IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 10, 2020, accepted April 20, 2020, date of publication April 30, 2020, date of current version May 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2991605

GANcon: Protein Contact Map Prediction With Deep Generative Adversarial Network

HANG YANG[1], MINGHUI WANG[1,2], (Member, IEEE), ZHENHUA YU[3], XING-MING ZHAO[4,5], (Senior Member, IEEE), AND AO LI[1,2], (Member, IEEE)

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei AH230027, China
[2]Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH230027, China
[3]Department of Software Engineering, Ningxia University, Yinchuan, China
[4]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China
[5]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, Beijing, China

Corresponding author: Minghui Wang (mhwang@ustc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871361, Grant 61971393, Grant 61471331, Grant 61571414, Grant 61901238, Grant 61932008, Grant 61772368, and Grant 61572363, in part by the Science and Technique Research Foundation of Ningxia Institutions of Higher Education under Grant NGY2018-54, in part by the National Key Research and Development Program of China under Grant 2018YFC0910500, in part by the Natural Science Foundation of Shanghai under Grant 17ZR1445600, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2018SHZDZX01, and in part by the ZJLab.

ABSTRACT Accurate protein contact map prediction is essential for *de novo* protein structure prediction. Over the past few years, deep learning has brought a significant breakthrough in protein contact map prediction and optimized deep learning architectures are highly desired for performance improvement. As an emerging deep learning architecture, the generative adversarial network (GAN) has shown the powerful capability of learning intrinsic patterns, which inspires us to comprehensively exploit GAN for predicting accurate protein contact maps. In this study, we present GANcon, a novel GAN-based deep learning architecture for protein contact map prediction, which to the best of our knowledge is the first GAN-based approach in this field. Instead of using a single neural network, GANcon is composed of two competitive networks that are evolving through adversarial learning. The generator network employs a dedicated encoder-decoder architecture that can efficiently capture the underlying contact information from versatile protein features to generate contact maps, while the discriminator network learns the differences between generated contact maps and real ones and promotes the generator network to produce more accurate contact maps. Moreover, to deal with the imbalance problem and take into account the symmetry of contact maps, we also propose a novel symmetrical focal loss, which can further enhance the effectiveness of adversarial learning for better performance. The experimental results on several datasets demonstrate that GANcon outperforms many state-of-the-art methods, indicating the effectiveness of our method for predicting protein contact maps. GANcon is freely available at https://github.com/melissaya/GANcon.

INDEX TERMS Protein contact map prediction, deep learning, generative adversarial network, adversarial learning.

I. INTRODUCTION

Proteins are crucially important macromolecules in an organism and play a fundamental role in almost all biological processes. In order to carry out the essential cellular function, proteins fold into specific three-dimensional structures, which are driven and stabilized by the interactions between protein residues, i.e., protein residue contacts. For a protein sequence, all contacts of residue pairs can be encoded into a binary matrix named 'contact map', which has been regarded as a critical contributor for accurate *de novo* protein structure prediction [1], [2]. In recent Critical Assessment of protein Structure Prediction (CASP) experiments, many excellent *de novo* protein structure prediction methods have benefited much from the incorporation of predicted contact maps [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Hualong Yu.

Due to the importance of contact map in protein structure prediction, researches on predicting protein contact maps have been booming in the past decade. For example, the evolutionary coupling analysis (ECA) methods predict contacts by capturing co-evolved residues from protein multiple sequence alignments (MSAs), such as CCMpred [5] and FreeContact [6]. These methods are effective for predicting contacts in proteins with a large number of high-quality MSAs, while their predictive performance is limited if the proteins have few or low-quality MSAs [7], [8]. In contrast, machine learning methods can learn complex relationships from all kinds of information, including the co-evolutionary information estimated by ECA methods, and therefore have been more successful in contact map prediction [9].

Over the past few years, as a powerful machine learning technique, deep learning has brought a significant breakthrough in protein contact map prediction [8], [10], [11]. A typical deep learning method usually adopts a multilayer convolutional neural network (CNN) architecture to learn inherent patterns in contact maps automatically. For example, DNCON2 is composed of six CNN blocks [12], RaptorX-Contact uses deep residual convolutional network (ResNet) [11], and SPOT-Contact combines ResNet with two-dimensional residual bidirectional recurrent long short-term memory networks [8]. These carefully designed deep learning architectures have shown remarkable predictive power in recent CASP experiments, which inspires us to further explore more optimized deep learning architectures for performance improvement.

Most recently, as an emerging deep learning architecture, the generative adversarial network (GAN) [13] has received considerable attention due to its powerful capability of learning intrinsic patterns in diverse fields, such as image classification [14] and gene expression inference [15]. Instead of using a single deep neural network, GAN is composed of two competitive networks, namely a generator network and a discriminator network, which are evolving in an adversarial learning strategy: the generator network produces fake samples and tries to fool the discriminator network into believing the generated samples are real, while the discriminator network tries to distinguish the generated samples from the real ones and guides the generator network to produce more realistic samples. Although it is of great interest to comprehensively exploit GAN for predicting protein contact maps, there are still several issues to address. First, despite that contact map prediction can be interpreted as image classification problem at pixel level where each pixel represents one residue pair [16], the input features adopted in contact map prediction include versatile protein features such as the GaussDCA scores [17], Atchley factors [18] and log number of sequences in the alignment, which are usually more complex than the input of image classification [11]. Therefore, to produce accurate contact maps, the generator network of GAN is required to efficiently capture underlying contact information from these complex features. Second, the contact map is a binary matrix and the ratio of contact and non-contact residue pairs is extremely low ($<2\%$) [19], which leads to a severe imbalance problem [7], [9] especially for GAN models optimized by commonly-used binary cross-entropy (BCE) loss [20]. Third, symmetry is an important and unique property of contact map [21], which, however, is absent in almost all other prediction tasks and therefore not considered in existing GAN models.

In this work, we present a novel GAN-based deep learning architecture called GANcon for protein contact map prediction, which to the best of our knowledge is the first GAN-based approach in this field. The generator network of GANcon employs a dedicated encoder-decoder architecture to efficiently capture the underlying contact information from versatile protein features. Meanwhile, through adversarial learning, the discriminator network of GANcon learns the differences between generated contact maps and real ones and promotes the generator network to produce more accurate contact maps. Moreover, to cope with the imbalance problem and take into account the symmetry of contact maps, we also propose a novel symmetrical focal (SF) loss in this study, which can further enhance the effectiveness of adversarial learning for better prediction results. We assess the prediction performance of GANcon on an independent test dataset of 360 proteins, and CASP12 and CASP13 datasets. The experimental results demonstrate that GANcon shows very promising performance for all length cutoffs and sequence separations.

II. METHODS AND MATERIALS

A. THE ARCHITECTURE OF GANcon

The overall architecture of GANcon including a generator network and a discriminator network is depicted in Figure 1. The generator network of GANcon takes the given protein features as input to produce contact maps, while the discriminator network of GANcon works as a classifier to discriminate generated contact maps from real contact maps. With both the adversarial loss and the SF loss, adversarial learning promotes the generator network to produce accurate contact maps that approximate the distribution of real contact maps. After training, the generator network will then be adopted for contact map prediction.

To efficiently capture the underlying contact information from versatile protein features, the generator network of GANcon employs a dedicated encoder-decoder architecture. As shown in Figure 1, the encoder path consists of a series of residual blocks (ResBlocks) [22], 3×3 convolutional layers with rectified linear unit (ReLU), and 2×2 max-pooling layers with stride of 2. ResBlocks utilize short skip connections to skip the block input to its output, and therefore learn a residual representation of the protein features that is helpful to capture complex relationships between residue pairs. Max-pooling layers down-sample the output of the previous layer and two dropout layers are added into the last two steps of the encoder path to prevent the network from overfitting. The decoder path consists of a series of

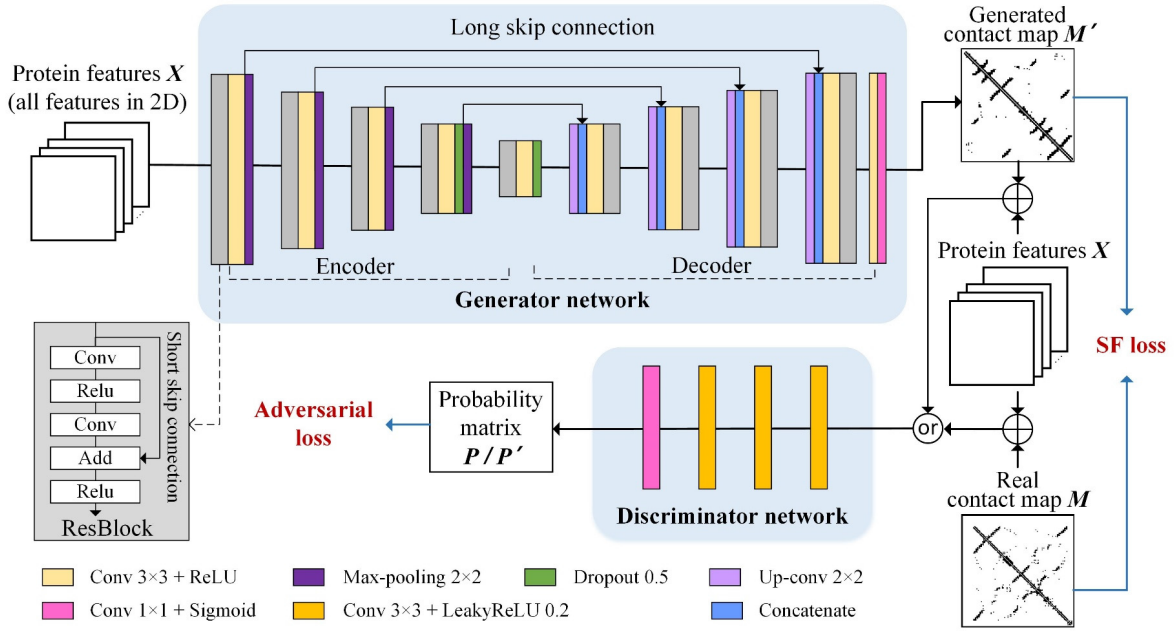


FIGURE 1. The overall architecture of GANcon. GANcon is composed of a generator network and a discriminator network. The generator network produces contact maps that are highly similar to real contact maps while the discriminator network tries to distinguish generated contact maps from real contact maps. With both the adversarial loss and the SF loss, adversarial learning promotes the generator network to produce accurate contact maps that the discriminator network cannot distinguish from the real ones.

2×2 up-convolutional layers, 3×3 convolutional layers with ReLU and ResBlocks. Moreover, long skip connections are added between the encoder path and the decoder path to provide more different-level details about contact information. Finally, a 1×1 convolutional layer with a sigmoid activation function is used to produce pixel-level contact map prediction.

The discriminator network of GANcon extensively plays an adversarial role to promote the generator network to produce accurate contact maps, which comprises three 3×3 convolutional layers with a leaky rectified linear unit (LeakyReLU) and a 1×1 convolutional layer with a sigmoid activation function. Similar to the previous study [23], the discriminator network receives the concatenated pair of the generated or real contact map and the corresponding input protein features. The outputs are the pixel-level probabilities of real or generated contact residue pairs in a contact map.

B. ADVERSARIAL LEARNING

We denote the input protein features as X of size $L \times L \times N$, where L is the length of protein sequence and N is the number of protein features. The corresponding real contact map is denoted as M of size $L \times L \times 1$. The generator network of GANcon is denoted as $G(\cdot)$ that outputs a generated contact map $M' = G(X)$ of size $L \times L \times 1$. The discriminator network of GANcon is denoted as $D(\cdot)$ that outputs a probability matrix $P = D(X, M)$ or $P' = D(X, M')$ of size $L \times L \times 1$, which includes pixel-level probabilities of contact pairs coming from a real contact map M or a generated contact map M'.

During the adversarial learning process, the adversarial loss used for the discriminator network is based on BCE loss:

$$min\ L_D^{adv} = -\sum_i \sum_j z \log P_{ij} + (1 - z) log\left(1 - P'_{ij}\right), \quad (1)$$

where $z = 0$ if the input includes generated contact maps and $z = 1$ if the input includes real contact maps. P_{ij} and P'_{ij} are the values of the i-th row and j-th column in P and P', respectively. The first term of (1) is used to classify M as real at pixel level, while the second term is used to make M' to be classified as fake at pixel level.

To train the generator network, GANcon uses a loss function that is a weighted sum of an adversarial loss based on P' and an SF loss between M and M', which is defined as follows:

$$min\ L_G = \lambda L_G^{adv} + L_G^{SF}, \quad (2)$$

where λ is set to 1.0 to maintain the balance of adversarial learning. The adversarial loss used for generator network aims to fool the discriminator network through maximizing the probability of M' being considered as real and is defined as:

$$L_G^{adv} = -\sum_i \sum_j \log P'_{ij}. \quad (3)$$

Although BCE loss between M and M' is commonly used in existing GAN models for almost all other prediction tasks, it is not suitable for protein contact map prediction as it fails to deal with the imbalance problem caused by the low rate of contact and non-contact residue pairs and ignores the symmetry of contact maps. In order to solve these problems,

we propose a novel SF loss to amend BCE loss by using both focal loss introduced by Lin *et al.* [20] and symmetrical loss:

$$L_G^{SF} = \beta L_G^F + L_G^S, \tag{4}$$

where β is set to 1.0 to balance the role of two loss terms. L_G^F is the focal loss that is effective for the imbalance problem and is defined as:

$$L_G^F = -\sum_i \sum_j \alpha \left(1 - M'_{ij}\right)^\gamma M_{ij} \log M'_{ij} + (1-\alpha) M'^{\gamma}_{ij} \left(1 - M_{ij}\right) \log\left(1 - M'_{ij}\right), \tag{5}$$

where $\alpha \in [0, 1]$ is a weighting factor to adjust the importance of contact and non-contact residue pairs and γ is a parameter that puts the focus on hard and misclassified residue pairs and reduces the loss contribution of easy-to-classify residue pairs. In this study, α is set to 0.25 and γ is set to 2.0. In order to keep the symmetry of the generated contact maps as much as possible, the symmetrical loss L_G^S is defined as follows:

$$L_G^S = -\sum_i \sum_j \left(M'_{ij} - M'_{ji}\right)^2. \tag{6}$$

C. IMPLEMENTATION DETAILS

GANcon is implemented using the Keras library (https://keras.io) along with Tensorflow (https://www.tensorflow.org). We use the Adam optimization method with the initial learning rate as 1E-4 in the generator network and 1E-5 in the discriminator network. In each epoch, a mini-batch size of 1 is used for both networks due to GPU memory limitation and we train the discriminator network 3 times while training the generator network once. GANcon takes approximately 15 hours (20-30 epochs) to converge with an Nvidia 1080 Ti GPU.

D. PROTEIN FEATURES

As shown in Supplementary Table S1, the protein features used in GAN-con include various two-dimensional, one-dimensional and scalar features, which are consistent with those used by many other methods [24], [25]. To derive these features from MSAs, by following a similar procedure in previous methods [12], [25], we first run HHblits [26] with an E-value threshold of 1E-3 to search the Uniclust30 database [27] to generate alignments. If the alignment found by HHblits has fewer than 2000 homologous sequences, we then run JackHMMER [28] with E-value thresholds of 1E-20, 1E-10, 1E-4 and 1 to search the UniRef90 database [29]. After that, we use these alignments to generate other protein features, e.g., GaussDCA scores [17]. Finally, both one-dimensional and scalar protein features are duplicated to form two-dimensional matrixes, which are used together with the two-dimensional protein features as the inputs of GANcon [12].

E. DATASETS

The dataset used in this study consists of SCOPe 2.07 subsets filtered for sequences with less than 30% sequence identity (based on PDB SEQRES records) and sequence lengths between 50 and 500 [30]. Meanwhile, the dataset is divided into three non-overlapping sets for training, validation and independent test, which is a commonly-used performance evaluation strategy in deep learning methods [31], [32]. In this way, 7192 proteins from SCOPe 2.06 are allocated to the training and validation datasets (90% and 10%, respectively) [31], while 360 proteins newly released in the SCOPe 2.07 are allocated to the independent test dataset. Moreover, we also carry out additional testing on the publicly available targets in recent CASP experiments including 22 CASP12 free modeling (FM) targets and 15 CASP13 FM targets, for an objective comparison with state-of-the-art methods.

F. EVALUATION CRITERIA

According to the standard CASP definition [33], protein residues are defined as in contact when the Euclidean distance between two C_β atoms (C_α for Glycine) falls within 8 Å. All contacts are divided into three groups depending on the sequence separation, including long-range (sequence separation ≥ 24), medium-range ($12 \leq$ sequence separation < 24) and short-range ($6 \leq$ sequence separation < 12). Following the CASP routine, we take the top L/k ($k = 5, 2, 1$) predicted contacts, where L is protein sequence length, to calculate the precision, recall, and F1 score. These three metrics are defined as:

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}}, \tag{7}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}}, \tag{8}$$

and

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

where True Positive is the number of correctly predicted contacts, False Positive is the number of contacts falsely predicted as non-contacts and False Negative is the number of non-contacts falsely predicted as contacts.

To further analyze all the predictions from GANcon, we provide Precision-Recall (PR) curves and Receiver Operator Characteristic (ROC) curves. The corresponding area under the PR curve (AUPRC) and area under the ROC curve (AUC) scores for long-, medium- and short-range are also provided. In consistent with previous studies [16], [34], we use P-values in the Student's t-test to compare these metrics of other methods with those of GANcon, which is a measure of statistical significance of the difference between two methods' results.

III. RESULTS

A. THE EFFECTIVENESS OF ADVERSARIAL LEARNING

In order to measure the impact of the adversarial learning in protein contact map prediction, in Table 1 we first compare

TABLE 1. Precisions of different models on the validation dataset.

	Models	$L/5$ (%)	$L/2$ (%)	L (%)
Long	Baseline	77.76	67.23	53.74
	Baseline+GAN	83.03	72.62	58.03
	Baseline+GAN+SF (GANcon)	**87.07**	**77.24**	**62.06**
Medium	Baseline	67.01	47.15	29.95
	Baseline+GAN	71.69	50.20	31.54
	Baseline+GAN+SF (GANcon)	**74.89**	**52.46**	**32.17**
Short	Baseline	64.05	42.71	26.30
	Baseline+GAN	68.60	45.63	27.47
	Baseline+GAN+SF (GANcon)	**72.67**	**47.60**	**27.99**

Note: The baseline is the model without adversarial learning and using BCE loss.

the prediction precisions of models with or without adversarial learning on the validation dataset for both top-ranking prediction length cutoffs ($L/5$, $L/2$ and L) and sequence separations (long-, medium- and short-range). The model without adversarial learning is treated as the baseline, which only uses the generator network of GANcon and is trained with the commonly used BCE loss. And the model with adversarial learning (i.e., GAN) uses both the generator network and the discriminator network of GANcon and is trained with both adversarial loss and BCE loss. As shown in Table 1, the prediction performance is greatly improved for all levels of contact precisions with the help of adversarial learning. For example, for top $L/5$ long-, medium- and short-range contacts, the model with adversarial learning has a precision of 83.03%, 71.69% and 68.60%, respectively, which is 5.27%, 4.68% and 4.55% higher than that without adversarial learning (77.76%, 67.01% and 64.05%), respectively. The corresponding P-value in the Student's t-test is 2.56E-61, 1.02E-117 and 1.52E-30 (Supplementary Table S2), respectively, indicating that the improvement is statistically significant. In addition to precisions, we also show the F1 scores, AUPRC scores and AUC scores of different models in Supplementary Tables S3-S5. From these results, we find that adversarial learning leads to significant improvements for all length cutoffs and sequence separations. For example, the model without adversarial learning obtains an AUPRC score of 49.66%, 57.94% and 56.57% for long-, medium- and short-range contacts, respectively, which is 6.18%, 5.65% and 5.62% lower than the model with adversarial learning (55.84%, 63.59% and 62.19%), respectively. Taken together, these results indicate the effectiveness of adversarial learning in protein contact map prediction.

Furthermore, to explore the optimal loss function yielding better performance during adversarial learning, we train GANcon model using adversarial learning with the proposed SF loss. As shown in Table 1, the proposed SF loss consistently brings additional performance improvements for all length cutoffs and sequence separations. For example, the precision for top $L/5$, $L/2$ and L long-range contacts are 87.07%, 77.24% and 62.06% by SF loss, respectively, compared to 83.03%, 72.62% and 58.03% by BCE loss, respectively. Also, the corresponding P-values shown in Supplementary Table S2 suggest the performance improvements obtained by using SF loss are all statistically significant. Similar results in F1 scores, AUPRC scores and AUC scores can also be observed in Supplementary Tables S3-S5, which further demonstrates that the proposed SF loss is indeed effective for protein contact map prediction. Overall, by jointing SF loss with adversarial learning, GANcon can successfully boost the precision by 9.31%, 7.88% and 8.62% for top $L/5$ long-, medium- and short-range contacts.

B. COMPARISONS OF GANcon WITH EXISTING METHODS

We compare GANcon on the independent test dataset with four state-of-the-art deep learning methods, including DNCON2 [12], DeepContact [7], PconsC4 [24] and DeepCov [34], and two well-known ECA methods including CCMpred [5] and FreeContact [6]. All of these compared methods are downloaded and implemented in our local computers with default settings. Among these methods, DeepCov, PconsC4, CCMpred and FreeContact are fed with the same MSAs used in GANcon since they do not have a built-in pipeline to generate MSAs, while other methods are fed with protein sequences directly, which is consistent with previous studies [12], [16].

The comparison results in Table 2 for the precisions clearly show that the deep learning methods significantly outperform the ECA methods, which is also corroborated by previous studies [16], [35]. For example, DeepContact achieves a precision of 86.66%, 75.52% and 60.20% for top $L/5$, $L/2$ and L long-range contacts, respectively (Table 2), which is 24.85%, 25.88% and 22.92% higher than FreeContact, respectively. At the same time, GANcon performs consistently better than other deep learning methods and the corresponding precision reaches 89.93%, 80.84% and 65.87%, respectively. And the P-values shown in Supplementary Table S6 suggest that the improvement is significant. We also train GANcon with other training-validation dataset ratio (80%-20% and 70%-30%), and the precision results in Supplementary Tables S7 show that there is no obvious difference in the performance. Moreover, as shown in Supplementary Table S8, the F1 score of GANcon is 45.91% for top $L/5$ long-range contacts, while the next-best deep learning method has the F1 score of 43.01%. All these results indicate that with the novel deep learning architecture, GANcon has a very competitive performance for contact map prediction. In addition, we also provide the PR and ROC curves with the corresponding AUPRC and AUC scores of different methods for long-range contacts in Figure 2 and Figure 3. As shown in Figure 2, the PR curve for long-range contacts confirms GANcon has a better precision under a given level of recall than other methods and the corresponding AUPRC score is 63.57%, which is at least 7% better than other methods investigated in this study. Meanwhile, the PR curves in Supplementary Figure S1 demonstrate the

TABLE 2. Comparison of precision with state-of-the-art methods on the independent test dataset.

Method	Long			Medium			Short		
	L/5 (%)	*L*/2 (%)	*L* (%)	*L*/5 (%)	*L*/2 (%)	*L* (%)	*L*/5 (%)	*L*/2 (%)	*L* (%)
DeepContact	86.66	75.52	60.20	74.11	51.30	31.67	69.02	44.61	26.95
DeepCov	81.73	67.20	50.30	67.79	45.95	28.92	67.86	44.29	27.02
DNCON2	85.81	74.94	59.67	74.32	51.04	31.42	72.99	47.92	28.38
PconsC4	85.67	74.88	59.41	74.60	51.04	31.15	73.10	46.90	27.90
FreeContact	61.81	49.64	37.28	46.40	29.25	18.95	34.83	22.95	15.79
CCMpred	66.15	55.43	42.20	54.48	33.31	20.47	46.54	27.42	17.32
GANcon	**89.93**	**80.84**	**65.87**	**77.89**	**54.13**	**32.83**	**74.06**	**48.46**	**28.46**

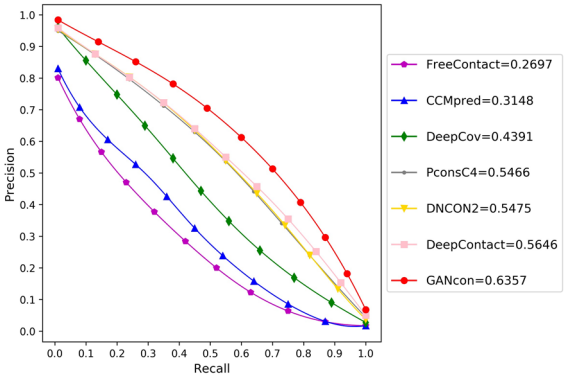


FIGURE 2. PR curves of different methods for long-range contacts on the independent test dataset.

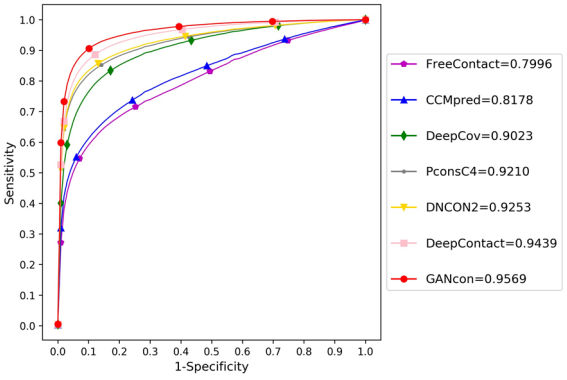


FIGURE 3. ROC curves of different methods for long-range contacts on the independent test dataset.

advantage of GANcon for medium- and short-range contacts. Also, the ROC curves with the corresponding AUC scores in Figure 3 and Supplementary Figure S2 show similar results when long-, medium- and short-range contacts are evaluated.

To explore the effect of the number of homologous sequences on the performance of computational methods, we present the precisions of the top *L*/5 long-range contacts as a function of the maximum log Neff scores in Figure 4, where Neff is defined as the number of effective sequences in MSAs and a higher score implies more homologous sequences in the reference database. As shown in Figure 4, in general, all

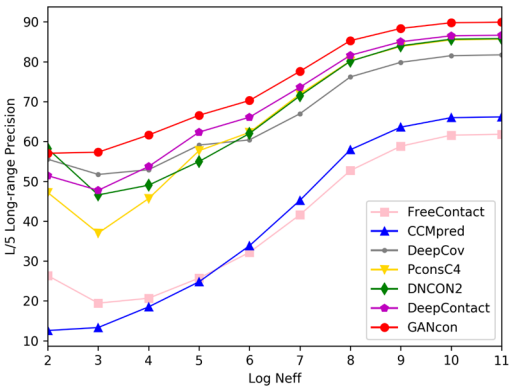


FIGURE 4. Precisions of top L/5 long-range contacts from proteins from the independent test dataset grouped by a maximum log Neff score.

methods have the lower performance for lower Neff score and the precisions of all methods are increasing as the Neff score increases, indicating the significance of a large number of homologous sequences in MSAs for contact map prediction. Across all Neff scores, GANcon achieves comparable or better precisions than other methods.

C. BENCHMARK RESULTS ON CASP DATASETS

Finally, the performance of GANcon is evaluated on 22 CASP12 FM targets and 15 CASP13 FM targets. To compare with well-performing methods in recent CASP experiments, we evaluate the prediction results of RaptorX-Contact [11] from CASP website (http://predictioncenter.org/) and the prediction results of SPOT-Contact [8] (http://sparks-lab.org/jack/server/SPOT-Contact/) from its webserver. The comparison results with respect to precisions can be found in Table 3 and Supplementary Tables S9-S10. The baseline model of GANcon has in general comparable performance to most of the investigated methods except the state-of-the-art RaptorX-Contact and SPOT-Contact. Meanwhile, we also observe that using adversarial learning and SF loss brings significant performance improvements for all length cutoffs and sequence separations. For example, there are more than 23%, 15% and 8% improvements in precision for top *L*/5, *L*/2 and *L* long-range contacts on 15 CASP13 FM targets

TABLE 3. Comparison of precision for long-range contacts with state-of-the-art methods on CASP12 and CASP13 datasets.

Method	CASP12 dataset			CASP13 dataset		
	L/5 (%)	*L*/2 (%)	*L* (%)	*L*/5 (%)	*L*/2 (%)	*L* (%)
DeepContact	43.96	36.26	28.91	24.73	21.41	17.46
DeepCov	48.76	40.16	32.29	32.99	28.20	23.15
DNCON2	53.51	42.60	33.78	28.36	19.86	15.41
PconsC4	44.71	39.88	32.15	35.81	25.47	20.34
FreeContact	32.45	25.02	18.72	17.10	12.44	10.01
CCMpred	35.36	29.66	20.94	12.08	8.79	6.73
RaptorX-Contact	48.59	39.55	31.48	**62.14**	**53.99**	**41.00**
SPOT-Contact	**64.30**	**54.58**	**44.69**	49.22	41.76	31.98
GANcon (Baseline)	47.67	40.65	33.46	31.33	25.84	22.09
GANcon (Baseline+GAN+SF)	57.40	45.37	37.22	54.63	41.05	30.15

Note: The baseline is the model without adversarial learning and using BCE loss.

(Table 3). In addition, similar results in F1, AUPRC and AUC scores of GANcon and other methods are shown in Supplementary Tables S11-S14, which further demonstrate that GANcon can be used as a complementary method for protein contact map prediction.

IV. CONCLUSION

Accurate prediction of the protein contact map is of great significance in *de novo* protein structure prediction. As many carefully designed deep learning architectures have shown remarkable prediction power in many areas of bioinformatics [36]–[38], especially in contact map prediction [8], [11], further exploration of more optimized deep learning architectures for performance improvement is highly desired. In this study, we propose a novel GAN-based architecture, GANcon, for contact map prediction. Different from previous deep learning methods training a single network in protein contact map prediction, GANcon incorporates a discriminator network to promote the generator network to achieve accurate contact map prediction. During the adversarial learning process, the generator network of GANcon captures the underlying contact information from versatile protein features by employing a dedicated encoder-decoder architecture, while the discriminator network learns the differences between generated contact maps and real ones and automatically transfers them back to the generator network. Meanwhile, to deal with the imbalance problem and consider the symmetry of contact maps, a novel SF loss is proposed in this study that together with the adversarial loss, can further enhance the adversarial learning of GANcon for better prediction results. Notably, jointing adversarial learning with SF loss brings consistent improvements in prediction performance across all the datasets assessed in this study, indicating adversarial learning and SF loss might be adopted as a general learning strategy for the task of protein contact map prediction.

Although GANcon shows a promising performance of protein contact map prediction, there is still room for further improvement. The adversarial loss of GANcon is based on pixel level probabilities in the output matrix of discriminator, while the adversarial loss based on the whole contact map level output is also very useful to training GAN model, which can be adopted in our future work. Besides, a well-known problem is that the training of GAN sometimes suffers from instability [39], which also occurs during the training process of GANcon in this study. Therefore, some advanced GAN training methods, such as WGAN [39], can improve training stability and will be explored in our future study. Also, it would be interesting to integrate GAN with other popular deep learning modules, such as long short-term memory (LSTM) that is confirmed to be effective in contact map prediction [8], to boost the predictive power of GAN-based architectures. Moreover, in addition to the protein features adopted in this study, other important features, e.g. predictions of third-party predictors such as CCMpred and FreeContact, may also be used by GANcon to enhance prediction performance. In conclusion, we propose a novel GAN-based deep learning architecture for contact map prediction, which can efficiently improve the overall performance and serves as an alternative tool for contact map prediction.

REFERENCES

[1] M. Michel, S. Hayat, M. J. Skwark, C. Sander, D. S. Marks, and A. Elofsson, "PconsFold: Improved contact predictions improve protein models," *Bioinformatics*, vol. 30, no. 17, pp. 1482–1488, Sep. 2014, doi: 10.1093/bioinformatics/btu458.

[2] M. L. Tress and A. Valencia, "Predicted residue–residue contacts can help the scoring of 3D models," *Proteins, Struct., Function, Bioinf.*, vol. 78, no. 8, pp. 134–1980, 2010.

[3] S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, and D. Baker, "Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta," *Proteins, Struct., Function, Bioinf.*, vol. 84, pp. 67–75, Sep. 2016, doi: 10.1002/prot.24974.

[4] W. Zhang, J. Yang, B. He, S. E. Walker, H. Zhang, B. Govindarajoo, J. Virtanen, Z. Xue, H.-B. Shen, and Y. Zhang, "Integration of QUARK and I-TASSER for *ab initio* protein structure prediction in CASP11," *Proteins, Struct., Function, Bioinf.*, vol. 84, pp. 76–86, Sep. 2016, doi: 10.1002/prot.24930.

[5] 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, Nov. 2014, doi: 10.1093/bioinformatics/btu500.

[6] L. Kaján, T. A. Hopf, M. Kalas, D. S. Marks, and B. Rost, "FreeContact: Fast and free software for protein contact prediction from residue co-evolution," *BMC Bioinf.*, vol. 15, no. 1, p. 85, 2014, doi: 10.1186/1471-2105-15-85.

[7] Y. Liu, P. Palmedo, Q. Ye, B. Berger, and J. Peng, "Enhancing evolutionary couplings with deep convolutional neural networks," *Cell Syst.*, vol. 6, no. 1, pp. 65–74, Jan. 2018, doi: 10.1016/j.cels.2017.11.014.

[8] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. 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, Dec. 2018, doi: 10.1093/bioinformatics/bty481.

[9] J. Yang, Q.-Y. Jin, B. Zhang, and H.-B. Shen, "R_2C: Improving *ab initio* residue contact map prediction using dynamic fusion strategy and Gaussian noise filter," *Bioinformatics*, vol. 32, no. 16, pp. 2435–2443, Aug. 2016, doi: 10.1093/bioinformatics/btw181.

[10] W. Ding, W. Mao, D. Shao, W. Zhang, and H. Gong, "DeepConPred2: An improved method for the prediction of protein residue contacts," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 503–510, 2018, doi: 10.1016/j.csbj.2018.10.009.

[11] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLOS Comput. Biol.*, vol. 13, no. 1, Jan. 2017, Art. no. e1005324, doi: 10.1371/journal.pcbi.1005324.

[12] B. Adhikari, J. Hou, and J. Cheng, "DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks," *Bioinformatics*, vol. 34, no. 9, pp. 1466–1472, May 2018, doi: 10.1093/bioinformatics/btx781.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengi "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2014, pp. 2672–2680.

[14] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," 2016, *arXiv:1611.08408*. [Online]. Available: http://arxiv.org/abs/1611.08408

[15] X. Wang, K. Ghasedi Dizaji, and H. Huang, "Conditional generative adversarial network for gene expression inference," *Bioinformatics*, vol. 34, no. 17, pp. 603–611, Sep. 2018, doi: 10.1093/bioinformatics/bty563.

[16] Y. Li, J. Hu, C. Zhang, D.-J. Yu, and Y. Zhang, "ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks," *Bioinformatics*, vol. 35, no. 22, pp. 4647–4655, Nov. 2019, doi: 10.1093/bioinformatics/btz291.

[17] 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, Mar. 2014, Art. no. e92721, doi: 10.1371/journal.pone.0092721.

[18] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Druke, "Solving the protein sequence metric problem," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 18, pp. 6395–6400, May 2005, doi: 10.1073/pnas.0408677102.

[19] J. Bacardit, P. Widera, A. Marquez-Chamorro, F. Divina, J. S. Aguilar-Ruiz, and N. Krasnogor, "Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features," *Bioinformatics*, vol. 28, no. 19, pp. 2441–2448, Oct. 2012, doi: 10.1093/bioinformatics/bts472.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007, doi: 10.1109/ICCV.2017.324.

[21] Q. Wu, Z. Peng, I. Anishchenko, Q. Cong, D. Baker, and J. Yang, "Protein contact prediction using metagenome sequence data and residual neural networks," *Bioinformatics*, vol. 36, no. 1, pp. 41–48, Jan. 2020, doi: 10.1093/bioinformatics/btz477.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[23] J. Pan, C. Canton Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*. [Online]. Available: http://arxiv.org/abs/1701.01081

[24] M. Michel, D. Menéndez Hurtado, and A. Elofsson, "PconsC4: Fast, accurate and hassle-free contact predictions," *Bioinformatics*, vol. 35, no. 15, pp. 2677–2679, 2018, doi: 10.1093/bioinformatics/bty1036.

[25] 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, Apr. 2015, doi: 10.1093/bioinformatics/btu791.

[26] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, Feb. 2012, doi: 10.1038/Nmeth.1818.

[27] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger, "Uniclust databases of clustered and deeply annotated protein sequences and alignments," *Nucleic Acids Res.*, vol. 45, no. 1, pp. D170–D176, Jan. 2017, doi: 10.1093/nar/gkw1081.

[28] L. S. Johnson, S. R. Eddy, and E. Portugaly, "Hidden Markov model speed heuristic and iterative HMM search procedure," *BMC Bioinf.*, vol. 11, no. 1, p. 431, Dec. 2010, doi: 10.1186/1471-2105-11-431.

[29] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium, "UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, Mar. 2015, doi: 10.1093/bioinformatics/btu739.

[30] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, "SCOPe: Structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures," *Nucleic Acids Res.*, vol. 42, no. 1, pp. D304–D309, Jan. 2014, doi: 10.1093/nar/gkt1240.

[31] D. Xiong, J. Zeng, and H. Gong, "A deep learning framework for improving long-range residue–residue contact prediction using a hierarchical strategy," *Bioinformatics*, vol. 33, no. 17, pp. 2675–2683, Sep. 2017, doi: 10.1093/bioinformatics/btx296.

[32] F. Luo, M. Wang, Y. Liu, X.-M. Zhao, and A. Li, "DeepPhos: Prediction of protein phosphorylation sites with deep learning," *Bioinformatics*, vol. 35, no. 16, pp. 2766–2773, Aug. 2019, doi: 10.1093/bioinformatics/bty1051.

[33] I. Ezkurdia, O. Graña, J. M. G. Izarzugaza, and M. L. Tress, "Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8," *Proteins, Struct., Function, Bioinf.*, vol. 77, no. 9, pp. 196–209, 2009, doi: 10.1002/prot.22554.

[34] 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, Oct. 2018, doi: 10.1093/bioinformatics/bty341.

[35] B. Adhikari, "DEEPCON: Protein contact prediction using dilated convolutional neural networks with dropout," *Bioinformatics*, vol. 36, no. 2, pp. 470–477, 2020, doi: 10.1093/bioinformatics/btz593.

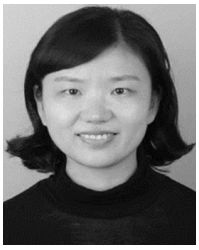
[36] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, and S. Peng, "Deep learning in omics: A survey and guideline," *Briefings Funct. Genomics*, vol. 18, no. 1, pp. 41–57, Feb. 2019, doi: 10.1093/bfgp/ely030.

[37] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *J. Parallel Distrib. Comput.*, vol. 117, pp. 212–217, Jul. 2018, doi: 10.1016/j.jpdc.2017.08.009.

[38] P. Wang, R. Ge, X. Xiao, Y. Cai, G. Wang, and F. Zhou, "Rectified-linear-unit-based deep learning for biomedical multi-label data," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 9, no. 3, pp. 419–422, Sep. 2017, doi: 10.1007/s12539-016-0196-1.

[39] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: https://arxiv.org/abs/1701.07875

HANG YANG received the B.S. degree in automation from the School of Information Science and Technology, University of Science and Technology of China (USTC), in 2017, where she is currently pursuing the M.S. degree. Her research interests include deep learning and bioinformatics.



MINGHUI WANG (Member, IEEE) received the B.S. degree from the School of Gifted Youth, University of Science and Technology of China (USTC), and the Ph.D. degree in biomedical engineering from the School of Information Science and Technology, USTC, in 2006. She is an Associate Professor with the School of Information Science and Technology and the Centers for Biomedical Engineering, USTC. Her research interests include bioinformatics, biostatistics, and machine learning.

ZHENHUA YU received the B.S. degree in electronic science and technology and the Ph.D. degree in biomedical engineering from the School of Information Science and Technology, University of Science and Technology of China, in 2011 and 2016, respectively. He is currently an Associate Professor with the School of Information Engineering, Ningxia University. He has authored over eight research articles. His current research interests include computational cancer genomics, bioinformatics, statistical machine learning, and artificial intelligence.

XING-MING ZHAO (Senior Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China. He is currently a Professor with the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. He has published over 70 journal articles. His current research interests include data mining and computational systems biology. He is an editorial board member of several journals.

AO LI (Member, IEEE) received the B.S. degree in biophysics from the School of Life Science, University of Science and Technology of China (USTC), in 2000, and the Ph.D. degree in biomedical engineering from the School of Information Science and Technology, USTC, in 2005. He is currently an Associate Professor with the School of Information Science and Technology and the Centers for Biomedical Engineering, USTC. His research interests include biomedical information processing and brain-inspired computing.