Contact-Distil: Boosting Low Homologous Protein Contact Map Prediction by Self-Supervised Distillation

Qin Wang[134†], Jiayang Chen [2†], Yuzhe Zhou [134], Yu Li [2], Liangzhen Zheng [5], Sheng Wang [5], Zhen Li[134*], Shuguang Cui [134]

[1] The Chinese University of Hong Kong(Shenzhen) [2] The Chinese University of Hong Kong
[3] The Future Network of Intelligence Institute (FNii) [4] Shenzhen Research Institute of Big Data [5] Shanghai Zelixir Biotech
{qinwang1@link., lizhen@}cuhk.edu.cn, realbigws@gmail.com

Abstract

Accurate protein contact map prediction (PCMP) is essential for precise protein structure estimation and further biological studies. Recent works achieve significant performance on this task with high quality multiple sequence alignment (MSA). However, the PCMP accuracy drops dramatically while only poor MSA (e.g., absolute MSA count less than 10) is available. Therefore, in this paper, we propose the **Contact-Distil** to improve the low homologous PCMP accuracy through knowledge distillation on a self-supervised model. Particularly, two pre-trained transformers are exploited to learn the high quality and low quality MSA representation in parallel for the teacher and student model correspondingly. Besides, the co-evolution information is further extracted from pure sequence through a pretrained ESM-1b model, which provides auxiliary knowledge to improve student performance. Extensive experiments show Contact-Distil outperforms previous state-of-the-arts by large margins on CAMEO-L dataset for low homologous PCMP, i.e., around **13.3%** and **9.5%** improvements against Alphafold2 and MSA Transformer respectively when MSA count less than 10.

Introduction

Protein structure estimation shows the great importance of drug design, vaccine development and other biological function studies. The problem of obtaining highly accurate 3D protein folding remains challenging. Nowadays, there are two major routes to obtain the structure prediction which are experimental method and computer-aid algorithm. For instance, in experimental approaches, (Noble, Endicott, and Johnson 2004; Wuthrich 1989; Wang et al. 2015) utilize X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy (cryo-EM) to solve this issue. However, those biology practical methods are either time consummation or labor-intensive. Tremendous economic cost further limits the development of those approaches especially for cryo-EM. Therefore, to get rid of those limitations, computer-aid approach (Mandell and Kortemme 2009) is attracted by researchers' attention. Take advantage of deep learning algorithm development, (Li and Yu 2016; Wang et al. 2016; Zhou and Troyanskaya 2014) resolve protein structure by exploiting finely designed neural networks. Nevertheless, the accurate protein contact map prediction (PCMP) is the key to massive improvement of the folding accuracy in (Källberg et al. 2014). Particularly, PCMP represents whether contact or not for any two-residue pair in the 3D protein folding space. In (Tegge et al. 2009), NNcon is proposed to resolve PCMP by 2D-recursive neural networks. To further improve PCMP performance, (Wang et al. 2017) proposes a very deep residual model to achieve the protein contact map which exploits multiple 2D residual convolution layers to learn from 1D sequential features.

Recently, MSA transformer (Rao et al. 2021) extracts the embedding features from representation learning as the auxiliary features to improve the downstream PCMP accuracy.

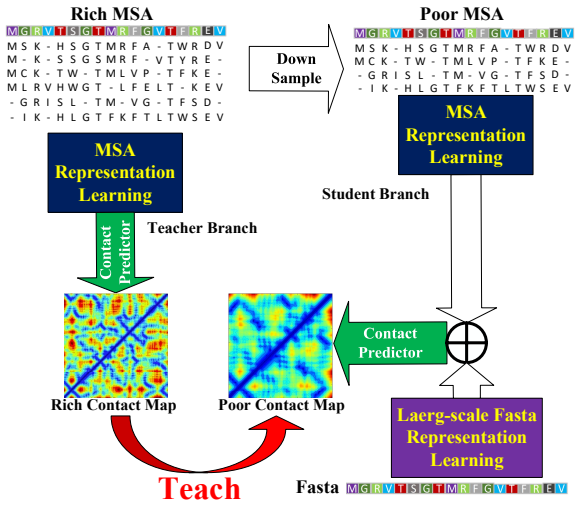

Figure 1: Contact-Distil for low homologous PCMP. The model is trained in a self-supervised manner. The teacher branch accepts rich MSA and student branch takes down-sampled poor MSA as the input. Each branch adopts a pretained representation learning model to provide co-evolution features. Knowledge distillation loss aims to align the poor features from the student to the teacher rich ones. Besides, the student branch exploits pseudo knowledge from large-scale pretrained ESM-1b to compensate for the extremely low homologous proteins with poor MSAs.

*Corresponding author. † Equal first authorship.

But it simply utilizes embedding knowledge from the transformer as an input to optimize deep residual networks regardless of jointly optimizing bert with PCMP improvement. Alphafold2 (Jumper et al. 2021) achieves state-of-the-art performance and largely improves the accuracy of supervised protein structure prediction and PCMP. However, all those approaches require highly rich multiple sequence alignment(MSA) as additional features *i.e.* , input sequences with MSA count > 500 for their model. Nevertheless, low homologous proteins usually have poor MSA features whose MSA count < 10 and cannot provide enough rich features to predict accurate PCMP. Hence, the current state-of-the-art methods are failed to cope with those proteins with low quality MSA *e.g.* , merely 50% accuracy for Alphafold2 and low homologous PCMP still remains challenging.

Therefore, in this paper, we propose a framework with knowledge distillation to tackle this issue by exploiting teacher and student modules to learn the high and low homologous MSA features separately as shown in Fig. 1. More specifically, we utilize a MSA transformer as the feature extractor to provide pseudo homologous features for downstream PCMP task which is pretrained on 26 million MSAs in (Rao et al. 2021). Two successive contact predictors are followed behind MSA transformers to predict the PCMP outputs in teacher and student modules respectively. Therefore, we jointly optimize the downstream PCMP task and upstreaming representation learning models, which aims to improve the final PCMP accuracy. A self-supervised manner is adopted in the training phase. Particularly, the teacher module accepts high homologous MSAs as the input and student module is fed with low homologous ones which are downsampled from rich MSAs. A knowledge distillation loss is adopted between the teacher module and student module to teach the low homologous features towards the high homologous ones and minimize the domain gap between two kinds of MSAs. It is worth to mention that another large-scale dataset pretrained ESM-1b model (Rives et al. 2021) is exploited to provide fake homologous knowledge to improve extremely low homologous PCMPs *i.e.* , proteins with MSA count $= 1$. Different from MSA transformer, ESM-1b is trained on a more larger dataset and takes pure sequence as input rather than MSA input in (Rao et al. 2021). Therefore, for proteins with extremely low quality MSAs, it can produce pseudo co-evolution knowledge which can compensate for the input poor MSA to boost the performance.

Through extensive comparison experiments on two public available datasets trRosetta (Yang et al. 2020) and CAMEO-L, our Contact-Distil achieves state-of-the-art performance on low homologous PCMPs which surpasses previous best methods Alphafold2 and MSA transformer with a large margin *i.e.* , $\sim 10\%$. Benefit from pseudo co-evolution knowledge of ESM-1b, we achieve 9.9% and 37.4% improvement against MSA transformer and Alphafold2 respectively on extremely low homologous proteins whose MSA count is equal to 1. Moreover, finely detailed ablation studies are conducted to examine each proposed component is necessary.

In summary, our main contributions can be concluded into 3 foldings.

- We propose Contact-Distil to solve the low homologous protein contact map prediction (PCMP) which consists of a teacher module and a student module. Each module contains one pretrained MSA transformer to provide co-evolution features and is jointly optimized with contact predictor towards PCMP accuracy improvement.
- To compensate for the extremely low homologous MSA, we import prior knowledge from a pretrained ESM-1b model which is optimized on 0.25 billion proteins for representation learning. To evaluate the performance of proposed approach we further release a new low homologous protein dataset CAMEO-L which is the first dataset for low homologous PCMP evaluation.
- Extensive experiments demonstrate the superiority of proposed Contact-Distil which achieves 13.3% improvement against previous state-of-the-art model Alphafold2 (Jumper et al. 2021) and 9.5% gains against MSA transformer (Rao et al. 2021) on low homologous proteins in CAMEO-L dataset.

Related Works

MSA for Protein Structure Prediction

Given a target protein sequence, multiple sequence alignment (MSA) is a batch of sequences which are homologous with the target sequence and obtained by searching on the protein cluster database such as Uniref90 and Uniref50 (Consortium 2010). Exploiting MSA to boost the protein structure prediction is commonly used in recent researches (Li and Yu 2016; Wang et al. 2017; Wang and et al. 2021b,a; Rao et al. 2019) and achieves significant accuracy improvement.

Therefore, the quality of MSA plays an essential role for protein structure estimation. For instance, rich MSAs with count > 2000 usually perform well and poor MSAs with count < 10 cannot achieve satisfactory results. The profile and PSSM are two kinds of most frequently used features to quantize the MSA which converts a batch of sequences (*i.e.* , MSA) to a statistical matrix with the fixed shape $L \times 20$ where L indicates the sequence length and 20 means the number of amino acid categories. (Zahiri et al. 2013) is firstly to introduce PSSM from MSA to improve the protein structure prediction performance. Profile can be calculated by Eq. 1 where F_i is a vector with length 20 to represent the frequency of each amino acid occurrence at residue i.

$$profile_i = \frac{F_i}{\sum F_i} \tag{1}$$

Hence, profile can be regarded as normalization of frequency map which is counted from MSA. In this paper, we utilize profile to extract homologous knowledge from MSA.

Low Homologous Protein Structure Prediction

Recently, Alphafold (AlQuraishi 2019) and Alphafold2 (Jumper et al. 2021) largely improve the folding accuracy of protein structures. However, those approaches

still require rich MSA features to provide co-evolution knowledge. (Guo and et al. 2020) firstly proposes a self-supervised approach to predict an enhanced PSSM from low quality PSSM to improve the protein secondary structure prediction. (Wang and et al. 2021b) further introduces knowledge distillation and contrastive learning to jointly optimize the enhanced network and secondary structure predictor. Prior knowledge from ESM-1b is firstly utilized in (Wang and et al. 2021a) and aggregated with original low quality MSA to further boost the low homologous secondary structure accuracy. However, those approaches only focus on protein secondary structure regardless of more essential structures such as protein distance and contract maps. MSA transformer is proposed in (Rao et al. 2021) to learn the co-evolution knowledge through representation learning on MSAs by row and column attention which achieves relatively good quality contact maps. Nevertheless, MSA transformer fails on the extremely low quality MSAs *i.e.* , MSA count equal to 1 which cannot obtain satisfactory performance. Besides, for supervised PCMP part, MSA transformer only utilizes embedding features from pretrained BERT as the input for downstream PCMP task without jointly optimize the transformer towards PCMP performance improvement.

Method

Contact Distillation

To tackle the low homologous PCMP issue, we present Contact-Distil which consists of a teacher module and a student module as shown in green and blue part respectively of Fig. 2. The teacher module aims to learn high quality features from high homologous MSA with corresponding high quality transformer and profile embedding features. The knowledge distillation loss is applied at the tail of two branches by utilizing high quality prediction of teacher module to teach student module which learns from downsampled poor MSA. In inference phase, student module will be only exploited to predict contact maps from real natural poor MSAs.

More specifically, the teacher module accepts rich MSA as input and utilizes a pretrained MSA transformer (Rao et al. 2021) which is optimized on 26 million MSAs to provide co-evolution knowledge. A profile net is exploited to transform the input profile to profile embedding features which can suppress the redundant residue column feature and amplify the essential features adaptively as shown in yellow part of Fig. 2 where T means the transform function as shown in Eq. 2. And N is a predicted vector by profile net as shown in yellow part of Fig. 2.

$$Prof.embedding = \frac{1 - profile^N}{N} \tag{2}$$

As shown in the blue part in Fig. 2, the student module takes same architecture with teacher module except introducing pseudo co-evolution knowledge from ESM-1b model as auxiliary information to compensate for the poor MSA especially for the extremely low quality cases *i.e.* , MSA count=1. ESM-1b (Rives et al. 2021) is a large-scale dataset pretrained BERT which utilizes 0.25 billiton proteins for the training. In practice, we extract pseudo profile from ESM-1b to represent the pseudo co-evolution knowledge to help the latter PCMP predictor. Similar with teacher module, a profile net is adopted to transform the pseudo profile to pseudo profile embedding feature which will be concatenated with transformer features from downsampled poor MSAs for final contact map prediction.

The contact predictor is utilized to predict the final contact map in both two modules whose architecture is shown as the blue part of Fig. 2. Particularly, the transformer embedding with shape $L \times 768$ is first reduced into a $L \times 128$ features by a 1-D convolution layer which will be inner concatenated with profile embedding to form a $L \times 148$ feature as shown with green and yellow arrows in the blue part of Fig. 2. Then the fused feature will outer concatenate with itself to form $L^2 \times 296$ feature maps for the input of Resnet32 as illustrated with green arrows of blue part in Fig. 2.

Pseudo Profile Generation

Pseudo profile is generated from ESM-1b (Rives et al. 2021) to provide auxiliary knowledge for compensation of poor MSAs towards PCMP performance improvement. We utilize the same approach with (Wang and et al. 2021a) to produce pseudo profile which will mask the each residue in the protein sequence and exploit remain tokens to predict the categories of masked reside. Here the prediction vector with length 20 after softmax layer can be regarded as one column of the pseudo profile for the masked residue position. Therefore, for a protein with length L, by masking each residue and predicting its soft label (*i.e.* ,the prediction vector with length 20), we can obtain a probability matrix with shape $20 \times L$ to represent the pseudo profile and boost PCMP performance with contact predictor.

Model Optimization and Loss Function

First, two MSA transformers will load the pretrain weights in (Rao et al. 2021). The parameters of ESM-1b are fixed to extract features to generate pseudo profile, while another two pretrained MSA transformers in teacher and student module will be jointly optimized with PCMP task in an end-to-end manner by smaller learning rates. We first utilize the samples with rich MSA to train the teacher module F_t for PCMP with cross-entropy (CE) loss. To train the student module F_s, we fix the weights of whole teacher module and extract the final probability maps from the contact predictor of the teacher module to teach the student module by a knowledge distillation loss $\mathcal{L}$ as shown in Eq. 3 where KL is Kullback-Leibler Divergence to distil the knowledge from teacher to student. M_h is the high quality MSA whose downsampled version is denoted as M_l. P_h indicates the high quality profile which can be counted from high quality MSA M_h while P_b is the pseudo profile which can be generated from sequence S by ESM-1b as shown in previous section. The contact map ground truth is denoted as Y in Eq. 3. More specific descriptions can be found in Algorithm. 1.

$$\mathcal{L} = KL(F_s(M_l, P_p), F_t(M_h, P_h)) + CE(F_s(M_l, P_p), Y) \tag{3}$$

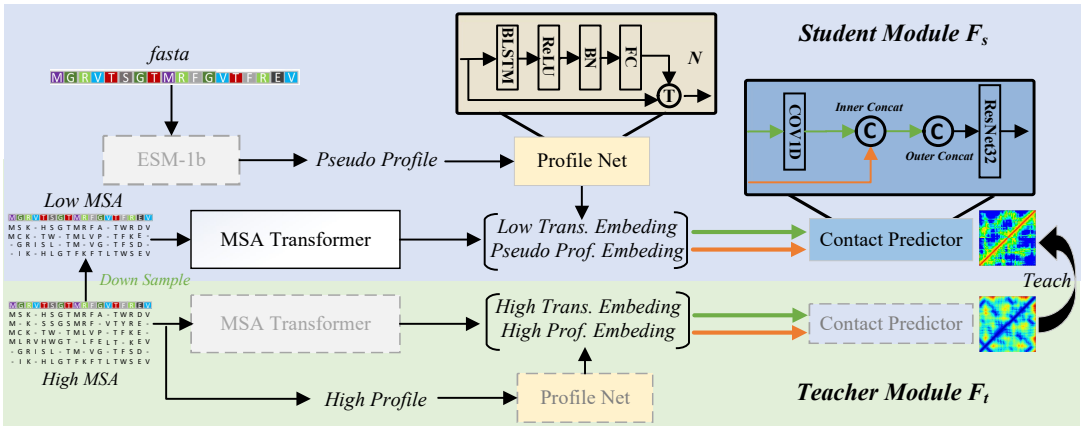

Figure 2: The overview of proposed Contact-Distil framework. The teacher module is illustrated in green color which aims to learn features from high quality MSA input. Once teacher module is well optimized, its knowledge can be transferred to student module which accepts low quality MSA as input as shown in blue color. Specifically, a knowledge distillation loss is well designed for the knowledge transfer from teacher module to student module. To compensate for the extremely poor MSA *i.e.* , MSA count=1, the pseudo profile is extracted from a pure protein sequence input with a ESM-1b model which is pretrained on 0.25 billion proteins for co-evolution learning. Moreover, a profile net is exploited to transform original profile to profile embedding which can amplify the discriminative features and suppress poor ones. Given the profile and transformer embedings, the contact predictor accepts those features and aims to predict the final protein contact map as shown in blue part.

Algorithm 1: Contact-Distil for PCMP

Input: Protein Sequence S; High-quality Profile $\mathbf{P}_h$; Low-quality MSA $\mathbf{M}_l$, High-quality MSA $\mathbf{M}_h$, Label Y;

Parameters : Student Module F_s; Teacher Module F_t;

`/* Training Phase */`

Teacher Module $F_t \xleftarrow{Pretrain} M_h, Y$;

$\mathbf{M}_l \xleftarrow{MSADownsample} M_h$;

$\mathbf{P}_b \xleftarrow{ESM-1b}$ Get Pseudo Profile From S;

$\mathbf{F}_t(M_h, P_h) \xleftarrow{Predict}$ High MSA M_h, P_h;

$\mathbf{F}_s(M_l, P_b) \xleftarrow{Predict}$ Low MSA M_l, P_b;

`/* Minimize KD loss Eq. 3 on` F_s `*/`

$F_s \xleftarrow{Minimize\ \mathcal{L}} \mathbf{F}_s(M_l, P_b), \mathbf{F}_t(M_h, P_h), Y$;

`/* Inference Phase */`

$\mathbf{P}_b \xleftarrow{ESM-1b}$ Get Pseudo Profile From S;

$X \leftarrow F_s(M_l, P_b)$;

$Contact \leftarrow Argmax(X)$;

Output: Parameters of F_t, F_s

Experiment

Implementation Details

The Contact-Distil is implemented by Pytorch[1] and Ignite. The source code and dataset are released [2]. 4 Nvidia V100 GPU cards are utilized to optimize the model. The pretrained parameters are loaded from the released models[3] in (Rives et al. 2021) and (Rao et al. 2021). In practice, due to GPU memory limitation, we randomly select up to 2^{14} tokens from high-MSA M_h to train teacher. To train student with teacher of the best performance, we select top 100 sequences in M_h with max-hamming distance strategy to forward teacher. The initial learning rates (LR) for MSA transformer and contact predictor are 10^{-5} and 10^{-4} respectively, and a cosine learning rate schedule with 2-epoch warming up steps. To evaluate Alphafold2 with low quality MSAs, we load pretrained weights from (Jumper et al. 2021) and turn off the MSA template search which only utilizes the given poor MSA as input. The Alphafold2 PCMP is extracted from its output PDB file by biopython[4].

Network Architecture

The student module contains several components including two pretrained transformers, a profile net and a contact predictor which are shown in Fig. 2. Only MSA transformer is optimized in the training phase while ESM-1b is only utilized to generate pseudo profile. Two transformers take the same network design as (Rives et al. 2021; Rao et al. 2021) and load pretrained weights from them. The profile net adopts a BiLSTM (Wang, Zhang, and Wang 2017) to predict transform parameter N in Eq. 2 to refine the original profile by a residue-wise transformation whose architecture is shown in yellow part of Fig. 2. In contact predictor, a 1D-CNN layer is utilized in contact predictor to reduce the channel size of transformer embedding from 768 to 128 which can be concatenated with profile embedding to form 148-channel features. As same in (Rives et al. 2021), outer concatenation is performed to obtain 2D features with shape $L^2 \times 296$. Finally, a ResNet32 is applied to predict the final PCMP output as shown in the blue part of Fig. 2. The

[1]https://www.pytorch.org

[2]https://github.com/qinwang-ai/Contact-Distil

[3]https://github.com/facebookresearch/esm

[4]https://biopython.org

contact predictor of teacher module has the same architecture with student one while teacher module only contains MSA transformer and contact predictor without profile net and ESM-1b.

Dataset

Two public-available datasets are utilized to examine the performance of Contact-Distil with other approaches.

trRosetta It is first proposed in (Yang et al. 2020) for protein structure prediction evaluation which consists of 15051 proteins. In (Rao et al. 2021), MSA transformer also utilizes this dataset to evaluate the supervised PCMP performance. Therefore, for a fair comparison, we exploit this dataset to conduct the comparison experiments as well. We extract the contact map ground truth from its PDB files on its website [5]. We randomly divide the 15051 proteins into training set and validation set according to the ratio 8:2 respectively. Uniref90 (Consortium 2010) cluster with date July 2019 is utilized for MSA searching. We randomly downsample the searched MSAs to form the low homologous validation set. The training set is utilized to optimize the proposed Contact-Distil. Once the model is well optimized, we will evaluate PCMP performance on the low homologous validation set.

CAMEO-L We randomly downsample proteins of previous half-year in 2021 on CAMEO dataset (Haas et al. 2013) to construct CAMEO-L for the evaluation of low quality contact map prediction. To the best of our knowledge, this is the first public-available dataset to evaluate low homologous PCMP. It contains 339 proteins and each of them has no more than 10 homologous sequences which are randomly selected from its original MSA. Particularly, the distribution of MSA count is uniform and it contains 52 proteins with extremely low quality MSA *i.e.* , count=1. The original MSAs of those proteins are obtained by searching on Uniref90 (Consortium 2010) cluster with date July 2019.

Method	MSA.C	Num	Top L/2	Top L/5	Top L
AF2			0.192	0.213	0.140
MSA.T	≤ 1	304	0.205	0.270	0.161
Ours			**0.546**	**0.683**	**0.414**
AF2			0.388	0.400	0.330
MSA.T	≤ 3	922	0.344	0.441	0.263
Ours			**0.630**	**0.754**	**0.488**
AF2			0.608	0.618	0.558
MSA.T	≤ 10	3011	0.543	0.663	0.416
Ours			**0.718**	**0.828**	**0.571**

Table 1: The comparison with previous best methods at various MSA count partitions on trRosetta dataset. By the comparison, Contact-Distil significantly surpasses previous best methods on different metrics for low homologous PCMP especially for extremely low quality MSAs whose sequence count is less equal than 1.

[5]https://yanglab.nankai.edu.cn/trRosetta/benchmark/

Result

Extensive experiments are conducted to evaluate the performance of proposed Contact-Distil which consists of comparison experiments and ablation studies.

Comparison Experiment The comparison experiment is implemented to compare Contact-Distil with previous state-of-the-art methods such as MSA transformer (Rao et al. 2021) and Alphafold2 (Jumper et al. 2021) for low homologous PCMP on two public-available datasets: tRosetta and CAMEO-L. Tab. 1 illustrates the comparison between Contact-Distil with MSA transformer and Alphafold2 for each MSA count partition *e.g.* ,$\leq 1, \leq 3, \leq 10$ on validation set of trRosetta. 'Top-L/x' indicates metric filtering which only counts on reside pairs within distance ratio L/x. 'AF2' is Alphafold2 (Jumper et al. 2021) and 'MSA.T' is MSA transformer (Rao et al. 2021). The proposed Contact-Distil shows superiority against previous state-of-the-art methods MSA transformer and Alphafold2 in each partition as shown in Tab. 1. The evaluation result of CAMEO-L dataset shows the same facts in Tab. 2 It demonstrates a great importance of exploiting knowledge distillation to transfer high quality knowledge from teacher module to help the student module to predict accurate PCMP with low homologous proteins. Especially, for the extremely low homologous proteins *i.e.* , MSA count=1 in CAMEO-L dataset, Contact—Distil surpasses the previous best approaches MSA transformer and Alphafold2 by 9.9% and 37.4% respectively which exactly proves the effectiveness of co-evolution knowledge compensation from pseudo profile with a pretained ESM-1b model. As same in (Wang and et al. 2021b), we further demonstrate the comparison with Meff score partitions and the results on the validation set of trRosetta are show in Tab. 3. On CAMEO-L dataset, the same evidence can be found in Tab. 4, Contact-Distil achieves the highest PCMP accuracy among each partition on two datasets regardless of MSA tranformer and Alphafold2, which shows the superior performance of proposed Contact-Distil against state-of-the-art methods.

Method	MSA.C	Num	Top L/2	Top L/5	Top L
AF2			0.247	0.251	0.190
MSA.T	≤ 1	52	0.593	0.696	0.465
Ours			**0.691**	**0.773**	**0.564**
AF2			0.363	0.382	0.309
MSA.T	≤ 3	155	0.643	0.754	0.504
Ours			**0.730**	**0.814**	**0.594**
AF2			0.514	0.532	0.461
MSA.T	≤ 10	396	0.645	0.768	0.499
Ours			**0.737**	**0.837**	**0.594**

Table 2: The comparison with previous best methods at various MSA count partitions on CAMEO-L dataset. Contact-Distil achieves state-of-the-art performance on CAMEO-L with all metrics*i.e.* , 'Top L/2', 'Top L/5' and 'Top L'.

Ablation Study The ablation study is conducted on the trRosetta validation set in Tab. 5 to demonstrate the each

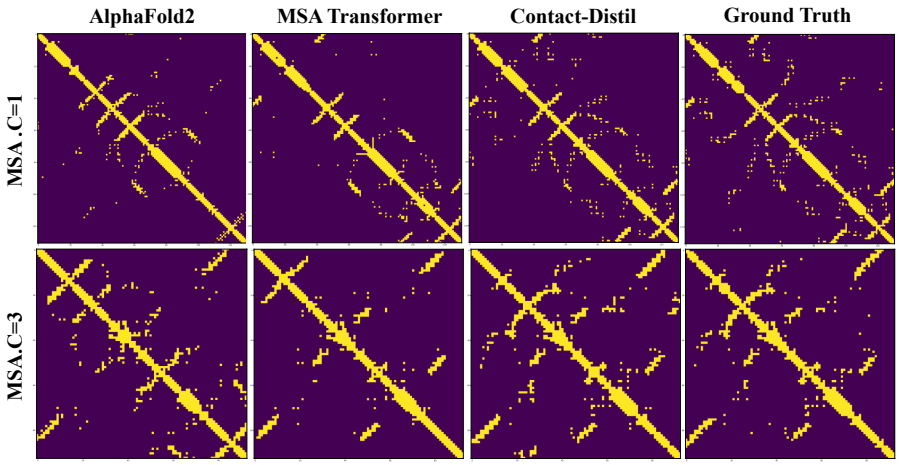

Figure 3: The visualization comparison of PCMP for Contact-Distil, Alphafold2 and MSA Transformer. The last column is ground truth. The comparison qualitatively proves Contact-Distil significantly outperforms the Alphafold2 and MSA Transformer for low homologous proteins.

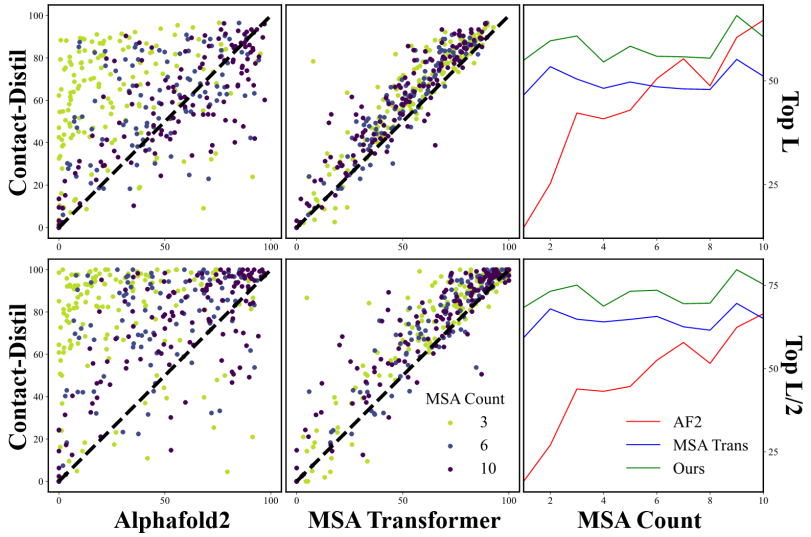

Figure 4: Comparison visualization of Contact-Distil with Alphafold2 and MSA Transformer. The previous two columns show the detailed comparison between Contact-Distil with state-of-the-art methods where the horizontal axis indicates PCMP accuracy and vertical axis represents other methods. Each point indicates a specific protein whose color represents the MSA depth, the brighter the lower. We can clearly observe that most of the points are positioned to left side especially for the extremely low quality cases, which exactly proves the performance superiority of Contact-Distil. The last column shows the accuracy comparison between different methods with various MSA counts. Contact-Distil surpasses other two methods among different MSA counts and is obviously more stable with various MSAs qualities.

component gains against the full model which is denoted as 'Full'. 'wo ESM' indicates the ablation of ESM-1b module which can provide the pseudo co-evolution knowledge by generating pseudo profile to facilitate the student PCMP with poor MSA input. The result in Tab. 5 shows the accuracies of extremely low homologous proteins drop from 0.414 down to 0.243 for 'wo ESM' which exactly shows the great importance of pseudo co-evolution knowledge compensation for extremely low quality MSAs. Nevertheless, in Tab. 5, the full model shows the best performance compare with all ablation models on all MSA count partitions, which exactly illustrates the proposed component in Contact-Distil is necessary and effectiveness.

Qualitative Visualization

More detailed visualization comparisons are shown in Fig. 3 to examine the performance improvement of Contact-Distil against state-of-the-art methods such as Alphafold2 (Jumper

Method	Meff.S	Num	Top L/2	Top L/5	Top L
AF2			0.419	0.422	0.352
MSA.T	≤ 2.8	321	0.364	0.471	0.276
Ours			**0.639**	**0.756**	**0.499**
AF2			0.528	0.537	0.471
MSA.T	≤ 4	963	0.444	0.562	0.337
Ours			**0.671**	**0.787**	**0.526**
AF2			0.656	0.665	**0.606**
MSA.T	≤ 10	2701	0.581	0.708	0.445
Ours			**0.739**	**0.845**	0.589

Table 3: The comparison with previous methods at various meff score partitions on trRosetta validation set and Contact-Distil achieves the highest performance.

Method	Meff.S	Num	Top L/2	Top L/5	Top L
AF2			0.364	0.385	0.304
MSA.T	≤ 2.8	69	0.633	0.733	0.499
Ours			**0.703**	**0.781**	**0.574**
AF2			0.447	0.465	0.395
MSA.T	≤ 4	150	0.638	0.751	0.494
Ours			**0.714**	**0.807**	**0.579**
AF2			0.561	0.579	0.509
MSA.T	≤ 10	337	0.658	0.784	0.510
Ours			**0.750**	**0.852**	**0.604**

Table 4: The comparison at various meff score partitions on CAMEO-L dataset. Across all the subsets, our method outperforms other methods.

et al. 2021) and MSA Transformer (Rao et al. 2021). From the comparison in Fig. 3, we can observe that our prediction is most similar with the ground truth regardless of previous methods for low homologous PCMP, which indicates the Contact-Distil significantly improves PCMP performance through knowledge distillation and exploiting pseudo co-evolution knowledge to boost PCMP performance with extremely low quality MSAs. Anther detailed comparisons are shown in Fig. 4. The horizontal axis indicates other methods and the vertical axis represents Contact-Distil. Each point is a sample whose color means the MSA depth the deeper the higher. From the comparison in Fig. 4, we can clearly notice that most of points are positioned on the right side which shows that Contact-Distil surpasses the other methods on PCMP performance for those proteins. Especially, for the points with shallow color, almost all those samples are positioned in the left part of the graph and exactly examined the effectiveness of knowledge distillation and introducing pseudo profile to improve the extremely low homologous PCMP. The last column also proves the superior performance of proposed Contact-Distil thought comparison with other methods on different MSA counts.

Conclusion

In this paper, we propose a refinement approach for protein contact map prediction with low homologous MSA which consists of knowledge distillation and representation learning. Two pretrained MSA transformers are exploited in the

Method	MSA.C	Num	Top L/2	Top L/5	Top L
wo ESM			0.314	0.408	0.243
wo Distil	≤ 1	304	0.531	0.661	0.407
Full			**0.546**	**0.683**	**0.414**
wo ESM			0.470	0.581	0.361
wo Distil	≤ 3	922	0.626	0.750	0.488
Full			**0.638**	**0.759**	**0.496**

Table 5: The ablation study on trRosetta validation set. As shown in the table, ESM-1b plays an important role in the whole framework. For instance, after removal of ESM-1b, the PCMP accuracy drops from 0.414 down to 0.243 which exactly proves the effectiveness of pseudo knowledge from a large-scale dataset pretrained model.

teacher and student module to extract co-evolution features. The teacher module aims to learn high quality features with high homologous MSAs for PCMP while student module only accepts low homologous MSAs as input. Therefore, a well designed knowledge distillation loss is adopted to transfer the knowledge from teacher module to student module. Different from MSA transformer (Rao et al. 2021), our method jointly optimizes the contact predictor with transformer part rather than only extracting embedings from a fixed BERT. Here, the optimization of transformer is towards PCMP performance improvement instead of merely representation learning. Moreover, a large-scale dataset pretrained ESM-1b model is applied to provide pseudo co-evolution knowledge to compensate for the extremely low homologous protein*i.e.* , MSA count=1. To evaluate the Contact-Distil performance, we further release a public-available dataset CAMEO-L which is the first dataset for low homologous PCMP evaluation. Through extensive comparison experiments, Contact-Distil achieves new state-of-the-art performance among previous methods such as MSA transformer and Alphafold2 for low homologous PCMP*i.e.* , $\sim 10\%$ improvement for MSA count ≤ 10. A detailed ablation study further shows each component of Contact-Distil is necessary. More comprehensive qualitative visualizations further demonstrate the effectiveness of Contact-Distil. Future work should focus on more precise tertiary structures for low homologous proteins such as protein distance map.

Acknowledgements

This work was supported in part by NSFC-Youth 61902335, by Key Area R&D Program of Guangdong Province with grant No.2018B030338001, by the National Key R&D Program of China with grant No.2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No.2017ZT07X152, by Guangdong Regional Joint Fund-Key Projects 2019B1515120039, by the NSFC 61931024&81922046, by zelixir biotechnology company Fund and CCF-Tencent Open Fund, by HPCP of ITSO (CUHSKZ).

References

AlQuraishi, M. 2019. AlphaFold at CASP13. *Bioinformatics*, 35(22): 4862–4865.

Consortium, U. 2010. The universal protein resource (UniProt) in 2010. *Nucleic acids research*, 38(suppl_1): D142–D148.

Guo, Y.; and et al. 2020. Bagging MSA Learning: Enhancing Low-Quality PSSM with Deep Learning for Accurate Protein Structure Property Prediction. In *International Conference on Research in Computational Molecular Biology*, 88–103. Springer.

Haas, J.; Roth, S.; Arnold, K.; Kiefer, F.; Schmidt, T.; Bordoli, L.; and Schwede, T. 2013. The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, 2013.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 1–11.

Källberg, M.; Margaryan, G.; Wang, S.; Ma, J.; and Xu, J. 2014. RaptorX server: a resource for template-based protein structure modeling. In *Protein structure prediction*, 17–27. Springer.

Li, Z.; and Yu, Y. 2016. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *arXiv preprint arXiv:1604.07176*.

Mandell, D. J.; and Kortemme, T. 2009. Computer-aided design of functional protein interactions. *Nature chemical biology*, 5(11): 797–807.

Noble, M. E.; Endicott, J. A.; and Johnson, L. N. 2004. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303(5665): 1800–1805.

Rao, R.; Bhattacharya, N.; Thomas, N.; Duan, Y.; Chen, P.; Canny, J.; Abbeel, P.; and Song, Y. 2019. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, 9686–9698.

Rao, R.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J. F.; Abbeel, P.; Sercu, T.; and Rives, A. 2021. Msa transformer. *bioRxiv*.

Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15).

Tegge, A. N.; Wang, Z.; Eickholt, J.; and Cheng, J. 2009. NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic acids research*, 37(suppl_2): W515–W518.

Wang, J.; Zhang, J.; and Wang, X. 2017. Bilateral LSTM: A two-dimensional long short-term memory model with multiply memory units for short-term cycle time forecasting in re-entrant manufacturing systems. *IEEE Transactions on Industrial Informatics*, 14(2): 748–758.

Wang, Q.; and et al. 2021a. Adaptive Residue-wise Profile Fusion for Low Homologous Protein Secondary Structure Prediction Using External Knowledge. In *Proceedings of the Thirtieth IJCAI*.

Wang, Q.; and et al. 2021b. PSSM-Distil: Protein Secondary Structure Prediction (PSSP) on Low-Quality PSSM by Knowledge Distillation with Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 617–625.

Wang, R. Y.-R.; Kudryashev, M.; Li, X.; Egelman, E. H.; Basler, M.; Cheng, Y.; Baker, D.; and DiMaio, F. 2015. De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nature methods*, 12(4): 335–338.

Wang, S.; Peng, J.; Ma, J.; and Xu, J. 2016. Protein secondary structure prediction using deep convolutional neural fields. *Scientific reports*, 6(1): 1–11.

Wang, S.; Sun, S.; Li, Z.; Zhang, R.; and Xu, J. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, 13(1): e1005324.

Wuthrich, K. 1989. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science*, 243(4887): 45–50.

Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; and Baker, D. 2020. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3): 1496–1503.

Zahiri, J.; Yaghoubi, O.; Mohammad-Noori, M.; Ebrahimpour, R.; and Masoudi-Nejad, A. 2013. PPIevo: Protein–protein interaction prediction from PSSM based evolutionary information. *Genomics*, 102(4): 237–242.

Zhou, J.; and Troyanskaya, O. G. 2014. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. *arXiv preprint arXiv:1403.1347*.