Zhang Chengxin (Orcid ID: 0000-0001-7290-1324)
Zheng Wei (Orcid ID: 0000-0002-2984-9003)
Li Yang (Orcid ID: 0000-0003-2480-1972)
Zhang Yang (Orcid ID: 0000-0002-2739-1916)

Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14

Short title: Protein distance map prediction in CASP14

Yang Li[1,2], Chengxin Zhang[2], Wei Zheng[2], Xiaogen Zhou[2], Eric W. Bell[2], Dong-Jun Yu[1,*], Yang Zhang[2,*]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China,
[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America

*Correspondence should be addressed to
Dong-Jun Yu (njyudj@njust.edu.cn) and Yang Zhang (zhng@umich.edu)

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/prot.26211

Abstract

This article reports and analyzes the results of protein contact and distance prediction by our methods in the 14th Critical Assessment of techniques for protein Structure Prediction (CASP14). A new deep learning-based contact/distance predictor was employed based on the ensemble of two complementary coevolution feature coupling with deep residual networks. We also improved our Multiple Sequence Alignment (MSA) generation protocol with wholesale meta-genome sequence databases. On 22 CASP14 Free modeling (FM) targets, the proposed model achieved a top-$L/5$ long-range precision of 63.8% and a mean distance bin error of 1.494. Based on the predicted distance potentials, 11 out of 22 FM targets and all of the 14 FM/TBM targets have correctly predicted folds (TM-score > 0.5), suggesting that our approach can provide reliable distance potentials for *ab initio* protein folding.

Keywords: CASP, contact-map prediction, deep learning, protein structure prediction

Introduction

Ab initio protein structure prediction has been a longstanding challenge in the field of computational biology[1]. We have witnessed evident signs of progress[2,3] in the recent Critical Assessment of techniques for protein Structure Prediction (CASP) experiments with respect to protein structure prediction based on long-range predicted geometric potentials, usually from supervised machine learning methods. To accurately predict the long-range geometric restraints, e.g., inter-residue contacts, early methods derive correlations between positions of MSAs inspired by the coevolution phenomenon[4]. Those coevolution analysis methods can be classified into local[5-7] and global[8-11], i.e., direct coupling analysis (DCA), approaches, according to whether all other positions are considered when computing the coupling between a residue pair. The residue-wise correlation maps are later used by supervised machine learning methods as features[12-15], and have been incorporated into multiple successful methods[16-18] when coupled with deep convolutional networks. An improved feature extraction strategy[19-23] is employed by feeding raw coevolution to a deep ResNet[24] with features to avoid possible information leakage during post-processing. Recent approaches further extend the pipeline for contact map prediction to distance[25-27] and orientations[28] and provide more precise restraints to assist protein folding.

In this article, we introduce DeepPotential, which participated in CASP14 for protein contact/distance prediction. DeepPotential collects both local and global raw coevolution matrices with post-processing and derives sequence-specific descriptors using a set of candidates from a progressive MSA construction pipeline. By coupling features with deep dilated residual convolutional networks[29], DeepPotential is capable of predicting high precision contact maps superior to our previous approach in CASP13[21]. We also have found significant improvements by searching against multiple gigantic databases. High-quality protein 3D structure models can be obtained based on the low-error distance prediction by DeepPotential, without the need for template information.

Materials and methods

In CASP14, we tested DeepPotential, and the pipeline can be broken down into a procedure of steps as shown in Figure 1. The prediction contains steps of multi-MSA construction and selection, complementary coevolution feature extraction, and neural network prediction. We also build tertiary protein structures from the predictions of DeepPotential.

Datasets. DeepPotential was trained on 26,151 experimentally solved structures collected from PDB[30], with the latest structure being timestamped as 2019.11.12. We first set the maximum sequence length to 1000 and kept the representative sequences after a 35% sequence identity clustering procedure with CD-HIT[31].

Collection of Multiple sequence alignments. In CASP14, the 6 candidate MSAs for each target sequence were generated by searching against two whole-genome databases and four metagenome sequence databases (Metaclust[32], BFD[33], Mgnify[34] and IMG/M[35]). Following our previous work[36], a progressive strategy, i.e., stopping the search if the target Neff (Equation 1) threshold is satisfied, was implemented. This early stopping criterion is found to reduce database search time without sacrificing MSA quality for targets with a great deal of homologs[36]. The Neff value can be calculated by

$$\mathrm{Neff} = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1}^{N} \mathbb{I}\left[S_{m,n} \geq 0.8\right]} \tag{1}$$

where N is the total number of sequences in the MSA; $\mathbb{I}$ is the indicator function; $S_{m,n}$ is the sequence identity between sequence n and sequence m. Figure 2 summarizes the entire MSA generation pipeline. We first search against Uniclust30 (version 2020_01) and UniRef90 using HHblits[37] and Jackhmmer[38] and obtain MSAs for Stage 1 and 2, respectively. The resulting Stage 2 MSA will be used as the initial profile for HMMsearch and HHblits to search against the Metaclust (Stage 3 MSA) and BFD (Stage 4 MSA) databases, respectively. When searching against the Mgnify and IMG/M databases in Stage 5 and Stage 6, the MSA from Stage 4 is considered as the initial MSA profile. Note that custom HHblits databases will be constructed from raw hits to ensure that the output MSAs come from the same algorithm (Hhblits2) from Stage 2 to Stage 6. For training sequences, the ordinarily HHblits MSA (Stage 1 MSA) was employed with Uniclust30 (version 2017_04). Once the desired MSAs are obtained, the distance map with the highest confidence score will be considered as the final prediction. In CASP14, we explored two confidence score configurations. For Group 010, the confidence score is defined as the mean of the cumulative probability under 12 Å of the top 10*L predicted C_β-C_β distance distributions for all residue pairs, while for Group 024, the corresponding threshold is 8Å.

Feature extraction. DeepPotential extracts a set of complementary features as the input of the deep learning model. Coevolution analysis represents conditional or marginal correlations between residues and thus can be critical discriminative features. In DeepPotential, two types of coevolutionary features, the PseudoLikelihood Maximized Potts model (PLM)[11] and Mutual Information (MI), are extracted. Given the input MSA (X) with N alignments and L sequence length, the PLM feature can be obtained by minimizing the following loss function:

$$\mathcal{L}_{PLM} = -\sum_{l=1}^{L}\sum_{n=1}^{N} \log \frac{\exp\left(e_i\left(X_{n,i}\right) + \sum_{j=1,j\neq i}^{L} P_{i,j}\left(X_{n,i}, X_{n,j}\right)\right)}{\sum_{q=1}^{Q} \exp\left(e_i(q) + \sum_{j=1,j\neq i}^{L} P_{i,j}\left(q, X_{n,j}\right)\right)} + \lambda_{single} \sum_{i=1}^{L} \|e_i\|_2^2 + \lambda_{pair} \sum_{\substack{i,j=1 \\ i\neq j}}^{L} \left\|P_{i,j}\right\|_2^2 \tag{2}$$

where $Q = 22$ represents 20 regular amino acid types, an undetermined amino acid type state and a gap state; $e \in R^{L\times Q}$ and $P \in R^{L\times L\times Q\times Q}$ are field and coupling parameters of Potts model, respectively; $\lambda_{single} = 1$ and $\lambda_{pair} = 0.2 \times (L-1)$ are the regularization coefficients for e and P. The parameter $P_{i,j}(q_1, q_2)$ measures the linear coefficient of the q_1 state of residue i and the q_2 state of residue j conditioned on other residues and states, which can eliminate transitive interactions in the observed interactions.

The PLM feature eliminates transitional noise in marginal correlations, which should be more relevant to the structural interaction terms between residue pairs. However, the optimization of PLM could be ill-posed when there are no sufficient aligned sequences in the MSA. In this regard, we utilize a raw marginal correlation measurement, i.e., Mutual Information, as another pairwise feature. The Mutual Information feature of residue i and j is defined as:

$$M_{i,j}(q_1, q_2) = f_{i,j}(q_1, q_2) \ln \frac{f_{i,j}(q_1, q_2)}{f_i(q_1) f_j(q_2)} \quad (3)$$

here, $f_i(q_1)$ is the frequency of a residue type q_1 at position i of the MSA, $f_{i,j}(q_1, q_2)$ is the co-occurrence of two residue types q_1 and q_2 at positions i and j. Note that the raw Mutual Information matrix $M_{i,j} \in R^{Q \times Q}$ for residue pair i and j will be kept as features without summation over residue types. Such a formula can capture residue type pair-specific information and provide the fine-grained feature to deep leaning model. Compared to the regular Pearson correlation, Mutual information is capable of measuring the non-linear relationships between variables. For each residue pair, another three post-processed residue-wise features will be extracted from the coevolution residue pair matrix ($P_{i,j}$ and $M_{i,j}$), i.e., Frobenius norms of (1) the whole residue pair matrix, (2) residue pair matrix excluding the gap state, and (3) residue pair matrix excluding both the gap and undetermined amino acid state. The above coevolutionary feature set records the features for the whole MSA, while $P_{i,j}(X_{1,i}, X_{1,j})$ and $M_{i,j}(X_{1,i}, X_{1,j})$ record the target sequence-specific parameters in the PLM and MI matrices. Thus, they are considered as extra 2-D features.

In addition to the 2-D features, there are four 1-D features collected. The first two are the field parameter e of the Potts model and the mutual information matrix where $j = i$. HMM profile features (30 descriptors for each position) are also considered as the third component by building a profile hidden Markov models from the input alignment using the hhmake program in the hhsuite package[39]. The last feature encodes the categorical target sequence with one-hot-encoding.

Deep ResNet for multiple geometric map prediction. As shown in Figure 1, the above 1-D and 2-D input features are fed into 10 1D residual blocks[24] and 10 2D residual blocks, respectively. A general structure of a residual block is shown in Figure 1, where a shortcut link is added from previous layers to the output, compared to the traditional convolutional neural networks. Here one weight layer contains two layers, i.e., one convolutional layer and an instance normalization layer[40], collected sequentially. The transformed 1D features with 32 channels will be tiled vertically and horizontally and concatenated with transformed 2D features (64 channels). The composited 2D features (32*2+64=128 channels) will further go through 40 2D residual blocks. The prediction layer for each potential term performs a simple pixel-wise linear transformation to the desired channel size, prior to a softmax layer. Dilation is applied for both 1D and 2D convolutional layers with cycling of 1, 2, 4, 8, and 16. The padding size is then set accordingly to ensure the consistency of feature signal spatial shapes. Dropout is used in all residual blocks and the dropout rate was set to 0.2 globally.

The model was trained by the supervision of distance between inter-residue C_β atoms (C_α for glycine). The distance value is discretized into 36 equal-width bins from 2 Å to 20 Å, with additional two bins representing distance less than 2 Å and over 20 Å. Thus, the model can be trained with cross-entropy loss over all residue pairs. In addition, some auxiliary tasks, i.e., inter-C_α atom distance, inter-residue orientation angles[28] and H-bond geometry terms defined on long-range neighboring C_α atoms[41], are also considered in a multi-task learning strategy[42].

Template Free structure modeling built on DeepPotential restraints. To explore the effectiveness of predicted inter-residue distance for protein structure prediction, we also

utilized the predicted C_β distance distribution, along with the outputs of auxiliary tasks (predicted C_α distance and inter-residue orientation angles) as restraints for *ab initio* protein structure prediction in the form of potentials. The negative log of geometric distributions were smoothed by a cubic spline into smooth potentials so that they could be optimized by gradient-descent based methods, e.g., L-BFGS implemented by the PyRosetta package[43]. Starting from a random initial conformation, the structure was iteratively optimized using L-BFGS. To find the global lowest energy conformation, we used an iterative strategy, and at each iteration, random noise (± 20 degrees) in the backbone torsion angle space was added to the previous conformation for further energy minimization. The conformation with the lowest energy value was kept for final submission.

Results

We analyzed the performance of DeepPotential on 22 FM (T1070-D1 was excluded since there were not sufficient long-range contacts in this target) and 14 TBM/FM targets from CASP14 (T1085-D2 was excluded because of the unavailability of its experimental structure). The performance was evaluated mainly in two tasks, i.e., inter-residue contact and distance prediction. For the evaluation of predicted contacts, the conventional long-range Top-*N* (*N*=*L*/10, *L*/5, *L*/2, and *L*) precision is reported. For predicted discrete distance distribution, we report the results with two different evaluation indexes. The first index is the long-range Top-*N* bin classification ACCuracy (BACC), considering that the distance prediction is formulated as a multi-class classification problem:

$$BACC = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[argmax(P_i) == k(D_i)] \quad (4)$$

Here, P_i is the distance distribution of *i*-th residue pair ranked by the probability of the distance less than 20 Å. $argmax(P_i)$ finds the predicted distance bin of residue pair *i*; $k(D_i)$ returns the bin of experimental distance D_i. In CASP14, there are 10 bins for distance prediction, representing < 4Å, 4-6Å, 6-8Å, …, 16-18Å, 18-20Å, and >20Å. It should be noted that there is a natural order among distance bins in such a formulation. Such order sorts the distance bins according to the definition of bins in distance. Thus, we also utilize a second index, long-range Top-*N* Mean Bin Error (MBE), to provide a closer evaluation of distance prediction by considering the closeness in distance bins:

$$MBE = \frac{1}{N}\sum_{i=1}^{N} |argmax(P_i) - k(D_i)| \quad (5)$$

A lower MBE index means better prediction, and the ideal case of the MBE value should be 0. For distance prediction, a larger threshold (20Å) was considered in CASP14, compared to the threshold of 8Å for contacts. In this regard, we set *N*=*L*/2, *L*, 2*L*, and 5*L* when evaluating distance prediction.

Contact and distance prediction performance in CASP14. The accuracy of long-range contact prediction on CASP14 FM and FM/TBM targets for Group 010 and 024 are listed in Table 1. On average, the Top-*L*/5 precisions were 0.638 and 0.615, and 16 and 15 out of all 22 FM targets have a precision ≥ 0.5 for Group 010 and 024, respectively. On 14 FM/TBM targets, the average Top-*L*/5 precisions increased to 0.868 and 0.862, and 13 out of 14 have a precision ≥ 0.5 for both of the two Groups. Such a difference is likely because FM/TBM targets have more related samples in the training set of DeepPotential than FM

targets, because by definition they are evolutionarily closer than the FM targets to the PDB structure by which DeepPotential was trained. To verify, we searched each of CASP14 domains against the training structures of DeepPotential by TM-align. The average TM-score between FM/TBM targets and their best template is 0.587, significantly higher than 0.497 for FM targets, with an unpaired t-test *p*-value of 8e-04. FM/TBM targets also, on average, have 258 related structures (TM-score > 0.5) in the training set. Meanwhile, for FM targets, the corresponding number is only 28. Such data suggest very different relationships to training samples of DeepPotential for FM and FM/TBM targets in CASP14. In addition, FM/TBM targets also have more sequence homologs than FM targets (167.8 versus 79.7 of average Neff), which provide more reliable features, especially coevolutionary features. In Table 1, we also compare the performance of two Groups based on the DeepPotential pipeline. Although Group 010 had slightly higher precisions for all evaluation indexes, the differences are not statistically significant, with two-tailed test *p*-values of 0.196, 0.410, 0.350, and 0.787, respectively, for both FM and FM/TBM targets. Figure S1 shows the correlation between the confidence score of selected MSAs for Group 010 and Group 024 respectively. Both two configurations of confidence score have high correlations with their Top-*L* long-range precisions, with Pearson correlation coefficient (PCC) of 0.743 and 0.752 for Group 010 and Group 024 respectively. Such data suggest that MSA selection based on the mean Top-*N* probability scores has robust performance regardless of the specific choice of the threshold for confidence score definition.

In Table 2, we report the two metrics for distance prediction, i.e., BACC and MBE. A similar pattern can be observed that FM/TBM targets achieved higher accuracy or lower bin error relative to FM targets. Interestingly, unlike contact precisions that drop sharply when evaluating Top-*L*/10 to Top-*L* contacts, we observe a relatively gentle slope even evaluating from Top-*L*/2 to Top-5*L* ranked residue pairs. Taking the performance of Group 010 on TBM/FM targets as an example, the accuracy of predicted contacts, with a threshold of 8 Å, drops from 0.898 to 0.620 (31.0% of decrease) when evaluating from Top-*L*/10 to Top-*L* contacts. Meanwhile, for distance, the Bin accuracy only drops from 0.562 to 0.456 (18.9% of reduction) when selecting from Top-*L*/2 to Top-5**L* distance predictions, sorted by the cumulative probability less than 20 Å. Such an observation suggests a stable accuracy for distance prediction when sorting according to $P(d < 20\text{Å})$. In addition to the bin accuracy, Table S1 also shows other multi-class classification evaluation indexes, i.e., Precision, Recall and F1-score, from Top-*L*/2 to Top-5**L* distance predictions. The evaluation indexes for each target are the average over all classes (bins). Similar to the precision of contact-maps, Group 024 tends to have more accurate distance prediction for FM while Group 010 favors FM/TBM targets, respectively. For the average Top-*N* Bin Error presented in Table 2, it can be observed that both of the two Groups maintain relatively low values. For example, the long-range Top-*L* average bin errors on FM and FM/TBM targets are 1.336 (1.415) and 0.593 (0.532) for Group 010 (024), respectively. It is notable that both Groups successfully provide 11 FM and 10 FM/TBM targets that have a mean bin error lower than 1.0, even up to Top-5*L* selected residue pairs. Precise continuous distance prediction, utilized as potentials, will be critical to guide high-accuracy protein structure prediction.

We address the improvements of DeepPotential compared to our previous model, TripletRes, in CASP13 (denoted as TripletRes_CASP13) by head-to-head comparisons in Figure 3. We compare the long-range Top-*L*/5 TripletRes from their own default MSAs

(Figure 3A) and from the same MSAs by DeepPotential (Figure 3B), with the results of Group 010 in CASP14 as an illustration. With the original pipeline, including the previous MSA construction pipeline, TripletRes_CASP13 achieved a Top-*L*/5 precision of 0.310 on 36 FM and FM/TBM targets; when the newly constructed MSAs were applied, the precision rose to 0.398, an improvement of 28.7%, suggesting the new pipeline generates higher-quality MSAs over multiple sequence databases. Meanwhile, the precision for Group 010 in CASP14 on the same dataset was 0.483, 56.1% and 21.3% higher than the control TripletRes_CASP13 from the default and the same MSAs, respectively. The corresponding *p*-values are 3.58e-08 and 2.35e-07 respectively. Besides, a head-to-head comparison of Top-*L* long-range contact precision between TripletRes (Post-CASP13), introduced in our previous research[22], and Group 010 is presented in Figure S2. The TripletRes (Post-CASP13) was trained using the same dataset with DeepPotential. The Top-*L* precision increases from 0.398 to 0.418 after the re-training, but still lower than the precision of 0.483 (*p*-value = 1.0e-05) obtained by DeepPotential. These significant improvements could be attributed the complete set of features coupled with multi-task learning framework and the larger amount of training data. However, the results show that for those targets that DeepPotential and TripletRes models had similar performance, where limited superiority could be found for DeepPotential. Our next work should be focusing on proper training strategies for those extremely hard targets.

Constructing MSAs from gigantic databases improves performance. In CASP14, we utilized a progressive pipeline for MSA construction from whole-genome and meta-genome databases with an enormous number of sequences. The employment of those sequence databases becomes a critical element to achieve higher quality contact/distance prediction, especially for FM targets. In Figure 4A, we show the distributions of long-range Top-*L*/5 precisions of 22 FM targets by gradually adding the considered sequence databases. Other factors, e.g., MSA selection score configuration, are fixed. Comparable performance, i.e., precisions of 0.402 versus 0.416, can be observed with two whole-genome databases. In contrast, significant improvements can be observed when meta-genome databases are utilized by our pipeline. When Metaclust and BFD databases were added, the precision increased to 0.495 and 0.610, 19.0% and 46.6% higher than the best results of whole-genome databases, respectively. Additionally, when two extra databases, i.e., the Mgnify and IMG/M databases, are considered, mean precisions of 0.647 and 0.660 can be observed respectively. There is also a consistent improvement in median precision from 29.2% to 77.2% with the consideration of more databases. Such data suggest the fundamental role of the MSA construction pipeline and the great usefulness of the utilization of meta-genome databases. It can be observed in Figure 4B that the Neff of selected MSAs is continuously increasing along with the addition of the corresponding databases. The detailed mean (median) logarithm Neff values are 0.21 (-0.11), 0.54 (0.06), 1.21 (0.52), 1.9 (2.07), 2.19 (2.41), and 2.69 (3.35) when adding the corresponding databases. The rapid improvements of contact-map precision are likely highly driven by the MSAs with continuously increasing Neff values at each MSA stage.

A more quantitative analysis about the impact of MSA quality on the performance of contact/distance prediction of DeepPotential is shown in Figure 5A-C. In Figure 5A, the long-range Top-*L*/5 precision of Group 024 in CASP14 versus the Neff values of the MSAs is presented. The Pearson correlation coefficient (PCC) between precision and the common logarithm of Neff is 0.364, indicating a modest correlation. The precision of DeepPotential

is less dependent on Neff than TripletRes_CASP with the same input MSAs, whose corresponding PCC is 0.419. As shown in the left-upper block in Figure 5A, 10 out of 14 targets with a Neff value lower than 10 have a precision ≥ 0.5, which is 3 more than that of TripletRes_CASP; this may help explain the lower correlation coefficient seen for DeepPotential than TripletRess_CASP.

In addition, we also found that the PCCs are very different for FM and FM/TBM targets. When only FM targets are considered, the correlation is 0.382, slightly higher than the combination of FM and FM/TBM targets. The reason for higher correlation coefficient for FM targets should be the extreme cases (red dots in the lower-left block in Figure 5A) where their Neff values are all less than 1, and the corresponding precisions are < 0.3. In contrast, for FM/TBM targets, the correlation became 0.181 with a *p*-value of 0.535 where a statistically significant correlation could not be detected. Still, we can find one FM/TBM target that has a high Neff value and precision < 0.5, T1080-D1. Nevertheless, the 3D model built based on the geometric restraints predicted by DeepPotential eventually had the correct fold, with a TM-score[44] of 0.503, possibly because of the modest Top-*L* bin error of 1.37.

For FM targets with Neff ≥ 10, 10 out of 13 had precisions over 0.5, and the 3 exceptions are T1029-D1, T1093-D1 and T1093-D3. Interestingly, for T1093-D1 and T1093-D3, if we re-run DeepPotential based on MSAs built with domain sequences, the corresponding precisions will be boosted to 0.821 and 1.0, even with lower Neff values of 10.91 and 17.13. We simply visualize the MSA consensus using Sequence logos[45] with MSAs generated with full-length sequence and only domain sequence in Figure S3 and Figure S4 for T1093-D1 and T1093-D3, respectively. The graphical characters displayed in Figures S3 and S4 represent significant residues and subtle sequence patterns in the corresponding MSAs. The height of each letter is made proportional to its frequency, and the letters are sorted so the most common one is on top. We found that the patterns of consensus sequences are visibly different for domain and full-length MSAs. The average Kullback-Leibler divergencies over positional amino acid compositions between domain and full-length MSAs are 1.30 and 1.19 for T1093-D1 and T1093-D3 respectively. For MSAs from the full-length sequence, the average sequence identity between aligned sequences and query sequence are 0.098 and 0.118 for T1093-D1 and T1093-D3, respectively, which are much lower than those of the domain sequence MSAs (0.202 and 0.232). The differences in distribution and sequence identity suggest that there could be noisy alignments in our selected MSAs, despite their high Neff value.

Figure 5B and 5C show the long-range Top-*L* bin classification accuracy and bin error versus Neff respectively. We found a relatively higher PCC of 0.465 between classification accuracy and the common logarithm of Neff, compared to that between contact precision and the latter. The correlation indicates that there is much room for improvement of the current model for further higher resolution distance prediction, especially for FM targets (PCC = 0.474) with limited sequence homologs. For Top-*L* bin error, we can observe a modest correlation; the corresponding PCCs are -0.363 and -0.386 for all targets and FM targets, respectively.

Structure modeling based on DeepPotential in CASP14. The final goal of contact/distance prediction is to assist protein structure prediction, and the best way to evaluate the quality of distance prediction is to analyze the quality of predicted protein structure purely based on predicted distance. On 36 FM and FM/TBM target domains, the

3D models based on predicted geometric descriptors by DeepPotential achieved a mean TM-score of 0.591. When FM domains are evaluated, the mean TM-score becomes 0.514, where 12 have correctly predicted global folds (TM-score ≥ 0.5). Surprisingly, our submissions had all 14 FM/TBM folds correctly predicted, with a mean TM-score of 0.712, ranging from 0.503 to 0.868, even though structural templates were not used. Figure 5D shows the TM-score and the Neff of 36 FM and FM/TBM targets. Similar to the previous observations, there is a modest correlation between the TM-score and the common logarithm of Neff (PCC = 0.482), but for FM/TBM targets, the correlation is weak, even not statistically significant (PCC = 0.158, *p*-value=0.61). Figure S5 plots the correlation between the TM-score of predicted structures and the precision of long-range Top-*L*/2 predicted contacts. There is a significant correlation between TM-score and precision (PCC = 0.788), indicating that the accuracy of predicted structures is highly depending on the quality of deep learning predictions. There are hardly any cases with low contact precision but high TM-score, which is, however, reasonable since we used a basic protein folding procedure without other information sources, e.g., template information.

In Figure 6, we present an example FM domain, T1042-D1, to show that DeepPotential can predict reliable distance potentials with limited sequence alignments (Neff=8.85) for accurate protein structure prediction. T1042-D1 is the 8th domain of a viral protein[46] (PDB ID: 6VR4) officially partitioned by CASP. In this case, the long-range Top-*L*/5 precision was 0.818 and for contact map prediction, the Top-*L* bin accuracy and bin error were 0.464 and 0.601, respectively. As shown in Figure 6A, the predicted distance bin map in general successfully recovered the distance patterns of the experimental structure except one region with some geometric interactions the between N- and C- terminals. Nevertheless, the predicted structure based on the distance map has a TM-score of 0.725.

Despite this successful prediction, we found that domain T1047s1-D1 has an unexpectedly low TM-score. The contact Top-*L*/5 precision, Top-*L* bin accuracy and Top-*L* bin error are 1.0, 0.597, and 0.502 respectively. However, we observed some noisy distance predictions between a beta-sheet region (residue 50-125) and the C-terminal structure region (residue 125-211) in Figure S6B. Those noisy predictions pulled the two regions together and destroyed the structure of the beta-sheet region (Figure S6D). Thus, our 3D model had only a TM-score of 0.416, although the TM-score of the region near the C-terminal achieved 0.632 (Figure S6E). In Figure S6A, we plot the residue-wise prediction based on Potts model coupling parameters and found observable noisy signals. Thus, the noisy prediction partly came from the MSA and its features.

This example also exposed one weakness of our protein contact/distance and structure prediction pipeline. The current distance prediction is only formulated to predict distance under a fixed threshold, i.e., 20 Å. Thus, the subsequent restraint-based protein folding strategy will be influenced by possibly noisy distance (or other geometric descriptors) potentials. If a region has sparse connections with other parts, the noise could completely mislead the folding. Seeking better formulations, e.g., real-distance, to obtain reliable geometric restraints without the limitation of a threshold should help build better 3D models, not only for targets similar to T1047s1, but also for the modeling of inter-domain or inter-chain structures. One feasible way could be predicting the parameters of distribution (e.g., the expected value (or mean) and standard deviation of the variable's natural logarithm for Log-Normal Distribution) for the inter-atom distance modelling. The predicted distribution map could be easily converted to smooth potentials for protein

folding. Whether such prediction would result in more accurate protein folding still requires further examination.

Discussion

We have introduced DeepPotential, which participated in CASP14 for contact and distance prediction. Our model takes an ensemble of complementary features directly extracted from selected MSAs constructed by progressive searching against multiple sequence databases. Detailed analysis showed that the proposed method can produce relatively high precision contact maps that significantly outperform our previous method in CASP13. We also showed that the distance predictions can be used as reliable restraints for protein structure prediction.

What went right? The results in CASP14 confirm the conclusion of our previous strategy for contact/distance prediction in CASP13, i.e., constructing and selecting MSAs from multiple protein sequence databases can significantly improve performance, especially for FM targets. In CASP14, we further extended our MSA construction pipeline by the utilization of large-scale meta-genome databases, which brought a further boost in contact/distance and structure prediction. In addition, the ensemble of multiple raw coevolution features which extract complementary information from MSAs, together with a multi-task learning scheme, contributed to the advantage of DeepPotential over previous approaches.

What went wrong? The DeepPotential model predicts distance and other geometric terms marginally, ignoring the inherent relationships among geometric terms and residue pairs in the loss function. Designing a formulation to effectively model the joint distributions between residue pairs or geometric terms should be a necessity in our future work. Also, the convolutional neural networks used in DeepPotential have a relatively long max path between features of residue pairs. Considering the revolutionary results obtained by AlphaFold2[47] in CASP14 and the recent success of Transformer[48] applied in protein sequence modelling[49-51], the Transformer framework with a direct max path should be considered in our future work to model direct long-range interactions in space.

Acknowledgements

We thank Dr. Jianyi Yang and Dr. Ivan Anishchenko for their insightful discussions. This work is supported in part by the National Institute of General Medical Sciences (GM136422, S10OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.), the National Science Foundation (IIS1901191, DBI2030790, MTM2025426 to Y.Z.), the National Natural Science Foundation of China (62072243, 61772273, to D.Y.) and the Natural Science Foundation of Jiangsu (BK20201304 to D.Y.) .

DeepPotential was trained using the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (ACI-1548562). The work was done when Y.L. visited the University of Michigan.

Data Availability

All relevant data are within the manuscript and its Supporting Information files.

References

1. Zhang Y. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology.* 2008;18(3):342-348.
2. Abriata LA, Tamò GE, Dal Peraro M. A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(12):1100-1112.
3. Abriata LA, Tamò GE, Monastyrskyy B, Kryshtafovych A, Dal Peraro M. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins: Structure, Function, and Bioinformatics.* 2018;86(S1):97-112.
4. Fitch WM, Markowitz E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics.* 1970;4(5):579-593.
5. Korber BT, Farber RM, Wolpert DH, Lapedes AS. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences.* 1993;90(15):7176.
6. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics.* 2008;24(3):333-340.
7. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics.* 1994;18(4):309-317.
8. Morcos F, Pagnani A, Lunt B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences.* 2011;108(49):E1293.
9. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics.* 2012;28(2):184-190.
10. Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics.* 2015;31(21):3506-3513.
11. Ekeberg M, Hartonen T, Aurell E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *Journal of Computational Physics.* 2014;276:341-356.
12. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics.* 2015;31(7):999-1006.
13. He B, Mortuza SM, Wang Y, Shen H-B, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics.* 2017;33(15):2296-2306.
14. Buchan DWA, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics.* 2018;86(S1):78-83.
15. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics.* 2013;29(13):i266-i273.

16. Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology.* 2017;13(1):e1005324.
17. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics.* 2017;34(9):1466-1472.
18. Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems.* 2018;6(1):65-74.e63.
19. Golkov V, Skwark MJ, Golkov A, et al. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. Paper presented at: NIPS2016.
20. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics.* 2018;34(19):3308-3315.
21. Li Y, Zhang C, Bell EW, Yu D-J, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(12):1082-1091.
22. Li Y, Zhang C, Bell EW, et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLOS Computational Biology.* 2021;17(3):e1008865.
23. Li Y, Hu J, Zhang C, Yu D-J, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics.* 2019;35(22):4647-4655.
24. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2016.
25. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020;577(7792):706-710.
26. Xu J. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences.* 2019;116(34):16856-16865.
27. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications.* 2019;10(1):3977.
28. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences.* 2020;117(3):1496.
29. Yu F, Koltun V, Funkhouser T. Dilated residual networks. Paper presented at: Proceedings of the IEEE conference on computer vision and pattern recognition2017.
30. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Research.* 2000;28(1):235-242.
31. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics.* 2012;28(23):3150-3152.
32. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nature communications.* 2018;9(1):2542.

33. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods.* 2019;16(7):603-606.
34. Mitchell AL, Almeida A, Beracochea M, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research.* 2020;48(D1):D570-D578.
35. Chen IMA, Chu K, Palaniappan K, et al. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Research.* 2021;49(D1):D751-D763.
36. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics.* 2020;36(7):2105-2112.
37. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods.* 2012;9(2):173.
38. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics.* 2010;11(1):431.
39. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics.* 2019;20(1):473.
40. Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:160708022.* 2016.
41. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nature Methods.* 2015;12(1):7-8.
42. Thrun S. Is learning the n-th thing any easier than learning the first? Paper presented at: Advances in neural information processing systems1996.
43. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics.* 2010;26(5):689-691.
44. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics.* 2010;26(7):889-895.
45. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 1990;18(20):6097-6100.
46. Drobysheva AV, Panafidina SA, Kolesnik MV, et al. Structure and function of virion RNA polymerase of a crAss-like phage. *Nature.* 2021;589(7841):306-309.
47. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021.
48. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv preprint arXiv:170603762.* 2017.
49. Elnaggar A, Heinzinger M, Dallago C, et al. ProtTrans: towards cracking the language of Life's code through self-supervised deep learning and high performance computing. *bioRxiv.* 2020.
50. Rao R, Liu J, Verkuil R, et al. Msa transformer. *bioRxiv.* 2021.
51. Rao R, Meier J, Sercu T, Ovchinnikov S, Rives A. Transformer protein language models are unsupervised structure learners. *bioRxiv.* 2020:2020.2012.2015.422761.

Tables

Table 1. Mean accuracy and standard error (in brackets) of long-range predicted contacts for DeepPotential's two Groups in CASP14.

Group	FM				TBM/FM			
	L/10	L/5	L/2	L	L/10	L/2	L/5	L
010	0.686 (0.086)	0.638 (0.081)	0.534 (0.068)	0.396 (0.051)	0.898 (0.057)	0.868 (0.061)	0.771 (0.062)	0.620 (0.057)
024	0.653 (0.090)	0.615 (0.084)	0.502 (0.068)	0.386 (0.051)	0.872 (0.064)	0.862 (0.059)	0.784 (0.061)	0.629 (0.059)

Table 2. Performance of long-range Top-*N* distance prediction for DeepPotential's two Groups in CASP14 with standard error in brackets.

Index	Group	FM				TBM/FM			
		L/2	L	2L	5L	L/2	L	2L	5L
Bin acc	010	0.387 (0.048)	0.369 (0.044)	0.370 (0.041)	0.338 (0.037)	0.562 (0.039)	0.535 (0.037)	0.502 (0.038)	0.456 (0.039)
	024	0.374 (0.047)	0.373 (0.045)	0.362 (0.043)	0.336 (0.039)	0.594 (0.034)	0.581 (0.040)	0.526 (0.039)	0.477 (0.044)
Bin error	010	1.341 (0.258)	1.336 (0.248)	1.343 (0.248)	1.495 (0.229)	0.505 (0.057)	0.593 (0.083)	0.717 (0.155)	0.835 (0.162)
	024	1.384 (0.268)	1.415 (0.266)	1.441 (0262)	1.559 (0.241)	0.453 (0.052)	0.532 (0.084)	0.699 (0.156)	0.826 (0.174)

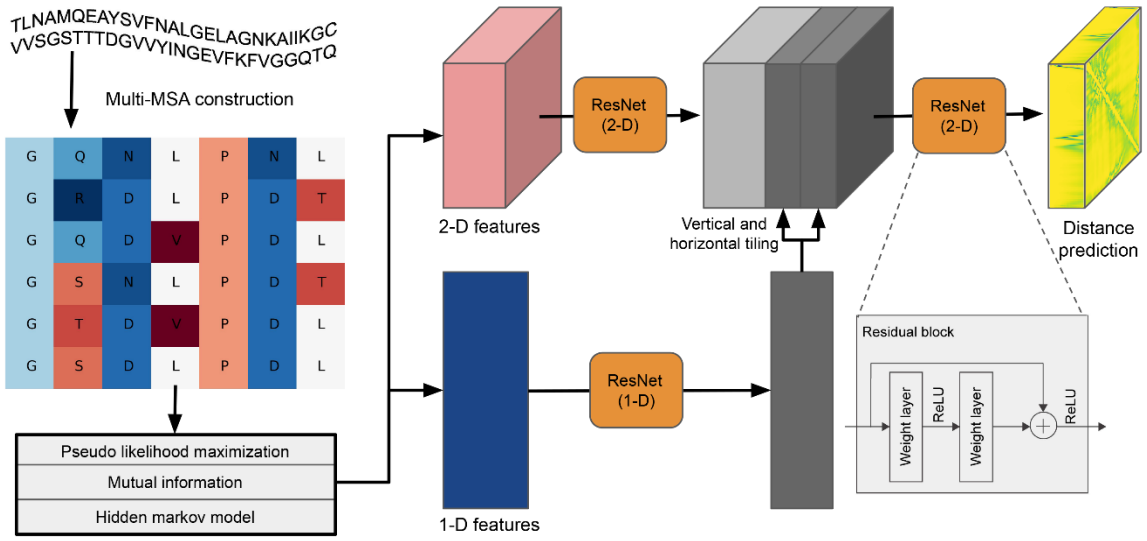


Figure 1. The pipeline of DeepPotential. Starting from a sequence, multiple MSAs will be constructed. Three models, i.e., the pseudolikelihood maximization of the Potts model, Mutual information, and a hidden markov model will be used to extract 2-D and 1-D features. The two features will go through two sets of residual convolutional layers and tiled together. The combined hidden features will go through another set of residual blocks for the final prediction.

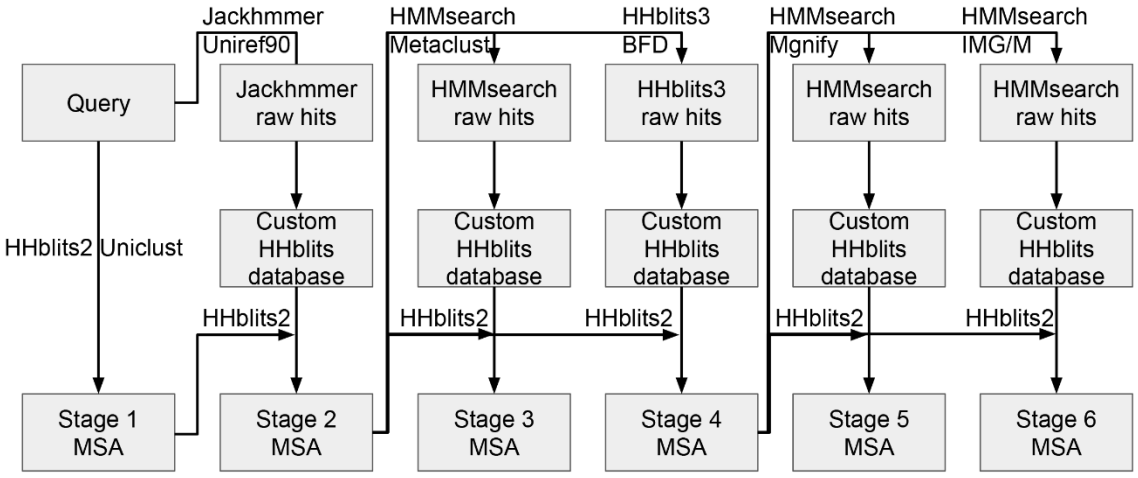


Figure 2. The multi-stage progressive MSA construction pipeline in DeepPotential.

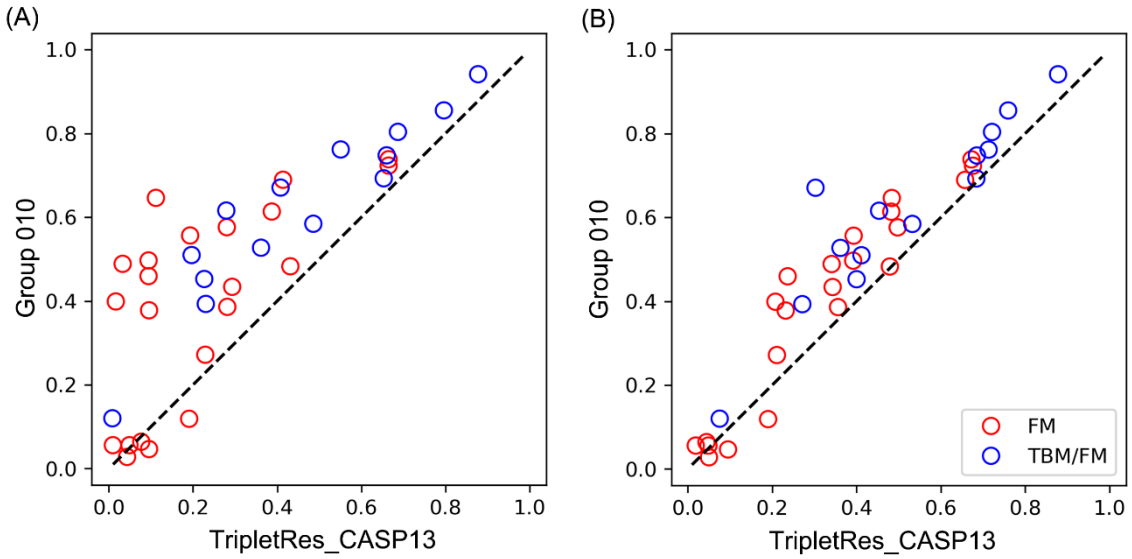


Figure 3. Head-to-head comparison of long-range Top-*L* precision between our pipeline in CASP13 and CASP14. (A) Contacts predicted using their own MSAs. (B) Contacts predicted using same MSAs generated by the newer pipeline.

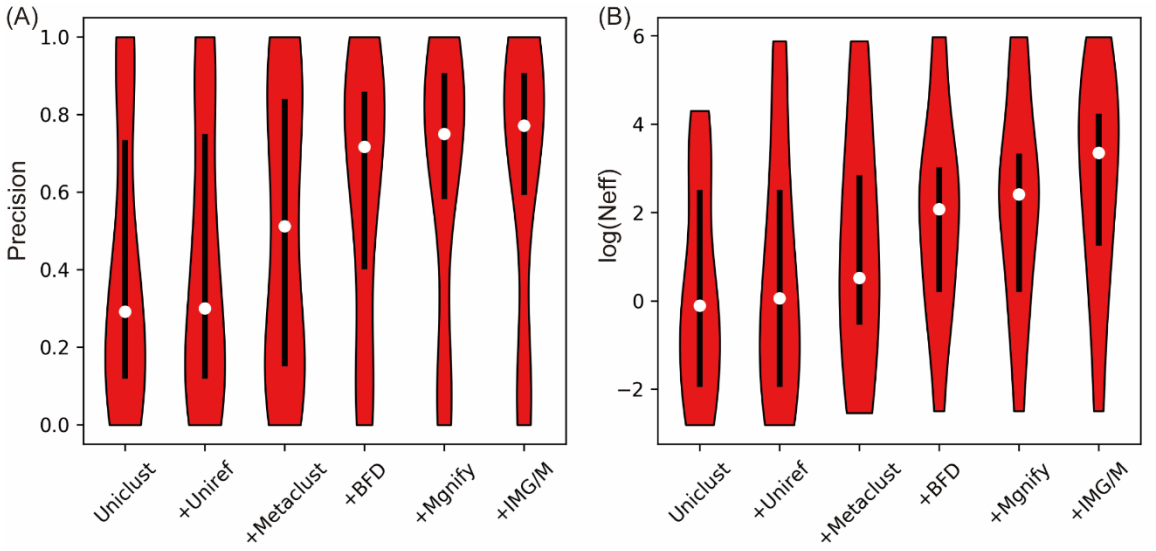


Figure 4. Increases of contact long-range Top-*L*/5 precision and average log(Neff) with the consideration of sequence databases illustrated in violin plots. White dots indicate the median value. Vertical black lines indicate 25% to 75% percentile. (A) Precision changes as more databases are used. (B) log(Neff) changes when more databases are employed.

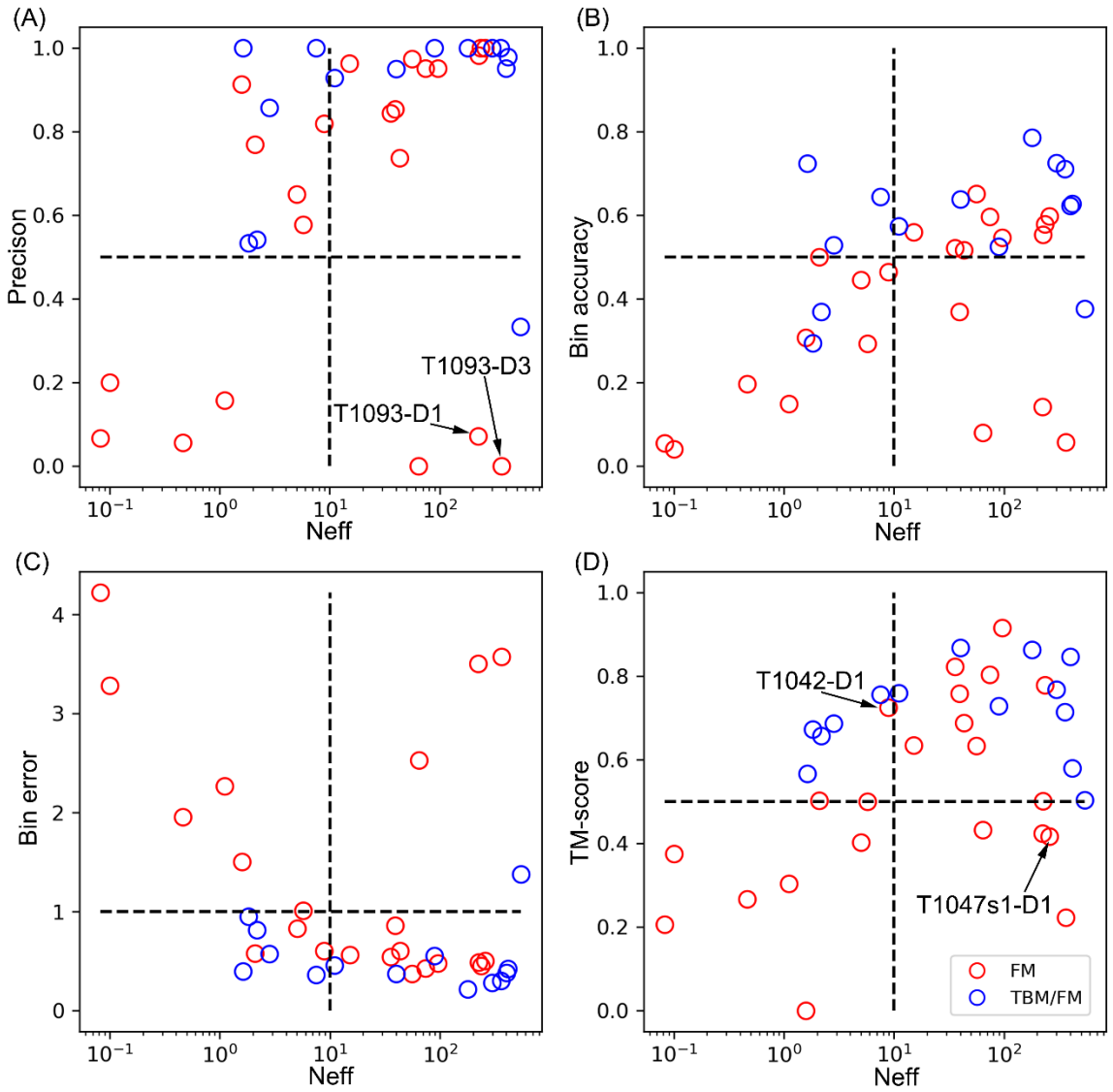


Figure 5. Illustration of the effect of MSAs on the performance DeepPotential. (A) The precision of long-range Top-*L*/5 contact prediction vs. Neff of MSAs. (B-C) Mean bin accuracy and Mean bin error of long rang Top-*L* distance prediction vs. Neff of MSAs. (D) TM-score of 3D models based on the prediction of DeepPotential vs. Neff of MSAs.

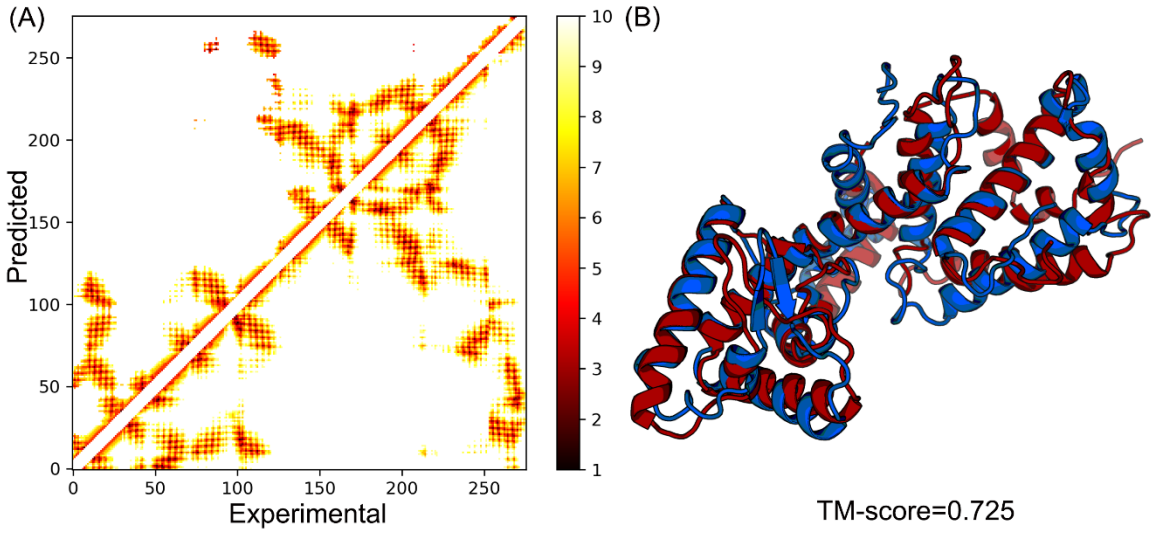
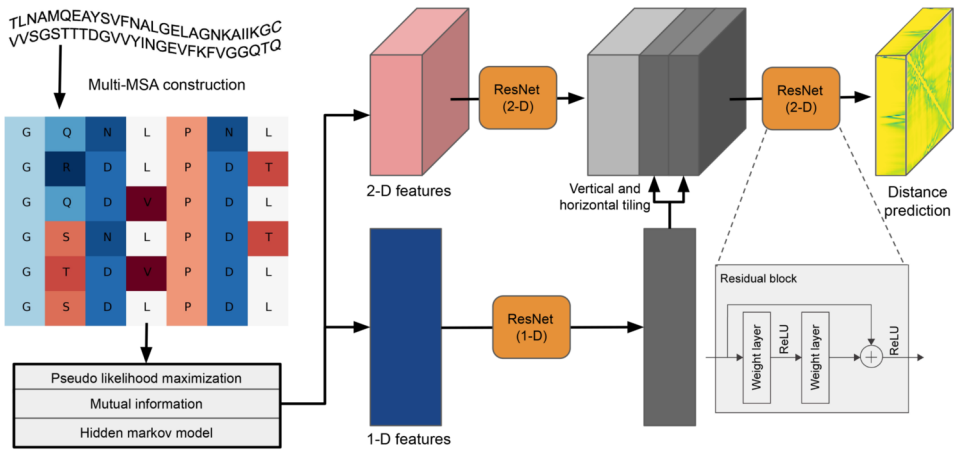
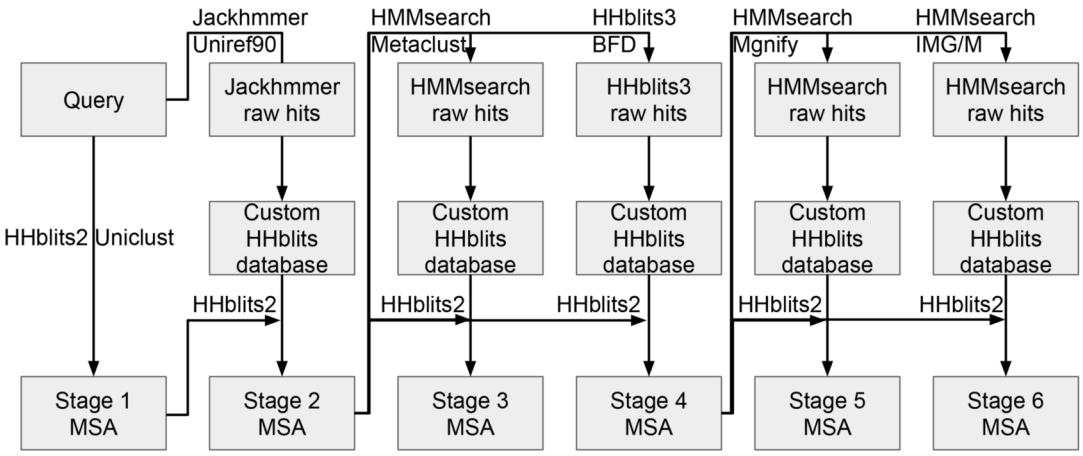


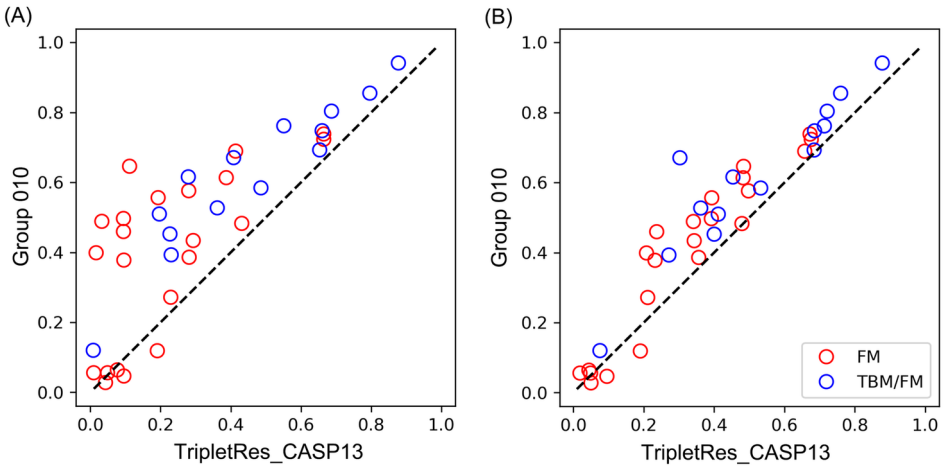
Figure 6. An illustrative example of CASP14 domain T1042-D1. (A) Comparison of predicted discrete distance map by DeepPotential and the distance map of the experimental structure. The distance bins are defined according to CASP format. (B) Superposition of submitted first model (blue) and experimental structure (red).



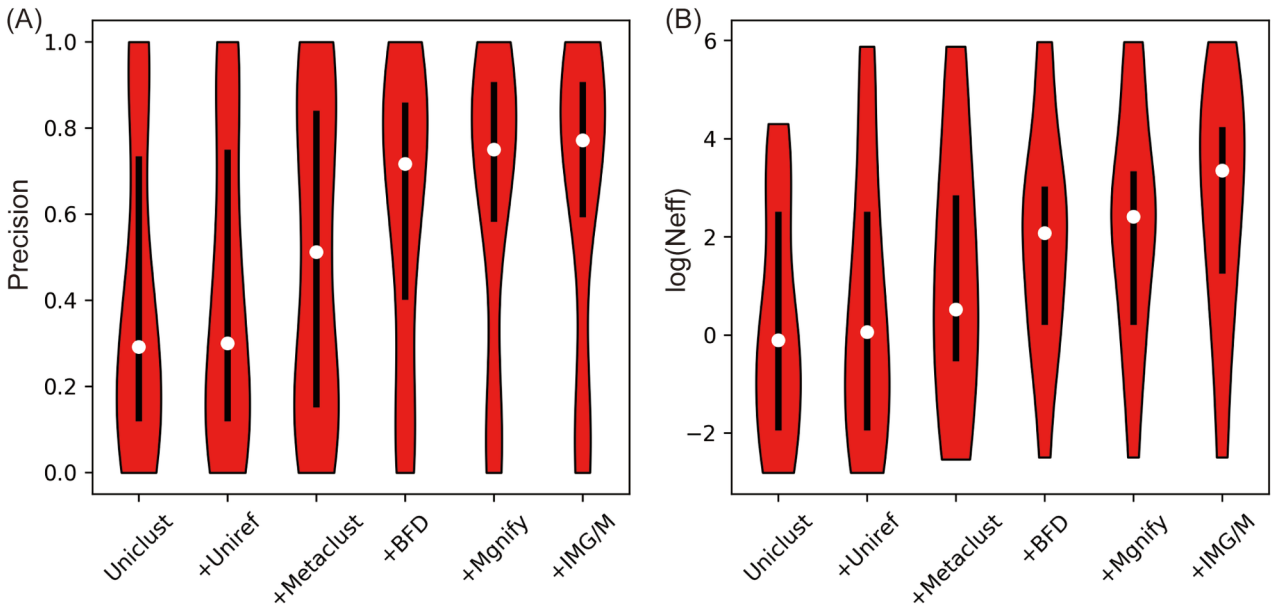
PROT_26211_1.tif



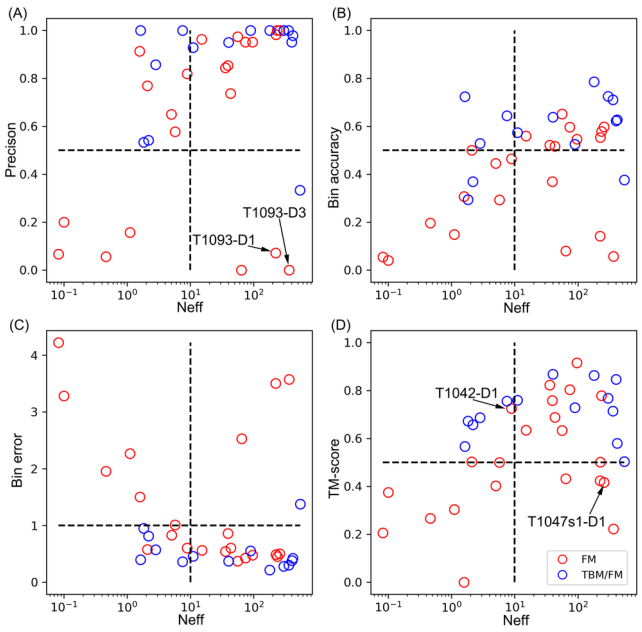
PROT_26211_2.tif



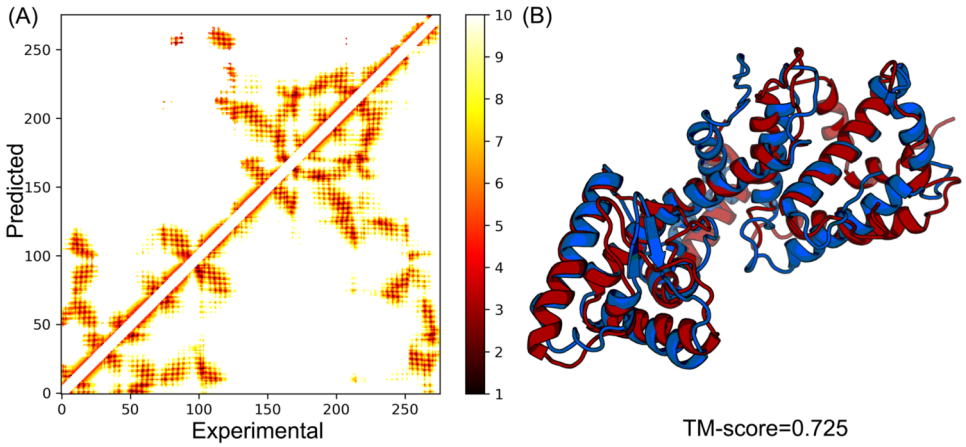
PROT_26211_3.tif



PROT_26211_4.tif



PROT_26211_5.tif



PROT_26211_6.tif