRE		TripletRes		SSCpred	
Target	Top-*L*/5	Top-*L*/2	Top-*L*/5	Top-*L*/2	Top-*L*/5	Top-*L*/2	Top-*L*/5	Top-*L*/2
T0761	0.116	0.065	0.047	0.037	0.279	0.168	0.093	0.084
T0763	0.077	0.108	0.000	0.015	0.231	0.185	0.308	0.215
T0771	0.083	0.146	0.278	0.146	0.194	0.157	0.361	0.281
T0777	0.087	0.104	0.145	0.133	0.174	0.127	0.174	0.104
T0781	0.052	0.068	0.065	0.058	0.117	0.089	0.182	0.115
T0785	0.217	0.214	0.044	0.071	0.000	0.018	0.217	0.250
T0820	0.037	0.03	0.000	0.015	0.000	0.000	0.111	0.045
T0832	0.119	0.086	0.000	0.019	0.000	0.000	0.024	0.038
T0859	0.087	0.035	0.000	0.070	0.348	0.246	0.652	0.491
T0860	0.036	0.029	0.536	0.294	0.000	0.015	0.750	0.559
T0862	0.368	0.340	0.263	0.255	0.579	0.383	0.421	0.255
T0867	0.286	0.250	0.048	0.096	0.000	0.096	0.095	0.115
T0880	0.103	0.072	0.103	0.072	0.000	0.062	0.564	0.412
T0896	0.044	0.036	0.033	0.022	0.011	0.045	0.100	0.080
T0897	0.038	0.038	0.132	0.122	0.094	0.084	0.453	0.214
T0898	0.152	0.148	0.273	0.161	0.121	0.136	0.364	0.210
T0900	0.143	0.196	0.286	0.216	0.048	0.059	0.762	0.784
T0904	0.095	0.071	0.048	0.090	0.127	0.083	0.159	0.147
T0928	0.391	0.322	0.319	0.292	0.188	0.304	0.754	0.374
T0941	0.015	0.035	0.029	0.029	0.014	0.018	0.044	0.035

method used here is the same as we used for the generation of MSA, i.e., using HHblits version 3.0.3 with default parameters to search on the Uniprot20_2016_02 database. In the next, the average $\log_{10}N_{eff}$ value of a randomly chosen PDB target is calculated as 1.696, significantly higher than 1.059 as shown in Table3. It leads to the conclusion that a contact prediction method's performance may be overestimated when evaluated on the PDB data set. The difference between the predictors' performance on the PDB1000 data set and on the HARD90 data set or CASP67 data set also appears to provide further evidence. Considering the selection method of CASP FM targets, the distribution of N_{eff} in the CASP FM data set can reflect the practical situation to a large extent. As shown in Table 3, our CASP67 data set contains a large proportion of hard targets, with average $\log_{10}N_{eff}$ value 1.059 and median $\log_{10}N_{eff}$ value 1.116.

To intuitively present the difference between SSCpred and the other three, here we use a CASP12 target as an example and show four methods' predictions: T0880 with length 193 residues categorized as FM targets, as shown in Figure 5(a). HHblits[38] finds that no sequence has more than 25% identity with it on Uniprot20_2016_02, so TripletRes, DeepCov, and ResPRE take an alignment containing only one sequence as input and have 6.2%, 7.2%, and 8.2% Top-*L*/2 long-range accuracy, respectively, which are merely better than random guesses. SSCpred, on the contrary, achieves 42.2% on this one. As we can see in the lower triangle, not only does SSCpred perform accurate prediction but also the predicted pairs cover the entire sequence. Such contact prediction can provide accurate and diverse constraints along the sequence.

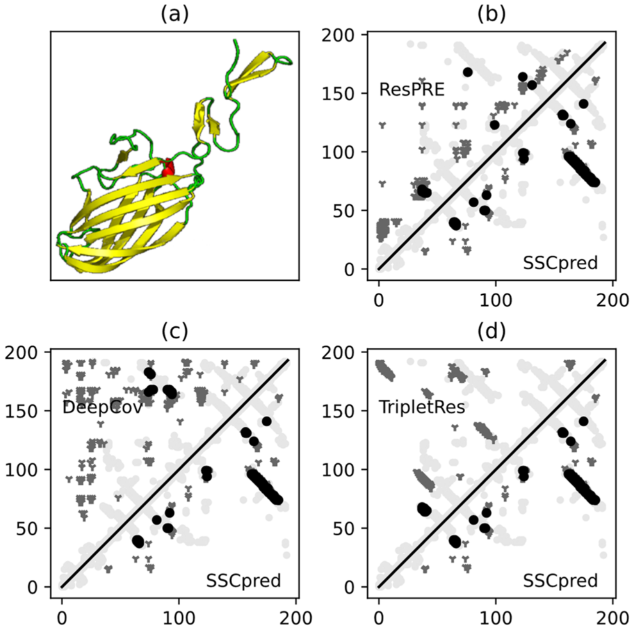


Figure 5. Target T0880: (a) structure of T0880, (b) predictions of SSCpred and ResPRE, (c) predictions of SSCpred and DeepCov, (c) predictions of SSCpred and DeepCov, and (d) predictions of SSCpred and TripletRes. In panels (b), (c), and (d), gray triangles represent wrong predictions, black dots right predictions, and light-gray dots ground-truth contacts.

CONCLUSIONS

We developed a novel pipeline, SSCpred, for homology information-independent contact prediction. From a query sequence, pair-wise encoding is used to keep sequence information in an integral state, and we then utilize Deep FCN to learn the relationship between sequence and structure. The main uniqueness of SSCpred is that it no longer struggles for extracting statistical features from few homologous sequences. Instead, it focuses on exploring the hidden information in the single target sequence. By doing so, better performance and robustness are achieved for hard targets. At the same time, SSCpred avoids a redundant feature set and the time-consuming process for database search.

In this study, the experiments on the large PDB database show that SSCpred is capable of generating stable results on a large scale with average Top-*L*/5 long-range accuracy reaching

25.9%. The experiments on hard targets show that SSCpred outperforms other most state-of-the-art methods on proteins with low-homology information. The encoding methods analysis demonstrates that the pair-wise encoding technique is more favorable to residue-wise encoding. Moreover, the use of predicted structure properties improves the performance of the model too.

Beyond the usefulness of SSCpred, there are still problems that need to be addressed. For example, the pipeline still contains an intermediate step to obtain extra structure properties using SPIDER3-single. Possible future work would be using multitask learning to predict the contact map and secondary structure properties at the same time. Also, due to the limitation of time and computing resources, hyperparameters of SSCpred are determined roughly based on limited experiments, so there certainly is still room for improvement.

Over the past several years, there has been a big step forward in contact prediction; however, the prediction results of proteins with low-homology information are still unsatisfactory. The fact that SSCpred achieved better performance with less information raises a question: is focusing on MSA the only effective way to inference residues' contacts? SSCpred gives another possibility, and the studies along these lines are under progress.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.9b01207.

Information as mentioned in the text. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Dong-Jun Yu − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P. R. China*; orcid.org/0000-0002-6786-8053;
Email: njyudj@njust.edu.cn

Authors

Ming-Cai Chen − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P. R. China*

Yang Li − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P. R. China; Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109-2218, United States*

Yi-Heng Zhu − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P. R. China*

Fang Ge − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P. R. China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.9b01207

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61772273, 61373062, and 61876072) and the Natural Science Foundation of Anhui Province of China (KJ2018A0572).

■ ABBREVIATIONS

NMR, Nuclear Magnetic Resonance; CASP12, 12th Critical Assessment of protein Structure Prediction; MSA, Multiple Sequence Alignments; FCN, Fully Convolutional Network; FM, Free-Modeling; N_{eff}, Number of effective sequences; ELU, Exponential Linear Unit.

■ REFERENCES

(1) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by x-Ray Analysis. *Nature* **1958**, *181* (4610), 662−666.

(2) Kühlbrandt, W. The Resolution Revolution. *Science* **2014**, *343* (6178), 1443−1444.

(3) Abola, E. E.; Sussman, J. L.; Prilusky, J.; Manning, N. O. Protein Data Bank Archives of Three-Dimensional Macromolecular Structures. *Methods Enzymol.* **1997**, *277*, 556−571.

(4) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical Assessment of Methods of Protein Structure Prediction: Progress and New Directions in Round XI. *Proteins: Struct., Funct., Genet.* **2016**, *84* (April), 4−14.

(5) Bateman, A.; Martin, M. J.; O'Donovan, C.; Magrane, M.; Alpi, E.; Antunes, R.; Bely, B.; Bingley, M.; Bonilla, C.; Britto, R.; Bursteinas, B.; Bye-AJee, H.; Cowley, A.; Da Silva, A.; De Giorgi, M.; Dogan, T.; Fazzini, F.; Castro, L. G.; et al. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45* (D1), D158−D169.

(6) Bateman, A. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506−D515.

(7) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Costanzo, L. Di; Christie, C.; Duarte, J. M.; Dutta, S.; Feng, Z.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Guzenko, D.; Hudson, B. P.; Liang, Y.; Lowe, R.; et al. Protein Data Bank: The Single Global Archive for 3D Macromolecular Structure Data. *Nucleic Acids Res.* **2019**, *47* (D1), D520−D528.

(8) Floudas, C. A.; Fung, H. K.; McAllister, S. R.; Mönnigmann, M.; Rajgaria, R. Advances in Protein Structure Prediction and de Novo Protein Design: A Review. *Chem. Eng. Sci.* **2006**, *61* (3), 966−988.

(9) Lee, J.; Freddolino, P. L.; Zhang, Y. Ab Initio Protein Structure Prediction. In *From Protein Structure to Function with Bioinformatics*, Second ed.; Springer: Dordrecht, The Netherlands, 2017; pp 3−35.

(10) Kryshtafovych, A.; Fidelis, K. Protein Structure Prediction and Model Quality Assessment. *Drug Discovery Today* **2009**, *14* (7−8), 386−393.

(11) Prentiss, M. C.; Hardin, C.; Eastwood, M. P.; Zong, C.; Wolynes, P. G. Protein Structure Prediction: The next Generation. *J. Chem. Theory Comput.* **2006**, *2* (3), 705−716.

(12) Carnevali, P.; Tóth, G.; Toubassi, G.; Meshkat, S. N. Fast Protein Structure Prediction Using Monte Carlo Simulations with Modal Moves. *J. Am. Chem. Soc.* **2003**, *125* (47), 14244−14245.

(13) Skolnick, J.; Zhou, H. Why Is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? *J. Phys. Chem. B* **2017**, *121* (15), 3546−3554.

(14) Buchan, D. W. A.; Jones, D. T. EigenTHREADER: Analogous Protein Fold Recognition by Efficient Contact Map Threading. *Bioinformatics* **2017**, *33* (17), 2684−2690.

(15) Saitoh, S.; Nakai, T.; Nishikawa, K. A Geometrical Constraint Approach for Reproducing the Native Backbone Conformation of a Protein. *Proteins: Struct., Funct., Genet.* **1993**, *15* (2), 191−204.

(16) Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A. M. J. J. Assessment of Contact Predictions in CASP12: Co-

Evolution and Deep Learning Coming of Age. *Proteins: Struct., Funct., Genet.* **2018**, *86* (October 2017), 51−66.

(17) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated Mutations and Residue Contacts in Proteins. *Proteins: Struct., Funct., Genet.* **1994**, *18* (4), 309−317.

(18) Pollock, D. D.; Taylor, W. R. Effectiveness of Correlation Analysis in Identifying Protein Residues Undergoing Correlated Evolution. *Protein Eng., Des. Sel.* **1997**, *10* (6), 647−657.

(19) Vijayakumar, M.; Zhou, H. X. Prediction of Residue-Residue Pair Frequencies in Proteins. *J. Phys. Chem. B* **2000**, *104* (41), 9755−9764.

(20) Dekker, J. P.; Fodor, A.; Aldrich, R. W.; Yellen, G. A Perturbation-Based Method for Calculating Explicit Likelihood of Evolutionary Co-Variance in Multiple Sequence Alignments. *Bioinformatics* **2004**, *20* (10), 1565−1572.

(21) Jones, D. T.; Buchan, D. W. A.; Cozzetto, D.; Pontil, M. PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics* **2012**, *28* (2), 184−190.

(22) Jarmolinska, A. I.; Zhou, Q.; Sulkowska, J. I.; Morcos, F. DCA-MOL: A PyMOL Plugin To Analyze Direct Evolutionary Couplings. *J. Chem. Inf. Model.* **2019**, *59* (2), 625−629.

(23) Seemayer, S.; Gruber, M.; Söding, J. CCMpred - Fast and Precise Prediction of Protein Residue-Residue Contacts from Correlated Mutations. *Bioinformatics* **2014**, *30* (21), 3128−3130.

(24) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (39), 15674−15679.

(25) Wu, S.; Zhang, Y. A Comprehensive Assessment of Sequence-Based and Template-Based Methods for Protein Contact Prediction. *Bioinformatics* **2008**, *24* (7), 924−931.

(26) Ding, C. H. Q.; Dubchak, I. Multi-Class Protein Fold Recognition Using Support Vector Machines and Neural Networks. *Bioinformatics* **2001**, *17* (4), 349−358.

(27) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **2017**, *13* (1), No. e1005324.

(28) Friedman, J.; Hastie, T.; Tibshirani, R. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics.* **2008**, *9* (3), 432−441.

(29) Liu, Y.; Palmedo, P.; Ye, Q.; Berger, B.; Peng, J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Syst.* **2018**, *6* (1), 65−74.e3.

(30) Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks. *Bioinformatics* **2019**, *35*, 4647−4655.

(31) Li, Y.; Zhang, C.; Bell, E. W.; Yu, D. J.; Zhang, Y. Ensembling Multiple Raw Coevolutionary Features with Deep Residual Neural Networks for Contact-Map Prediction in CASP13. *Proteins: Struct., Funct., Genet.* **2019**, *87* (12), 1082−1091.

(32) Shrestha, R.; Fajardo, E.; Gil, N.; Fidelis, K.; Kryshtafovych, A.; Monastyrskyy, B.; Fiser, A. Assessing the Accuracy of Contact Predictions in CASP13. *Proteins: Struct., Funct., Genet.* **2019**, *87* (12), 1058−1068.

(33) Finn, R.; Pfam, D. Clans, Web Tools and Services. *Nucleic Acids Res.* **2006**, *34* (90001), D247−D251.

(34) Heffernan, R.; Paliwal, K.; Lyons, J.; Singh, J.; Yang, Y.; Zhou, Y. Single-Sequence-Based Prediction of Protein Secondary Structures and Solvent Accessibility by Deep Whole-Sequence Learning. *J. Comput. Chem.* **2018**, *39* (26), 2210−2216.

(35) Li, W.; Godzik, A. Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.

(36) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12* (1), 7−8.

(37) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−3402.

(38) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* **2012**, *9* (2), 173−175.

(39) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE, 2016; Vol. *2016-Decem*, pp 770−778.

(40) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **2017**, *13* (1), No. e1005324.

(41) Li, X.; Chen, S.; Hu, X.; Yang, J. Understanding the Disharmony Between Dropout and Batch Normalization by Variance Shift. In *2019 IEEE/CVF Conference on computer Vision and Pattern Recognition (CVPR)*; IEEE, 2019; pp 2677−2685.

(42) Nakai, K.; Kidera, A.; Kanehisa, M. Cluster Analysis of Amino Acid Indices for Prediction of Protein Structure and Function. *Protein Eng., Des. Sel.* **1988**, *2* (2), 93−100.

(43) Henikoff, S.; Henikoff, J. G. Ample Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89* (22), 10915−10919.

(44) He, B.; Mortuza, S M; Wang, Y.; Shen, H.-B.; Zhang, Y. NeBcon: Protein Contact Map Prediction Using Neural Network Training Coupled with Naïve Bayes Classifiers. *Bioinformatics* **2017**, *33* (15), 2296−2306.