Improving deep learning-based protein distance prediction in CASP14

Zhiye Guo[1&], Tianqi Wu[1&], Jian Liu[1], Jie Hou[2], Jianlin Cheng[1*]

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA

[2]Department of Computer Science, Saint Louis University, Saint. Louis, MO 63103, USA

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

[&]Joint first authors; *corresponding author

Abstract

Motivation: Accurate prediction of residue-residue distances is important for protein structure prediction. We developed several protein distance predictors based on a deep learning distance prediction method and blindly tested them in the 14th Critical Assessment of Protein Structure Prediction (CASP14). The prediction method uses deep residual neural networks with the channel-wise attention mechanism to classify the distance between every two residues into multiple distance intervals. The input features for the deep learning method include co-evolutionary features as well as other sequence-based features derived from multiple sequence alignments (MSAs). Three alignment methods are used with multiple protein sequence/profile databases to generate MSAs for input feature generation. Based on different configurations and training strategies of the deep learning method, five MULTICOM distance predictors were created to participate in the CASP14 experiment.

Results: Benchmarked on 37 hard CASP14 domains, the best performing MULTICOM predictor is ranked 5th out of 30 automated CASP14 distance prediction servers in terms of precision of top L/5 long-range contact predictions (i.e. classifying distances between two residues into two categories: in contact (< 8 Angstrom) and not in contact otherwise) and performs better than the best CASP13 distance prediction method. The best performing MULTICOM predictor is also ranked 6th among automated server predictors in classifying inter-residue distances into 10 distance intervals defined by CASP14 according to the precision of distance classification. The results show that the quality and depth of MSAs depend on alignment methods and sequence databases and have a significant impact on the accuracy of distance prediction. Using larger training datasets and multiple complementary features improves prediction accuracy. However, the number of effective sequences in MSAs is only a weak indicator of the quality of MSAs and the accuracy of predicted distance maps. In contrast, there is a strong correlation between the accuracy of contact/distance predictions and the average probability of the predicted contacts, which can therefore be more effectively used to estimate the confidence of distance predictions and select predicted distance maps.

Availability: The software package, source code, and data of DeepDist2 are freely available at https://github.com/multicom-toolbox/deepdist and https://zenodo.org/record/4712084#.YIIM13VKhQM.

1 Introduction

Accurate prediction of inter-residue distances (or its simplified representation - inter-residue contacts) is critical for template-free (ab initio) tertiary structure prediction, i.e., predicting the structure of a protein without using any known structure as templates (Kryshtafovych, et al., 2019). The predicted inter-residue distances can be translated into tertiary structures by off-shelf tools such as trRosetta (Yang, et al., 2020), CONFOLD2 (Adhikari and Cheng, 2018) built on top of CNS (Brünger, et al., 1998), and DMPfold (Greener, et al., 2019). In the 2018 CASP13 experiment, the top-ranked methods (Hou, et al., 2019; Kandathil, et al., 2019; Senior, et al., 2020; Xu and Wang, 2019; Zheng, et al., 2019) all used distance or contact predictions to guide template-free (FM) structure modeling to achieve significant success. Since then, the inter-residue distance prediction has become a focal point of protein structure prediction.

In the last several years, the advances in protein distance/contact prediction were mostly driven by two technologies: the residue-residue co-evolutionary analysis (Ekeberg, et al., 2013; Kamisetty, et al., 2013; Seemayer, et al., 2014) for generating informative features for prediction and various deep learning methods (Goodfellow, et al., 2013; He, et al., 2016) for effectively extracting protein distance/contact patterns from the features. Since classifying the distances between residues into multiple distance intervals (commonly called distance prediction) can provide more detailed information about residue-residue distances than classifying them into two binary categories - in contact or not in contact (commonly called contact prediction), recent methods such as AlphaFold and RaptorX focused on the distance prediction. The multi-classification or binary classification of distances produces a multi-class or binary-class distance probability map. Most recently, some methods such as DeepDist (Wu, et al., 2020) were developed to predict real-value inter-residue distances using deep learning regression methods, in addition to classifying the distances into multiple distance intervals. Moreover, the attention mechanism that can pick up relevant signals anywhere in the input features was also applied to predict protein contacts and explain the predictions (Chen, et al., 2020). In the CASP14 experiment, the attention mechanism was also used by AlphaFold2, tFold, and our MULTICOM distance predictors to improve distance/structure prediction.

In this work, we describe the design and implementation of our MULTICOM distance predictors based on our DeepDist2 distance prediction method and analyze their results and performance in CASP14. Following the CASP14 norm, the analysis is focused on hard template-free modeling (FM) target domains instead of template-based modeling (TBM) domains that have recognizable known template structures in the Protein Data Bank (PDB) (Berman, et al., 2000). The FM/TBM domains that might have very weak templates that cannot be recognized by existing sequence alignment methods are also used in the evaluation.

2 Materials and Methods

The overall pipeline of the MULTICOM distance predictors based on our latest deep learning method - DeepDist2 is shown in Fig.1. Three methods are used to generate multiple sequence alignments (MSAs) for a target protein in parallel, including our in-house tool – DeepAln (Wu, et al., 2020), DeepMSA (Zhang, et al., 2019), and HHblits (Remmert, et al., 2012). DeepAln and DeepMSA are also used in the original DeepDist method. In CASP14, MULTICOM predictors added the HHblits search against the Big Fantastic Database (BFD) (Steinegger, et al., 2019) (denoted as HHblits_BFD) to generate MSAs when the number sequences in MSAs generated by DeepAln and DeepMSA was less than 10L (L: sequence length).

Each MSA is used to produce multiple co-evolutionary features such as covariance matrix (Jones and Kandathil, 2018), precision matrix (Li, et al., 2019), and pseudolikelihood maximization matrix (Seemayer, et al., 2014). The quality of the co-evolutionary features depends on the depth of MSA (i.e. the number of sequences) as well as the quality of the MSA (e.g., the proportion of true homologous sequences in MSA). For instance, when the number of effective sequences (Neff) in an MSA is too small, the co-evolutionary scores tend to be noisy and less informative (Wu, et al., 2020). To complement the co-evolutionary features, the non-coevolutionary features such as position-specific scoring matrix (PSSM) generated by PSI-BLAST (Bhagwat and Aravind, 2007) and secondary structures predicted by PSIPRED (Jones, 1999) are also used.

Different kinds of co-evolutionary features are combined with non-co-evolutionary features to generate the four sets of features (COV_Set, PRE_Set, PLM_Set, and OTHER_Set; see details in Section 2.2). Each of four sets of features derived from the same MSA is used by a deep residual network with a channel-wise attention mechanism to predict a distance map. The average of the four predicted distance maps is the predicted distance map for the MSA. Different from DeepDist that uses four different deep architectures for different sets of features, DeepDist2 uses the same network architecture for all the feature sets. For most CASP14 targets, the distance maps predicted from the features generated from DeepAln's MSA and DeepMSA's MSA were averaged as the final prediction. When the number of sequences in the combination of MSAs generated by DeepAln and DeepMSA was less than 10 L, the distance map predicted from the MSA of HHblits_BFD was averaged with the distance maps predicted from MSAs of DeepAln and DeepMSA as the final prediction.

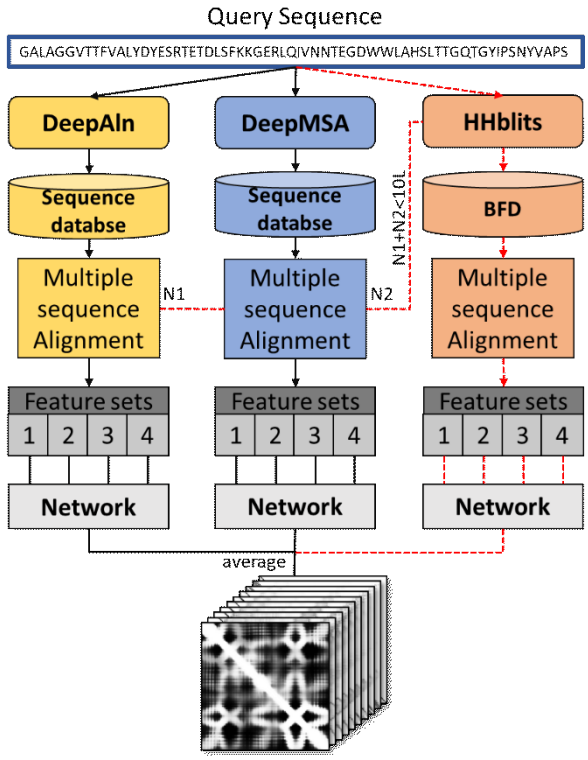

Fig.1. The overall pipeline of the MULTICOM distance predictors based on DeepDist2. The two data flows (branches) applied to all the targets are connected by the black solid line, while the optional flow (branch) is connected by the red dotted line, which is only invoked when the MSAs are produced by DeepMSA and DeepAln are not sufficiently deep. Each flow (branch) produces four sets of features (COV_Set, PRE_Set, PLM_Set, and OTHER_Set; see details in Section 2.2), each of which is used as input for a deep network to predict a distance map. The four distance maps predicted from the four sets of features of each branch are averaged as the predicted distance map of the branch. The final prediction is the average of the predicted distance maps of the first two or all the three branches.

Based on the same protocol above, four automated MULTICOM distance predictors MULTICOM-CONSTRUCT, MULTICOM-AI, MULTICOM-DIST, MULTICOM-HYBRID were trained with different labelings of distance intervals. MULTICOM-DEEP used the average of the four predictors as its prediction. The distance intervals (or bins) of MULTICOM-CONSTRUCT are 0 to 4 Å, 4 to 6 Å, 6 to 8 Å, …, 18 to 20 Å, and > 20 Å. MULTICOM-DIST uses 42 bins, i.e. dividing 2 to 22 Å into 40 bins with a bin size of 0.5 Å, plus 0 - 2 Å bin and > 22 Å bin. MULTICOM-HYBRID shares the same distance segmentation strategy as MULTICOM-DIST, except that it starts with an interval 0 - 3.5 Å and its last interval is set to > 19 Å. MULTICOM-AI has 37 equally spaced intervals of 0.5 Å between 0 to 20 Å and the > 20 Å interval. Though the

predicted multi-class distance prediction maps of the five predictors are based on the different distance intervals, they are converted into the 10-bin classification maps required by CASP14. The 10 bins defined CASP14 are bin1: d≤4Å, bin2: 4<d≤6Å, bin3: 6<d≤8Å, ..., bin10: >20Å, which are the same as MULTICOM-CONSTRUCT.

2.1 Deep residual neural networks with channel-wise attention mechanism for inter-residue distance prediction

The architecture of the deep residual network with the attention mechanism is shown in supplemental **Fig. S1**. The input features (a tensor of L * L * N dimension; L: sequence length; N: number of channels) are first fed into an instance normalization layer (Ulyanov, et al., 2016), followed by a convolutional layer and a Maxout layer (Goodfellow, et al., 2013). The convolutional layer reduces the number of channels to 128 and then the Maxout layer halves it to 64.

Following the Maxout layer are 20 residual blocks with the same input and output dimension of 64. Each residual block starts with a normalization block (called RCIN) that includes three different kinds of normalization layers and one ReLU (Nair and Hinton, 2010) activation function. The three normalization layers of RCIN are row normalization layer (RN), column normalization layer (CN) (Mao, et al., 2019), and instance normalization (IN) layer. The output of the three normalization layers is concatenated as input for a ReLU activation function. Through this operation, the information in multiple directions can be effectively integrated to better capture contacts/distances between residues. The RCIN block is followed by a convolutional layer, an RCIN block, three convolutional layers, an RCIN block, and a convolutional layer. The final part of the residual block is the squeeze-and-excitation block (SE) (Hu, et al., 2018), which is a channel-wise attention method popular in the computer vision field. This block has good adaptability and can be embedded into different deep network architectures. It has two parts: one is the squeeze operation that can collect the global information between all the feature channels and another is the excitation operation that can boost the impact of relevant features by two fully connected layers with the ReLU activation function. The SE block recalibrates the feature channels through learning so that the network can assign more attention to more essential feature channels. We apply a softmax activation to classify inter-residue distances between residues into multiple intervals (bins), i.e. predict the probability distribution of inter-residue distances.

2.2 Multiple sequence alignments and input features

DeepAln and DeepMSA use HHblits and jackhmmer to search several protein sequence datasets to generate MSAs (Wu, et al., 2020). During the CASP14 experiment, all the databases (i.e. UniRef90 (2020-04) (Mirdita, et al., 2017), Uniclust30 (2020-03), Metaclust50 (2018-06) (Steinegger and Söding, 2018), and Myg_UniRef100) used for MSA generation were updated to their latest version. The BFD used by HHblits search was released by March 2019.

The residue-residue co-evolutionary features including covariance matrix (COV), precision matrix (PRE), and pseudolikelihood maximization matrix (PLM) calculated from MSAs are two-dimensional (2D) features with multiple channels, and have a dimension of L×L×441. PSSM generated from PSI-BLAST search against UniRef90 is also a useful feature. Other features like the Pearson's correlation between columns of PSSM, the co-evolutionary contact scores produced by CCMpred, the Shannon entropy sum, mean contact potential, normalized mutual information, and mutual information from DNCON2 are generated. These features are combined to generate four sets of features as follows. (1) COV_Set includes COV, PSSM, Pearson correlation, and CCMpred contact scores; (2) PLM_Set contains PLM, PSSM, and Pearson's correlation; (3) PRE_Set has PRE, PSSM, and entropy scores (joint entropy, Shannon entropy sum); and (4) OTHER_Set has PSSM, CCMpred contact scores, Pearson correlation, solvent accessibility, mean contact potential, normalized mutual information, and mutual information.

2.3 Datasets and evaluation metrics

11,234 proteins used by RaptorX (Xu and Wang, 2019) were employed to train the MULTICOM distance predictors. The proteins may have a single domain or multiple domains. The sequence identity between any two proteins in the dataset is less than 25%. Also, the proteins in the training dataset have less than 25% sequence identity with the proteins in the three test datasets: 43 CASP13 FM and FM/TBM domains, 37 CASP12 FM domains, and 268 CAMEO targets (released between 08/31/2018 and 08/24/2019). The predictors were trained and internally tested on the test datasets before they were blindly tested in CASP14 from May to July 2020.

The evaluation of the MULTICOM distance predictors is based on 37 hard FM and FM/TBM domains of CASP14 (i.e. 23 FM domains and 14 FM/TBM domains). To be consistent with the analysis of CASP14, the evaluation is carried out at the domain-level. The distance predictions are evaluated by three metrics: (1) the precision of top L/5, L/2, or L long-range contact prediction after the multi-class distance predictions are converted to binary contact predictions at 8 Å threshold (L: sequence length), (2) mean absolute error (MAE) between predicted distances and true distances; and (3) the average precision, recall, and F-measure of multi-classification of distances between long-range residue pairs over 10 distance bins.

Two residues are considered in contact if the distance between their β-carbon atoms (α-carbon for the glycine amino acid) is less than 8 Å. A contact map can be obtained by summing up the probability values of the intervals within 0-8 Å in a predicted multi-classification distance map. We use ConEVA (Adhikari, et al., 2016) to calculate the precision of predicted contacts. The CASP14's assessment results at https://predictioncenter.org/casp14/rrc_avrg_results.cgi are also used. A contact is considered long-range contact if the sequence separation between the two residues is >= 24 residues, medium-range if the sequence separation is within [12, 23], and short-range if the sequence separation is within [6, 11]. In this study, the evaluation is mostly focused on long-range residue-residue contact/distance predictions according to the CASP norm.

The real-value distance between two residues is estimated as the sum of the mean distance of each interval times the predicted probability of the interval (i.e. the weighted average). Because large distances contribute little to tertiary structure prediction, only predicted distances less than 16 Å are used for the MAE evaluation. The standard deviation of the MAE is also calculated. When the MAE is close, a smaller standard deviation is preferred.

For the multi-classification prediction, we apply the precision (denoted as Precision_m), recall (denoted as Recall_m) to evaluate the multi-classification of distances between long-range residue pairs. The precision and recall of each distance bin for a target are calculated first. The precision and recall of multiple distance bins is the arithmetic average of precision and recall of each bin over all the bins (see the detailed formula of Precision_m and Recall_m in the supplemental document). Therefore, the final

precision and recall (Precision_m and Recall_m) can evaluate the accuracy of the overall performance of multi-classification of distances for a target. We only calculate the precision and recall of the multi-classification prediction of the distances between long-range residue-residue pairs. The F1-measure is the geometric mean of Precision_m and Recall_m.

3 Results

3.1 Overall performance of distance prediction in CASP14

In this study, we only compare CASP14 server predictors, excluding CASP14 human predictors that had more prediction time and could use some server predictions as input. The performance of the top 20 out of 30 CASP14 automated server predictors on 37 FM and FM/TBM domains in terms of precision of top L/5 long-range contact predictions (called top L/5 precision) is shown in supplemental **Table S1**. The top L/2 precision of the predictors is also reported in the table. The result was compiled from the evaluation data at the CSAP website after excluding human distance predictors. Our best server predictor MULITCOM-CONSTRUCT has a top L/5 precision of 64.99% and is ranked no. 5 after TripletRes from Zhang Group and three tFold servers (tFold-CaT, tFold-IDT, and tFold) from tFold Group. Other MULTICOM predictors are also ranked among the top 20. Moreover, the top L/5 (or L/2) precision of the MULTICOM predictors is higher than RaptorX - the best contact predictor in CASP13, showing that multiple predictors including ours in the CASP14 experiment improve over the best CASP13 contact predictor.

Among the 30 server predictors, 19 of them submitted multi-class distance predictions, while the rest only submitted binary contact predictions. Three of the 19 groups missed some FM and/or FM/TBM targets. Supplemental **Table S2** reports the precision (Precision_m), recall (Recall_m), and F1-measure of multi-classification distance prediction of the 16 predictors that submitted predictions for all 37 FM and FM/TBM domains. Our best server predictor MULTICOM-DEEP is ranked no.6 after two tFold servers (tFold, tFold-CaT), TripletRes from Zhang group, and two servers (FoldX and TOWER) from Microsoft in terms of Precision_m.

The detailed results of the MULTICOM distance predictors (precision of top L/5, L/2, L long-range contact predictions, the mean absolute error and standard deviation of long-range distance predictions, and the precision_m and recall_m of multi-classification of distances) on 37 FM and FM/TBM domains are reported in supplemental **Table S3**. The MULTICOM distance predictors have similar performance. MULTICOM-CONSTRUCT performs best in terms of contact precision, MULTICOM-AI has the lowest MAE, and MULTICOM-DEEP has the highest multi-classification precision.

3.2 Comparison of different MSAs for distance prediction

The performance of deep learning distance predictors depends on the quality of the input features, particularly the most important co-evolutionary features whose quality is largely determined by the depth and quality of MSAs (Wu, et al., 2020).

The depth of an MSA is usually measured by the number of effective sequences (Neff) in the MSA. It is calculated using the formula $Neff = \sum_{i=1}^{N} \frac{1}{Sim_i}$, where N denotes the number of sequences in the MSA and Sim_i the sum of the identity between Sequence *i* and all the sequences in the MSA. Higher similarity (Sim_i), lower weight Sequence *i* contributes to the count of Neff.

Here we use the performance of MULTICOM-CONSTRUCT with three kinds of MSAs on the 37 FM and FM/TBM domains to compare their performance in distance prediction. Supplemental **Table S4** shows the performance of the long-range distance prediction of MULTICOM-CONSTRUCT with MSAs of DeepAln, DeepMSA, and HHblits_BFD according to multiple metrics, including Top L/2 and Top L precisions of long-range contact predictions, mean absolute error of long-range predicted distances < 16 Å (MAE_16) and their standard deviation (STD_16), the accuracy and recall of multi-classification of distances (Precision_m and Recall_m). HHblits_BFD performs best among the three according to all the metrics, DeepMSA works better than DeepAln. For instance, the top L/2 precision of HHbits_BFD is 51.33%, higher than DeepMSA's 46.18% and DeepAln's 43.87%. The reason is that the BFD database (released in April 2019) contains the hidden Markov model (HMM) profiles for the proteins in both UniProt and the metagenomics databases, which enables HHblits to generate high-quality alignments with the sequences in the databases. In contrast, DeepMSA or DeepAln uses HHblits to search the HMM profiles in UniProt and Jackhmmer to the sequences in the metagenomics database. Because Jackhammer's alignment quality and sensitivity are lower than HHblits, even though DeepMSA and DeepAln search a target against a newer version of UniProt and metagenomics databases than the BFD database, the quality gain of HHblits search on the BFD still outweighs the increase of the size of databases used by DeepMSA and DeepAln, leading to the better distance predictions with HHblits_BFD.

To further quantitatively analyze the impact of different MSA generation pipelines on the performance of the distance prediction, we study the relationship between the accuracy of distance prediction and the logarithm of the number of effective sequences (Neff) in the MSAs generated by DeepAln, DeepMSA, and HHblits_BFD in supplemental **Fig. S2**. Because our automatic domain parsing did not predict domains accurately in some cases during CASP14 where the predicted domain boundaries were different from the ground truth, here we only analyze the 31 full-length hard targets in which the 37 FM and FM/TBM domains are located. The Neff and prediction accuracy are calculated on the 31 full-length hard targets.

The correlation coefficients between top L/2 precision and the common logarithm of Neff for DeepAln and DeepMSA are 0.417 and 0.462, respectively. The correlation between the two is not very strong, mainly because some targets have a large Neff but low prediction accuracy due to the existence of the false-positive sequences in their MSAs. 10 (or 9) out of 31 hard targets that have a Neff > 10 for DeepAln (or DeepMSA) have the precision of < 50%. Interestingly, the correlation coefficient between the top L/2 precision and the common logarithm Neff is 0.357 for the HHblits_BFD on all the 31 hard targets, which is even lower than DeepAln and DeepMSA. The correlation coefficients between the precision of the multi-class classification of distances and the logarithm of Neff are 0.373, 0.414, and 0.295 for DeepAln, DeepMSA, and HHblits_BFD, respectively, which is lower than the correlation for the binary contact prediction. The correlation coefficients between the MAE of multi-classification of distances and the common logarithm Neff are -0.488, -0.546, and -0.370 for the DeepAln, DeepMSA, and BFD, respectively. These results show that there is only a weak correlation between Neff and the accuracy of distance predictions for the three MSA generation pipelines (DeepAln, DeepMSA, Hblits_BFD), while the correlation is weakest for HHblits_BFD that generates the MSAs of the best quality.

Therefore, we conclude that both the quality and depth of MSAs impact the accuracy of distance predictions, and the depth measured by Neff is only a weak indicator of the accuracy of distance prediction. Indeed, some CASP14 targets (e.g., T1093) have deep MSAs with a large Neff but get a low distance prediction accuracy. Different from the depth of MSAs that can be measured by a single quantity - Neff, the quality of MSA depends

on alignment accuracy, and relationships between sequences (homologous or not) in MSA are hard to quantify.

3.3 The strong correlation between distance prediction accuracy and predicted probability scores and its application to select/combine predicted distance maps

According to the analysis above, different MSAs generated by different methods may work well on different sets of targets. Therefore, there is a need to find good metrics to select or combine MSAs or distance maps to improve prediction. However, since Neff of MSAs has only a weak correlation with the accuracy of distance/contact prediction, it cannot accurately select MSAs or predicted distance maps. In order to find better metrics to select MSAs and predicted distance maps, we calculate the correlation between the precision of top L/2 long-range contact predictions and the average probability of the top L/2 contact predictions (**Fig. 2**) as folllows. The multi-class distance predictions for a target are converted into binary contact probability predictions by summing up the probabilities of all the bins falling in the interval [0, 8 Å] as contact probability at 8 Å threshold. Top L/2 long-range contact predictions with highest probabilities are selected, and their probabilities are averaged. The correlation between the average probability of top L/2 long-range contact predictions and the precision of top L/2 long-range contact predictions is then calculated. The correlation between the average probability of top L/2 long-range contact predictions and other merics (e.g., the precision of multi-class distance prediction, and the mean absolute error of the real-value distance prediction) can be calculated in the same way. The correlation between the precision of top L/2 long-range contact predictions and their average contact probability is 0.819. Moreover, the average probability also has a relatively strong correlation with the precision of multi-class classification of distances (correlation = 0.654) and the mean absolute error of the real-value distance prediction (correlation = -0.790). The precision of top L/2 long-range contact predictions, the precision of multi-class distance prediction, the mean absolute error of the real-value distance predictiion, and the average probability of top L/2 long-range contact predictions for 31 hard targets and their correlation coefficients are reported in **Table S5**. These correlations are much stronger than that between Neff and contact/distance prediction accuracy.

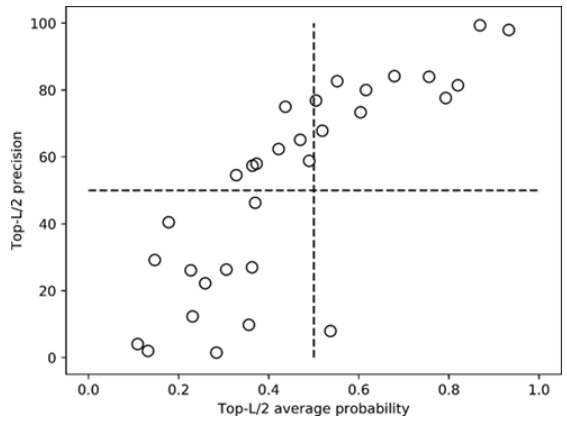

Fig.2. A plot of precisions of top L/2 long-range contact predictions against the average probabilities of the top L/2 predicted contacts. MULTICOM-CONSTRUCT with HHblits_BFD alignments were used to predict the distance maps.

The relatively strong correlation between the predicted contact probabilities and the accuracy of predicted distance maps provides a better approach to select distance maps predicted from different MSAs than Neff. To analyze the effectiveness of this approach for improving distance/contact predictions, we compare it with two approaches of combining MSAs or predicted distance maps: Combine_MSA_Map and Average_Map. Combine_MSA_Map merges the three MSAs generated by DeepAln, DeepMSA, and HHblits_BFD into one MSA file and uses CD-HIT (Li and Godzik, 2006) and HH-filter to do two rounds of redundancy filtering to generate a final MSA for MUlTICOM-CONSTRUCT to predict a distance map. Average_Map simply calculates the average of the distance maps predicted from the three MSAs as the final distance map prediction. We use Probability_Map to denote the approach of selecting a distance map whose corresponding/converted contact map has the highest average probability of top L/2 longe-range contact predictions from the three distance maps predicted from the three MSAs. Finally, Optimal_Map represents the ideal approach of always selecting the most accurate distance map in terms of evaluation metric (top L/2 precision, top L precision, Precision_m, Recall_m, and MAE_16) from the three maps predicted from the three MSAs, which is the upper limit that any distance map combination or selection methods can reach.

Supplemental **Table S6** reports the distance prediction results of using these approaches to select or combine the distance maps predicted from the three kinds of MSAs. Probability_Map works better than both Average_Map and Combine_MSA_Map in terms of almost all metrics and its performance is even close to Optimal_Map, indicating that the probability of top predicted contacts is a good metric to select distance maps predicted from different MSAs to improve distance prediction. In order to assess the significance of the difference in the performance of the three approaches, we apply the paired t-test compare their mean absolute errors (MAE) on the 31 hard targets in **Table S7**. The p-value between Probability_Map and Average_Map is 0.0129 (i.e., < 0.05), indicating that Probability_Map performs significantly better than Average_Map in terms of MAE. However, the p-value between Probability_Map and Combine_MSA_Map is 0.0942 (i.e., >0.05), indicating that there is no significant difference in their MAEs. We further use the paired t-test to compare the precisions of their top L/2 long-range contact predictions and get the p-value of 0.0242 (< 0.05), showing that Probability_Map performs significantly better than Combine_MSA in terms of this metric.

It is worth noting that Combine_MSA_Map performs worse than always selecting the distance maps predicted from the HHblits_BFD MSAs that work better than the MSAs of DeepAln and DeepMSA on average. The reason is that a simple combination of the MSAs from HHblits_BFD, DeepAln, and DeepMSA may introduce some noise (i.e., false positive - non-homologous sequences) into MSA, even though there are more sequences in the combined MSAs (higher depth).

3.4 Comparison of different feature sets on distance prediction

Each of the MULTICOM distance predictors uses four different sets of features derived from an MSA to predict distance maps and then average them as the final prediction from the MSA to improve the accuracy and stability of prediction. Supplemental **Table S8** summarizes the distance prediction performance of four different feature sets using MULTICOM-CONSTRUCT with HHblits_BFD alignments on 37 FM and FM/TBM domains in comparison with the ensemble approach of averaging the four predicted distance maps from the four sets of features as the prediction. The ensemble approach performs better than using each feature set alone

in terms of all evaluation metrics. Its mean precision of top L/2 long-range contacts is 50.18%, which is 3.33, 3.34, 3.47, and 6.19 percentage points higher than COV_set, PLM_set, PRE_set, and OTHER_set, respectively. The mean absolute error of the ensemble approach is 3.95 Å, lower than all the four feature sets. Also, the precision of the multi-class classification is 33.55%, higher than each feature set.

Although the average performance of the ensemble approach is better, it does not perform best on every individual target. Supplemental **Fig. S3** compares the max long-range top L/2 contact precision (diamond shape), average long-range top L/2 contact precision of four feature sets (square shape), and the long-range top L/2 contact precision of the ensemble approach (triangle shape). The results of the ensemble are not as good as the results of the best single feature set, especially for the target T1040-D1, T1047s1-D1, T1049-D1, T1082-D1, and T1096-D2 which are marked by red arrows. The gaps between the max precision of four feature sets and the precision of the ensemble approach on these targets are all greater than 8%, suggesting that there is still some room for improving the combination of features.

As a special case, Supplemental **Fig. S4** illustrates the top L/2 long-range contacts of T1047s1-D1 predicted by the ensemble approach and from the PRE_Set in comparison with the true contacts. The ensemble approach predicted more false positives marked in the eclipse than the PRE_Set.

After CASP14, we tried to ensemble the distance prediction of multiple deep learning models trained on a single feature set and found that the integration of the results of multiple models can improve the stability and accuracy of the prediction. Supplemental **Table S9** shows the comparison of a single deep learning model and the ensemble of four deep learning models that were trained on the COV_set and based on the approach similar to MULTICOM_CONSTRUCT. The performance of the ensemble of the four deep learning models using COV_set on CASP14 37 FM and FM/TBM domains is better than the single model in terms of all the evaluation metrics. The same phenomenon is also observed for the other three feature sets. Moreover, the ensemble of the four ensembles of the four feature sets obtains the long-range top L/2 contact prediction precision of 51.80%, the mean absolute error of 2.687Å, and the multi-classification precision of 34.17%, which is better than the ensemble of four single deep learning models trained on the four feature sets (i.e., 50.18%, 3.949Å, and 33.55% in **Table S8**).

3.5 Impact of the size of the training dataset on prediction accuracy

We investigated the impact of the size of training datasets on the accuracy of protein distance prediction using the deep learning model of MULTICOM-CONSTRUCT on CASP14 37 FM and FM/TBM domains. MULTICOM-CONSTRUCT was trained on two datasets of different sizes. Dataset_1 introduced in DeepDist1 has 6463 proteins. Dataset_2 has 11034 proteins. The precision of top L/2 long-range contact predictions for the deep learning model trained on Dataset_2 is 50.18%, nearly 3% percentage point higher than on Dataset_1. A target-to-target comparison of mean absolute error (MAE) on 37 domains for the two models is shown in supplemental **Fig. S5**. On almost all the domains, the model trained on Dataset_2 has a lower MAE than that on Dataset_1. In some cases, such as T1038-D2, the difference is substantial.

The comparison between the distance maps predicted by the deep learning models trained on Dataset_1 and Dataset_2 and the true distance map of T1038-D2 is illustrated in supplemental **Fig. S6**. The distance map predicted by the model trained on Dataset_2 is very similar to the true distance map, but the distance map predicted by the model trained on Dataset_1 is very different.

3.6 The study of good and bad CASP14 cases

The MULTICOM distance predictors performed very well on T1052-D3. The average precision of top L/2 long-range contact predictions of MULTICOM predictors is close to 100%, while the average top L/2 precision of all CASP14 server predictors is 58.13%. T1052 is a multi-domain protein that has 832 amino acids, Neff of the MSA of the full-length T1052 is less than 15. The domain parsing program of MULTICOM predictors was able to identify a hard modeling region [590, 688] covering the range ([589, 668]) of the third domain of the target (T01052-D3) well. The sequence of the region was used to search against the sequence databases to build deeper MSAs to predict distance maps for the region. The distance maps predicted for the regions were combined with the full-length distance maps as in DeepDist (Wu, et al., 2020). This domain-based distance map prediction substantially increased the quality of the distance prediction for T1052-D3.

Fig. 3 compares the domain-based distance map prediction and the full-length distance map prediction made by MULTICOM-CONSTRUCT with the true distance map of T1052-D3. The domain-based distance map prediction is much better and clearer than the full-length distance map prediction for T1052-D3. The results show that good domain parsing can improve the quality of MSAs and therefore the quality of distance prediction.

Usually, the poor prediction of protein distances is due to a lack of effective homologous sequences in MSAs (e.g., lower Neff on T1029, T1033, T1043, T1064) to generate good input features. The deep learning predictors cannot effectively extract distance patterns from them. However, in some cases, even though MSAs have high Neff, the accuracy of the distance prediction is still very low. For instance, the Neff of the MSAs generated by DeepAln for T1093 is 689.36 and that generated by DeepMSA is 425.12, which are high values. However, all of the MULTICOM predictors got 0% top L/2 contact prediction precision, even the domain of the target can be reasonably identified. **Fig. 4** compares the distance maps predicted by four different approaches with the ground truth: (1) the distance map predicted from MSAs generated from DeepAln and DeepMSA with the predicted domain information (our original CASP14 submission, denoted as Original_dm), (2) the distance map predicted from MSA generated by HHblits_BFD without utilizing the domain information (denoted as BFD_full), (3) the distance map predicted from MSA generated by HHblits_BFD with the predicted domain information (denoted as BFD_dm). All these four distance maps above were predicted by MULTICOM-CONSTRUCT to ensure consistency.

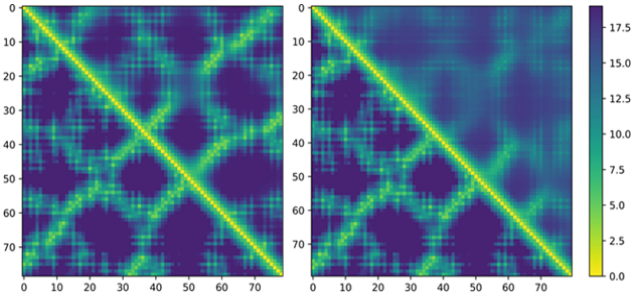

Fig. 3. Comparison of the domain-based distance prediction and the full-length distance prediction with true distance map of T1052-D3. In the subfigure on the left, the upper triangle denotes the domain-based distance prediction, and the lower triangle the true distance map. In the figure on the right, the upper triangle denotes the full-length distance

prediction, and the lower triangle the true distance map. The patterns in the domain-based distance prediction map are much clear and closer to the true distance map than the full-length distance prediction map.

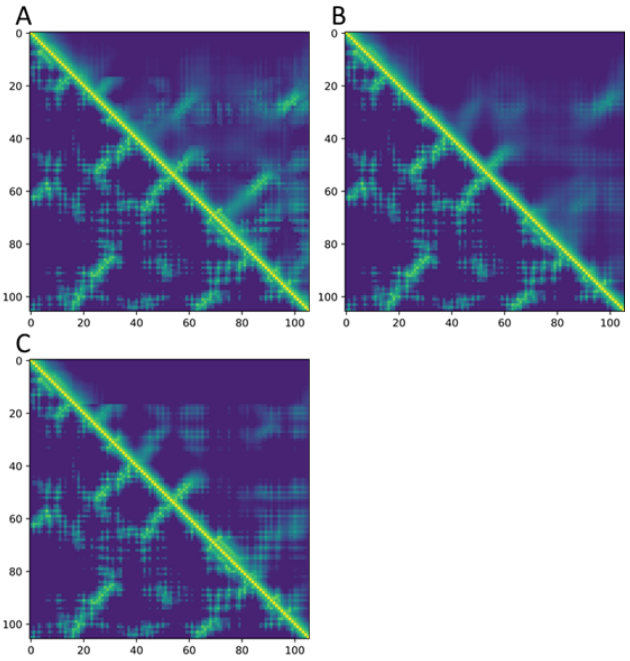

Fig.4. (A) The distanced map predicted from MSAs generated by DeepAln and DeepMSA with predicted domain information (upper triangle) versus the true distance map (lower triangle), (B) The distance map predicted from the HHblits_BFD MSA without domain information (upper triangle) versus true distance map (lower triangle), (C). The predicted distance map from the HHblits_BFD MSA with predicted domain information (upper triangle) versus the true distance map (lower triangle).

It can be seen that although MSAs generated by DeepAln and DeepMSA have a lot of sequences, most of them are false-positive positives leading to the prediction of many false-positive contact predictions (**Fig. 4A**). In **Fig. 4B**, the distance map predicted from the HHblits_BFD MSA without using predicted domain information is somewhat better, indicating that HHblits_BFD MSA (Neff = 133.0) has the better quality than MSAs of DeepAln and DeepMSA. If the predicted domain information is used, the distance prediction predicted from HHblits_BFD MSA is further improved in **Fig. 4C**, even though the Neff of the HHblits_BFD MSA for the domain is only 15, which is much lower than MSAs of DeepAln and DeepMSA. The long-range top L/2 contact prediction precision, the MAE of long-range distance prediction less than 16 Å, and the precision of multi-classification of distances using the different approaches for this domain are reported in supplemental **Table S10**. This case shows the quality of MSAs is important for distance prediction, and Neff is not always a good indicator of the quality of MSAs when there are false positives in MSAs.

4 Conclusion and future work

We developed several deep learning distance predictors and rigorously benchmarked them in CASP14. The predictors performed reasonably well in the highly competitive CASP14 experiment. The results demonstrate that MSAs generated from different alignment methods on different databases for distance prediction have different quality. The MSAs generated by HHblits on the BFD database lead to the most accurate distance prediction, but different MSAs are still complementary and can be combined to improve distance prediction. However, the number of effective sequences of MSAs has only a weak correlation with the quality of MSA and therefore is not a strong indicator of the quality of MSAs and the accuracy of the distance maps predicted from them because of the frequent existence of false positives (non-homolgous sequences) in some deep MSAs containing a lot of sequences. In contrast, the predicted probabilities of top long-range contact predictions have a strong correlation with the accuracy of distance map predictions, and therefore is a better metric to select or combine predicted distance maps to improve distance prediction. Moreover, we show that the distance maps predicted from different features generated from the same MSA are also complementary and can be integrated to improve prediction accuracy. Finally, using larger training datasets to train deep learning models, ensembling multiple deep learning models, or applying domain predictions to MSA generation of some multi-domain targets can also improve the accuracy of the distance prediction.

Acknowledgments

The project is partially supported by two NSF grants (DBI1759934 and IIS1763246), one NIH grant (GM093123), two DOE grants (DE-SC0020400 and DE-SC0021303), and the computing allocation on the Summit supercomputer provided by Oak Ridge Leadership Computing Facility (Project ID: BIF132).

References

Adhikari, B. and Cheng, J. CONFOLD2: improved contact-driven ab initio protein structure modeling. *BMC bioinformatics* 2018;19(1):22.

Adhikari, B., *et al.* ConEVA: a toolbox for comprehensive assessment of protein contacts. *BMC bioinformatics* 2016;17(1):1-12.

Berman, H.M., *et al.* The protein data bank. *Nucleic acids research* 2000;28(1):235-242.

Bhagwat, M. and Aravind, L. Psi-blast tutorial. In, *Comparative genomics*. Springer; 2007. p. 177-186.

Brünger, A.T., *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography* 1998;54(5):905-921.

Chen, C., *et al.* Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. *bioRxiv* 2020.

Ekeberg, M., *et al.* Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E* 2013;87(1):012707.

Goodfellow, I.J., *et al.* Maxout networks. *arXiv preprint arXiv:1302.4389* 2013.

Greener, J.G., Kandathil, S.M. and Jones, D.T. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature communications* 2019;10(1):1-13.

He, K., *et al.* Deep residual learning for image recognition. In, *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.

Hou, J., *et al.* Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics* 2019;87(12):1165-1178.

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

Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* 1999;292(2):195-202.

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

Kamisetty, H., Ovchinnikov, S. and Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences* 2013;110(39):15674-15679.

Kandathil, S.M., Greener, J.G. and Jones, D.T. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins: Structure, Function, and Bioinformatics* 2019;87(12):1092-1099.

Kryshtafovych, A.*, et al.* Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics* 2019;87(12):1011-1020.

Li, Y.*, et al.* ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* 2019;35(22):4647-4655.

Mao, W.*, et al.* AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. *Nature Machine Intelligence* 2019:1-9.

Mirdita, M.*, et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* 2017;45(D1):D170-D176.

Nair, V. and Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In, *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010. p. 807-814.

Remmert, M.*, et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* 2012;9(2):173.

Seemayer, S., Gruber, M. and Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* 2014;30(21):3128-3130.

Senior, A.W.*, et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020:1-5.

Steinegger, M., Mirdita, M. and Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature methods* 2019;16(7):603-606.

Steinegger, M. and Söding, J. Clustering huge protein sequence sets in linear time. *Nature communications* 2018;9(1):1-8.

Ulyanov, D., Vedaldi, A. and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* 2016.

Wu, T.*, et al.* DeepDist: real-value inter-residue distance prediction with deep residual network. *bioRxiv* 2020.

Wu, T.*, et al.* Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics* 2020;36(4):1091-1098.

Xu, J. and Wang, S. Analysis of distance‐based protein structure prediction by deep learning in CASP13. *Proteins: Structure, Function, and Bioinformatics* 2019;87(12):1069-1081.

Yang, J.*, et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* 2020;117(3):1496-1503.

Zhang, C.*, et al.* DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 2019.

Zheng, W.*, et al.* Deep‐learning contact‐map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics* 2019;87(12):1149-1164.