frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 16 May 2022
doi: 10.3389/fgene.2022.887491

Inter-Residue Distance Prediction From Duet Deep Learning Models

Huiling Zhang[1,2], *Ying Huang*[1,2], *Zhendong Bei*[1,2], *Zhen Ju*[1,2], *Jintao Meng*[1,2], *Min Hao*[3], *Jingjing Zhang*[1,2], *Haiping Zhang*[2] *and Wenhui Xi*[1,2]*

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]College of Electronic and Information Engineering, Southwest University, Chongqing, China

Residue distance prediction from the sequence is critical for many biological applications such as protein structure reconstruction, protein–protein interaction prediction, and protein design. However, prediction of fine-grained distances between residues with long sequence separations still remains challenging. In this study, we propose DuetDis, a method based on duet feature sets and deep residual network with squeeze-and-excitation (SE), for protein inter-residue distance prediction. DuetDis embraces the ability to learn and fuse features directly or indirectly extracted from the whole-genome/metagenomic databases and, therefore, minimize the information loss through ensembling models trained on different feature sets. We evaluate DuetDis and 11 widely used peer methods on a large-scale test set (610 proteins chains). The experimental results suggest that 1) prediction results from different feature sets show obvious differences; 2) ensembling different feature sets can improve the prediction performance; 3) high-quality multiple sequence alignment (MSA) used for both training and testing can greatly improve the prediction performance; and 4) DuetDis is more accurate than peer methods for the overall prediction, more reliable in terms of model prediction score, and more robust against shallow multiple sequence alignment (MSA).

Keywords: residue distance prediction, protein structure reconstruction, deep learning, residual network, multiple sequence alignment

OPEN ACCESS

Edited by:
Ruiquan Ge,
Hangzhou Dianzi University, China

Reviewed by:
Leyi Wei,
Shandong University, China
Jun Wang,
Nanjing University, China
Duolin Wang,
University of Missouri, United States

****Correspondence:***
Wenhui Xi
wh.xi@siat.ac.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *01 March 2022*
Accepted: *30 March 2022*
Published: *16 May 2022*

Citation:
Zhang H, Huang Y, Bei Z, Ju Z, Meng J, Hao M, Zhang J, Zhang H and Xi W (2022) Inter-Residue Distance Prediction From Duet Deep Learning Models. Front. Genet. 13:887491. doi: 10.3389/fgene.2022.887491

INTRODUCTION

Knowing the structure of a protein helps to understand the role of the protein, reveals how the protein performs its biological function, and also, sets the foundation for the protein's interaction with other molecules. Therefore, the knowledge of a protein's structure is very important for biology as well as for medicine and pharmacy. Since Anfinsen suggested that the advanced spatial structure of a protein is determined by its amino acid sequence (Anfinsen, 1973), it has been a "holy grail" for the computational biology community to develop an algorithm that can accurately predict a protein's structure from its amino acid sequence. Sequence-based residue contact/distance prediction plays a crucial role in protein structure reconstruction.

Residue–residue contacts refer to the residue pairs that are close within a specific distance threshold in the three-dimensional protein structure. The contact map of a protein tells the constraints between residues in a binary form. Unlike the contact map, the distance map of a protein contains fine-grained information and, thus, provides more physical constraints of a protein structure. Protein contact/distance maps are 2D representations of the 3D protein structure and are being considered as one of the most important components in modern protein structure prediction packages. The application of predicted contacts/distances has been extended to intrinsic disorder region recognition (Schlessinger et al., 2007; Shimomura et al., 2019), protein–protein interaction prediction (Vangone and Bonvin, 2015; Du et al., 2016; Cong et al., 2019), protein design (Anishchenko et al., 2021), etc.

Contact prediction methods in the early stage are mainly based on mutual information (MI) (Pollock and Taylor, 1997; Dunn et al., 2007; Lee and Kim, 2009), integer linear programming (ILP) techniques (McAllister and Floudas, 2008; Rajgaria et al., 2009; Rajgaria et al., 2010; Wei and Floudas, 2011), traditional machine learning (ML) algorithms (Cheng and Baldi, 2007; Wu and Zhang, 2008; Tegge et al., 2009), or techniques combining ILP with ML (Wang and Xu, 2013; Zhang et al., 2016). These methods are generally considered as local strategies since a residue pair is treated statistically independent of others (Zhang et al., 2020). Breakthroughs were achieved by capturing the correlated pattern of coevolved residues by global statistical inference methods such as direct coupling analysis (DCA) (Weigt et al., 2009) and sparse inverse covariance estimation (PSICOV) (Jones et al., 2012). Methods developed based on the ideas of DCA include EVfold (mfDCA) (Morcos et al., 2011), plmDCA (Ekeberg et al., 2013), GREMLIN (Kamisetty et al., 2013), CCMpred (Seemayer et al., 2014), gDCA (Baldassi et al., 2014), and Freecontact (Kaján et al., 2014). These methods emphasize the importance of distinguishing between directly and indirectly correlated residues. Consensus-predictors like PconsC (Skwark et al., 2013), MetaPSICOV (Jones et al., 2014), and NeBcon (He et al., 2017) combine the output of different DCA-based or ML-based contact predictors to create consensus predictions. In recent years, the introduction of deep learning (DL) techniques has made tremendous progress for residue contact prediction. The DL-based contact map prediction algorithms are mainly based on convolutional neural networks (CNN) (such as DeepCov (Jones and Kandathil, 2018), DeepContact (Liu et al., 2018), and DNCON2 (Adhikari et al., 2018)), Unet [such as PconsC4 (Michel et al., 2019)], residual networks (ResNet) [such as DeepConPred2 (Ding et al., 2018), ResPRE (Li et al., 2019), MapPred (Wu et al., 2020) and TripletRes (Li et al., 2021)], ResNet combined with long short-term memory (LSTM) [such as SPOT-Contact (Hanson et al., 2018)] and transformers [such as ESM (Malinin and Gales, 2021) and SPOT-Contact-LM (Singh et al., 2022)]. COMTOP (Reza et al., 2021) uses the mixed ILP technique to combine different contact predictors (including several DL predictors) to further improve the prediction performance.

Although the predicted contacts have been successfully applied to the protein structure prediction packages (Marks et al., 2012; Michel et al., 2014; Adhikari et al., 2015; Gao et al., 2019), contact maps are still insufficient for accurate structure prediction. The reason is twofold. Most contact prediction methods use a cutoff of 8 Å between Cβ-Cβ atoms to determine whether two residues are in contact or not, resulting a contact/non-contact ratio of less than 0.1 for globular proteins and a ratio of around 0.02 for alpha-helical transmembrane proteins (Zhang et al., 2016). The definition of contacts means that the native distance information is insufficiently being distinguished. Furthermore, contact-assisted conformation sampling may be misguided by several wrongly predicted contacts and needs a long time to generate good conformations for large proteins (Xu, 2019). In this context, inter-residue distance maps are more informative than residue–residue contact maps since distances are fine-grained or real numbers, while contacts are binary values.

The methods for inter-residue distance prediction can be roughly categorized into two groups, those based on multiclass classification with discrete values and those based on regression with continuous values. Early distance maps are mainly predicted from homologous proteins (Aszódi and Taylor, 1996) or from traditional machine learning techniques (Walsh et al., 2009; Zhao and Xu, 2012; Kukic et al., 2014). The introduction of deep learning technology has injected new life into distance prediction. Wang et al. (2017) pioneered the study of introducing residual network to multiclass distance prediction. The success of this approach can be partially attributed to the ability of deep learning to simultaneously consider the global set of pair-wise interactions instead of considering only one interaction at a time, thereby leading to more accurate discrimination between direct and indirect contacts. TripletRes (Li et al., 2021), which uses a similar deep learning architecture but with a unique set of features that include multiple coevolutionary coupling matrices directly deduced from deep multiple sequence alignment (MSA) without post-processing. GANProDist (Ding and Gong, 2020) predicts real value distance as a regression problem by generative adversarial network. PDNET (Adhikari, 2020), DeepDist (Wu et al., 2021), SDP (Rahman et al., 2022), and Li et al. (2021) (Li and Xu, 2021) predict both real-valued and binned distances from residual networks. DL-based distance prediction has recently demonstrated unprecedented ability to assist protein structure reconstruction such as DMPFold (Greener et al., 2019), RaptorX (Xu, 2019), trRosetta (Yang et al., 2020), and AlhpaFold (Senior et al., 2020). However, further progress needs more accurate inter-residue distance prediction since the quality of a predicted protein structure highly depends on the accuracy of the distance prediction.

Shimomura et al. (2019) introduced a technique for predicting structurally disordered regions in proteins through average distance maps (AMD) based on statistics of average distances between residues. AMD first divides the residue pairs into different ranges according to their sequence separations, and calculates the distances of residue pairs within each range. AMD contact density maps were plotted against distance thresholds in different ranges. AMD technology detects the boundaries of structurally compact regions and finally predicts structurally disordered regions by calculating differences in density maps. The accuracy of AMD technology is comparable to the leading methods in the CASP competition such as PrDOS, DISOPRED, and Biomine. Protein domains are subunits that can fold and function independently. Therefore, correct domain boundary assignment is a critical step to achieve accurate protein structure and function analysis. Zheng et al. (2020) proposed FUPred to detect protein domains based on contact maps predicted by deep learning. The core idea of this method is to retrieve domain boundary locations by maximizing the number of intra-domain contacts while minimizing the number of interdomain contacts from the contact map. FUpred was tested on a large-scale dataset consisting of 2,549 proteins and achieved a Matthews correlation coefficient (MCC) of 0.799 for single domain and multi-domain classification, which is 19.1% higher than the best machine learning-based method. For proteins with discontinuous domains, FUPred domain boundary detection and normalized domain overlap scores were 0.788 and 0.521, which were 17.3% and 23.8% higher than the best peer method. The results demonstrate that residue contact prediction provides a new way to accurately detect domains, especially discontinuous multi-domains. Cong et al. (2019) first compared the contact prediction methods based on mutual information, evolutionary coupling analysis, and deep learning in the prediction of residue contacts between protein complex chains and found that although the deep learning methods are outstanding for monomer contact prediction, they fail to outperform methods based on mutual information and evolutionary coupling analysis in inter-chain contact prediction. By identifying coevolving residue pairs between protein chains based on mutual information and evolutionary coupling analysis methods, 1,618 protein interactions (682 of which were unexpected) in *Escherichia coli*, and 911 protein interactions in *M. tuberculosis* (most of which were not identified in previous studies) were detected. The expected false positive rate for this study is between 10% and 20%, and the predicted interactions and networks provide a good starting point for further research. Anishchenko et al. (2021) investigated whether the residue distance information captured by deep neural networks is rich enough to generate new folded proteins. The study generated random amino acid sequences that were completely unrelated to the sequences of the native proteins used in the trRosetta training model, and fed them into the trRosetta structure prediction network to predict the starting residue distance map. Monte Carlo sampling is then performed in the amino acid sequence space to optimize the contrast between the network-predicted distribution of inter-residue distances and the background distribution averaged across all proteins. Optimization from different random starting points yields novel proteins spanning a broad range of sequences and predicted structures. Synthetic genes encoding 129 of the 'network-hallucinated' sequences were obtained, and the proteins were expressed and purified in *E. coli*; 27 of the proteins yielded monodisperse species with circular dichroism spectra consistent with the hallucinated structures. Three of the three-dimensional structures of the hallucinated proteins were determined by experiments, and these closely matched the hallucinated models. We can see that residue distance-assisted protein structure prediction methods can be inverted to *de novo* protein design.

In this study, we develop a method based on deep residual convolutional neural network, named DuetDis, to predict the full-length multiclass distance map from a sequence. DuetDis uses a modified ResNet module to build the network, and adopts two sets of complementary feature sets to further improve the prediction accuracy. The results by DuetDis suggest that prediction results from different feature sets show obvious differences and ensembles of different feature sets can improve the prediction performance. DuetDis is also evaluated together with 11 widely used contact/distance prediction methods, and the results show that DuetDis is more accurate for the overall prediction, more reliable in terms of model prediction score, and more robust against shallow MSA. DuetDis is available at http://hpcc.siat.ac.cn/hlzhang/DuetDis/.

MATERIALS AND METHODS

Datasets

The test set is obtained from our previous work, containing 610 highly non-redundant protein chains (Zhang et al., 2021). The training set is obtained through culling from the whole PDB with the following criteria: 1) with maximum sequence identity of 30% against each chain in the training set and test set; 3) with structure resolutions better than 2.5 Å; 4) released before 1 May 2018 (before the beginning of CASP13). Finally, we get a non-redundant training set with 13,069 protein chains.

Definition of Contact and Distance

In this study, the definition of contacts is directly taken from the CASP experiments. A pair of residues in the experimental structure is considered to be in contact if the distance between their Cβ atoms (Cα for Gly) is less than or equal to 8 Å. For direct comparison, the multiclass distance definition is taken directly from trRosetta (Adhikari, 2020). The Cβ–Cβ distance of every pair of residues in a target protein is treated as a vector of probabilities. The distance range (2–20 Å) is binned into 36 equally spaced segments, 0.5 Å each, and one bin indicating that residues are not in contact, generating a distance vector of 37 bins for each residue pair.

Depending on the separation of two residues along the sequence (*seq_sep*), the contacts are classified into four classes: all-range (*seq_sep* ≥6), short-range (6≤ *seq_sep* <12), medium-range (12≤ *seq_sep* <24), and long-range (*seq_sep* >24).

Multiple Sequence Alignment Generation for Training and Test

Generating high-quality MSA is the first step for protein structure prediction based on the fact that interacting residue pairs are under evolutionary pressure to maintain the structure. The MSA used for model training is obtained as indicated in **Figure 1**. The target sequence in the training set is searched against NCBI-nr (Jackhmmer), MetaClust (Jackhmmer), and BFD (HHblits) respectively, with *E*-values of 1e−10 and 1e−3. The search will stop if the target MSA has N_{seq}> 25*L (L is the sequence length) and N_{eff} > 8*L, where N_{seq} is the number of sequences (with sequence coverage >50%) and N_{eff} [defined in (Zhang et al., 2021)] is the number of effective sequences in the MSA. After the search, the final MSA is obtained through sequence clustering (with sequence identity of 95%) using our in-house software nGIA (Ju et al., 2021).

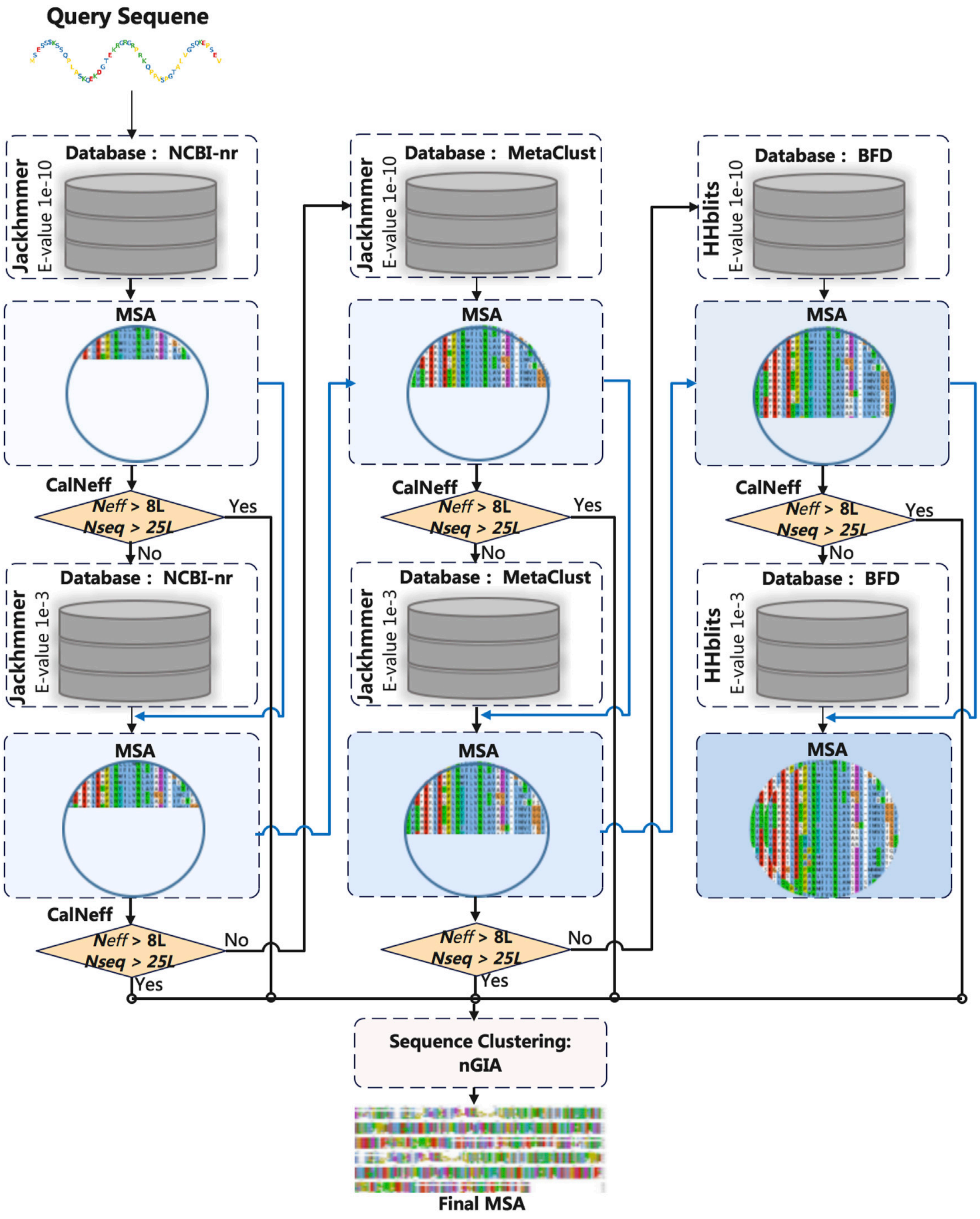

FIGURE 1 | The flowchart of MSA generation for the training set.

The MSA used for testing is obtained through searching JackHMMER (Johnson et al., 2010) against the NCBI-nr database with iteration = 3 and *E*-value = 0.0001.

Input Features

We used two subsets of features as the inputs for the deep residual network of DuetDis. The first feature set contains 526 feature channels: one-hot-encoder of the target sequence (1D features, 20*2 channels); position-specific frequency matrix (1D features, 21*2 channels, considering gap) and positional entropy (Yang et al., 2020) (1D features, 1*2 channels); and coupling features (Yang et al., 2020) (2D features, 441 channels) derived from the inverse of the shrunk covariance matrix of MSA. The second feature set contains 151 feature channels: one-hot-encoder of the

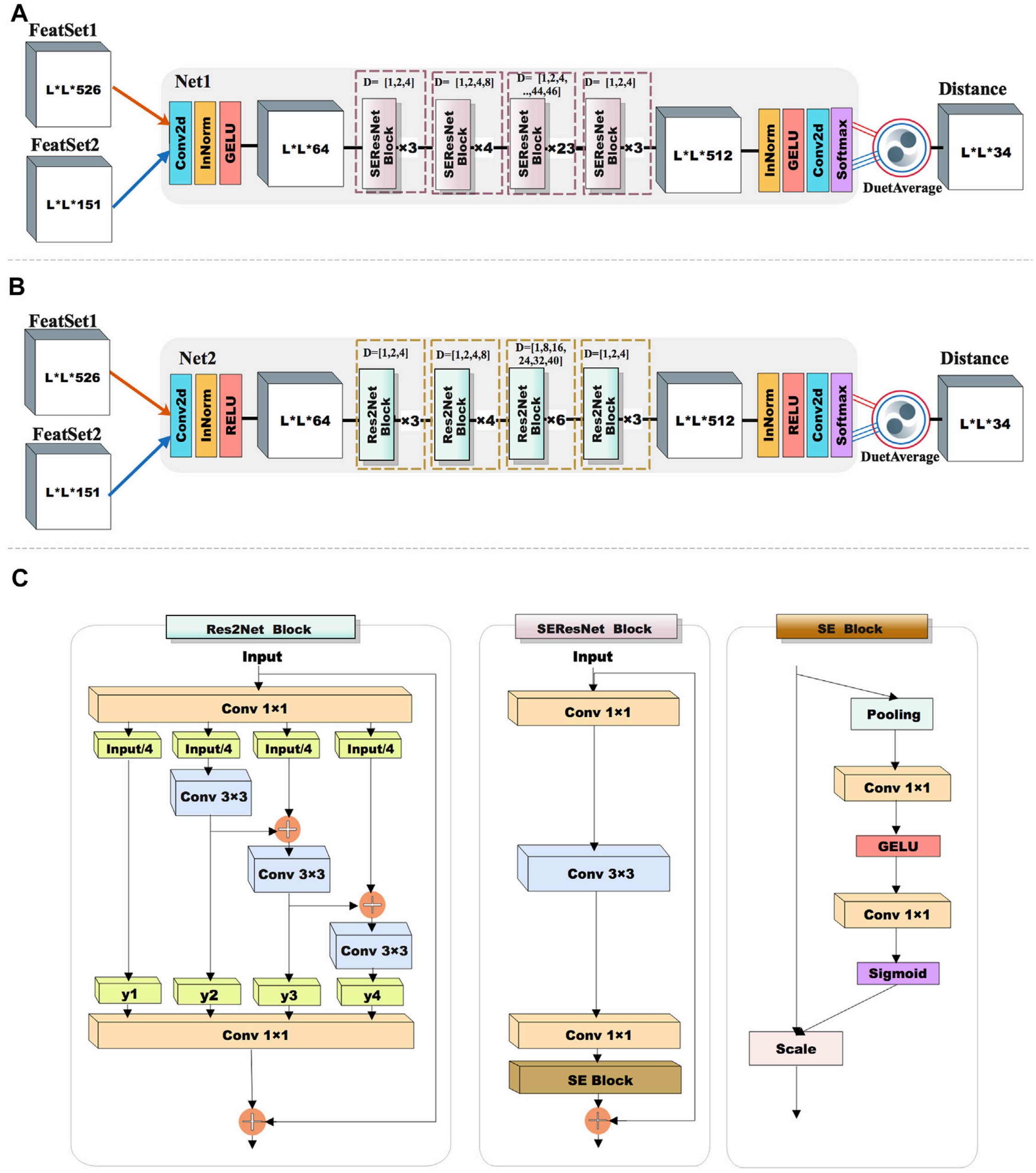

FIGURE 2 | The network architecture used in this work. **(A)** The network used by DuetDis; **(B)** the reference network; **(C)** basic modules used in the networks; dilated convolution.

target sequence (1D features, 20*2 channels), position-specific scoring matrix (Altschul et al., 1997) (1D features; 20*2 channels; not considering gap), HMM profile (Remmert et al., 2012) (1D features, 30*2 channels), secondary structure from SPOT-1D (Hanson et al., 2019) (1D features, 3*2 channels), solvent accessible surface area from SPOT-1D (Hanson et al., 2019) (1D features, 1*2 channels), CCMPRED score (Seemayer et al., 2014) (2D features, 1 channel), mutual information (Zhang et al., 2022) (2D feature, 1 channel), and statistical pair-wise contact potential (Betancourt and Thirumalai, 1999) (2D feature, 1 channel). The first feature set, indicated as FeatSet1, is mainly composed of 2D direct coupling features (441 out of 526 total features) from the MSA, while the second feature set, indicated as FeatSet2, is mainly composed of 1D sequence-based features (148 out of 151 total features). Most of the features except the one-hot-encoder features in FeatSet1 and FeatSet2 are different, so the prediction results from the two feature sets can be complementary in a duet way (as indicated in the results).

Both FeatSet1 and FeatSet2 are widely used by previous works (Hanson et al., 2018; Yang et al., 2020; Jain et al., 2021; Su et al.,

TABLE 1 | The strategies used for the training of sub-models (N1_M1/N1_M2/N1_M3/N1_M4/N1_M5 are used for DuetDis).

Sub-models	Network	Feature set	MSA	MSA shuffle
N1_M1	Net1	FeatSet1	MSA_All	Yes
N1_M2	Net1	FeatSet1	MSA_Top	No
N1_M3	Net1	FeatSet2	MSA_Top	No
N1_M4	Net1	FeatSet2	MSA_1	No
N1_M5	Net1	FeatSet2	MSA_2	No
N2_M1	Net2	FeatSet1	MSA_All	Yes
N2_M2	Net2	FeatSet1	MSA_Top	No
N2_M3	Net2	FeatSet2	MSA_Top	No
N2_M4	Net2	FeatSet2	MSA_1	No
N2_M5	Net2	FeatSet2	MSA_2	No

2021), showing their great efficacy in contact/distance prediction. The aim of DuetDis is not to design new feature types, but to evaluate the performance of previously widely used feature sets under the situation of unified input and identical network, as well to study how to complement the advantages of different types of features for better prediction performance.

Deep Network Architectures and Model Training for Distance Prediction

The proposed method DuetDis implements residual neural networks (ResNet) (He et al., 2016) as the deep learning model. Compared to traditional convolutional networks, ResNet adds feedforward neural networks to an identity map of input, which helps enable the efficient training of extremely deep neural networks. ResNet has shown its power in successful residue contact/distance prediction (Xu, 2019; Li et al., 2021). The deep residual network of DuetDis is shown in **Figure 2A**. The basic module of DuetDis network is a combination of squeeze-and-excitation and ResNet (SEResNet). The DuetDis network is composed of 33 SEResNet modules. In order to observe the impact of different networks and features on the prediction performance, we also designed another reference network (**Figure 2B**), which has very different basic modules and backbones from **Figure 2A**. The reference network is composed of 16 Res2Net modules. In this work, both SEResNet and Res2Net use dilation convolutions, while SEResNet use gelu and Res2Net use relu as the activation functions. The networks in **Figures 2A and B** are indicated as Net1 and Net2, respectively. The final MSA obtained in **Figure 1** is indicated as MSA_All, and a subset with top 10 L sequences (ranked with sequence identity against the target sequence) selected from MSA_All is indicated as MSA_Top, and two disjoint subsets with each containing 10 L sequences randomly selected from MSA_All are indicated as MSA_1 and MSA_2, respectively. As described in **Table 1**, 10 sub-models are trained based on Net1 (the DuetDis network) and Net2 (the reference network) with different feature sets from different MSAs. "MSA Shuffle" in **Table 1** means that the MSA are constructed through randomly selecting 10 L sequences in MSA_All. For each epoch, N1_M1/N2_M1 are trained through "MSA Shuffle" strategy, N1_M2/N1_M3/N2_M2/N2_M3 are trained with MSA_Top, N1_M4/N2_M4 are trained with MSA_1, and N1_M5/N2_M5 are trained with MSA_2. The outputs of five sub-models are averaged to produce the final distance map, indicated as "DuetAverage" in **Figures 2A,B**.

The sub-models are generated by independent training branches. AdamW optimizer is performed with an initial learning rate of 0.0001 (multi-step decay is adopted as the learning rate decay strategy). Cross-entropy is used as the loss-function, and L2 regularization is used during the training process to correct overfitting. The training set is split into two parts: 600 protein chains are used as the validation set and the rest are used for training. The precision of top-L long-range contact predictions (multiclass distance map is converted to the binary contact map according to the definition in **Section 2.2**) on the validation dataset is calculated at each epoch, and the training process will stop when there is no update of the validation precision for 10 epochs. The training processes are implemented in Pytorch on TeslaV100 SMX2, and each independent training generally takes 5–10 days.

Evaluation Metrics

1) The predicted distance map is a matrix of probability estimates. We analyze the performance of predictors on reduced lists of distances/contacts (sorted by the probability estimates) selected by either the probability threshold or the top-L/n (L is the sequence length, and n = 1, 2, 5) criteria. The prediction performance is assessed using precision (accuracy in some references), coverage (recall in some references), and Matthew's Correlation Coefficient (MCC), defined as follows:

$$Precision = \frac{TP}{TP + FP}, \quad (1)$$

$$Coverage = \frac{TP}{TP + FN}, \quad (2)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (3)$$

where *TP*, *FP*, *TN*, and *FN* are the number of true positive, false positive, true negative, and false negative contacts, respectively.

2) Standard deviation reflects the degree of dispersion among individuals within the group, which is defined as

$$STD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \quad (4)$$

where $\bar{x}$ is the mean of the variable x. The standard deviation can be used to evaluate the dispersion of *Precision*, *Coverage*, and *MCC*.

3) Jaccard index (Jaccard similarity coefficient) measures the similarities between sets. It is defined as the size of the intersection divided by the size of the union of two sets.

$$J(X, Y) = |X \cap Y| / |X \cup Y|, \quad (5)$$

where X and Y are the set of predicted contacts from two different predictors, $|X \cap Y|$ is the number of elements in the intersection of X and Y and the $|X \cup Y|$ represents the number of elements in

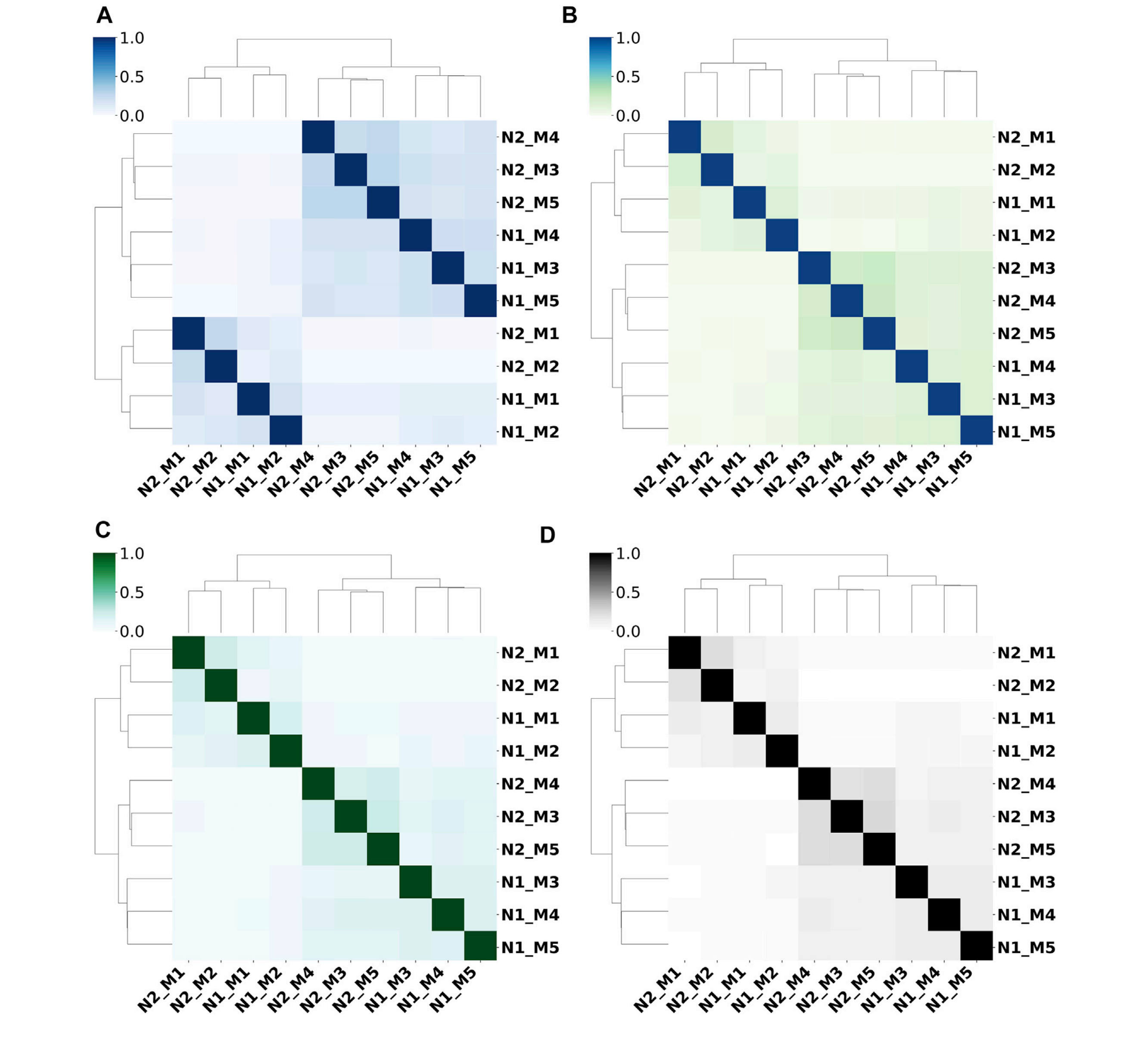

FIGURE 3 | Prediction similarities between different sub-models for **(A)** all-range, **(B)** short-range, **(C)** mid-range, and **(D)** long-range contacts/distances.

the union of *X* and *Y*. The Jaccard index has values in the range of [0,1], with the value of 0 for completely dissimilar ones and 1 for identical predictors.

RESULTS

In this section, we assess the performance of DuetDis from different perspectives. **Section 3.1**, **3.2** study the performance of sub-models, while **Section 3.3–3.5** focus on the comparison between DuetDis and peer methods. The peer methods used in this work are 4 DCA-based contact predictors (EVfold, FreeContact, gDCA, and CCMpred), 4 DL-based contact predictors (DeepCov, PconsC4, DNCON2, and SPOT-Contact), and 3 DL-based distance predictors (TripletRes, trRosetta, and RaptorX). **Section 3.1–3.3** and **Section 3.5** use the results of top-L/*n* (*n* = 1, 2, 5) predictions, while **Section 3.4** considers the results given by specific probability/score threshold. All sub-models and peer-methods use the same MSA as input.

Prediction Results From Different Feature Sets Show Obvious Differences

We use the Jaccard indices of prediction results from 10 sub-models (as described in **Table 1**) to study their prediction similarities. **Figure 3** shows the dendrogram heatmap of Jaccard indices using Ward's hierarchical clustering method on the independent test set. The Jaccard index between two methods

TABLE 2 | The prediction precisions of N1_M1/N1_M2/N1_M3/N1_M4/N1_M5/ N1_Ensemble for different sequence separations.

Range	Method	Top-L	Top-L/2	Top-L/5
All	N1_M1	0.7769	0.8717	0.9206
	N1_M2	0.7587	0.8475	0.8941
	N1_M3	0.7491	0.846	0.9027
	N1_M4	0.7256	0.8266	0.8888
	N1_M5	0.7319	0.8328	0.8942
	N1_Ensemble	0.7896	0.8786	0.9266
Short	N1_M1	0.2955	0.481	0.7389
	N1_M2	0.2928	0.4754	0.7287
	N1_M3	0.2948	0.4757	0.7374
	N1_M4	0.2824	0.4588	0.7109
	N1_M5	0.2947	0.473	0.7219
	N1_Ensemble	0.2988	0.4918	0.7633
Medium	N1_M1	0.3512	0.5477	0.7725
	N1_M2	0.3422	0.5336	0.7514
	N1_M3	0.342	0.5329	0.7533
	N1_M4	0.3306	0.5135	0.7275
	N1_M5	0.3371	0.5209	0.7352
	N1_Ensemble	0.3537	0.5592	0.7895
Long	N1_M1	0.6245	0.7696	0.865
	N1_M2	0.6062	0.7411	0.8273
	N1_M3	0.594	0.7308	0.8246
	N1_M4	0.5695	0.7091	0.8088
	N1_M5	0.5742	0.712	0.8121
	N1_Ensemble	0.6416	0.7797	0.8626

TABLE 3 | The prediction precisions of N2_M1/N2_M2/N2_M3/N2_M4/N2_M5/ N2_Ensemble for different sequence separations. -80

Range	Method	Top-L	Top-L/2	Top-L/5
All	N2_M1	0.7532	0.8562	0.9103
	N2_M2	0.7435	0.839	0.8938
	N2_M3	0.7148	0.8188	0.8828
	N2_M4	0.7091	0.8119	0.8768
	N2_M5	0.7071	0.8121	0.879
	N2_Ensemble	0.7590	0.8579	0.9153
Short	N2_M1	0.2864	0.4654	0.7172
	N2_M2	0.2901	0.4647	0.71
	N2_M3	0.2852	0.4583	0.7014
	N2_M4	0.2831	0.4547	0.6982
	N2_M5	0.2825	0.4548	0.7002
	N2_Ensemble	0.3449	0.5396	0.7367
Medium	N2_M1	0.3413	0.5325	0.755
	N2_M2	0.3428	0.5267	0.7395
	N2_M3	0.3298	0.5082	0.7206
	N2_M4	0.3281	0.5042	0.7152
	N2_M5	0.3283	0.5057	0.7159
	N2_Ensemble	0.3449	0.5396	0.7602
Long	N2_M1	0.6035	0.746	0.8473
	N2_M2	0.5997	0.7361	0.828
	N2_M3	0.5638	0.7022	0.8066
	N2_M4	0.5525	0.6877	0.7913
	N2_M5	0.5548	0.6917	0.7941
	N2_Ensemble	0.6136	0.7508	0.8473

is calculated by averaging the Jaccard index value of each protein on the whole test set. According to the clustering results, these 10 sub-models can be roughly divided into two categories, and each category contains two sub-categories. N1_M1/ N1_M2 and N2_M1/ N2_M2 trained by FeatSet1 are clustered into one category (Category_1), while N1_M3/ N1_M4/ N1_M5 and N2_M3/ N2_M4/ N2_M5 trained by FeatSet2 form another category (Category_2). N1_M1/ N1_M2 trained by Net1 and N2_M1/ N2_M1 trained by Net2 form two sub-categories in Category_1, while N1_M3/ N1_M4/ N1_M5 trained by Net1 and N2_M3/ N2_M4/ N2_M5 trained by Net2 form two sub-categories in Category_2. So, we can draw the conclusion that prediction results from different feature sets show obvious differences, and the conclusion is true for all-range, short-range, mid-range, and long-range contacts/distances. The feature set decides the similarity between models for typical architectures of networks.

Ensembling Different Feature Sets Improves Prediction Performance

The prediction accuracies of N1_M1/ N1_M2/ N1_M3/ N1_M4/ N1_M5/ N1_Ensemble (obtained by averaging the five Net1 sub-models) and N2_M1/ N2_M2/ N2_M3/ N2_M4/ N2_M5/ N2_Ensemble (obtained by averaging the five Net2 sub-models) are listed in **Tables 2**, **3**, respectively.

As we can see from **Table 2**, N1_M1 trained through randomly shuffling MSA_All can obtain the best performance, which is 1.8%/ 0.3%/ 0.9%/ 1.8%, 2.8%/ 0.1%/ 0.9%/ 3.0%, 5.1%/ 1%/ 2.1%/ 5.5%, and 4.5%/ 0.1%/ 2.1%/ 5% higher than N2_M2/ N2_M3/ N2_M4/ N2_M5 for top-L all-/ short-/ medium-/ long-range predictions. Although using the same network and feature set, N1_M1 shows superior prediction precisions than N1_M2, implying that randomly shuffling MSA_All in each epoch enables augmentation of the training set and thus, a better model can be obtained. N1_M3 uses the same network and feature set as N1_M4 and N1_M5, but the prediction precisions of N1_M3 are higher than N1_M4 and N1_M5, indicating that high-quality MSA used for training helps to boost the model performance. N1_Ensemble outperforms the individual sub-models N1_M1/ N1_M2/ N1_M3/ N1_M4/ N1_M5 by 1.3%/ 3.1%/ 4.0%/ 6.4%/ 5.8%, 0.3%/ 0.6%/ 0.4%/ 1.6%/ 0.4%, 0.3%/ 1.2%/ 1.2%/ 2.3%/ 1.7%, and 1.7%/ 3.5%/ 4.8%/ 7.2%/ 6.7% for top-L all-/ short-/ medium-/ long-range predictions, suggesting that ensembles of models trained on different feature sets can improve the overall prediction performance. Similar phenomenon can be observed and consistent conclusions can be drawn from the results in **Table 3**.

The Overall Performance of DuetDis

The prediction precisions of all-/ short-/ medium-/ long-range contacts for DuetDis and other 11 peer methods on the independent test set are shown in **Figure 4**. In general, DL methods, which can capture the higher-order residue correlations and use nonlinear models with fewer parameters to be estimated from thousands of protein families (Rajgaria et al., 2010), significantly outperform DCA methods. Specifically, DuetDis shows the best overall performance. Compared with DeepCov/ PconsC4/ DNCON2/ SPOT/ TripletRes/ trRosetta/ RaptorX, DuetDis obtains 22.1%/ 18.8%/ 17.2%/ 3.5%/ 6.3%/ 3.8%/ 2.4%, 5.2%/ 5.2%/ 3.9%/ 1.0%/ 1.4%/ 0.8%/ 2.2%, 8.7%/ 7.5%/ 6.2%/ 1.4%/ 1.8%/ 1.4%/ 1.4%, and 2.4%/ 1.9%/ 3.8%/

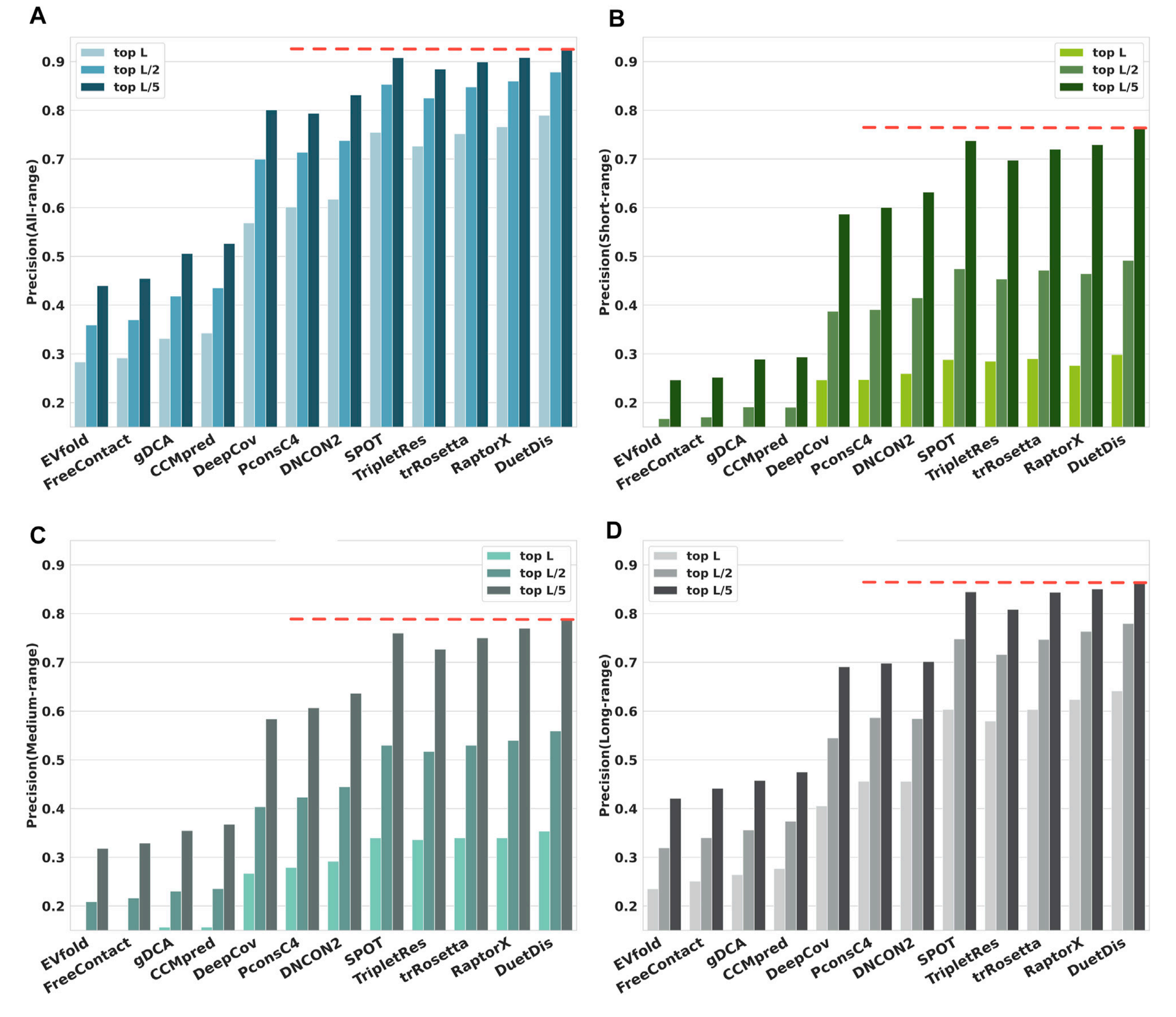

FIGURE 4 | The overall prediction precisions for **(A)** all-range, **(B)** short-range, **(C)** medium-range, and **(D)** long-range contacts/distances.

6.2%/ 3.8%/ 1.8% higher precisions for all-range, short-range, medium-range, and long-range top-L predictions, as well as 12.5%/ 13.3%/ 9.5%/ 1.9%/ 4.2%/ 2.7%/ 1.9%, and 17.6%/ 16.3%/ 13.1%/ 2.6%/ 6.6%/ 4.3%/ 3.4% higher precisions for all-range, short-range, medium-range, and long-range top-L/5 predictions, respectively. The better performance of DuetDis is probably due to the high-quality MSAs used for training, the delicately designed deep residual network, and the effective integration of different features.

DuetDis Embraces High Model Reliability in Terms of Prediction Score

The confidence of the probability (score) given by a DCA or DL model can greatly reflect the reliability of the corresponding model. The prediction probabilities (scores) given by EVfold, FreeContact, gDCA, CCMpred, DeepCov, PconsC4, DNCON2, SPOT, TripletRes, trRosetta, RaptorX, and DuetDis are distributed at (0.000,1.309), (−2.537,17.931), (−1.243, 6.564), (0.000, 5.270), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), (0.0, 1.0), and (0.0, 1.0), respectively. For machine learning (both traditional and deep learning) applications, people usually use 0.5 as a threshold for classification. However, the threshold may be inaccurate for a complex problem like contact/distance prediction. Therefore, studying the scoring trend and the reliability of the model is of great benefit to understand the model performance.

Figure 5 illustrates the prediction performance in terms of precision/ coverage/ MCC with the increase in probability (score) threshold given by DuetDis and the peer methods. With the increase of the probability (score) threshold, the prediction coverages decrease monotonically for all methods. As the threshold increases, their precision curves go down at some

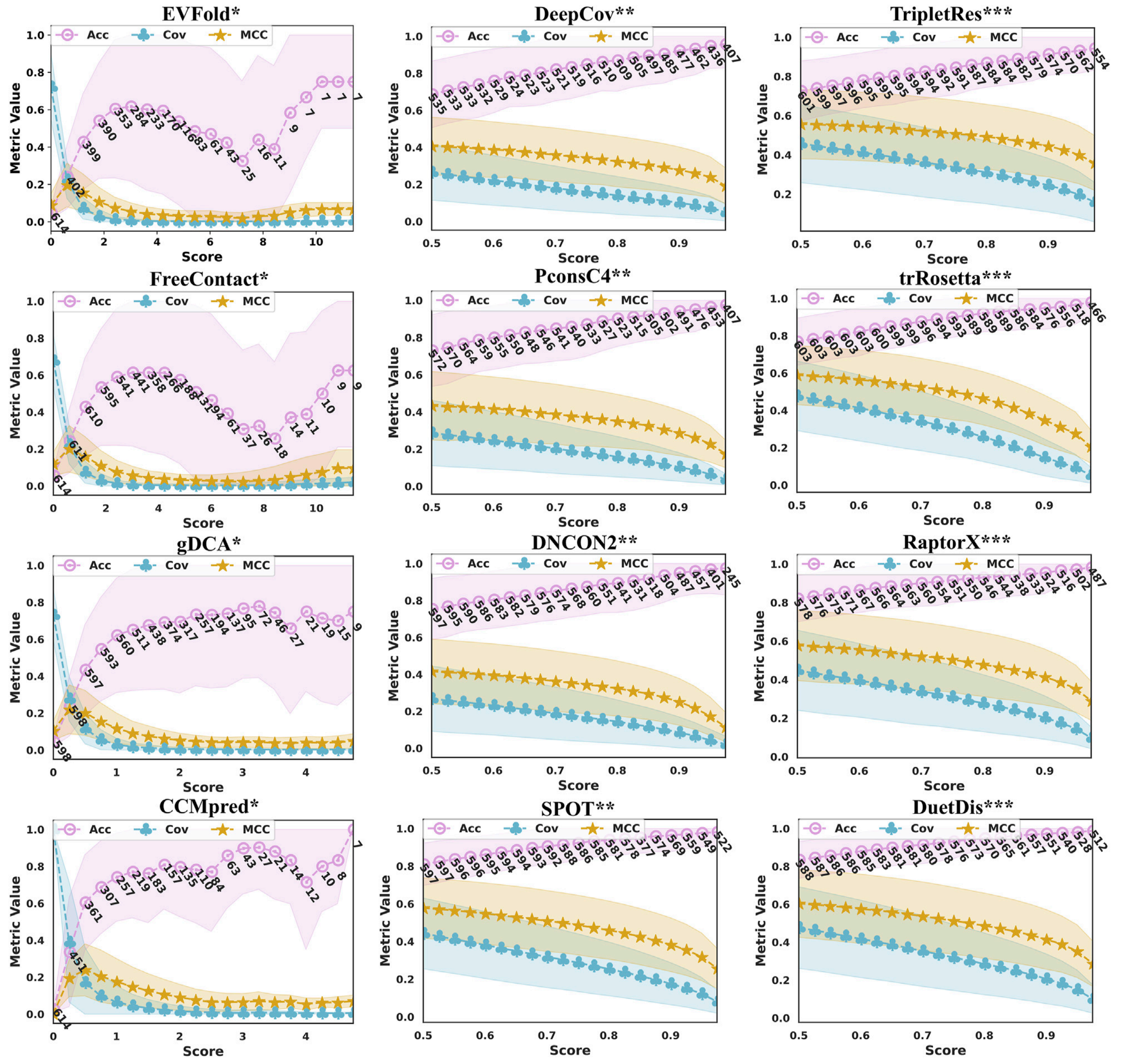

FIGURE 5 | Prediction performance in terms of precision, coverage, MCC, and the corresponding standard deviation (the shaded area around the curves) with the increasing probability (score) threshold given by the predictors. The numbers under the precision curve (blue) are the numbers of proteins with predictions returned using the corresponding (score) threshold on the x-axis.

probability (score) value. The prediction precisions of all DL methods (DeepCov/ PconsC4/ DNCON2/ SPOT/ TripletRes/ trRosetta/ RaptorX) increase monotonically with the probability (score) threshold. However, the precision curves of DCA methods (EVfold/ FreeContact/ gDCA/ CCMpred) show turning points at some probability (score) values. Meanwhile, DCA methods also show much larger STDs on precisions and relatively lower coverages/MCCs compared with DL methods. The numbers under the precision curve in **Figure 4** are the numbers of proteins with predictions returned using the corresponding probability (score) threshold on the x-axis. It is obvious that, as the probability (score) threshold increases, there are more proteins being predicted by DL methods than by DCA methods. Specifically, DuetDis achieves prediction precisions/ coverages/ MCCs of 98.1%/ 15.0%/ 0.352 (calculated on the 523 proteins with prediction scores higher than 0.95) at the (score) threshold of 0.95, which are higher than that by DeepCov (94.7%/ 7.4%/ 0.240: 431 proteins), PconsC4 (96.3%/ 6.5%/ 0.228: 448 proteins), DNCON2 (96.8%/ 4.4%/ 0.173: 396 proteins), SPOT (97.5%/ 12.3%/ 0.318: 544 proteins), TripletRes (93.0%/ 19.6%/ 0.399: 557 proteins), trRosetta (96.5%/ 9.4%/ 0.276: 513 proteins), and RaptorX (97.2%/ 14.5%/ 0.352: 497 proteins). In summary,

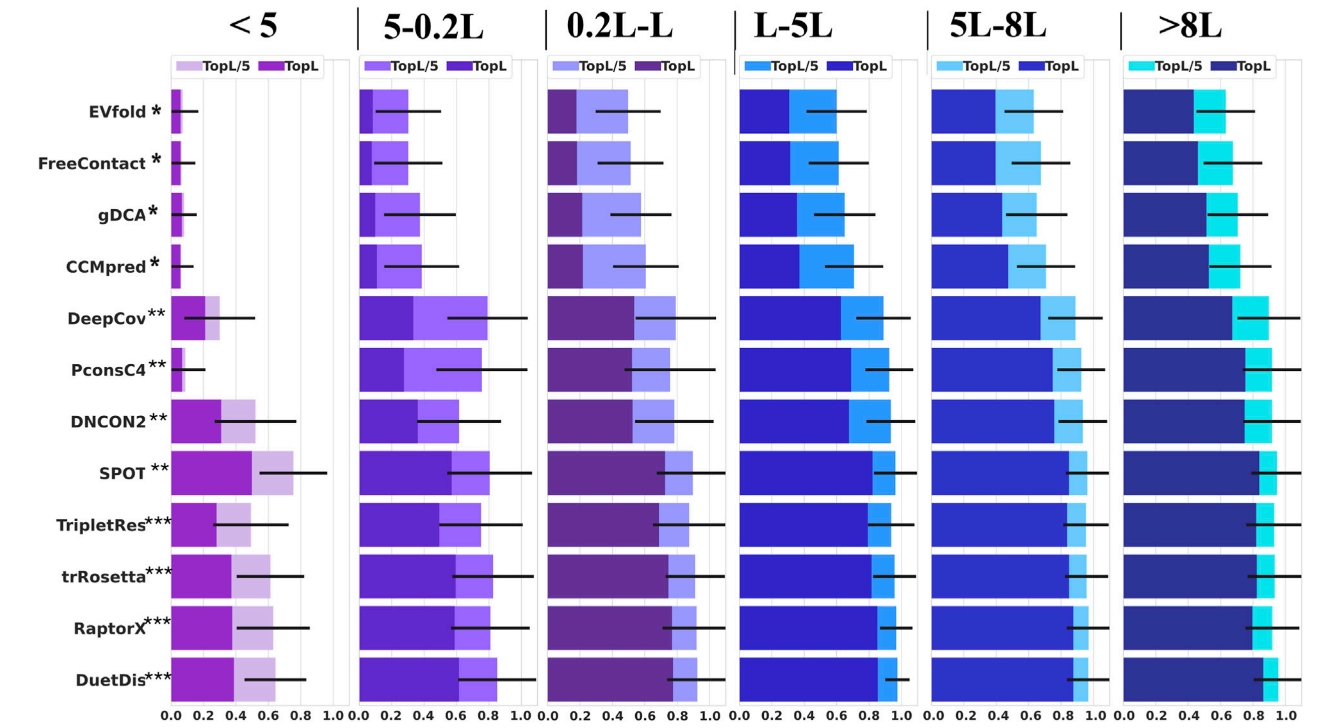

FIGURE 6 | Prediction precisions of different methods for all-range, top-L, and top-L/5 predictions with the variation of N_{eff}. The error bar is the standard deviation of all precisions (for top L/5 predictions) in each sub-test set.

DuetDis shows higher reliability in model probability (score) compared with peer methods.

DuetDis Is Robust Against Shallow Multiple Sequence Alignment

Coevolutionary coupling signals extracted from MSA play central role in most modern contact/distance prediction methods. In this study, the independent test set is divided into six groups according to N_{eff} (<5, 5–0.2 L, 0.2 L–L, L–5 L, 5–8 L, and >8 L). The performance of different methods on these sub-groups of the test set is shown in **Figure 6**. DuetDis achieves prediction precisions of 64.4% for N_{eff} <5, 85.1% for N_{ef} = 5–0.2 L (2.5% higher than the second), 92.5% for N_{eff} = 0.2 L–L (0.5% higher than the second), 97.5% for N_{eff} = L–5 L (0.8% higher than the second), 96.9% for N_{eff} = 5–8 L (0.2% higher than the second), and 95.6% for N_{eff} = 5–8 L (0.9% higher than the second). For N_{eff} <5 L, DuetDis ranks the second in prediction precision; while for N_{eff} = 5–0.2 L, 0.2 L–L, L–5 L, 5–8 L and >8 L, DuetDis is in the leading position of prediction precision. For Neff <5 L, PconsC4 shows a STD of 0.125 which is smaller than DuetDis, however, the smaller STD is because of lower overall precision by PconsC4 (the average prediction precisions are 8.7% for PconsC4 and 64.4% for DuetDis). Hence, DuetDis obtains the least STD among all DL methods for all sub-groups of the test set. In general, DuetDis shows leading precisions and the smallest STD for most ranges of N_{eff}, especially highlights its robustness in shallow MSA-based distance prediction.

CONCLUSION

Proteins are considered as the molecular machines and perform many important functions of life (Zhang et al., 2017). Knowing the structure of a protein helps to understand the role of the protein, how the protein performs its biological function, and the interaction between the protein and the protein (or other molecules), which is very important for biology as well as for medicine and pharmacy. Residue distance prediction from the sequence is critical for many biological applications such as protein structure reconstruction. However, prediction of large distances and distances between residues with long sequence separation length still remains challenging.

In this paper, we propose DuetDis, which uses duet deep learning models for distance prediction. DuetDis adopts two complementary feature sets, one set is mainly composed of 2D coevolutionary couplings, and another set contains mainly 1D sequence-based features. We trained 10 sub-models using two different networks (Net1 and Net2), two different sets of features (FeatSet1 and FeatSet2), and four different MSAs (MSA_All, MSA_Top, MSA_1, MSA_2). By evaluating 10 sub-models based on the large-scale test set, we found that: 1) prediction results from different feature sets show obvious differences; 2) ensembling different feature sets can improve the prediction performance; and 3) high-quality MSA used for both training and testing can greatly improve the prediction performance. DuetDis is also compared with 11 widely used contact/distance predictors. The experimental results show that DuetDis outperforms the peer methods in terms of overall prediction precisions, model reliability, and robustness against shallow MSA.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HZ, YH, and ZB conducted the experiments; all authors analyzed the data; HZ and WX wrote the manuscript.

FUNDING

This work was partly supported by the National Key Research and Development Program of China under Grant No. 2018YFB0204403, Strategic Priority CAS Project XDB38050100, the Key Research and Development Project of Guangdong Province under Grant No. 2021B0101310002, National Science Foundation of China under Grant No. U1813203, the Shenzhen Basic Research Fund under Grant Nos. RCYX2020071411473419, JCYJ20200109114818703, and JSGG20201102163800001, and CAS Key Lab under Grant No. 2011DP173015.

REFERENCES

Adhikari, B. (2020). A Fully Open-Source Framework for Deep Learning Protein Real-Valued Distances. *Sci. Rep.* 10 (1), 13374. doi:10.1038/s41598-020-70181-0

Adhikari, B., Bhattacharya, D., Cao, R., and Cheng, J. (2015). CONFOLD: Residue-Residue Contact-Guidedab Initioprotein Folding. *Proteins* 83 (8), 1436–1449. doi:10.1002/prot.24829

Adhikari, B., Hou, J., and Cheng, J. (2018). DNCON2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *Bioinformatics* 34 (9), 1466–1472. doi:10.1093/bioinformatics/btx781

Altschul, S., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a New Generation of Protein Database Search Programs. *Nucleic Acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science* 181 (4096), 223–230. doi:10.1126/science.181.4096.223

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., et al. (2021). De Novo protein Design by Deep Network Hallucination. *Nature* 600 (7889), 547–552. doi:10.1038/s41586-021-04184-w

Aszódi, A., and Taylor, W. R. (1996). Homology Modelling by Distance Geometry. *Folding Des.* 1 (5), 325–334.

Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., et al. (2014). Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PloS one* 9 (3), e92721. doi:10.1371/journal.pone.0092721

Betancourt, M. R., and Thirumalai, D. (1999). Pair Potentials for Protein Folding: Choice of Reference States and Sensitivity of Predicted Native States to Variations in the Interaction Schemes. *Protein Sci.* 8 (2), 361–369. doi:10.1110/ps.8.2.361

Cheng, J., and Baldi, P. (2007). Improved Residue Contact Prediction Using Support Vector Machines and a Large Feature Set. *Bmc Bioinformatics* 8 (1), 113. doi:10.1186/1471-2105-8-113

Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein Interaction Networks Revealed by Proteome Coevolution. *Science* 365 (6449), 185–189. doi:10.1126/science.aaw6718

Ding, W., and Gong, H. (2020). Predicting the Real-Valued Inter-Residue Distances for Proteins. *Adv. Sci.* 7 (19), 2001314. doi:10.1002/advs.202001314

Ding, W., Mao, W., Shao, D., Zhang, W., and Gong, H. (2018). DeepConPred2: An Improved Method for the Prediction of Protein Residue Contacts. *Comput. Struct. Biotechnol. J.* 16, 503–510. doi:10.1016/j.csbj.2018.10.009

Du, T., Liao, L., Wu, C. H., and Sun, B. (2016). Prediction of Residue-Residue Contact Matrix for Protein-Protein Interaction with Fisher Score Features and Deep Learning. *Methods* 110, 97–105. doi:10.1016/j.ymeth.2016.06.001

Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2007). Mutual Information without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction. *Bioinformatics* 24 (3), 333–340. doi:10.1093/bioinformatics/btm604

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 87 (1), 012707. doi:10.1103/PhysRevE.87.012707

Gao, M., Zhou, H., and Skolnick, J. (2019). DESTINI: A Deep-Learning Approach to Contact-Driven Protein Structure Prediction. *Sci. Rep.* 9 (1), 3514. doi:10.1038/s41598-019-40314-1

Greener, J. G., Kandathil, S. M., and Jones, D. T. (2019). Deep Learning Extends De Novo Protein Modelling Coverage of Genomes Using Iteratively Predicted Structural Constraints. *Nat. Commun.* 10 (1), 3977. doi:10.1038/s41467-019-11994-0

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2018). Accurate Prediction of Protein Contact Maps by Coupling Residual Two-Dimensional Bidirectional Long Short-Term Memory with Convolutional Neural Networks. *Bioinformatics* 34 (23), 4039–4045. doi:10.1093/bioinformatics/bty481

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., and Zhou, Y. (2019). Improving Prediction of Protein Secondary Structure, Backbone Angles, Solvent Accessibility and Contact Numbers by Using Predicted Contact Maps and an Ensemble of Recurrent and Residual Convolutional Neural Networks. *Bioinformatics* 35 (14), 2403–2410. doi:10.1093/bioinformatics/bty1006

He, B., Mortuza, S. M., Wang, Y., Shen, H.-B., and Zhang, Y. (2017). NeBcon: Protein Contact Map Prediction Using Neural Network Training Coupled with Naïve Bayes Classifiers. *Bioinformatics* 33 (15), 2296–2306. doi:10.1093/bioinformatics/btx164

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep Residual Learning for Image Recognition,"in Proceedings of the IEEE conference on computer vision and pattern recognition. 17-19 June 1997. Juan, PR, USA. (IEEE). doi:10.1109/cvpr.2016.90

Jain, A., Terashi, G., Kagaya, Y., Venkata Subramaniya, S. R. M., Christoffer, C., and Kihara, D. (2021). Analyzing Effect of Quadruple Multiple Sequence Alignments on Deep Learning Based Protein Inter-residue Distance Prediction. *Scientific Rep.* 11 (1), 1–13. doi:10.1038/s41598-021-87204-z

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure. *BMC bioinformatics* 11 (1), 431. doi:10.1186/1471-2105-11-431

Jones, D. T., Buchan, D. W. A., Cozzetto, D., and Pontil, M. (2012). PSICOV: Precise Structural Contact Prediction Using Sparse Inverse Covariance Estimation on Large Multiple Sequence Alignments. *Bioinformatics* 28 (2), 184–190. doi:10.1093/bioinformatics/btr638

Jones, D. T., and Kandathil, S. M. (2018). High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* 34 (19), 3308–3315. doi:10.1093/bioinformatics/bty341

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2014). MetaPSICOV: Combining Coevolution Methods for Accurate Prediction of Contacts and Long Range Hydrogen Bonding in Proteins. *Bioinformatics* 31 (7), 999–1006. doi:10.1093/bioinformatics/btu791

Ju, Z., Zhang, H., Meng, J., Zhang, J., Li, X., Fan, J., et al. (2021). "An Efficient Greedy Incremental Sequence Clustering Algorithm," in *International Symposium on Bioinformatics Research and Applications* (Springer, Cham). doi:10.1007/978-3-030-91415-8_50

Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S., and Rost, B. (2014). FreeContact: Fast and Free Software for Protein Contact Prediction from Residue Co-evolution. *BMC bioinformatics* 15 (1), 85. doi:10.1186/1471-2105-15-85

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and

Structure-Rich Era. *Proc. Natl. Acad. Sci. U.S.A.* 110 (39), 15674–15679. doi:10.1073/pnas.1314045110

Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri, P., and Pollastri, G. (2014). Toward an Accurate Prediction of Inter-residue Distances in Proteins Using 2D Recursive Neural Networks. *BMC bioinformatics* 15 (1), 6–15. doi:10.1186/1471-2105-15-6

Lee, B.-C., and Kim, D. (2009). A New Method for Revealing Correlated Mutations under the Structural and Functional Constraints in Proteins. *Bioinformatics* 25 (19), 2506–2513. doi:10.1093/bioinformatics/btp455

Li, J., and Xu, J. (2021). Study of Real-Valued Distance Prediction for Protein Structure Prediction with Deep Learning. *Bioinformatics* 37 (19), 3197–3203. doi:10.1093/bioinformatics/btab333

Li, Y., Hu, J., Zhang, C., Yu, D.-J., and Zhang, Y. (2019). ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks. *Bioinformatics* 35 (22), 4647–4655. doi:10.1093/bioinformatics/btz291

Li, Y., Zhang, C., Bell, E. W., 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. doi:10.1371/journal.pcbi.1008865

Liu, Y., Palmedo, P., Ye, Q., Berger, B., and Peng, J. (2018). Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Syst.* 6 (1), 65–74. e3. doi:10.1016/j.cels.2017.11.014

Malinin, A., and Gales, M. J. F. (2021). Uncertainty Estimation in Autoregressive Structured Prediction. 9th International Conference on Learning Representations, {ICLR} 2021, Virtual Event, Austria, May 3-7, 2021. Available at: https://openreview.net/forum?id=jN5y-zb5Q7m.

Marks, D. S., Hopf, T. A., and Sander, C. (2012). Protein Structure Prediction from Sequence Variation. *Nat. Biotechnol.* 30 (11), 1072–1080. doi:10.1038/nbt.2419

McAllister, S. R., and Floudas, C. A. (2008). α-Helical Topology Prediction and Generation of Distance Restraints in Membrane Proteins. *Biophysical J.* 95 (11), 5281–5295. doi:10.1529/biophysj.108.132241

Michel, M., Hayat, S., Skwark, M. J., Sander, C., Marks, D. S., and Elofsson, A. (2014). PconsFold: Improved Contact Predictions Improve Protein Models. *Bioinformatics* 30 (17), i482–i488. doi:10.1093/bioinformatics/btu458

Michel, M., Menéndez Hurtado, D., and Elofsson, A. (2019). PconsC4: Fast, Accurate and Hassle-free Contact Predictions. *Bioinformatics* 35 (15), 2677–2679. doi:10.1093/bioinformatics/bty1036

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., et al. (2011). Direct-coupling Analysis of Residue Coevolution Captures Native Contacts across many Protein Families. *Proc. Natl. Acad. Sci. U S A.* 108 (49), E1293–E1301. doi:10.1073/pnas.1111471108

Pollock, D. D., and Taylor, W. R. (1997). Effectiveness of Correlation Analysis in Identifying Protein Residues Undergoing Correlated Evolution. *Protein Eng. Des. Selection* 10 (6), 647–657. doi:10.1093/protein/10.6.647

Rahman, J., Newton, M. A. H., Islam, M. K. B., and Sattar, A. (2022). Enhancing Protein Inter-residue Real Distance Prediction by Scrutinising Deep Learning Models. *Sci. Rep.* 12 (1), 787. doi:10.1038/s41598-021-04441-y

Rajgaria, R., McAllister, S. R., and Floudas, C. A. (2009). Towards Accurate Residue-Residue Hydrophobic Contact Prediction for α Helical Proteins via Integer Linear Optimization. *Proteins* 74 (4), 929–947. doi:10.1002/prot.22202

Rajgaria, R., Wei, Y., and Floudas, C. A. (2010). Contact Prediction for Beta and Alpha-Beta Proteins Using Integer Linear Optimization and its Impact on the First Principles 3D Structure Prediction Method ASTRO-FOLD. *Proteins* 78 (8), 1825–1846. doi:10.1002/prot.22696

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment. *Nat. Methods* 9 (2), 173–175. doi:10.1038/nmeth.1818

Reza, M. S., Zhang, H., Hossain, M. T., Jin, L., Feng, S., and Wei, Y. (2021). COMTOP: Protein Residue-Residue Contact Prediction through Mixed Integer Linear Optimization. *Membranes* 11 (7), 503. doi:10.3390/membranes11070503

Schlessinger, A., Punta, M., and Rost, B. (2007). Natively Unstructured Regions in Proteins Identified from Contact Predictions. *Bioinformatics* 23 (18), 2376–2384. doi:10.1093/bioinformatics/btm349

Seemayer, S., Gruber, M., and Söding, J. (2014). CCMpred-fast and Precise Prediction of Protein Residue-Residue Contacts from Correlated Mutations. *Bioinformatics* 30 (21), 3128–3130. doi:10.1093/bioinformatics/btu500

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., et al. (2020). Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* 577, 706–710. doi:10.1038/s41586-019-1923-7

Shimomura, T., Nishijima, K., and Kikuchi, T. (2019). A New Technique for Predicting Intrinsically Disordered Regions Based on Average Distance Map Constructed with Inter-residue Average Distance Statistics. *BMC Struct. Biol.* 19 (1), 3–12. doi:10.1186/s12900-019-0101-3

Singh, J., Litfin, T., Singh, J., Paliwal, K., and Zhou, Y. (2022). SPOT-Contact-LM: Improving Single-Sequence-Based Prediction of Protein Contact Map Using a Transformer Language Model. *Bioinformatics*. doi:10.1093/bioinformatics/btac053

Skwark, M. J., Abdel-Rehim, A., and Elofsson, A. (2013). PconsC: Combination of Direct Information Methods and Alignments Improves Contact Prediction. *Bioinformatics* 29 (14), 1815–1816. doi:10.1093/bioinformatics/btt259

Su, H., Wang, W., Du, Z., Peng, Z., Gao, S. H., Cheng, M. M., et al. (2021). Improved Protein Structure Prediction Using a New Multi-Scale Network and Homologous Templates. *Adv. Sci.* 8, 2102592. doi:10.1002/advs.202102592

Tegge, A. N., Wang, Z., Eickholt, J., and Cheng, J. (2009). NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks. *Nucleic Acids Res.* 37, W515–W518. doi:10.1093/nar/gkp305

Vangone, A., and Bonvin, A. M. (2015). Contacts-based Prediction of Binding Affinity in Protein-Protein Complexes. *elife* 4, e07454. doi:10.7554/eLife.07454

Walsh, I., Baù, D., Martin, A. J., Mooney, C., Vullo, A., and Pollastri, G. (2009). Ab Initio and Template-Based Prediction of Multi-Class Distance Maps by Two-Dimensional Recursive Neural Networks. *BMC Struct. Biol.* 9 (1), 5–20. doi:10.1186/1472-6807-9-5

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate De Novo Prediction of Protein Contact Map by Ultra-deep Learning Model. *Plos Comput. Biol.* 13 (1), e1005324. doi:10.1371/journal.pcbi.1005324

Wang, Z., and Xu, J. (2013). Predicting Protein Contact Map Using Evolutionary and Physical Constraints by Integer Programming. *Bioinformatics* 29 (13), i266–i273. doi:10.1093/bioinformatics/btt211

Wei, Y., and Floudas, C. A. (2011). Enhanced Inter-helical Residue Contact Prediction in Transmembrane Proteins. *Chem. Eng. Sci.* 66 (19), 4356–4369. doi:10.1016/j.ces.2011.04.033

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009). Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing. *Proc. Natl. Acad. Sci. U.S.A.* 106 (1), 67–72. doi:10.1073/pnas.0805923106

Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., and Yang, J. (2020). Protein Contact Prediction Using Metagenome Sequence Data and Residual Neural Networks. *Bioinformatics* 36 (1), 41–48. doi:10.1093/bioinformatics/btz477

Wu, S., and Zhang, Y. (2008). A Comprehensive Assessment of Sequence-Based and Template-Based Methods for Protein Contact Prediction. *Bioinformatics* 24 (7), 924–931. doi:10.1093/bioinformatics/btn069

Wu, T., Guo, Z., Hou, J., and Cheng, J. (2021). DeepDist: Real-Value Inter-Residue Distance Prediction with Deep Residual Convolutional Network. *BMC Bioinform.* 22, 30. doi:10.1186/s12859-021-04269-3

Xu, J. (2019). Distance-based Protein Folding Powered by Deep Learning. *Proc. Natl. Acad. Sci. U.S.A.* 116 (34), 16856–16865. doi:10.1073/pnas.1821309116

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved Protein Structure Prediction Using Predicted Interresidue Orientations. *Proc Natl Acad Sci U S A.* 117(3), 1496-1503. doi:10.1073/pnas.1914677117

Zhang, H., Bei, Z., Xi, W., Hao, M., Ju, Z., Saravanan, K. M., et al. (2021). Evaluation of Residue-Residue Contact Prediction Methods: From Retrospective to Prospective. *Plos Comput. Biol.* 17 (5), e1009027. doi:10.1371/journal.pcbi.1009027

Zhang, H., Wu, H., Ting, H. F., and Wei, Y. (2020). "Protein Interresidue Contact Prediction Based on Deep Learning and Massive Features from Multi-Sequence Alignment," in International Conference on Parallel and Distributed Computing: Applications and Technologies, Shenzhen, China, December 28-30 (Shenzhen: Springer).

Zhang, H., Hao, M., Wu, H., Ting, H.-F., Tang, Y., Xi, W., et al. (2022). Protein Residue Contact Prediction Based on Deep Learning and Massive Statistical Features from Multi-Sequence Alignment. *Tsinghua Sci. Technol.* 27 (5), 843–854. doi:10.26599/tst.2021.9010064

Zhang, H., Huang, Q., Bei, Z., Wei, Y., and Floudas, C. A. (2016). COMSAT: Residue Contact Prediction of Transmembrane Proteins Based on Support Vector Machines and Mixed Integer Linear Programming. *Proteins* 84 (3), 332–348. doi:10.1002/prot.24979

Zhang, H., Xi, W., Hansmann, U. H. E., and Wei, Y. (2017). Fibril-Barrel Transitions in Cylindrin Amyloids. *J. Chem. Theor. Comput.* 13 (8), 3936–3944. doi:10.1021/acs.jctc.7b00383

Zhao, F., and Xu, J. (2012). A Position-specific Distance-dependent Statistical Potential for Protein Structure and Functional Study. *Structure* 20 (6), 1118–1126. doi:10.1016/j.str.2012.04.003

Zheng, W., Zhou, X., Wuyun, Q., Pearce, R., Li, Y., and Zhang, Y. (2020). FUpred: Detecting Protein Domains through Deep-Learning-Based Contact Map Prediction. *Bioinformatics* 36 (12), 3749–3757. doi:10.1093/bioinformatics/btaa217

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.