PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks

**Yang Li[1,2‡], Chengxin Zhang[2‡], Eric W. Bell[2], Wei Zheng[2], Xiaogen Zhou[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

‡ These authors are first senior authors on this work.
* njyudj@njust.edu.cn (DY); zhng@umich.edu (YZ)





**Citation:** Li Y, Zhang C, Bell EW, Zheng W, Zhou X, Yu D-J, et al. (2021) Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLoS Comput Biol 17(3): e1008865. https://doi.org/10.1371/journal.pcbi.1008865

**Editor:** Rachel Kolodny, University of Haifa, ISRAEL

**Received:** September 21, 2020

**Accepted:** March 10, 2021

**Published:** March 26, 2021



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files.

**Funding:** 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 to Y.Z.), the Natural Science Foundation of Jiangsu (BK20201304 to D.Y.) and the National Natural

## Abstract

The topology of protein folds can be specified by the inter-residue contact-maps and accurate contact-map prediction can help *ab initio* structure folding. We developed TripletRes to deduce protein contact-maps from discretized distance profiles by end-to-end training of deep residual neural-networks. Compared to previous approaches, the major advantage of TripletRes is in its ability to learn and directly fuse a triplet of coevolutionary matrices extracted from the whole-genome and metagenome databases and therefore minimize the information loss during the course of contact model training. TripletRes was tested on a large set of 245 non-homologous proteins from CASP 11&12 and CAMEO experiments and outperformed other top methods from CASP12 by at least 58.4% for the CASP 11&12 targets and 44.4% for the CAMEO targets in the top-*L* long-range contact precision. On the 31 FM targets from the latest CASP13 challenge, TripletRes achieved the highest precision (71.6%) for the top-*L*/5 long-range contact predictions. It was also shown that a simple re-training of the TripletRes model with more proteins can lead to further improvement with precisions comparable to state-of-the-art methods developed after CASP13. These results demonstrate a novel efficient approach to extend the power of deep convolutional networks for high-accuracy medium- and long-range protein contact-map predictions starting from primary sequences, which are critical for constructing 3D structure of proteins that lack homologous templates in the PDB library.

## Author summary

*Ab initio* protein folding has been a major unsolved problem in computational biology for more than half a century. Recent community-wide Critical Assessment of Structure Prediction (CASP) experiments have witnessed exciting progress on *ab initio* structure

Science Foundation of China (62072243, 61772273, to D.Y.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

prediction, which was mainly powered by the boosting of contact-map prediction as the latter can be used as constraints to guide *ab initio* folding simulations. In this work, we proposed a new open-source deep-learning architecture, TripletRes, built on the residual convolutional neural networks for high-accuracy contact prediction. The large-scale benchmark and blind test results demonstrate competitive performance of the proposed methods to other top approaches in predicting medium- and long-range contact-maps that are critical for guiding protein folding simulations. Detailed data analyses showed that the major advantage of TripletRes lies in the unique protocol to fuse multiple evolutionary feature matrices which are directly extracted from whole-genome and metagenome databases and therefore minimize the information loss during the contact model training.

This is a *PLOS Computational Biology* Methods paper.

## Introduction

Protein structure prediction represents an important unsolved problem in computational biology, with the major challenge on distant-homology modeling (or *ab initio* structure prediction) [1,2]. Recent CASP experiments have witnessed encouraging progress in protein contact predictions, which have been proven to be helpful to improve accuracy and success rate for distant-homologous protein targets [3–6].

The idea of developing sequence-based contact-map prediction to assist *ab initio* protein structure prediction is, however, not new, which can be traced back to at least 25 years ago [7,8]. In general, the methods for sequence-based protein contact-map prediction can be classified into two categories: coevolution analysis methods (CAMs) and machine learning methods (MLMs). In CAM, the predictors try to predict inter-residue contacts by analyzing evolutionary correlations of the target residue pairs from multiple sequence alignments (MSAs), under the assumption that correlated mutations in evolution usually correspond to spatial contacts of residue pairs. The CAMs can be further divided into local and global approaches. The local approaches use correlation coefficient, e.g., mutation information [9] and covariance [10], to predict contacts; these approaches are "local" because they predict contact between two residue positions regardless of other positions. In contrast, the global approaches, also called direct coupling analysis (DCA) methods, consider effects from other positions to better quantifying the strength of direct relationship between two residue positions. DCA models demonstrated significant advantage over the local approaches, and essentially re-stimulated the interest of the field of protein structure prediction in contact-map predictions. However, the success of most DCA methods [11–16] is still limited for the proteins with few sequence homologs, because a shallow MSA significantly reduces the accuracy of DCA to derive the inherent correlated mutations. In addition, DCA models only capture linear relationships between residues on MSA data (S1 Text) while residue-residue relationships in proteins are inherently non-linear.

As a more general approach, MLMs intend to learn the inter-residue contacts from sequential information and coevolution analysis features with supervised machine learning models trained with known structures from the PDB. Early attempts utilized support vector machines (SVMs) [17,18], random forests (RFs) [13,19], artificial neural networks (NNs) [20–23] etc., to

model the complex relationships between residues. Recently, great improvements have been achieved by the application of convolutional neural networks (CNNs) in several predictors, including DNCON2 [24], DeepContact [25] and RaptorX-Contact [26]. Most of the predictors were however trained on the final contact-map confidence scores [24–26], which may suffer coevolutionary information loss in data post-processing. In a recent study, we proposed ResPRE [27] which directly utilized the ridge-regularized precision matrices calculated from raw alignments without post-processing in regular coevolution analysis features. Although it uses the evolutionary matrix as the only input feature, the performance of ResPRE was comparable to many state-of-the-art methods that combine additional one-dimensional features, such as solvent accessibility, predicted secondary structure and physicochemical properties. Despite the success, ResPRE still bears several shortcomings. First, ResPRE lacks consideration for multiple coevolutionary matrices as features, which could provide complementary information. Second, it was trained by the supervision of binary protein contact-maps that lack continuous inter-residue distance information. Finally, the coevolution features were derived from a somewhat simplified HHblits [28] MSA collection procedure, which did not always include sufficient homologous sequences for meaningful precision matrix generation.

In this work, we proposed a new deep learning architecture, TripletRes, built on a residual neural network protocol [29] to integrate a triplet of coevolutionary matrices features from pseudolikelihood maximization of Potts model, precision matrix and covariance matrix for high-accuracy contact-map prediction (Fig 1). The model was trained on a non-redundant subset of sequences with known PDB structures supervised by discretized inter-residue distance-maps in order to capture the inherent distance information between residues, where a previously introduced deep MSA generation protocol [30] was employed to derive the coevolutionary matrices. The benchmark results on the public CASP and CAMEO targets, along with the community-wide blind tests in the CASP13 experiment, show that the new approach is capable of creating contact-maps with high precision. Although the TripletRes does not outperform the state-of-the-art methods trained after CASP13, the precision is higher than previous methods based on the same training set up to CASP13. An improvement of 9.2% in mean precision can be further observed based on an augmented training set after CASP13. Thus, TripletRes provides an alternative approach to protein contact-map prediction using multiple coevolution ensembles and is capable of achieving comparable performance to other available leading methods. The TripletRes server is available at https://zhanglab.ccmb.med.umich.edu/TripletRes/.

## Results

To examine the contact prediction pipelines, we collected two independent sets of test targets, including 50 non-redundant free-modeling (FM) domains from the CASP11 and CASP12 and 195 non-redundant targets assigned as *hard* by CAMEO [31]. TripletRes was trained on 7,671 non-redundant domains collected from SCOPe-2.07 (downloaded in March 2018) [32]. Here, non-redundancy is defined by setting the maximum pairwise sequence identity to 30%. Detailed procedures to obtain the training and testing datasets are described in S2 Text.

### Overall performance of TripletRes

Following the CASP criterion [4], two residues are defined as in contact if the Euclidian distance between their Cβ atoms (or Cα in case of Glycine) is below 8.0 Å. In this study, the accuracies, or mean precisions, of the top $L/10$, $L/5$, $L/2$, and $L$ of medium- ($12 \leq |i-j| \leq 23$) and long-range ($|i-j| \geq 24$) contacts are evaluated, where $i$ and $j$ are sequential indexes for the pair of considered residues and $L$ is the sequence length of the target. We focus on the performance

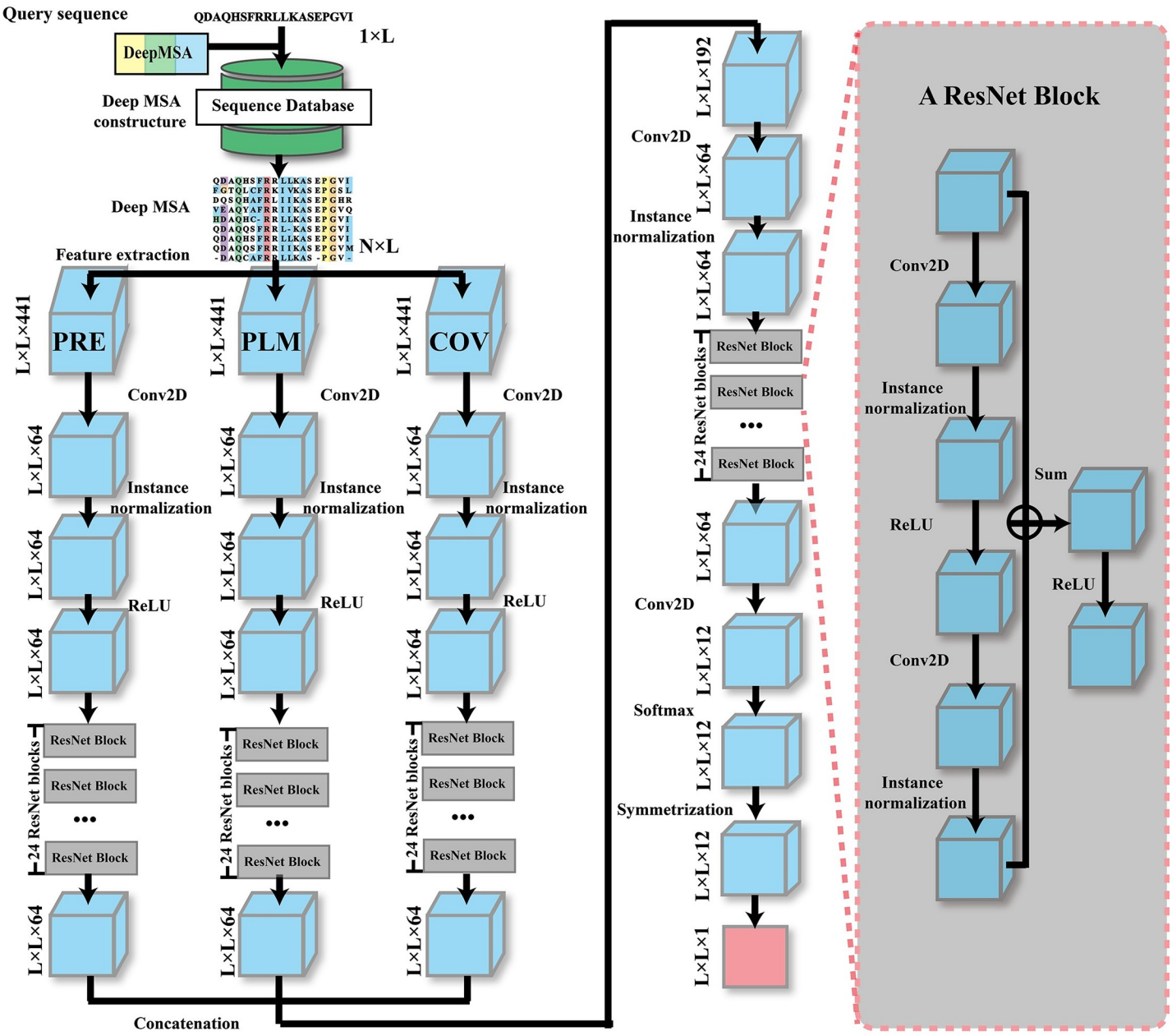


**Fig 1. The architecture of TripletRes, which formulates the contact-map prediction as a pixel-level labeling problem, where a pixel in the image represents a pair of residue positions in the contact-map of the query protein.** Starting from the MSA generated for the query sequence, three $L\times L\times 441$ feature matrices (also called tensors) are computed for the three sets of coevolutionary features (PRE, PLM, COV). Here, $L$ is the length of the query sequence while $441 = 21\times 21$ is the combination of all 21 amino acid types (including the gap) for two positions in the MSA. Each tensor is input to a separate ResNet, where the first layer reduces the number of feature channels from 441 to 64, followed by instance normalization and 24 consecutive residual blocks to get an $L\times L\times 64$ tensor. Details of a residual block are shown on the right-hand side inset. The three tensors from the three ResNets are concatenated into an $L\times L\times 192$ tensor to feed into a final ResNet. In this ResNet, the first layer again reduces the feature channels from 192 to 64, followed by instance normalization, and 24 residual blocks to get an $L\times L\times 64$ tensor, which is further reduced to $L\times L\times 12$. Finally, a softmax layer is used to scale the values in the tensor between 0 and 1 and to make the sum of all values for each pixel (i.e. residue pair) equal to one. Since a protein contact/distance map is symmetric, TripletRes averages the corresponding softmax output of residue pair $(i,j)$ and $(j,i)$ to get the final $L\times L\times 12$ distance-map prediction, where 12 stands of the number of distance bins. The contact-map is obtained by summing up the first 4 distance bin.

https://doi.org/10.1371/journal.pcbi.1008865.g001

of FM targets (or hard targets in CAMEO) and on long-range contacts for evaluation, since the metric is most relevant for assisting the prediction of the tertiary structure of non-homologous proteins [6,33].

**Table 1. Summary of long-range contact precision by TripletRes and control methods on 50 CASP11&12 FM targets and 195 CAMEO hard targets, sorted in ascending order of top-*L* precision.** *p*-values in parenthesis are from a Student's t-test between TripletRes and each of the control methods, where bold fonts highlight the best performer in each category.

| Methods | 50 CASP FM targets | | | | 195 CAMEO hard targets | | | |
|---|---|---|---|---|---|---|---|---|
| | *L*/10 | *L*/5 | *L*/2 | *L* | *L*/10 | *L*/5 | *L*/2 | *L* |
| CCMpred | 0.416 (1.0e-11) | 0.374 (3.2e-13) | 0.264 (2.6e-16) | 0.187 (4.5e-17) | 0.451 (1.0e-50) | 0.411 (5.7e-56) | 0.314 (2.8e-66) | 0.229 (4.6e-67) |
| DNCON2 | 0.482 (3.8e-08) | 0.446 (8.8e-09) | 0.369 (3.3e-10) | 0.286 (2.4e-11) | 0.635 (4.1e-17) | 0.574 (1.7e-23) | 0.453 (1.3e-33) | 0.339 (1.9e-39) |
| MetaPSICOV2 | 0.522 (2.0e-07) | 0.467 (2.2e-08) | 0.368 (2.1e-11) | 0.283 (9.8e-13) | 0.585 (8.8e-23) | 0.528 (7.7e-28) | 0.415 (3.2e-36) | 0.313 (1.5e-39) |
| DeepContact | 0.497 (4.1e-08) | 0.466 (2.5e-07) | 0.380 (4.4e-10) | 0.293 (1.1e-11) | 0.633 (1.3e-14) | 0.579 (1.7e-18) | 0.455 (7.9e-31) | 0.340 (1.1e-35) |
| ResPRE | 0.580 (1.7e-06) | 0.535 (7.3e-07) | 0.439 (3.0e-07) | 0.339 (1.9e-07) | 0.756 (1.2e-04) | 0.703 (9.2e-08) | 0.573 (4.0e-17) | 0.436 (3.8e-20) |
| TripletRes | **0.771** | **0.714** | **0.597** | **0.464** | **0.801** | **0.756** | **0.637** | **0.491** |

https://doi.org/10.1371/journal.pcbi.1008865.t001

Table 1 summarizes the overall performance of long-range contact prediction on the two test datasets by TripletRes, in control with five state-of-the-art methods which are available for free-download and run with default setting (see S3 Text for introduction of the control methods). The results show that TripletRes creates contact models with a higher accuracy than the control methods in all separation ranges for both test datasets. For example, on the 50 FM CASP targets, the average precision of the long-range top *L*/10, *L*/5, *L*/2, and *L* predicted contacts by TripletRes is 55.1%, 53.2%, 57.1%, and 58.4% higher, respectively, than the precision achieved by DeepContact, the most accurate third-party program in this comparison, which correspond to statistically significant *p*-values of 4.1e-08, 2.5e-07, 4.4e-10, and 1.1e-11 in the Student's t-test. Notably, TripletRes only uses coevolutionary features, which is a subset of the diverse features employed by DeepContact. The better performance is also probably due to the more effective integration of raw coevolutionary information in the TripletRes neural-network training.

TripletRes also outperforms ResPRE, an in-house program previously trained on precision matrix [27], by a large margin. The long-range top-*L* precision of TripletRes is 36.9% higher than that of ResPRE with a *p*-value of 1.9e-07 on 50 FM targets. ResPRE achieved a significantly higher precision on CAMEO than the FM dataset, but its precision is still lower than that of TripletRes. For example, the mean precision of the top-*L* long-range contacts by TripletRes is 12.6% higher than that of ResPRE on the CAMEO targets. Given that both programs utilized the same precision matrix feature, the superiority of TripletRes is mainly attributed to the integrations of triplet coevolutionary features. In addition, as examined in detail below, the supervision of the distance predictions and the new deep MSA constructions also helped improve the accuracy of the TripletRes models.

The proposed TripletRes pipeline's performance could be overrated since more data have been used compared to those methods in CASP11&12. To reduce the bias, we have ensured that the maximum pairwise sequence identity is 30% between training and test set. In addition, we have tweaked those control methods by replacing their MSAs with DeepMSA and S1 Table presents the performance of TripletRes and control methods after the tweaks. The use of DeepMSA improves all control methods, including ResPRE, for which the top-*L* precision increases from 33.9% to 42.9% on 50 CASP FM targets. Nevertheless, TripletRes still take the lead over the control methods and the top-*L* precisions on CASP and CAMEO targets are 28.8% and 27.9% higher than those of the best third-party programs, DeepContact.

## Feature extraction based on raw potentials outperforms that with post-processing

Feature extraction is essential for all machine-learning based modeling approaches. To quantitatively examine the effectiveness of the feature extraction strategy and the contribution of

different feature types in TripletRes, we compare in Fig 2A–2C the performance of two feature extraction strategies, based on three component features from covariance (COV), precision (PRE), and pseudolikelihood maximization (PLM) analyses (see "Coevolutionary feature extraction" in Methods and Materials), respectively. The first feature extraction strategy, which was used by TripletRes, uses the raw coevolution potentials as input features, while the second strategy, which was commonly employed in many state-of-the-art predictors [22,24,25,34], employs a specific post-processing procedure as described in Supplementary Eqs A and B in S4 Text. Since the traditional coevolutionary features can also be used to predict contacts directly without using supervised training, we list their performance as baselines (see dotted lines in Fig 2A–2C). Here, a total of 767 sequences are randomly selected from 7,671 non-redundant SCOPe proteins as the validate set, while the remaining 6,904 sequences are used as the training set for feature extraction strategy selection in TripletRes. All experiments are performed by keeping other elements (e.g., MSA generation, neural network structure and its hyper-parameters) fixed.

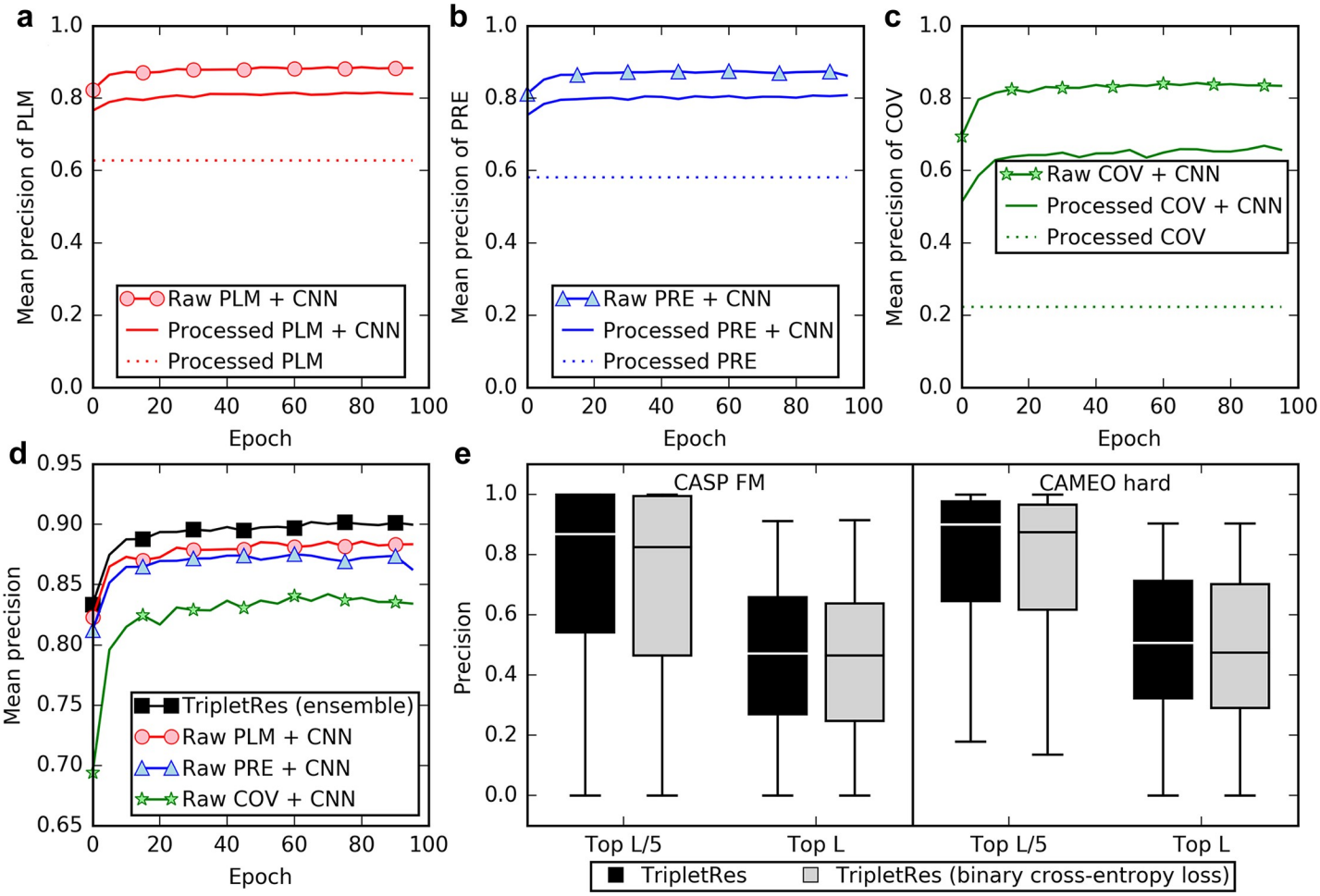


**Fig 2. Comparisons of different strategies used to train TripletRes. (a-c)** Comparisons of the average long-range top-$L/5$ precisions over training epochs using different feature extraction strategies but trained with the same deep neural-network structure on three different coevolutionary analysis methods: **(a)** DCA based on pseudolikelihood maximization (PLM), **(b)** DCA based on the precision matrix (PRE), **(c)** Covariance analysis (COV) for contact-map prediction, on the validation set. "Processed" means the coevolutionary features are post-processed by Eqs A and B in S4 Text. **(d)** Comparison of the average long-range top-$L/5$ precisions over training epochs of individual coevolutionary features and the TripletRes model that ensembles all three sets of features, on the validation set. Each curve is for the training of a single model. **(e)** Comparison of long-range top-$L/5$ and top-$L$ precisions with different loss functions on the CASP FM and CAMEO hard targets.

https://doi.org/10.1371/journal.pcbi.1008865.g002

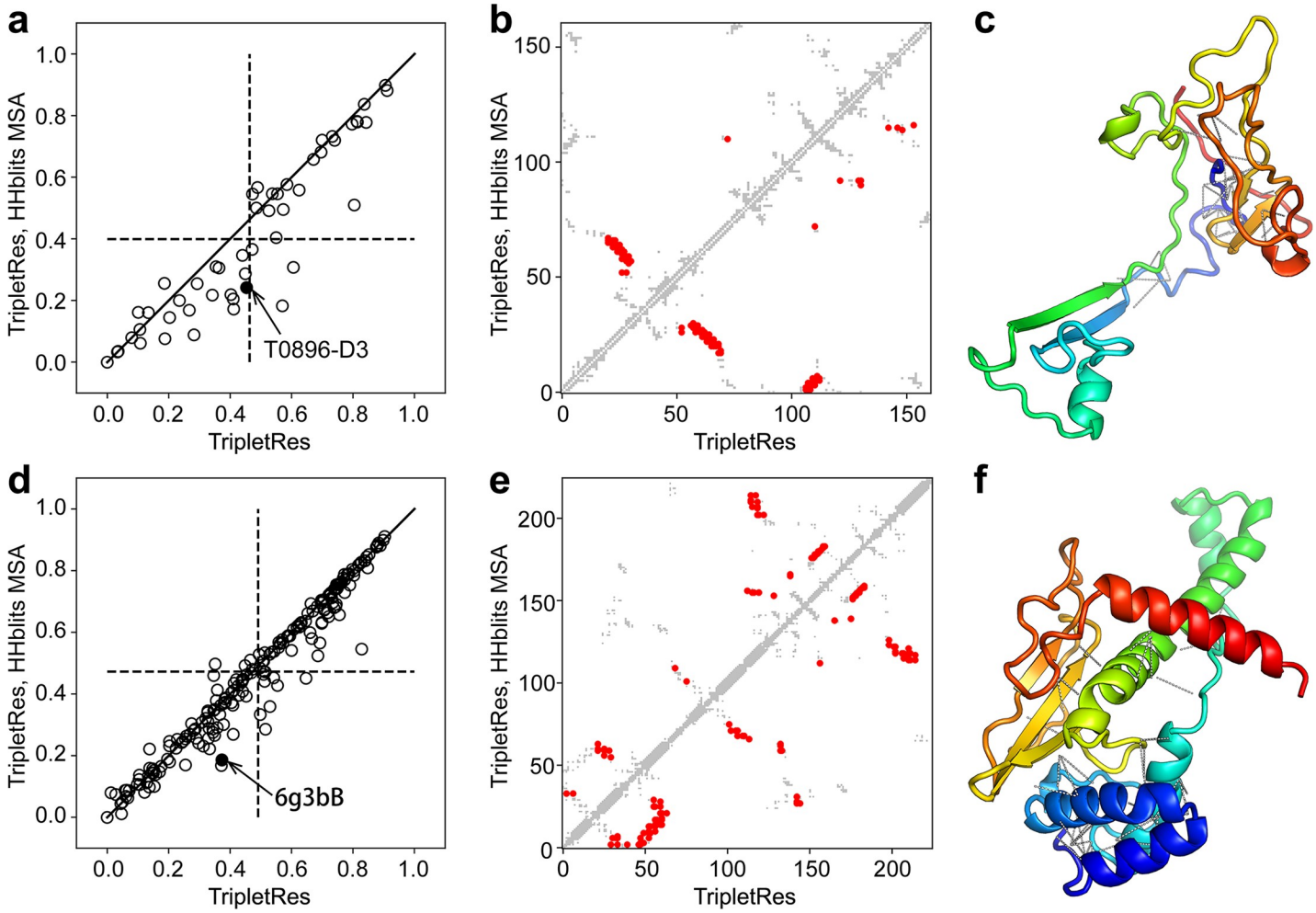


**Fig 3. Long-range top-*L* precision of contact-maps predicted by TripletRes with deep MSAs versus that without deep MSAs.** **(a)** overall results on 50 CASP FM targets; (**b,c**) illustrative example of contact-map and the native structure of the T0896-D3 domain in CASP12; (**d**) Overall results on 195 CAMEO hard targets; **(e,f)** illustrative example of the contact-map and the native structure of the PDB ID 6g3bB in CAMEO. In (**a**) and (**d**), dashed lines mark the average precision of the top-L long-range contact prediction. In (**c**) and (**f**), dashed lines label the additional contacts predicted due to the employment of deep MSA.

https://doi.org/10.1371/journal.pcbi.1008865.g003

It can be observed from Fig 3A–3C that the new feature extraction strategy achieves a better contact prediction performance compared to the traditional feature extraction for all three considered matrix features. The highest mean precisions of the new feature extraction strategy on the long-range top-*L*/5 contact prediction are 84.2%, 87.5%, and 88.6%, respectively, for COV, PRE, and PLM features. If the post-processed features of Eqs A and B in S4 Text are used, the mean precisions are reduced to 66.8%, 80.8%, and 81.6%, which represent a precision drop by 20.7%, 7.7%, and 7.9%, respectively, compared to the TripletRes feature extraction strategy. On the other hand, the mean precisions of both feature extraction strategies are consistently higher than the baseline through the training epochs, indicating the necessity of supervised training.

One reason for the performance degradation by the post-processing approach is that the potential score for different types of residue-pairs have been treated equally and the sign of these potential scores is thus completely ignored in Eq A in S4 Text, when the post-processed coevolutionary features are fed to the supervised models. In contrast, the approach in TripletRes can keep detailed score information of different residue-pair types from the coevolutionary

analyses for each residue pair, and thus allow for deep residual neural networks to automatically learn the inter-residue interactions not only on the spatial information but also on the residue pair-specific scores of different residue-pair types, while the traditional supervised machine learning models can only learn the spatial information of each residue pair during the training.

## Ensembling different component features improves contact-map prediction

Compared to ResPRE [27], a major new development in TripletRes is on the integration of multiple coevolutionary feature extractions. To examine the efficiency of ensembled feature collection on the contact predictions, Fig 2D presents the average long-range top $L/5$ precisions of the predictors trained by three individual component features and their ensemble on the validation set containing 767 sequences. Note that the feature ensemble model in Fig 2D was trained by keeping other elements (e.g., MSA generation, training set, neural network structure and its hyper-parameters) identical to other individual feature models. All models shown in Fig 2D become stable after 40 rounds of training, and obtain a precision of 88.4%, 86.3%, 83.4%, and 90.0% when using PLM, PRE, COV and an ensemble of all three features, respectively, after 100 epochs of training. In general, the COV-based model has the lowest precision among the three individual feature models, probably due to the translational noise in the covariance matrix [27]. The performance of the two DCA-based features by PRE and PLM are comparable and both consistently outperform COV by a large margin. TripletRes ensembles three features that can obtain more comprehensive coevolutionary information from the deep MSAs. As a result, the ensemble model has a higher precision than all models from the component features, demonstrating the effectiveness of multiple feature integration.

To perform a critical analysis of individual features' contributions, S1 Fig compares the precision of TripletRes against the feature sets without corresponding particular features on the validation set. For both top-$L/5$ and top-$L$ precisions, excluding the PLM feature has the lowest values during the training process, indicating that the PLM feature makes significant contributions. Interestingly, feature sets without PRE or COV feature and feature ensemble seem to be indistinguishable for top-$L/5$ precision. While for top-$L$ precision, the full TripletRes with the triplet features ensemble stands out, achieving a precision of 68.2%, higher than the precisions of 67.3%, 67.5%, and 67.4% without COV, PRE, and PLM feature respectively. Surprisingly, COV and PRE seem to have similar contributions to the TripletRes model, even though the model using only PRE feature is previously shown to significantly outperform the model using only COV feature (Fig 2). The reason could be that COV and PLM are two different kinds of co-evolutionary features, i.e., local and global, providing complementary information when ensembled by TripletRes. In other words, all considered features make contributions, and a combination of all three feature generates the most robust contact models.

In CASP13, in addition to TripletRes_CASP13 that used an ensemble of PLM, COV and PRE features, the individual raw PLM and COV features have also been utilized by AlphaFold [35] and DMP [36], respectively. The inverse of the covariance matrix, i.e., the PRE feature (with a different derivative) has also been considered by trRosetta [37] afterward. Thus, the introduction of the concept of multiple raw coevolutionary feature ensemble should help improve individual methods and push the boundary of inter-residue contact/distance prediction.

## Loss function with continuous distances outperforms that with binary contacts

The correct loss function selection plays an important role in the training of neural networks because it determines the performance metric of the model during training. The most

commonly used loss function for contact-map predictions is the binary cross-entropy loss function, which encodes each residue pair with 2 states (contact and not in contact). Typically, with a single distance threshold of 8 Å, such a loss function does not encode detailed distance information, e.g., residue pairs separated by 9 Å will be treated the same as those by 22 Å. Alternatively, recent methods [34,35,37] have considered predicting discretized distance distribution matrices rather than contact-maps, mostly for assisting 3D structure prediction. However, whether incorporating distance training could help contact-map prediction accuracy remains unstudied. Inspired by those works, the loss function in TripletRes (Eq 6 in Methods and Materials) considers a discrete representation of each residue pair's distance information. We then systematically evaluate the impact of adding distance information during the training on the accuracy of contact-map prediction.

Fig 2E compares the long-range top-$L$/5 and top-$L$ precisions between TripletRes programs trained with Eq 6 or a binary cross-entropy loss (see Eq A in S5 Text) on CASP FM and CAMEO hard targets, respectively. It can be observed that incorporating continuous distance information in training can lead to improvements in contact-map prediction, even though the contact-maps are not directly optimized. For example, the distance information in the loss function can improve the top-$L$/5 precision from 68.7% to 71.4%, and 74.1% to 75.6%, for the CASP set and CAMEO set, respectively, which correspond to a $p$-value of 2.1e-02 and 7.9e-03 in Student's t-test. Interestingly, when more top-ranked contacts are considered (i.e., top-$L$), the $p$-values become more significant and decrease to 1.6e-04 and 2.0e-07 on the two datasets, respectively, which means the distance information may have a stronger effect on improving the precision when more contacts are evaluated. Protein structure prediction methods can thus benefit more from TripletRes, which was trained with the discrete distance loss function because more predicted contacts can be reliably considered as restraints for protein folding.

To have a detailed analysis of the effect of discrete distance loss function on different fold types, S2 Fig presents the comparison of long-range top-$L$/5 and top-$L$ precisions with different loss functions on the different fold types, with median and mean precisions marked in solid and dash lines, respectively. Structures from 195 CAMEO set and CASP 11&12 set are classified into 63 alpha proteins, 24 beta proteins and 157 alpha&beta (alpha+beta and alpha/beta) proteins. For three fold types, consistent improvements can be observed with the distance loss function for all evaluation indexes. For example, for long rang top-$L$ predicted contacts, training with discrete distance loss function achieves precisions of 34.9%, 49.0%, and 54.0% for alpha, beta, and alpha&beta folds, which are slightly higher than the baselines, corresponding to $p$-values of 4.0e-03, 2.0e-02 and 9.4e-08, respectively. Among three types of fold types, alpha proteins have the lowest mean top-$L$/5 and top-$L$ precision, regardless of the loss function type; this may be due to the fact that contact patterns, including hydrogen-bonds, between alpha-helical segments are not as evident as those between beta-strand elements in proteins.

## Deep MSA search help create more comprehensive coevolutionary information

TripletRes utilizes MSAs as the only input and the quality of the latter is thus essential to the final contact prediction models. It is worth noting that the TripletRes model is trained on features extracted from MSAs generated by HHblits, but a deeper MSA generated by multiple databases has been used for test proteins (see "MSA generation" in Methods and Materials). We expect the strategy could reduce over-fitting between the training and test proteins.

To examine the impact of different MSA collections on the contact models, Fig 3 shows a comparison of TripletRes models with and without deep MSAs on the test proteins from

CASP FM targets (Fig 3A) and CAMEO hard targets (Fig 3D). Here, dashed lines mark the mean precision value of the long-range top-*L* prediction by each dataset. For the CASP FM targets, the usage of deep MSAs during testing significantly improves the mean precision of TripletRes from 40.0% to 46.4% with a *p*-value 1.9e-05 in Student's t-test, where 35 out of 50 FM targets (70%) achieve a higher precision with deep MSAs while only 8 targets (16%) do so when the HHblits MSAs are used. The same trend can be observed in the CAMEO targets, where the *p*-value of improvement in long-range top-*L* precision is 1.7e-06. This difference is mainly due to the higher number of homologous sequences collected in deep MSA search protocol, which allows the extraction of more reliable coevolutionary information. For example, the average number of effective sequences of MSAs, or Neff calculated by Eq 1 in Methods and Materials, generated by deep MSA is 85.4, which is 34.3% higher than that obtained by HHblits on CASP FM targets (63.6).

In Fig 3B, 3C, 3E and 3F, we select two illustrative cases from the CASP and CAMEO datasets respectively. The example in Fig 4B is from the third domain of CASP12 target T0896 with experimental structure presented in Fig 4C, where HHblits collects a relatively shallow MSA with a Neff = 0.94, which resulted in only 39 true positives in the 162 long-range top-*L* contact predictions. The deep MSA search increased the Neff value to 3.78, where the number of true contacts with the deeper MSA increases to 73, which is 87.2% higher than that with the HHblits MSA. In Fig 3E and 3F, the structure comes from the Type II site-specific deoxyribonuclease (PDBID: 6g3bB) with 225 residues, where HHblits creates an MSA with Neff = 2.0 and results in 42 true positives out of top-*L* long-range predictions; while 42 more contacts are detected by TripletRes through the deep MSA that has a Neff of 14.0. These examples highlight again the importance of using deep MSA pipeline for coevolutionary feature collection and the impacts on final contact-map prediction. The new contacts correctly predicted after performing deep MSA searching strategy are marked as dashed lines in Fig 3C and 3F; these contacts provide additional spatial restraints and have shown critical in creating correct global fold for the domain structures [38].

## Performance of TripletRes for blind prediction in CASP13

An early version of TripletRes, denoted as TripletRes_CASP13, participated in the 13th CASP experiment for inter-residue contact prediction [6,35]. It was ranked among the top two methods based on the mean precision score (http://www.predictioncenter.org/casp13/zscores_rrc.cgi), with another top method RaptorX-Contact which also ranked as the top method in previous CASPs. In Table 2, we list a summary of the average results by TripletRes and TripletRes_CASP13, along with three other top CASP13 predictors from RaptorX-Contact, DMP, and ZHOU-Contact. For the long-range top-*L*/5 contacts on the 31 FM targets, TripletRes_CASP13 achieved a mean precision of 64.6%, while the mean precision of RaptorX-Contact, DMP, and ZHOU-Contact are 69.4%, 60.2%, and 58.3%, respectively. TripletRes, however, achieves the highest precision of 71.6% for long-range top *L*/5 contacts. Here, TripletRes and TripletRes_CASP13 are based on the same input MSAs and the only difference between them is that TripletRes utilizes a new loss function (Eqs 6 and 7 in Methods and Materials) to integrate distance profiles for contact-maps, while TripletRes_CASP13 used a binary cross-entropy loss function (Eq A in S5 Text). These data demonstrate the validity of the distance-supervised training strategy.

In a recent study, trRosetta [37] reported an alternative MSA construction approach by performing HHblits and hmmsearch search through a much larger propriety database with ~7 billion sequences. In comparison, the Metaclust database used by DeepMSA only has 424 million sequences. Unfortunately, both the scripts and the database used in the trRosetta MSA

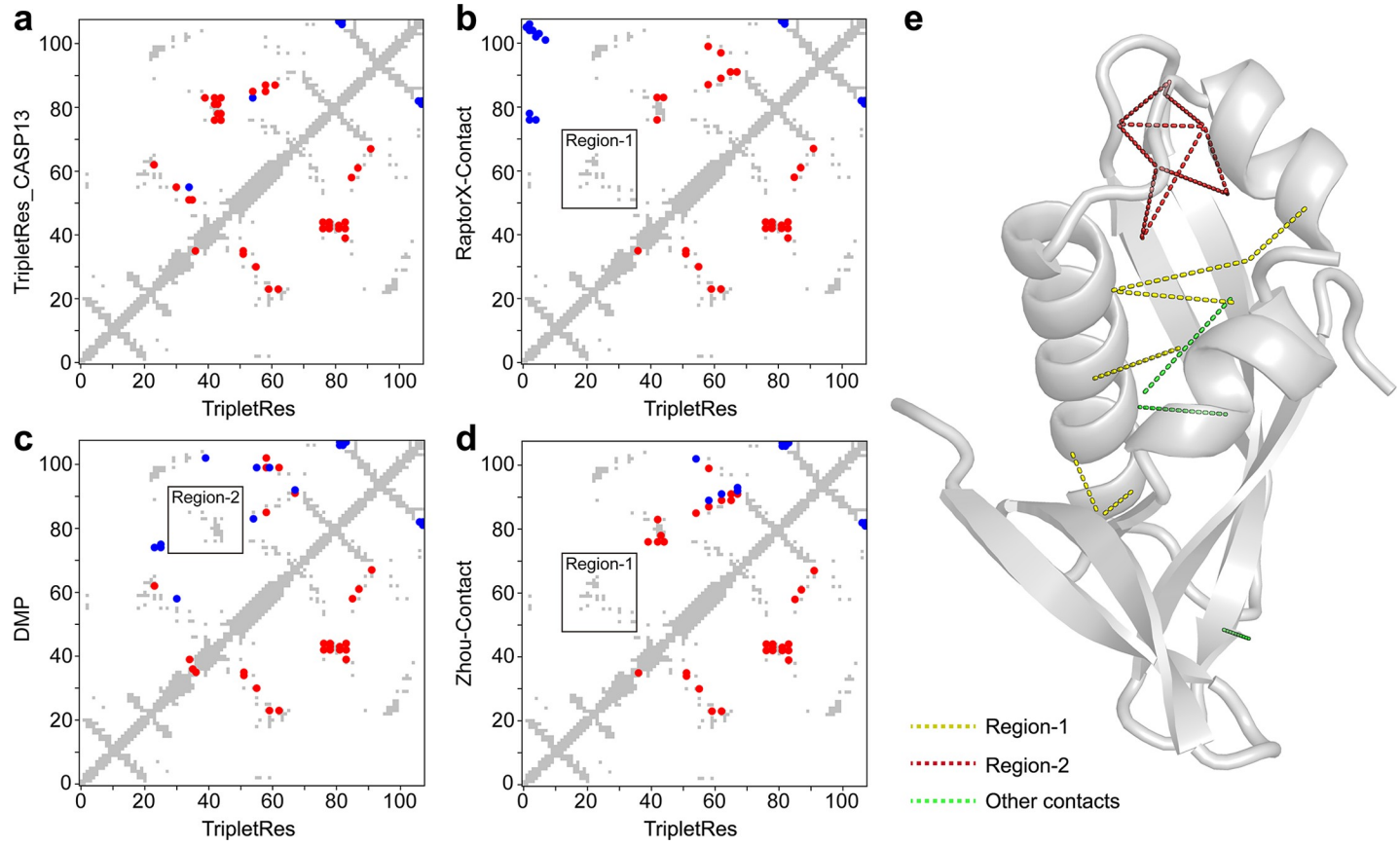


**Fig 4. An illustrative example of a CASP13 domain T0957s1-D1 showing a comparison of top-*L*/5 long-range contact prediction by TripletRes and the control methods.** In each map, the true contacts are marked in grey, true positives in red, and false positives in blue. **(a-d)** The comparison between TripletRes_CASP13, RaptorX-Contact, DMP, and ZHOU-Contact (in upper-left triangle) against TripletRes (in lower-right triangle). **(e)** Experimental structure of T0957s1-D1, with the long-range true positive prediction by TripletRes in Region 1, Region 2 and others marked in yellow, magenta and green dashed lines, respectively.

https://doi.org/10.1371/journal.pcbi.1008865.g004

construction are unavailable, preventing us from testing DeepMSA on the same database. Nonetheless, we observed that the top-*L*/10, *L*/5, *L*/2 and *L* precisions could be boosted to 84.1%, 78.4%, 62.0% and 47.1%, respectively, by simply feeding TripletRes model with pre-generated MSAs downloaded from the trRosetta [37] website. The average Neff value of

**Table 2. Performance comparisons on CASP13 FM targets between TripletRes and RaptorX-Contact, DMP, and ZHOU-Contact servers, sorted in ascending order of top *L* long-range contact precision.**

| Method | Medium range | | | | Long-range | | | |
|---|---|---|---|---|---|---|---|---|
| | *L*/10 | *L*/5 | *L*/2 | *L* | *L*/10 | *L*/5 | *L*/2 | *L* |
| ZHOU-Contact | 0.727 | 0.623 | 0.453 | 0.319 | 0.641 | 0.583 | 0.474 | 0.367 |
| DMP | 0.772 | 0.682 | 0.505 | 0.344 | 0.645 | 0.602 | 0.470 | 0.361 |
| TripletRes_CASP13* | 0.842 | 0.746 | 0.543 | 0.360 | 0.695 | 0.646 | 0.534 | 0.409 |
| RaptorX-Contact | 0.805 | 0.702 | 0.527 | 0.364 | 0.762 | 0.694 | 0.567 | 0.438 |
| TripletRes* | **0.865** | **0.770** | **0.562** | **0.367** | **0.775** | **0.716** | **0.573** | **0.440** |

* "TripletRes" is the current version of TripletRes trained using distance-based loss function (Eq 6). "TripletRes_CASP13" is the early version of TripletRes used in CASP13, trained using binary cross-entropy loss function (Eq A in S5 Text).

https://doi.org/10.1371/journal.pcbi.1008865.t002

trRosetta generated MSAs is 82.18, which is 2.6 times higher than that of DeepMSA. This data confirmed again the impact of the size of sequence databases on the contact prediction models.

In Fig 4, we present an example from the first domain of T0957s1 of CASP13 which is a contact-dependent growth inhibition toxin-immunity protein (PDB ID:6cp8) with an α+β fold and 108 residues. TripletRes collected a deep MSA with Neff = 6.7, significantly higher than the Neff value (1.3) by HHblits. This resulted in a mean precision of 86.4% for the top-*L/5* long-range contact predictions, compared to 40.9% by RaptorX-contact, 36.4% by DMP, and 54.5% by ZHOU-Contact, respectively. TripletRes also performed better than the CASP13 version in precision (77.3%), benefited from the distance information during the training. As shown in Fig 4B and 4D, RaptorX-Contact and ZHOU-Contact failed to hit any long-range contacts in Region 1 which is a critical loop-loop contact region. DMP, on the other hand, was not able to cover contacts in Regions 2 that are important to pack the core structure of the two helices with the center beta-sheet (Fig 4C). TripletRes can cover both Regions marked in yellow and magenta in Fig 4E, respectively. Among the top-*L/5* correctly predicted long-range contacts, 94.7% of them have the distance profile with a probability peak at <8Å and nearly 74% of the residue pairs have the accumulated probability >80% in the region below 15Å, indicating a high confidence of contact prediction on the residue pairs based on the distance profile.

Note that both TripletRes_CASP13 and TripletRes are trained on the same training set before CASP13. To examine the impact of the size of training dataset on the proposed framework, we re-train the TripletRes model with a dataset newly collected after CASP13 with 26,151 PDB sequences and perform the evaluation on a test set containing 37 sequences (S2 Text), with the re-trained model termed as TripletRes (Post-CASP13). S2 Table lists the overall performance of TripletRes (Post-CASP13) in comparison with TripletRes and trRosetta, considering that trRosetta is the representative method that predicts inter-residue geometric terms for protein folding after the CASP13 season. DeepMSA was employed to generate MSAs for the test set for its availability and all control methods are sharing the same MSAs. It is shown in S2 Table that the performance can be considerably improved by a simple employment of a larger training set. TripletRes (Post-CASP13) achieves a top-*L*/5 precision of 76.2% on the 37 test sequences, 9.2% higher than that of TripletRes with a p-value of 1.7e-03. Such differences in performance with different amounts of training data indeed demonstrate the importance of available dataset when training the model. Compared to TripletRes, trRosetta has a slightly higher prediction along all the cutoffs; the difference is however statistically insignificant, with the p-value equal to 0.68, 0.59, 0.15, and 0.20 for top-*L*/10, *L*/5, *L*/2 and *L* precisions, respectively. It is noted that the higher contact accuracy by trRosetta is mainly attributed to various auxiliary prediction tasks such as orientation prediction, while for TripletRes, the improvements mainly come from the ensemble of multiple co-evolutionary features. In this sense, the proposed TripletRes method should be considered complementary to the trRosetta.

Apart from trRosetta and DMP discussed above, AlphaFold [35] also perform contact/distance prediction by predicting discretized distance bins. While AlphaFold did not participate in the contact prediction category of CASP, its top-*L* long range contact precision has reportedly achieved 46.1% [35], which was higher than what TripletRes achieved in CASP13. In the MSA generation step, AlphaFold performs a routine HHblits search through the standard Uniclust database, which is equivalent to Stage 1 of our three-stage DeepMSA approach. The input features of AlphaFold is mainly derived from PLM, which is only a subset of our triplet features. Given its simple MSA and feature design, part of the advantage of AlphaFold over TripletRes is the complexity of neural network architecture. Since the DeepMind team has access to computational resources unattainable for most academic groups, it can train a neural

network with 220 residual blocks. In comparison, due to the resource limit, TripletRes can only be trained with 24 residual blocks for each of its three ResNet branches (corresponding to the three sets of input feature) and another 24 residual blocks for fusing the three branches. Meanwhile, the iterative 3D model construction and contact prediction procedures can further improve the contact prediction accuracy since the process of 3D structure construction can help filter out physically non-practical contacts.

The re-training of the TripletRes (Post-CASP13) model took up to 30 days on 4 Nvidia P100 GPUs from the public XSEDE Comet Cluster [39] due to the heavy I/O loads of pre-calculated feature data. However, the running time during the test should be theoretically comparable with regular methods, e.g., AlphaFold or RaptorX-Contact. The full 3-stage DeepMSA pipeline, on average, takes 1.32 hours [30] on its benchmark set. After that, the majority of the time would be spent on the calculation of the PLM matrix. To the best of our knowledge, the CCMpred program utilized by TripletRes to calculate the PLM matrix is one of the most efficient programs in the field.

## Conclusion

Protein contact-map prediction has been critical to assist protein folding in the form of spatial constraints. This work presented a new deep learning method for high-accuracy contact prediction by learning from raw coevolutionary features extracted with deep multiple sequence alignments. The method was tested on FM domains in CASP11-13 and hard targets from CAMEO experiments, which demonstrated the effectiveness of the proposed method.

Several factors were found to contribute to the success of the TripletRes pipeline. First, coupling deep residual convolutional networks directly with raw coevolutionary matrices can result in better performance than feeding neural networks with the post-processed features. Second, a triplet of coevolutionary features, from covariance matrix, inverse covariance matrix and the inverse Potts model approximated by pseudolikelihood maximization, are ensembled in TripletRes by a set of four neural networks constructed with residual blocks. This feature ensemble strategy was found to enable more accurate prediction than using the three sets of features individually. Third, including more discrete distance information into the network training was proven to be beneficial to the contact-map prediction compared to binary contact training, although the contact-map models are binary on their own. This is largely because the distance-based loss function enables the learning of detailed spatial features specified by the sequence profiles. Finally, a hierarchical sequence searching protocol was proposed to obtain deeper MSAs, which impact the performance of the final model prediction. A significant improvement of contact prediction precision can be achieved through MSAs generated by searching an enlarged protein sequence database. These data underscore the impact of the volume of the sequence database on contact/distance prediction. The studies extending the DeepMSA pipeline to utilize the enlarged databases are in progress.

It is worth noting that the major goal of contact-map prediction is for assisting *ab initio* 3D structure construction, where a significant amount of efforts has been made along this line in the past decades [8,33,40–42]. Although recent progress of the field has shown an advantage of distance predictions [34,35], contact-map can provide reliable information of short-distance residue-residue interactions that is critical to specify the global topology of the protein fold. In fact, our results showed that most of the accurately predicted distances in TripletRes are still on the residues pairs with a short distance below 9–10 Å, which is part of the reason that has motivated our idea of distance-supervised learning in TripletRes. In addition, the development of feature extraction for protein contact-map prediction has direct contributions to the prediction of other forms of long-range residue-residue interactions. Therefore, with the

development of new approaches and consistent improvement of the model accuracy, the advanced sequence-based contact-map predictions will continue to be an important driving force for template-free structure prediction of the field.

## Methods and materials

TripletRes is a deep-learning based contact-map prediction method consisting of three consecutive steps (Fig 1). It first creates a deep MSA and extracts three coevolutionary matrix features. Next, the feature sets are fed into three sets of deep ResNets and trained in an end-to-end fashion. Finally, a symmetric matrix distance histogram probability is created and binarized into the contact-map prediction.

### MSA generation

To help offset the overfitting effects, TripletRes creates MSAs using different strategies for training and testing protein sequences. For training proteins, MSAs are created by HHblits with an E-value threshold of 0.001 and a minimum sequence coverage of 40% to search through Uniclust30 (2017_10) [43] database with 3 iterations.

For test proteins, the DeepMSA pipeline [30] was utilized to generate MSAs. The initial MSA is created also by HHblits but followed-up with multiple iterations. If the Neff value of the initial MSA is lower than a given threshold (= 128 that was decided by trial and error), a second step will be performed using jackhmmer [44] through UniRef90 (release-2017_12) [45]. Here, Neff measures the number of effective sequences in the MSA and is defined as:

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

where $N$ is the total number of sequences in the MSA, $\mathbb{I}[S_{m,n} \geq 0.8] = 1$ if the sequence identity $S_{m,n}$ between sequences $m$ and $n$ is over 0.8; or = 0 otherwise. To assist the MSA concatenation, the jackhmmer hits are converted into an HHblits format sequence database, against which a second HHblits search was performed. In case that Neff is still below 128, a third iteration is performed by hmmsearch [44] through the MetaClust (2017_05) [46], where the final MSA is pooled from all iterations (see S3 Fig for the whole MSA construction pipeline).

### Coevolutionary feature extraction

Three sets of coevolutionary features are extracted from the deep MSAs. First, the covariance (COV) feature measures the marginal dependency between different sequential positions and is calculated by

$$S_{i,j}^{a,b} = f_{i,j}(a,b) - f_i(a) f_j(b) \tag{2}$$

where $f_i(a)$ is the frequency of a residue $a$ at position $i$ of the MSA, $f_{i,j}(a,b)$ is the co-occurrence of two residue types $a$ and $b$ at positions $i$ and $j$.

The COV feature captures marginal correlations among variables, which contains transitional correlations. The negative of the inverse of the covariance matrix, i.e., precision matrix, can be interpreted as the Mean-field approximation of Potts model [12] and thus can capture direct couplings. In this work, a ridge regularized precision matrix (PRE), $\Theta$, is estimated by minimizing the regularized negative log-likelihood function [27,47]

$$G = tr(S\Theta) - log|\Theta| + R(\Theta) \tag{3}$$

where the first two terms are the negative log-likelihood of Θ assuming that the data follows a multivariate Gaussian distribution; $tr(S\Theta)$ is the trace of matrix $S\Theta$; $log|\Theta|$ is the log determinant of Θ; and $R(\Theta) = \rho \sum \|\Theta_{i,j}\|_2^2$ is the regularization function of Θ to avoid over-fitting, with $\rho = e^{-6}$ being a positive regularization hyper-parameter.

The last feature, which was firstly introduced by plmConv [48], is the raw coupling parameter matrix of the inverse Potts model approximated by PLM. Instead of assuming the data follows a multivariate Gaussian distribution, PLM approximates the probability of a sequence for the Potts model with

$$P(\sigma^m) = \prod_{l=1}^{L} P(\sigma_l = \sigma_l^{(m)}) | \sigma_{\setminus l} = \sigma_{\setminus l}^{(m)}) \tag{4}$$

Here, $P(\sigma^m)$ is the probability model for the $m$-th sequence in the MSA and $P(\sigma_l = \sigma_l^{(m)} | \sigma_{\setminus l} = \sigma_{\setminus l}^{(m)})$ is the marginal probability of $l$-th position in the sequence by

$$P(\sigma_l = \sigma_l^{(m)} | \sigma_{\setminus l} = \sigma_{\setminus l}^{(m)}) = \frac{\exp(h_l(\sigma_l^{(m)}) + \sum_{k=1,k\neq l}^{L} J_{lk}(\sigma_l^{(m)}, \sigma_k^{(m)}))}{\sum_{q=1}^{21} \exp(h_l(q) + \sum_{k=1,k\neq l}^{L} J_{lk}(q, \sigma_k^{(m)}))} \tag{5}$$

where $h$ and $J$ are single site and coupling parameters, respectively. In TripletRes, the raw coupling parameter matrix $J$ is used as the PLM feature.

Thus, each feature is represented by a $21^*L$ by $21^*L$ matrix for a protein sequence with $L$ amino acids. The entries of the 21 by 21 sub-matrix of a corresponding amino acid pair are the descriptors, which are fed into a convolutional transformer as conducted by a fully convolutional neural network with residual architecture (Fig 1).

## Deep neural-network modeling

TripletRes implements residual neural networks (ResNets) [29] as the deep learning model. Compared to traditional convolutional networks, ResNets adds feedforward neural networks to an identity map of input, which helps enable the efficient training of extremely deep neural networks such as the one used in TripletRes. As illustrated in Fig 1, the neural network structure of TripletRes has four sets of residual blocks, where three of them are connected to the input layer for feature extraction. Each of the three ResNets has 24 basic blocks and can learn layered features based on the specific input. After transforming each input feature into a feature map of 64 channels, we concatenate the transformed features along the feature channel and employ another deep ResNet containing 24 residue blocks to learn the fused information from the three features.

The activation function of the last layer is a softmax function which outputs the probability of each residue pair belonging to specific distance bins. Here, the residue-residue distance is split into 10 intervals spanning 5-15Å with an additional two bins representing distance less than 5Å and more than 15Å, respectively. The whole set of deep ResNets are trained by the supervision of the maximum likelihood of the prediction, where the loss function is defined as the sum of the negative log-likelihood over all the residue pairs of the training proteins:

$$\mathcal{L} = -\sum_{t=1}^{T} \sum_{k=1}^{12} y_t^k \log(p_t^k) \tag{6}$$

Here, $T$ is the total number of residue pairs in the training set. $y_t^k = 1$ if the distance of $t$-th residue pair of native structures falls into $k$-th distance interval; otherwise $y^k = 0$. $p_t^k$ is the predicted probability that the distance of the $t$-th residue pair falls into the $k$-th distance interval.

The probability of the $t$-th residue pair forming a contact $P_t$ is the sum of the first 4 distance bins:

$$P_t = \sum_{k=1}^{4} p_t^k \tag{7}$$

The training process uses dropout to avoid over-fitting, where the dropout rate is set to 0.2. We use Adam [49], an adaptive stochastic gradient descent algorithm, to optimize the loss function. TripletRes implements deep ResNets using Pytorch [50] and was trained using the Extreme Science and Engineering Discovery Environment (XSEDE) [39].

## Supporting information

**S1 Fig. Comparison of the average long-range top-*L*/5 precisions over training epochs without individual coevolutionary features and the TripletRes model that ensembles all three sets of features, on the validation set. (a)** top-$L$/5 precision, **(b)** top-$L$ precision.
(PDF)

**S2 Fig. Comparison of long-range top-*L*/5 and top-*L* precisions with different loss functions on the different fold types, where median precision and mean precision are marked in solid and dash lines, respectively.**
(PDF)

**S3 Fig. The DeepMSA pipeline for generating deep multiple sequence alignments for TripletRes.**
(PDF)

**S1 Table. Summary of long-range contact precision by TripletRes and control methods tweaked with Deep MSAs on 50 CASP11&12 FM targets and 195 CAMEO *hard* targets, sorted in ascending order of top-*L* precision.** $p$-values in parenthesis are from a Student's t-test between TripletRes and each of the control methods, where bold fonts highlight the best performer in each category.
(PDF)

**S2 Table. Summary of long-range contact precision by TripletRes, TripletRes (Post-CASP13) and trRosetta based on the same MSAs on 37 hybrid test sequences.**
(PDF)

**S1 Text. Explanation that DCA models capture linear relationships between residues.**
(PDF)

**S2 Text. Detailed procedure to collect training and test datasets.**
(PDF)

**S3 Text. A brief introduction of control methods and other top participants in CASP13.**
(PDF)

**S4 Text. Traditional feature extraction strategy with post-processing.**
(PDF)

**S5 Text. Binary cross entropy loss function for training TripletRes in CASP13.**
(PDF)

## Acknowledgments

This work used the Extreme Scienceand Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation (ACI-1548562). The work was done when Yang Li visited at University of Michigan.

## Author Contributions

**Conceptualization:** Yang Zhang.

**Funding acquisition:** Dong-Jun Yu, Yang Zhang.

**Investigation:** Yang Li.

**Methodology:** Yang Li, Chengxin Zhang.

**Resources:** Chengxin Zhang.

**Software:** Yang Li, Chengxin Zhang.

**Supervision:** Dong-Jun Yu, Yang Zhang.

**Validation:** Yang Li, Chengxin Zhang, Eric W. Bell, Wei Zheng, Xiaogen Zhou.

**Writing – original draft:** Yang Li, Yang Zhang.

**Writing – review & editing:** Chengxin Zhang, Eric W. Bell, Wei Zheng, Xiaogen Zhou.

## References

1. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001; 294(5540):93–6. https://doi.org/10.1126/science.1065659 PMID: 11588250
2. Zhang Y. Progress and challenges in protein structure prediction. Current opinion in structural biology. 2008; 18(3):342–8. https://doi.org/10.1016/j.sbi.2008.02.004 PMID: 18436442
3. Abriata LA, Tamo 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. 2018; 86 Suppl 1:97–112. https://doi.org/10.1002/prot.25423 PMID: 29139163
4. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin A. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins. 2018; 86 Suppl 1:51–66. https://doi.org/10.1002/prot.25407 PMID: 29071738
5. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins. 2019. https://doi.org/10.1002/prot.25792 PMID: 31365149
6. Shrestha R, Fajardo E, Gil N, Fidelis K, Kryshtafovych A, Monastyrskyy B, et al. Assessing the accuracy of contact predictions in CASP13. Proteins. 2019; 87(12):1058–68. https://doi.org/10.1002/prot.25819 PMID: 31587357
7. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins. 1994; 18(4):309–17. https://doi.org/10.1002/prot.340180402 PMID: 8208723
8. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des. 1997; 2(5):295–306. https://doi.org/10.1016/S1359-0278(97)00041-2 PMID: 9377713
9. 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. https://doi.org/10.1073/pnas.90.15.7176 PMID: 8346232
10. Zhang H, Gao Y, Deng M, Wang C, Zhu J, Li SC, et al. Improving residue–residue contact prediction via low-rank and sparse decomposition of residue correlation matrix. Biochemical and biophysical research communications. 2016; 472(1):217–22. https://doi.org/10.1016/j.bbrc.2016.01.188 PMID: 26920058
11. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2011; 28 (2):184–90. https://doi.org/10.1093/bioinformatics/btr638 PMID: 22101153
12. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, 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–E301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

13. Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics. 2015; 31(21):3506–13. https://doi.org/10.1093/bioinformatics/btv472 PMID: 26275894

14. 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; 276:341–56.

15. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics. 2014; 30(21):3128–30. https://doi.org/10.1093/bioinformatics/btu500 PMID: 25064567

16. Kamisetty H, Ovchinnikov S, 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. 2013:201314045.

17. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC bioinformatics. 2007; 8(1):113. https://doi.org/10.1186/1471-2105-8-113 PMID: 17407573

18. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics. 2008; 24(7):924–31. https://doi.org/10.1093/bioinformatics/btn069 PMID: 18296462

19. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics. 2013; 29(13):i266–i73. https://doi.org/10.1093/bioinformatics/btt211 PMID: 23812992

20. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics. 2012; 28(19):2449–57. https://doi.org/10.1093/bioinformatics/bts475 PMID: 22847931

21. 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. 2014; 31(7):999–1006. https://doi.org/10.1093/bioinformatics/btu791 PMID: 25431331

22. Buchan DW, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins: Structure, Function, and Bioinformatics. 2017.

23. He B, Mortuza S, 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):2296–306. https://doi.org/10.1093/bioinformatics/btx164 PMID: 28369334

24. Adhikari B, Hou J, Cheng J. DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics. 2017; 34(9):1466–72.

25. 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. e3. https://doi.org/10.1016/j.cels.2017.11.014 PMID: 29275173

26. 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. https://doi.org/10.1371/journal.pcbi.1005324 PMID: 28056090

27. Li Y, Hu J, Zhang C, Yu DJ, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics. 2019:4647–55. https://doi.org/10.1093/bioinformatics/btz291 PMID: 31070716

28. 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.

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

30. 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):2105–12. https://doi.org/10.1093/bioinformatics/btz863 PMID: 31738385

31. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, et al. The Protein Model Portal—a comprehensive resource for protein structure and model information. Database: the journal of biological databases and curation. 2013; 2013:bat031. https://doi.org/10.1093/database/bat031 PMID: 23624946

32. Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic acids research. 2013; 42(D1):D304–D9. https://doi.org/10.1093/nar/gkt1240 PMID: 24304899

33. Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. Structure. 2011; 19(8):1182–91. https://doi.org/10.1016/j.str.2011.05.004 PMID: 21827953

34. Xu J. Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A. 2019; 116 (34):16856–65. https://doi.org/10.1073/pnas.1821309116 PMID: 31399549

35. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins. 2019; 87(12):1141–8. https://doi.org/10.1002/prot.25834 PMID: 31602685

36. 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. https://doi.org/10.1038/s41467-019-11994-0 PMID: 31484923

37. 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. https://doi.org/10.1073/pnas.1914677117 PMID: 31896580

38. Zhang C, Mortuza SM, He B, Wang Y, Zhang Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins: Structure, Function, and Bioinformatics. 2018; 86(S1):136–51. https://doi.org/10.1002/prot.25414 PMID: 29082551

39. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, et al. XSEDE: accelerating scientific discovery. Computing in Science & Engineering. 2014; 16(5):62–74.

40. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophys J. 2003; 85:1145–64. https://doi.org/10.1016/S0006-3495(03)74551-2 PMID: 12885659

41. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci. 2002; 11(8):1937–44. https://doi.org/10.1110/ps.3790102 PMID: 12142448

42. Nugent T, Jones DT. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. PLoS Comput Biol. 2010; 6(3):e1000714. https://doi.org/10.1371/journal.pcbi.1000714 PMID: 20333233

43. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research. 2016; 45(D1): D170–D6. https://doi.org/10.1093/nar/gkw1081 PMID: 27899574

44. Eddy SR. Accelerated profile HMM searches. PLoS computational biology. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361

45. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2014; 31(6):926–32. https://doi.org/10.1093/bioinformatics/btu739 PMID: 25398609

46. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nature communications. 2018; 9(1):2542. https://doi.org/10.1038/s41467-018-04964-5 PMID: 29959318

47. Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins: Structure, Function, and Bioinformatics. 2019. https://doi.org/10.1002/prot.25798 PMID: 31407406

48. Golkov V, Skwark MJ, Golkov A, Dosovitskiy A, Brox T, Meiler J, et al. Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. NIPS; 2016.

49. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

50. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in pytorch. 2017.