Alpha-Protein: Protein Contact-map Prediction Boosted by Attention

Hongjie Fang[1], Zhanda Zhu[1], Peishen Yan[1] and Hao Yin[1]

***Abstract*— Contact map is widely used in protein three-dimensional structure prediction, which denotes whether the distance between residue pairs is small enough. Many excellent researchers in this area have reached outperforming results. There are signs that deep-learning-based methods perform well even on low-quality feature sets. We select two representative models Deepcov [1] and ResPRE [6] to design our training and testing framework. It's worth mentioning that we introduced several kinds of attention mechanism into our proposed new models, including Cbam [15], SENet [3] and HaloNets [11]. In a word, we built an open source deep learning framework for contact map prediction and introduced attention mechanism into it. Last but not least, we have proved through ablation experiments that reducing the size of the last convolutional kernel from 3×3 to 1×1 is effective for performance improvement. All source codes are available at https://github.com/Galaxies99/alpha-protein.**

Keywords: Deep Learning, Attention, Protein contact-map prediction.

I. Introduction

Proteins are the focus of many areas of life science studies as they are responsible for most of the biological functions in living organisms. A protein is a long chain composed of multiple amino acids. The conformational space for directly modeling protein structure prediction is too large and often exceeds the upper limit of computing power. However, the contact-map reveals the contact status of protein residue pairs and thus it helps to predict the three-dimensional structure of the protein which determines the function of a protein.

In many existing studies, the predicted contact map is a two-dimensional Boolean matrix [5] [6], where each dimension represents the residue number, and the value represents whether the residue pairs are close enough. Further, some researches divide the distance between the residue pairs into more categories according to the distance range, *i.e*. three categories (long-range, medium-range and short range) [8]. In order to make the classification more refined, here we divide the distance range into ten categories.

In this area, many evidences have shown that deep learning can have great performance with even less quantity of feature sets. We applied many types of attention mechanism into our proposed models based on ResPRE model [6], considered the role of the attention mechanism in the model from a variety of perspectives. In our experiments, we compare our models' accuracy and efficiency with baseline models, which prove that our models have a better performance and the ability to solve these problems.

The contribution of this paper are listed as follows:

1) We build an open-source attention-based protein contact-map prediction framework, which provides some classic models and the models we proposed with attention mechanism.
2) We first introduce attention mechanism to protein contact-map prediction task systematically, and our models outperform the baseline models which do not apply attention mechanism.
3) We use ablation experiments to illustrate the effectiveness of the 1×1 kernel size in the last convolutional block proposed by FCN [9], then use this discovery to improve our models further.

II. Related Works

A. Protein Feature Extraction

Many state-of-the-art methods for contact prediction rely on additional sources of information, such as substitution frequency data from protein sequence alignments [5]–[7], [16], solvent accessibility [1], predicted secondary structure [8], and scores from other contact prediction methods.

Even though some studies only use amino acid sequence information, the input is quite different. Some directly take Multiple Sequence Alignment (MSA) as input, but some achieve covariance matrix from MSA and feed it into network, or some others get precision matrix from covariance matrix as input, and even use them all. All in all, these methods are all MSA-related.

B. Protein Contact-map Prediction

According to our knowledge, it's interesting that many researches like to use Residual neural networks to extract features from multiple sequence alignment.

Paper [13] takes sequential features and other pairwise features (e.g., coevolution information) as two different parts of input. Its model conducts a series of 1-dimensional convolutional transformations of sequential features to generate a 2-dimensional matrix and then concatenates it with coevolution information as the input of the second residual network. ResPRE model [6] predicts residue-level protein contacts using inverse covariance matrix (or precision matrix) of MSA through deep residual convolutional neural network training. DeepECA model [2] uses multilayer perceptron to calculate the weight for each sequence in an MSA with seven types of features generated from MSA. With such a lot of features and Residual neural networks, it has a pretty good performance on CASP dataset.

[1]Hongjie Fang, Zhanda Zhu, Peishen Yan and Hao Yin are students of Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Rd, Minhang District, Shanghai, China. {galaxies, dazz993, 1050335889, 1163706928}@sjtu.edu.cn

Some others which apply deep convolutional neural networks or GAN-based deep neural network on the input also show the good results. DeepCov model [5] uses fully convolutional neural networks operating on amino-acid pair frequency or covariance data derived directly from sequence alignments, without using global statistical methods such as sparse inverse covariance or pseudolikelihood estimation. GANcon model [16] uses a novel GAN-based deep learning architecture for protein contact map prediction, which is the first GAN-based approach in this field.

C. Attention

Attention mechanism is a data processing method in machine learning, which is widely used in many types of tasks such as natural language processing, image recognition and speech recognition. Traditionally, this mechanism is to hope that the neural network can automatically learning the areas that need attention in pictures.

The model structure in the attention mechanism for computer vision field is usually divided into three attention domains for analysis: spatial domain, channel domain and mixed domain.

1) **Spatial domain**: Apply spatial transformation to the spatial domain information of picture to extract the key information. Mask the space and score it. The masterpiece in this field is [4].
2) **Channel domain**: [3] applied attention of the channel domain in image processing. Similar to adding a weight to the signal on each channel to represent the correlation between the channel and the key information. The greater the weight, the higher the correlation. Generate a mask for the channel and score it.
3) **Mixed domain**: Mixed domain attention can make up for the shortcomings of the two methods mentioned before. By finding the corresponding attention weight for each feature element, [15] and [12] take the attention mechanism of spatial domain and channel domain into account at the same time.

We should notice that non-local neural network proposed by [14] takes both spatial domain and time domain into consideration. So we only need to remove the time dimension to apply it to image processing.

Recently, Google research team built multiscale self-attention models *HaloNets* [11] that are competitive with the best convolutional models. They developed two attention improvements: blocked local attention and attention downsampling. They also performed multiple ablations to understand how to improve the scaling of self-attention models.

III. Models

We choose ResPRE [6] as our base model since it is a well-known model that performs well on contact-map binary prediction task. We construct several attention-based models based on ResPRE model in order to make performance gain.

A. Cbam-FC-ResPRE

Cbam(Convolutional Block Attention Module) [15] is an important kind of attention mechanism modules which combines the spatial and channel features. That's the first attention mechanism we tried combining with existing ResPRE model seeking for performance improvement. It multiplies the residual module's output by the weight extracted by the Cbam attention mechanism to obtain the information containing the attention.

Fig. 1 shows the architecture of our Cbam-ResPRE model. We added an extra Cbam module in every basic block to extract global features for predicting. What's more, we substitute the last 3×3 convolutional block with 1×1 convolutional block, which is proposed in [9]. We believe it will speed up the convergence and improve the performance of the model.

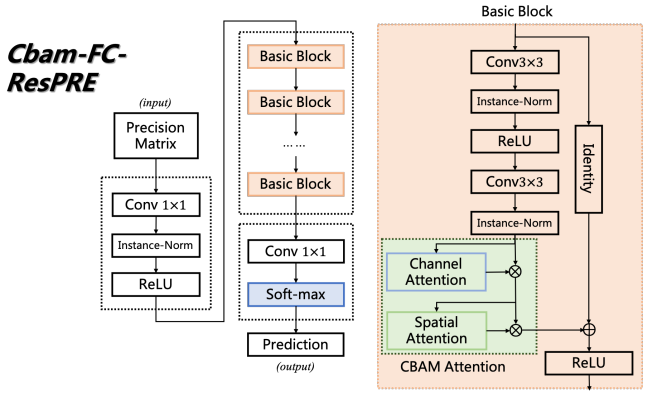


Fig. 1. Cbam-FC-ResPRE

Here we introduce the Cbam attention block. The Attention block takes the intermediate feature map $\mathbf{F} \in \mathbf{R}^{C \times H \times W}$ as input and processes this matrix as follows:

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F} \\ \mathbf{F}'' &= \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}' \end{aligned} \tag{1}$$

where $\otimes$ means element-wise multiplication, $\mathbf{M_c}(\mathbf{F})$ denotes the channel attention function and $\mathbf{M_s}(\mathbf{F})$ is the spatial attention function. They are computed as:

$$\begin{aligned} \mathbf{M_c}(\mathbf{F}) &= \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))) \\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{avg}})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{max}}))) \end{aligned} \tag{2}$$

$$\begin{aligned} \mathbf{M_s}(\mathbf{F}') &= \sigma(f^{7\times7}([AvgPool(\mathbf{F}'); MaxPool(\mathbf{F}')])) \\ &= \sigma(f^{7\times7}([\mathbf{F'^s_{avg}}; \mathbf{F'^s_{max}}])) \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is sigmoid function.

B. SE-FC-ResPRE

Except Cbam attention mechanism, we also apply a kind of channel domain attention adding weight on each channel and replace the residual basic block with Squeeze-and-Excitation block [3]. Similar to Cbam-attention module, it takes scale process with trained weight before adding the

identity shortcut. We also substitute the last 3×3 convolutional block with 1×1 convolutional block for performance considerations.

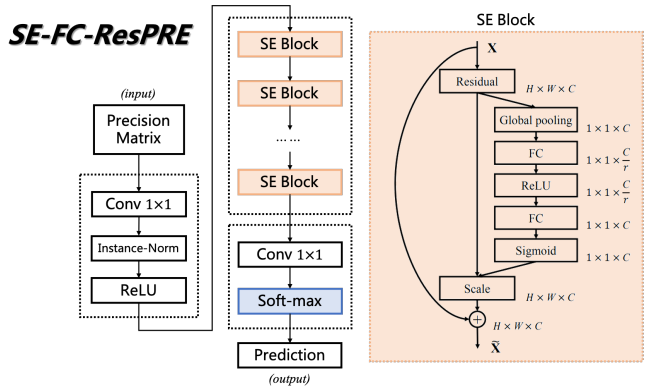


Fig. 2. SE-FC-ResPRE

As Fig. 2 shown, the first step of SE block is a transformation called $\mathbf{F}_{tr}$ mapping an input $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$ feature map to features map $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$. We use $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c]$ to denote the learned set of filter kernels, where $\mathbf{v}_c$ refers to the parameters of the c-th filter and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}]_c$ to denote the outputs where:

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s \tag{4}$$

where $*$ means convolution, $\mathbf{v} = [\mathbf{v}_c^1, \mathbf{v}_c^2, \ldots, \mathbf{v}_c^{C'}]$, $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2 \ldots, \mathbf{x}^{C'}]$ and $\mathbf{u}_c \in \mathbb{R}^{H \times W}$. $\mathbf{v}_c^s$ is is a 2D spatial kernel representing a single channel of $\mathbf{v}_c$ that acts on the corresponding channel of $\mathbf{X}$.

The second step is *squeeze* operation which aims to embed global spatial information. SENet block uses a global average pooling operation to :

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{5}$$

Then a statistic $\mathbf{z} \in \mathbb{R}^C$ can be generated by shrinking the $\mathbf{U}$ through its spatial dimensions $H \times W$ with formula (5).

Then an excitation operation is applied in SE block, aiming to fully capture channel-wise dependencies. SE block uses two fully connected layers and ReLU operations to calculate the final weights of different channels:

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{6}$$

where δ, σ refer to ReLU, sigmoid respectively. $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. It's worth mentioning that r is a "bottleneck" reducing the calculating complexity.

The final output of the block is obtained by rescaling $\mathbf{U}$ with the activations $\mathbf{s}$:

$$\tilde{\mathbf{x}}_c = \mathbf{F}_{scale} = (\mathbf{u}_c, s_c) = s_c \mathbf{u}_c \tag{7}$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_C]$ and $\mathbf{F}_{scale}$ refers to channel-wise multiplication between s_c and the feature map $\mathbf{u}_c \in \mathbb{R}^{H \times W}$.

C. Halo-ResPRE

Based on the idea of attention mechanism in Google research team's HaloNet [11], we propose our Halo-ResPRE model. After extracting features from the input using convolutional blocks, the data stream is divided into two branches. The first branch is the same as the residual sequence of ResPRE, and the second one extracts attention information through the Halo attention module.

We perform fusion operation to the outputs of the two branches, in order to obtain comprehensive information, and then the final prediction is obtained through the 1×1 convolutional layer and the softmax layer. The architecture of Halo-ResPRE model is shown in Fig. 3.

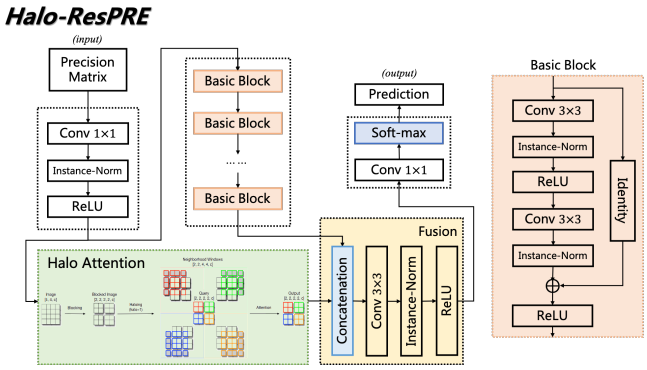


Fig. 3. Halo-ResPRE

D. NL-ResPRE

Inspired by the classic non-local mean method in computer vision, non-local operations [14] maintain more information because remote dependencies are directly captured by calculating the interaction between any two positions. So similar to the idea of the previous model, we apply this non-local operations in the second branch to take place of halo attention module.

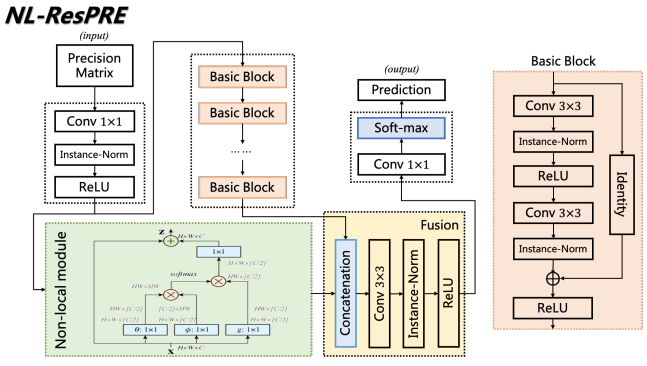


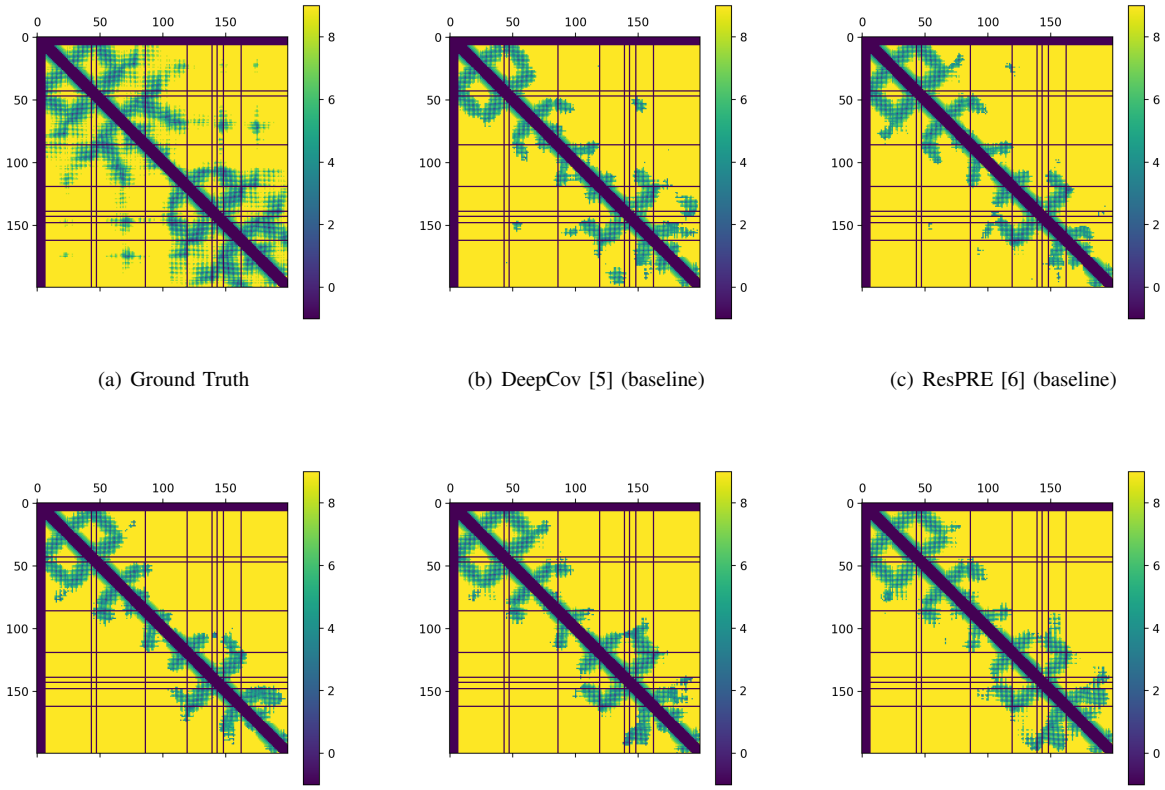
Fig. 4. NL-ResPRE

As shown in Fig. 4, non-local module takes the input with size of $H \times W \times C$, where C denotes channel number (proper reshaping is performed when necessary). $\otimes$ denotes matrix multiplication, and $\oplus$ denotes element-wise sum. The softmax operation is performed on each row. The blue

TABLE I

MODELS' PREDICTION ACCURACY AND SCORE

Models	All range				Long range				Score
	Top-L/10	Top-L/5	Top-L/2	Top-L	Top-L/10	Top-L/5	Top-L/2	Top-L	
DeepCov [5] (baseline)	0.6947	0.6763	0.6282	0.5747	0.4672	0.4453	0.4057	0.3710	16.29
ResPRE [6] (baseline)	0.6841	0.6731	0.6258	0.5670	0.4687	0.4499	0.4013	0.3577	16.19
Cbam-FC-ResPRE (ours)	0.7534	0.7269	0.6639	**0.5942**	0.4848	0.4576	0.4144	0.3775	16.97
SE-FC-ResPRE (ours)	0.7120	0.6928	0.6426	0.5839	0.5040	0.4741	**0.4329**	**0.3939**	17.06
Halo-ResPRE (ours)	**0.7590**	**0.7336**	**0.6680**	0.5922	**0.5054**	**0.4750**	0.4250	0.3797	**17.28**



(a) Ground Truth (b) DeepCov [5] (baseline) (c) ResPRE [6] (baseline)

(d) Cbam-FC-ResPRE (ours) (e) SE-FC-ResPRE (ours) (f) Halo-ResPRE (ours)

Fig. 5. Visualization of the prediction of different models on a sample in testing set

boxes with 1×1 denote 1×1 convolutions. Here we show the embedded Gaussian version, with a bottleneck of $C/2$ channels. The non-local neural network's structure can be discribed as following notations:

$$Z = h(y; W_z) + x \tag{8}$$

In our work, we calculate y as follows.

$$y = softmax(h^T(x; W_\theta) h(x; W_\phi)) h(x; W_g) \tag{9}$$

where function $h(x; W)$ is a convolution operation with input x and weight matrix W.

IV. EXPERIMENTS

A. Dataset preprocess

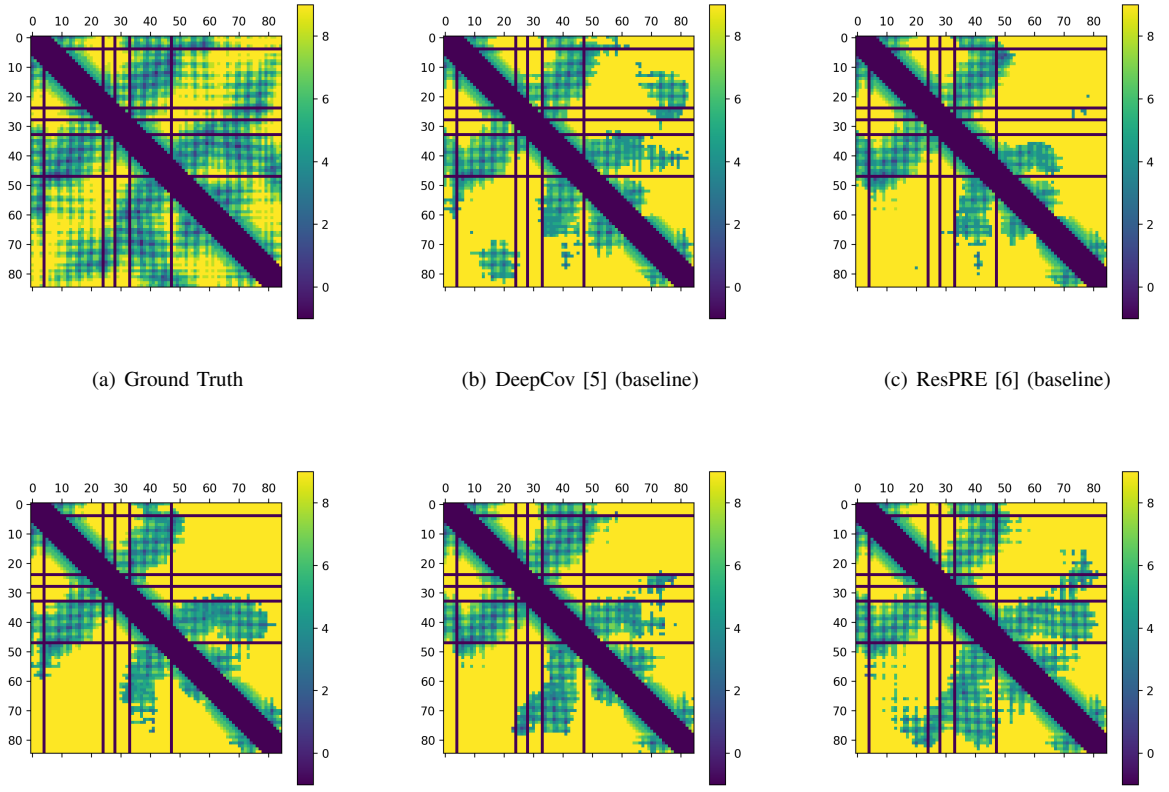
The given dataset contains the origin MSA dataset and the precision feature matrix (PRE) set of MSA. Due to insufficient computing power, we decide to use the precision feature matrix directly.

Here, although the data set is unbalanced, the number of long-range residue pairs is much higher than the number of short-range residue pairs, but the study [6] has verified that it's unnecessary to change the distribution or the weights of residual pairs. Therefore, we keep the original distribution and set the same weights.

We divide the dataset into 3 parts:

- Training set: 80%
- Validation set: 10%
- Testing set: 10%

Training set is used to train the models, while validation set is used to select the model, and testing set is used to evaluate the performance of the model.



(a) Ground Truth (b) DeepCov [5] (baseline) (c) ResPRE [6] (baseline)

(d) Cbam-FC-ResPRE (ours) (e) SE-FC-ResPRE (ours) (f) Halo-ResPRE (ours)

Fig. 6. The details in the bottom-right corner of different models' visualization results

B. Experiments

On the divided dataset, we train our proposed models with 50 epochs. We use AdamW optimizer [10] with a learning rate of 0.005. Multi-step learning rate scheduler is used with gamma$=0.2$ and milestones$=[15,25,35,45]$ to perform learning rate decay in order to improve the performance.

However, we notice that the complexity of NL-ResPRE is $o(N \cdot L^6)$. We don't have enough time and GPU memory to finish the entire training process, so in the experiments, we train the rest models except the NL-ResPRE model.

We evaluate both the all-range metrics and the long-range metrics, including top-L, top-L/2, top-L/5 and top-L/10. And a score is defined based on these metrics as follows.

$$\begin{aligned} S = &\left(3 \times T^1 + 2 \times T^2 + 5 \times T^5 + T^{10}\right) \\ &+ \left(3 \times LT^1 + 2 \times LT^2 + 5 \times LT^5 + LT^{10}\right) \end{aligned} \tag{10}$$

The quantitative results are shown in Tab. I. Compared with the baseline DeepCov model [5], our proposed models with the attention mechanism have improved prediction accuracy in almost all distance ranges. Cbam-ResPRE gains the best accuracy on all range top-L prediction and SE-ResPRE shows the best on long range top-L/2 and long range top-L predictions. Halo-ResPRE achieves the best results in all remaining places. It is worth noticing that the halo attention mechanism can effectively capture both short-range and long-range information at the same time, increasing the all-range accuracy by about 10%.

We visualize the prediction of our models on a sample in the testing set, as illustrated in Fig. 5. From the bottom-right corner of the contact-map, we can easily found the obvious differences between the models as shown in the zoomed-up images shown in Fig. 6. The prediction of the ResPRE model misses a big part of contaction, while our attention-based methods can fill in some of the missing parts here by observing the features in a wider inception field. It is worth mentioning that our Halo-ResPRE model can fill in most of the missing parts in the contact-map and achieve a satisfactory results comparing to the ground-truth, which shows the effectiveness of our attention-based method on protein contact-map prediction task.

C. Ablation study

We noticed that the last layer of ResPRE is a 3×3 convolutional layer. As we mentioned above, according to the idea of FCN [9], we replaced it with a 1×1 convolution and proposed our models. Now we are going to conduct the experiments to verify its effectiveness. We believe that although theorectically speaking, its representation ability

TABLE II
ACCURACY AND SCORE IN ABLATION EXPERIMENT

Models	All range				Long range				Score
	Top-L/10	Top-L/5	Top-L/2	Top-L	Top-L/10	Top-L/5	Top-L/2	Top-L	
ResPRE [6]	0.6841	0.6731	0.6258	0.5670	0.4687	0.4499	0.4013	0.3577	16.19
FC-ResPRE	**0.6999**	**0.6808**	**0.6331**	**0.5783**	**0.4990**	**0.4665**	**0.4147**	**0.3747**	**16.68**
Cbam-ResPRE	0.6749	0.6552	0.6198	0.5793	**0.4973**	**0.4714**	**0.4284**	**0.3878**	16.68
Cbam-FC-ResPRE	**0.7534**	**0.7269**	**0.6639**	**0.5942**	0.4848	0.4576	0.4144	0.3775	**16.97**
SE-ResPRE	0.7094	0.6819	0.6303	0.5793	0.4934	0.4676	0.4304	0.3938	16.86
SE-FC-ResPRE	**0.7120**	**0.6928**	**0.6426**	**0.5839**	**0.5040**	**0.4741**	**0.4329**	**0.3939**	**17.06**

might be slightly weakened, it maybe easier to reach the convergence of the network.

We use the same network structure as SE-FC-ResPRE and Cbam-FC-ResPRE, except changing the last convolution block from 1×1 to 3×3, to build the SE-ResPRE model and the Cbam-ResPRE models. We use the same network structure as ResPRE, except changing the last convolution block from 3×3 to 1×1, to build the FC-ResPRE model.

To verify our hypothesis, we conduct the following ablation experiments. As shown in Tab. II, the performances of kernel size 1×1 are generally much better than 3×3 in the last convolution block, which shows that our modification has a significant improvement in the accuracy of the models. Our explanation is that the last convolution block only acts as a role of merging features to predictions, which does not need wider inception fields; hence the 1×1 convolution block, which has fewer model parameters, is enough and also easy to converge, resulting in the performance improvement.

V. CONCLUSION

Our proposed protein contact map prediction models take advantage of attention mechanism to focus on the intrinsic related information of some long-range residue pairs. Based on the performance of the existing residual neural network, our models have considerable improvements in the prediction accuracy of both long-range and all-range. By taking ablation experiments, we verified the ability of the using of 1×1 kernel size in the last convolutional layer for better convergence and performance improvements. In practice, the loss of our proposed Halo-ResPRE model can drop quickly, that is, this model can converge more quickly on the same data set. We have also established complete documentation for our open-source attention-based protein contact-map prediction framework, which makes it easier for beginners. Our framework and models are published at https://github.com/Galaxies99/alpha-protein.

ACKNOWLEDGEMENTS

Here, we would like to express our deepest gratitude to Dr. Yang for the machine learning and computational biology knowledge she taught. And thanks to all teaching assistants, especially our mentor Peidong Yu. They put forward a lot of suggestions and support for our work.

REFERENCES

[1] P. Di Lena, K. Nagata, and P. Baldi. Deep architectures for protein contact map prediction. *Bioinformatics*, 28(19):2449–2457, 2012.

[2] H. Fukuda and K. Tomii. Deepeca: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC bioinformatics*, 21(1):1–15, 2020.

[3] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[4] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv:1506.02025*, 2015.

[5] D. T. Jones and S. M. Kandathil. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, 34(19):3308–3315, 2018.

[6] 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*, 35(22):4647–4655, 2019.

[7] Y. Li, C. Zhang, E. W. Bell, D.-J. Yu, and Y. Zhang. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in casp13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1082–1091, 2019.

[8] Z. Li, Y. Lin, A. Elofsson, and Y. Yao. Protein contact map prediction based on resnet and densenet. *BioMed research international*, 2020, 2020.

[9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[10] I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. 2018.

[11] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens. Scaling local self-attention for parameter efficient visual backbones. *arXiv preprint arXiv:2103.12731*, 2021.

[12] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[13] 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 computational biology*, 13(1):e1005324, 2017.

[14] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[16] H. Yang, M. Wang, Z. Yu, X.-M. Zhao, and A. Li. Gancon: protein contact map prediction with deep generative adversarial network. *IEEE Access*, 8:80899–80907, 2020.