

Structural bioinformatics

# Toward the assessment of predicted inter-residue distance

Zongyang Du<sup>1</sup>, Zhenling Peng<sup>2</sup> and Jianyi Yang<sup>1,\*</sup>

<sup>1</sup>School of Mathematical Sciences, Nankai University, Tianjin 300071, China and <sup>2</sup>Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao 266237, China

\*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 5, 2021; revised on October 7, 2021; editorial decision on November 9, 2021; accepted on November 10, 2021

## Abstract

**Motivation:** Significant progress has been achieved in distance-based protein folding, due to improved prediction of inter-residue distance by deep learning. Many efforts are thus made to improve distance prediction in recent years. However, it remains unknown what is the best way of objectively assessing the accuracy of predicted distance.

**Results:** A total of 19 metrics were proposed to measure the accuracy of predicted distance. These metrics were discussed and compared quantitatively on three benchmark datasets, with distance and structure models predicted by the trRosetta pipeline. The experiments show that a few metrics, such as *distance precision*, have a high correlation with the model accuracy measure TM-score (Pearson's correlation coefficient >0.7). In addition, the metrics are applied to rank the distance prediction groups in CASP14. The ranking by our metrics coincides largely with the official version. These data suggest that the proposed metrics are effective for measuring distance prediction. We anticipate that this study paves the way for objectively monitoring the progress of inter-residue distance prediction. A web server and a standalone package are provided to implement the proposed metrics.

**Availability and implementation:** <http://yanglab.nankai.edu.cn/APD>.

**Contact:** yangjy@nankai.edu.cn

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Significant progress has been achieved in protein structure prediction since the 13th community-wide experiment on the critical assessment of techniques for protein structure prediction (CASP13) (Kryshtafovych *et al.*, 2019). This is mostly attributed to the improved inter-residue distance predicted by deep learning, which can be used to guide de novo folding (Greener *et al.*, 2019; Hou *et al.*, 2019; Pearce and Zhang, 2021; Senior *et al.*, 2020; Xu, 2019; Yang *et al.*, 2020). As accurate distance prediction leads to accurate structure folding, many inter-residue distance prediction methods emerged in recent years (Adhikari, 2020; Ding and Gong, 2020; Ji *et al.*, 2019; Kukic *et al.*, 2014; Wu *et al.*, 2020, 2021; Xu, 2019; Yang *et al.*, 2020). Similar to trRosetta (Yang *et al.*, 2020), inter-residue distance was first predicted and then used to guide the subsequent folding in RoseTTAFold's two-track PyRosetta model (Baek *et al.*, 2021). Inter-residue distance was used as one of the terms in AlphaFold2's loss function (Jumper *et al.*, 2021). Due to the key role of distance prediction in protein folding, the recent CASP14 experiment added a new category for assessing predicted distance.

At its infant stage, it remains unclear about how to assess predicted distance objectively. For example, AlphaFold used distogram

LDDT (DLDDT) (Senior *et al.*, 2020), which is analogous to the model quality measure LDDT (Mariani *et al.*, 2013), to assess predicted distance. In RaptorX-Contact (Xu and Wang, 2019), a few metrics such as absolute error and relative error were used instead. The Pearson's correlation coefficient between the predicted and the distance in experimental structure was adopted in trRosetta (Yang *et al.*, 2020). Although predicted distance was evaluated in CASP14, the definitions of the metrics are still vague (no formal publications are available at the time of this work), making them hard to be implemented. Thus, a systematic investigation is desirable on the problem of predicted distance assessment, which can guide further development of methods for inter-residue distance prediction.

In this work, inspired by the previous studies, a total of 19 metrics were defined to evaluate predicted distance. These metrics were applied to the distance prediction on three benchmark datasets. To quantify the objectiveness of these metrics, their correlations with the TM-score of the predicted structure models are used. Analysis on the distance prediction in the recent CASP14 experiment shows that the group rankings based on the proposed metrics are largely consistent with those from the official version. To facilitate the implementation and usage of the proposed methods, a web server and a standalone package are provided.

## 2 Materials and methods

### 2.1 Benchmark datasets

Three datasets from previous studies were used in this work. The first one (denoted by D680) contains 680 non-redundant single-domain proteins from the work of Wuyun *et al.* (2018). The proteins in this dataset were divided into three subsets based on their difficulty: easy (344), medium (105) and hard (231). The remaining two are from the work of trRosetta (Yang *et al.*, 2020): CASP13-31 and CAMEO131, which include 31 and 131 hard targets, from the CASP13 and the CAMEO experiments (Haas *et al.*, 2018), respectively.

### 2.2 Distance prediction and *de novo* structure modeling

To enable the correlation analysis in Section 3, the trRosetta algorithm was used for integrated predictions of the inter-residue distance and distance-based structure folding. trRosetta predicts the inter-residue distance and orientations based on deep residual neural networks, which are used to guide Rosetta to build structure models with energy minimizations. trRosetta was shown to be one of the state-of-the-art algorithms for inter-residue distance prediction and structure folding. Nevertheless, other algorithms, such as RaptorX-Contact (Xu, 2019), can be used as well.

The distance predicted by trRosetta is represented in the form of a distribution. The distance (between 2 and 20 Å) was originally divided into 36 bins with a step size of 0.5 Å. Another non-contact bin was also included to accommodate distances higher than 20 Å. The recent CASP14 experiment divides the distance into 10 bins with a coarser step size (i.e. 2 Å): (0, 4 Å], (4, 6 Å], (6, 8 Å], ..., (18, 20 Å] and >20 Å. To keep consistent with this division, the predicted probability for each CASP bin is obtained by summing up the probabilities of the corresponding sub-bins from the trRosetta predictions.

To avoid the potential influence of inter-residue orientations, only the predicted distance by trRosetta is used to build structure models based on energy minimization. The scatter plot of the GDT-TS scores (Zemla, 2003) and TM-scores (Zhang and Skolnick, 2004) of the structure models for the targets in all datasets is shown in Supplementary Figure S1. As expected, a strong correlation was observed between GDT-TS score and TM-score. Thus TM-score was applied later for measuring the accuracy of predicted structure models. The average TM-scores of the trRosetta models on the D680, CAMEO131 and CASP13-31 datasets are 0.721, 0.575 and 0.545, respectively. Note that difference was observed between the TM-score data in the trRosetta paper (Yang *et al.*, 2020), which is mostly because the orientation restraints were excluded here. Besides, only one MSA was generated here while an optimal MSA was selected from multiple MSAs in the study by Yang *et al.* (2020).

### 2.3 Distribution of the inter-residue distance in experimental structure

Before designing metrics for assessing predicted distance, it is worthy of investigating the distribution of the distance in experimental structure (defined as native distance). Figure 1A shows the logarithm of the average number of residue pairs (with separation  $\geq 12$ ) in each distance bin. In general, the number of residue pairs in the distance bins increases as the distance gets bigger. Note that, the first and the last bins show different characteristics with others. The number of residue pairs in the first bin (native distance  $\leq 4$  Å) is  $< 10$  on average, which is much lower than other bins. This is because the distance considered here is for two C $\beta$  atoms (with van der Waals radius  $\sim 1.7$  Å) and too close distance (e.g.  $< 3.5$  Å) in 3D space will result into steric clash. In contrast, the last bin (native distance  $> 20$  Å) includes a significantly higher proportion of residue pairs, as it covers more distance ranges than other bins. Indeed, as shown in Figure 1B, the average percentages of residue pairs with native distance  $> 20$  Å are 69.1%, 67.6% and 81.6% for the datasets D680, CASP13-31 and CAMEO131, respectively. Proteins in CAMEO131 have more residues pairs with distance higher than 20 Å. This is probably because this dataset contains multi-domain targets while the other two contain single-domain targets only.

### 2.4 Assessed set of residue pairs

The assessment can be done in three different flavors: *prediction-oriented*, *native-oriented* and *full-list*, depending on the set of residue pairs being assessed. For all cases, we only consider residue pairs with sequence separation  $\geq 12$  (i.e.  $|i-j| \geq 12$ ), to remove the effect of short-range interactions. In the prediction-oriented assessment, the assessed residue pairs are those with high predicted probabilities, regardless of their native distance. While in the native-oriented assessment, the assessed residue pairs are those with native distance no more than a specified threshold (20 Å here), regardless of the predicted probabilities. The full-list case takes all the residue pairs (short-range residue pairs are excluded as well) into account, regardless of distance range.

As discussed in the previous section, residue pairs in the last distance bin account for more than a half of all residue pairs. In addition, even if a prediction for the last bin is very confident (e.g. with a probability close to 1), we still do not know what the exact distance is, due to the broad range of distance represented by this bin (i.e.  $> 20$  Å). Therefore, we think it is necessary to differentiate the 10th bin with others in the assessment of predicted distance, with details given below.

In the prediction-oriented case, all residue pairs are ranked based on the predicted probability  $P(d_{ij} \leq 20 \text{ Å})$ , which equals to the sum of the first nine bins. As only the top residue pairs (i.e. with high confidence) are assessed in this case, the predicted class label for a residue pair is from the first *nine* bins with the highest probability. In contrast, for the native-oriented/full-list case, it is inevitable to include residue pairs with very low probability  $P(d_{ij} \leq 20 \text{ Å})$  (e.g. close to 0). In this situation, we manually assign distance  $d_{ij} = 25 \text{ Å}$ , if  $P(d_{ij} > 20)$  equals to 1 or no prediction is made for the residue pair  $(i, j)$ . In addition, a cutoff of 0.5 is set for the native-oriented/full-list case to tell whether a prediction falls into (0, 20 Å]. When  $P(d_{ij} \leq 20 \text{ Å})$  is higher than or equal to 0.5, the predicted class label is the one with the highest probability in the first *nine* bins (though this probability might be lower than the probability for the 10th bin). Otherwise, the predicted class label will be 10.

### 2.5 Proposed metrics for the assessment of the predicted distance

Once the set of assessed residue pairs are fixed, the assessment can be formulated as either a regression or a classification problem. In the former case, we calculate the real-valued distance from the predicted distance distribution and compare it with that in the experimental structure. In the latter case, different distance bins are regarded as different classes and the native distances are discretized accordingly to evaluate the predicted classes.

To facilitate the description, the following variables are defined. Let  $L$  and  $S$  denote the length of a target protein and the set of assessed residue pairs, respectively.  $D_{ij}$  and  $d_{ij}$  denote the native and the predicted real-valued distance between the  $i$ th and the  $j$ th residue, respectively.  $d_{ij}$  is calculated as a weighted average of the predicted distribution, using probabilities of the first nine bins.  $I(\cdot)$  is an indicator function, whose value is 1 when the corresponding event happens, and 0 otherwise.

In the following, we will introduce nine metrics for the prediction-oriented case, six metrics for the native-oriented case and four metrics for the full-list case. The definitions of these metrics are presented below. They are also summarized in Supplementary Tables S1 for the sake of convenience. Note that, the proposed metrics below is based on the above prediction format (e.g. in the form of a distribution). However, they can be easily adapted to deal with real-valued distance prediction.

#### 2.5.1 Prediction-oriented assessment

**Contact precision (CP).** As an extension of the binary contact prediction, the predicted distance can be converted into contact by summing up the probabilities corresponding to the bins with distance  $\leq 8$  Å. After such conversion, the contact precision, one of the most widely used metrics for assessing predicted contacts, can be used to estimate the accuracy of predicted distance. The contact

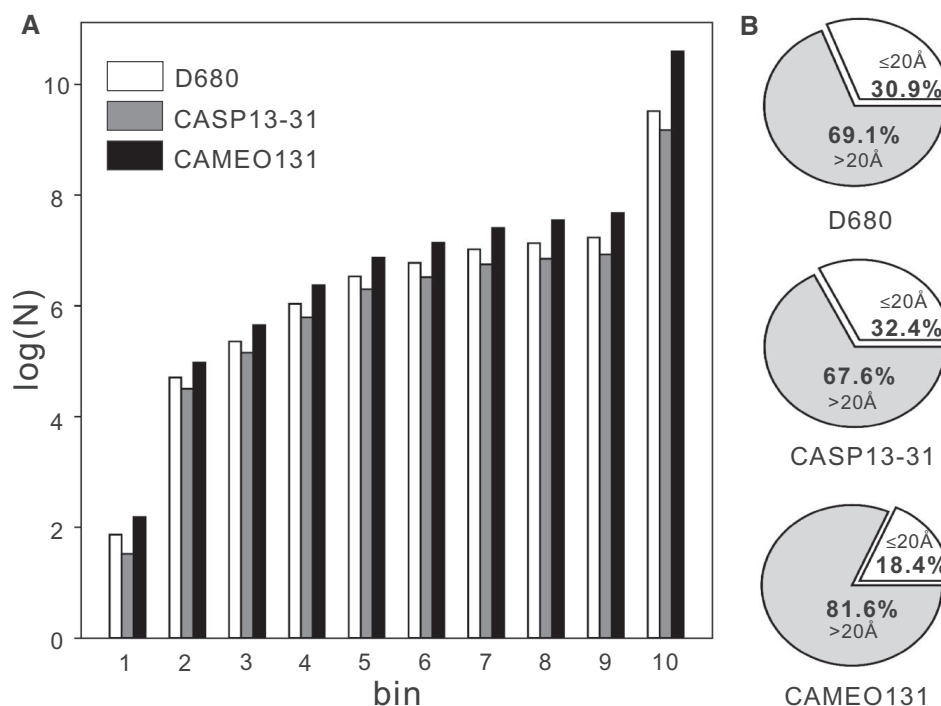


Fig. 1. Distribution of inter-residue distance in the experimental structures. (A) Logarithm of the average number of residue pairs (separation  $\geq 12$ ) in each distance bin on the three datasets. (B) Pie chart of the pair distribution on the three datasets

precision is defined as the number of correctly predicted residue pairs (i.e. with native distance  $\leq 8 \text{ \AA}$ ) in  $S$  divided by the size of  $S$ .

$$CP = \frac{1}{|S|} \sum_{(i,j) \in S} I(D_{ij} \leq 8). \quad (1)$$

**Absolute error (AE) and relative error (RE).** Absolute error and relative error are two intuitive metrics reflecting the deviation from the native distance. The absolute error is computed as the absolute difference between the native and the predicted distance averaged over the set  $S$  (Equation 2). The relative error is defined similarly but with a normalization by the native distance (Equation 3). Note that, similar metrics have been used in RaptorX-Contact (Xu and Wang, 2019).

$$AE = \frac{1}{|S|} \sum_{(i,j) \in S} |D_{ij} - d_{ij}|, \quad (2)$$

$$RE = \frac{1}{|S|} \sum_{(i,j) \in S} \frac{|D_{ij} - d_{ij}|}{D_{ij}}, \quad (3)$$

**Pearson's correlation coefficient (PCC).** We calculate the PCC between the native and the predicted distance. This measure has been used in the work trRosetta.

$$PCC = \frac{Cov(D_S, d_S)}{\sqrt{Var(D_S)Var(d_S)}}, \quad (4)$$

where  $D_S$  and  $d_S$  refer to the vectors containing the native and the predicted distances of the residue pairs in the set  $S$ , respectively.  $Cov(\cdot)/Var(\cdot)$  stands for the covariance/variance of the corresponding vectors. The PCC measures the linear relationship between the predicted and native distances.

**Distance precision (DP).** A residue pair  $(i, j)$  is defined as being 'correctly predicted' if the difference between  $D_{ij}$  and  $d_{ij}$  is less than a tolerance threshold ( $2 \text{ \AA}$  here).

$$DP = \frac{1}{|S|} \sum_{(i,j) \in S} P(d_{ij} \leq 20) I(|D_{ij} - d_{ij}| \leq 2), \quad (5)$$

where  $P(d_{ij} \leq 20)$  is the cumulative probability of all bins with

distance  $\leq 20 \text{ \AA}$ , reflecting the confidence of the predicted distance  $d_{ij}$ . The distance precision indicates the ratio of correctly predicted residue pairs in the set  $S$ .

**Fuzzy certainty (FC).** To effectively utilize the predicted probability, we define the fuzzy certainty of a predicted distance distribution as follows. Similar to the fuzzy analysis, besides the native distance bin, its adjacent bins are also considered but with a weight of 0.5, to reflect the dynamic feature of protein structure.

$$FC = \frac{1}{|S|} \sum_{(i,j) \in S, D_{ij} \leq 20} P(D_{ij}) + \frac{1}{2} P(D_{ij} - 2) I(D_{ij} > 2) + \frac{1}{2} P(D_{ij} + 2) I(D_{ij} \leq 18), \quad (6)$$

where  $P(\cdot)$  is the predicted probability of the corresponding distance bin. The fuzzy certainty reflects the probability of the predicted distance falling into the native bin and its adjacent bins.

**Macro fuzzy precision/recall/F1 (MFP, MFR, MFF).** For each of the first 9 distance bins (classes) with distance  $\leq 20 \text{ \AA}$ , fuzzy precision and fuzzy recall are defined. To define these metrics, the set  $S$  is first divided into a maximum of 9 subsets. For the residue pairs in each subset  $S_k$ , the predicted probability of the  $k$ th distance bin is the highest (among the first 9 distance bins). Here, the word fuzzy has similar meaning to the previous measure (i.e. fuzzy certainty), which means assigning a weight of 0.5 for the predicted class (i.e. distance bin) that is not correct but is adjacent to the native class (i.e. native distance bin).

$$fPRE_k = \frac{1}{|S_k|} \sum_{(i,j) \in S_k} [I(l_{ij} = k) + \frac{1}{2} I(|l_{ij} - k| = 1) I(D_{ij} \leq 20)], \quad k < 10, \quad (7)$$

$$fREC_k = \frac{1}{N_k} \sum_{(i,j) \in S_k} [I(l_{ij} = k) + \frac{1}{2} I(|l_{ij} - k| = 1) I(D_{ij} \leq 20)], \quad k < 10, \quad (8)$$

where  $N_k$  is the number of residue pairs that belong to the  $k$ th class (according to the experimental structure),  $l_{ij}$  is the real class label for the residue pair  $(i, j)$ . The fuzzy F1 score for each class is a harmonic

mean of the corresponding fuzzy precision and fuzzy recall. The macro fuzzy precision/recall/F1 are calculated as the average over all non-empty classes. These metrics are defined in a similar way as the classical macro precision/recall/F1 scores for the evaluation of multi-class predictions; but with additional consideration of the bins adjacent to the native bin.

### 2.5.2 Native-oriented assessment

For the native-oriented assessment, the five metrics defined above can be also calculated: distance precision, fuzzy certainty, macro fuzzy precision, macro fuzzy recall and macro fuzzy F1.

In addition, similar to AlphaFold (Senior *et al.*, 2020), we define the metrics distogram LDDT (DLDDT):

$$\text{DLDDT} = \frac{1}{4L} \sum_{t \in \{0.5, 1, 2, 4\}} \sum_{i=1}^L \frac{\sum_{j \in R_i} P(d_{ij} \leq 20) I(|D_{ij} - d_{ij}| \leq t)}{|R_i|} I(|R_i| > 0), \quad (9)$$

where  $R_i$  is the set of residues that are close to the  $i$ th residue within distance  $20 \text{ \AA}$  and with sequence separation not  $< 12$ . The DLDDT score evaluates the average local distance differences for all residues.

### 2.5.3 Full-list assessment

For full-list assessment, four metrics are defined. The first three are macro fuzzy precision/recall/F1 (MFP, MFR, MFF), which are defined similarly as in the previous cases. The only difference is the inclusion of the bin with native distances  $> 20 \text{ \AA}$ .

**Macro fuzzy certainty (MFC).** Macro fuzzy certainty is an extension of the previous measure fuzzy certainty, which takes the average on each class-specific subset instead of the whole set of assessed residue pairs. First, the fuzzy certainty for each distance class is calculated as below.

$$\text{FC}_k = \begin{cases} \frac{1}{|S_k|} \sum_{(i,j) \in S_k} P_k(i,j) + \frac{1}{2} P_{k-1}(i,j) I(k > 1) + \frac{1}{2} P_{k+1}(i,j) I(k < 9), & k < 10 \\ \frac{1}{|S_k|} \sum_{(i,j) \in S_k} P_k(i,j), & k = 10 \end{cases} \quad (10)$$

where  $S_k$  is the set of residue pairs in the  $k$ th distance class according to the native structure,  $P_k(i, j)$  is the predicted probability in the  $k$ th class for the residue pair  $(i, j)$ . The MFC is then calculated as the average of the fuzzy certainty over all non-empty classes.

## 3 Results and discussions

### 3.1 How many residue pairs should be taken into account?

To evaluate the predicted contacts for a target with  $L$  residues, the top  $L/5$ ,  $L/2$ ,  $L$  residue pairs ranked based on the predicted probabilities are often assessed. As regards to distance evaluation, it is apparent that a greater number of residue pairs should be involved because the distance considered has been extended from  $(0, 8 \text{ \AA}]$  to  $(0, 20 \text{ \AA}]$ . However, it is not advised to consider ‘too many’ residue pairs in the prediction-oriented case. This is because an accurate predictor is anticipated to have more residue pairs with distance  $\leq 20 \text{ \AA}$  in its list of the top-ranked predictions; while these pairs only account for  $< 1/3$  of the whole pair set (Fig. 1B). Thus, the first question in the prediction-oriented assessment of distance prediction is how many residue pairs should be considered?

To answer the above question, we count the number of residue pairs (denoted by  $N$ ) with native distance  $\leq 20 \text{ \AA}$  and compare it with a series of specified thresholds  $nL$  ( $n$  is an integer). Supplementary Figure S2 shows that the percentage of proteins satisfying  $N \geq nL$  decreases as  $n$  increases. For example, there are only about 40% targets having more than  $30L$  residue pairs with distance  $\leq 20 \text{ \AA}$ . To keep a balance between the number of eligible targets to be used and the number of residue pairs to be assessed, a maximum of  $15L$  residue pairs are assessed in this work, at which more than 80% targets have enough residue pairs to be assessed. We

carried out a similar analysis based on contacts. As is shown in Supplementary Figure S3, the percentage of targets satisfying  $N_c \geq L$  ( $N_c$  is the number of contacting residue pairs with distance  $\leq 8 \text{ \AA}$ ) is also close to 80%. Thus,  $L$  residue pairs are considered in the assessment of predicted contacts.

### 3.2 Distribution of the proposed metrics

The distributions of the proposed metrics are shown in Supplementary Figure S4 as boxplots (A: prediction-oriented, B: native-oriented, C: full-list). For better visualization, the absolute error has been divided by 10 and a few outlier targets were excluded. Overall, the dataset D680 has more outlier targets than the other two datasets, probably because its construction did not take into consideration of the target complexity. For example, there are some targets with simple topology (such as single alpha helix) or unstable structure that has binding partners.

In comparison with other metrics, the absolute error and the Pearson’s correlation coefficient have more outliers (the black dots outside the boxes) than other metrics. This may be explained by their definitions (Equations 2 and 4), which can be easily dominated/affected by a small portion of residue pairs that are predicted with high deviation to the native. Besides, the binary and multi-class metrics show different characteristics, though both are defined in the framework of classification. The binary classification measure, contact precision, has a wider range of variation than all the multi-class metrics. Specifically, the contact precisions for some targets reach to a perfect value of 1; while the maximum of other multi-class metrics on all datasets is  $< 0.9$ . This can be explained in two aspects: the calculation of distance metrics involves more residue pairs; and multi-class prediction is more difficult than binary prediction.

### 3.3 Relationship between the proposed metrics

The scatter plots and the correlation between the proposed metrics are presented in Figure 2, Supplementary Figures S5 and S6. As non-linear relationship is observed between some metrics, the Spearman’s rank correlation coefficient (SCC) is used for quantitative analysis.

The correlation between the prediction-oriented metrics is shown in Figure 2A. The distance precision and the fuzzy certainty, the only two metrics containing probability terms in their definitions, have a maximum SCC of 0.99. The hierarchical clustering of SCCs also suggests that they are grouped in the same branch (Fig. 2B). The next measure sharing high correspondence with these two metrics is the absolute error, which is the most intuitive measure to quantify the difference between the predicted and the native distances. The relative error and the absolute error have a high correlation (0.95), probably because they are defined analogously. Besides, the relative error also has a high correlation with the distance precision (0.95). As reflected in the scatter plots, the macro fuzzy precision correlates well with the above-mentioned metrics, with SCCs over 0.8. In contrast, the contact precision (CP) has a lower correlation with these metrics ( $< 0.8$ ). Furthermore, the remaining two multi-class metrics, the macro fuzzy recall and the macro F1 score, do not correlate well with other metrics. This is probably because the assessed top predictions ( $15L$  here) only cover part of the residue pairs that are close in the native structure. This can be seen from the improved correlation in the native-oriented and the full-list evaluations (Supplementary Figs S5 and S6), in which most metrics are highly correlated with over 0.9 SCC.

### 3.4 Correlation between the proposed metrics and TM-score

The ultimate goal of predicting inter-residue distance is to use them to guide tertiary structure prediction. For example, distance geometry has been used to fold protein structure from predicted distance in DMPfold (Greener *et al.*, 2019) and RaptorX-Contact (Xu, 2019). Inspired by AlphaFold (Senior *et al.*, 2020), constrained

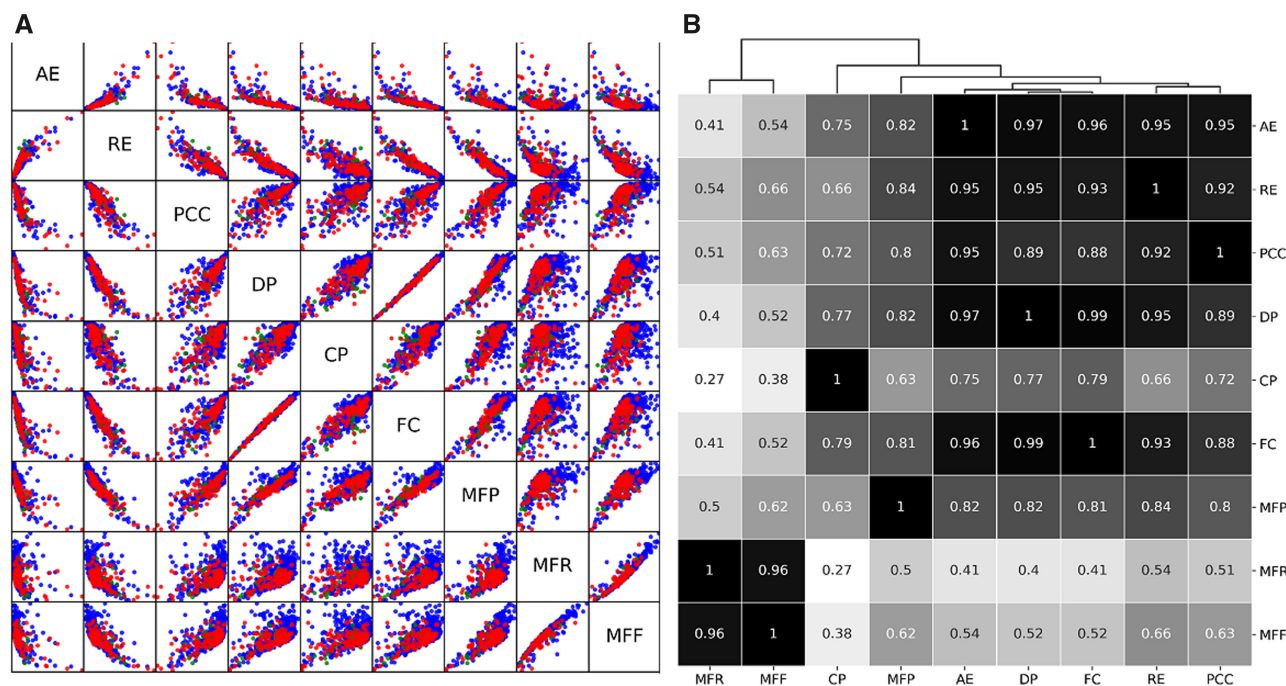


Fig. 2. Correlation between the prediction-oriented metrics. (A) Scatter matrix for nine distance evaluation metrics on D680 (blue dots), CASP13-31 (green dots) and CAMEO131 (red dots). (B) Spearman's rank correlation matrix (absolute value) for the metrics in (A), with hierarchical clustering result shown on the top. Note that, the correlation coefficients are calculated on the unified set of targets from all datasets

energy minimizations based on predicted distance and orientations were used to fold protein structure in trRosetta (Yang et al., 2020).

Accurate distance prediction is anticipated to lead to accurate structure models. There are many well-accepted metrics to measure the accuracy of predicted structure models, such as GDT-TS score (Zemla, 2003) and TM-score (Zhang and Skolnick, 2004). Thus, an intuitive way for assessing the objectiveness of the proposed distance evaluation metrics is to compare them with the metrics of structure model accuracy. TM-score, which is widely used by the community, is adopted here for model accuracy measurement. The intuition is that a good distance measure should correlate well with TM-score. To this end, we calculate the Pearson's correlation coefficient (PCC) between the proposed metrics and TM-score on the benchmark datasets, and compute the weighted sum of the Z-scores of the metrics on every dataset. The weight is 0.5, 0.2 and 0.3 for D680, CASP13-31 and CAMEO131, respectively, which was set empirically by considering the size and difficulty of each dataset.

The PCCs and the corresponding Z-scores on the three datasets are presented in Table 1 and Supplementary Tables S2 and S3, for the prediction-oriented, the native-oriented and the full-list case, respectively. The corresponding scatter plots are shown in Supplementary Figures S7–S9.

For the prediction-oriented case, several conclusions can be drawn from the data in Table 1. First, the two metrics, distance precision and fuzzy certainty, rank at the top on all datasets. Though both metrics have high correlation with TM-score, their definitions are quite different. The former is defined in a regression way, while the latter is a classification measure. Probability terms are used in both metrics, but the probabilities being used are different. More specifically, the distance precision uses the probability  $P(d_{ij} \leq 20\text{\AA})$  to represent the confidence of the predicted distance; while the fuzzy certainty only utilizes the probabilities of the native bin and its adjacent bins. In fact, according to previous analysis in Section 3.3, these two metrics are highly correlated with a SCC of 0.99. Thus, they can be used to reflect the quality of a protein structure model from different aspects, but with similar effect.

Second, the metrics that tolerate minor distance deviation seem to be more appropriate. For instance, among the top metrics, distance precision takes deviations within 2 Å as correctly predicted;

**Table 1.** Absolute values and summed Z-scores of the Pearson's correlation coefficient between the distance evaluation metrics and the TM-score (the prediction-oriented case)

	D680	CASP13-31	CAMEO131	Summed Z-score
DP	<b>0.783</b>	<b>0.838</b>	<b>0.709</b>	<b>0.94</b>
FC	0.774	0.835	0.707	0.897
MFP	0.733	0.761	0.698	0.635
RE	0.757	0.737	0.666	0.612
PCC	0.717	0.773	0.644	0.458
AE	0.595	0.786	0.558	−0.179
CP	0.624	0.613	0.567	−0.264
MFF	0.501	0.495	0.573	−0.837
MFR	0.348	0.334	0.301	−2.263

Note: The highest value in each column is highlighted in bold type. The metrics are ranked based on the summed Z-score.

the fuzzy certainty is a multi-classification measure which become higher when a prediction falls into the real or its adjacent bins. Other metrics that do not have such tolerance, such as Pearson's correlation coefficient and absolute error, tend to have lower correlations with TM-score.

Third, the contact precision (CP), a widely used measure, has a reasonable correlation with TM-score. Though only the top  $L$  pairs are considered in the calculation of CP, it has over 0.5 PCC with TM-score on all datasets. We note that the PCC is lower than that observed in trRosetta (Yang et al., 2020). This may be because that the orientation restraints were excluded in the modeling stage here. The correlation is improved when extending from contact to distance in a few metrics. For example, the PCC between the distance precision and TM-score increases to 0.783, 0.838 and 0.709 on the D680, CASP13-31 and CAMEO131, respectively. Thus, it becomes necessary to define distance-based metrics to objectively reflect the performance of inter-residue distance prediction methods.

The metrics defined in the native-oriented and the full-list cases seem to have a decreased correlation with TM-score compared with

prediction-oriented metrics. For example, the distance precision and the fuzzy certainty, which are the top two metrics in the prediction-oriented assessment, have lower correlation ( $<0.7$  for all three datasets) with TM-score in the native-orientated case. The reduced correlation with TM-score may be explained by the fact that some of the assessed residue pairs are not predicted well (i.e. with low probability), and do not have significant contribution to the structure folding. In contrast, for the prediction-oriented case, only confidently predicted residue pairs (i.e. the top 15L pairs with high probability) are considered in the assessment. Such residue pairs are anticipated to contribute more to the folding, thus a higher correlation was observed.

### 3.5 Accuracy estimation of predicted distance

In real-world application, such as protein structure prediction, experimental structures are not available for most proteins. Thus, it is desirable to estimate the accuracy of the predicted distance. For example, the average probability of the top-predicted contacts was used to estimate the precision of the predicted contacts in trRosetta. To estimate the accuracy of the predicted distance, two variables are defined here:  $P20$  and  $mP20$  by considering the distance below  $20 \text{ \AA}$ .

$$P20 = \frac{1}{|S|} \sum_{(i,j) \in S} \max_{1 \leq k \leq 9} P_k(i,j), \quad (11)$$

where  $S$  is the set of the top 15L residue pairs sorted by  $P(d_{ij} \leq 20 \text{ \AA})$ . To define  $mP20$ , the set  $S$  is divided into 9 subsets similarly as in Equation (7).

$$P20_k = \frac{1}{|M_k|} \sum_{(i,j) \in M_k} P_k(i,j), \quad k < 10 \quad (12)$$

$mP20$  is defined as the mean of  $P20_k$  over all non-empty subsets  $M_k$ , which consists of residue pairs with the maximum probability falls into the  $k$ th bin.

Supplementary Table S4 lists the Pearson's correlation coefficients (PCC) between the values of the above two variables and the distance evaluation metrics in predicted-orientated case on the D680 dataset. Fuzzy certainty and distance precision correlate well with  $P20$  and  $mP20$  ( $\text{PCC} \geq 0.9$ ). Since these two metrics have a high correlation with TM-score (see Table 1), we expect that they can be used to estimate the TM-score of the structure models generated

from the predicted distance. To study their relationship with TM-score, we draw the scatter plots on the D680 dataset in Figure 3. It shows that  $P20$  and  $mP20$  have high correlations with TM-score and the PCC is 0.71 and 0.72, respectively. Note that, a TM-score  $>0.5$  usually indicates that the model is in correct fold (Xu and Zhang, 2010). As illustrated in Figure 3, 0.3 seems to be an appropriate cutoff for both variables to differentiate whether a model can be predicted with correct fold or not. Indeed, 93% (91.2%) models have a TM-score  $>0.5$  when  $P20$  ( $mP20$ ) is  $>0.3$ .

### 3.6 Application to MSA selection

To improve the performance of distance prediction, multiple MSAs are often generated from different parameters and sequence databases, and then the optimal one is selected. As demonstrated in (Yang et al., 2020), the MSA selection contributes to an improved contact prediction by about 3% on the CASP13-31 dataset. trRosetta selects MSA based on the average probability of the top  $L$  predicted long+medium-range contacts (denoted by  $P_{con}$ ). In addition, the two distance-based variables  $P20$  and  $mP20$  are used to for MSA selection here.

To compare the performance in MSA selection, an average of 35 different MSAs were generated with different combinations of parameters. An independent CAMEO dataset was collected to conduct this experiment (161 targets between June 13, 2020 and September 5, 2020). The baseline is MSA generated by HHblits (Remmert et al., 2011) with an e-value 0.001 against the Uniclust30 database. The average contact precision and distance precision corresponding to different MSAs are summarized in Table 2. Compared with the baseline, the selected MSAs lead to  $>3\%$  increase in both contact precision and distance precision. Besides, for each target, we calculate the contact precision for each MSA and take the maximum as the upper bound. The average differences between the upper bound and the actual contact precision by the baseline and corresponding MSAs selected by  $P_{con}$ ,  $P20$  and  $mP20$  are 0.06, 0.026, 0.025 and 0.021, respectively. A head-to-head comparison between the upper bound and the actual contact precision by the three variables is shown in Supplementary Figure S10. These data suggest that the two newly defined variables can be utilized for MSA selection, which results in contact/distance predictions with higher precisions.

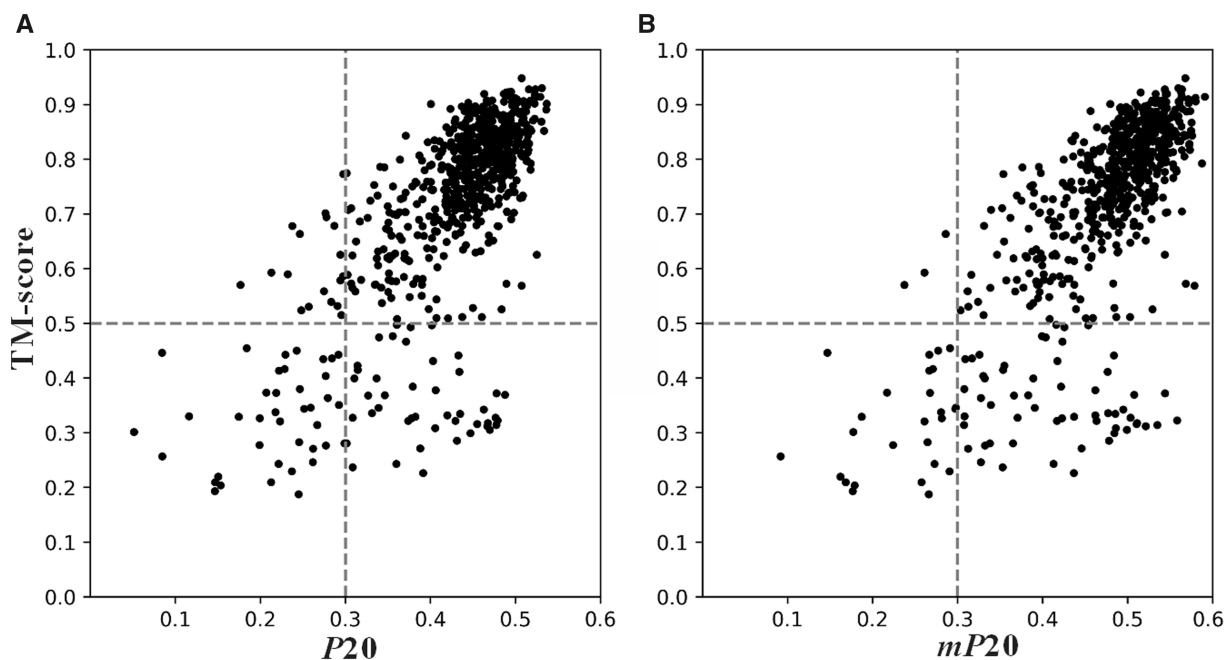


Fig. 3. Scatter plots between TM-score and the two variables  $P20$  and  $mP20$  on the D680 dataset

**Table 2.** Average contact precision (top  $L$ , separation  $\geq 12$ ) and distance precision (top 15 $L$ , separation  $\geq 12$ ) predicted from selected MSAs on an independent CAMEO dataset

	Baseline	<i>Pcon</i>	<i>P20</i>	<i>mP20</i>
Contact precision	0.727	0.761	0.762	<b>0.766</b>
Distance precision	0.627	0.66	0.661	<b>0.662</b>

Note: The highest value in each row is highlighted in bold type.

### 3.7 Application to the CASP14 distance prediction

As mentioned in Section 1, distance prediction was included in the recent CASP14 experiment. In the CASP14 meeting, the assessors introduced multiple metrics to evaluate the performance of each group, including five pair-based and four graph-based metrics. The final ranking of groups was based on the sum of Z-scores of five non-redundant metrics. The distance-based evaluation results are not listed on the CASP14 website. In addition, the definitions for some metrics (e.g. the graph-based metrics) have not been described in detail.

The metrics in this work are compared with the official metrics in terms of ranking groups. Following the procedure in CASP14, we calculate the summed Z-scores for each group on 37 FM and FM/TBM domains (38 in CASP14 but we do not have experimental structure for the target T1085-D2). The calculation of Z-score is the same as that described in the official result page. First, Z-scores are computed from the raw scores and predictions with Z-score  $< -2$  are removed as outliers. Then, new Z-scores are recalculated based on the reduced dataset; and those below 0 are assigned to 0. The CASP14 official ranking and the rankings by the summed Z-score of three representative metrics that have the highest correlation with TM-score in three different cases are listed in [Supplementary Table S5](#). The Spearman's correlation coefficients between our rankings and the official ranking are 0.977, 0.955 and 0.951, respectively. This suggests that the group rankings are largely consistent, though different sets of residue pairs and different metrics are used.

There are a few key differences between our metrics and the ones used in CASP14, though they result to similar group rankings. First, the CASP14 metrics only consider the full list while we define metrics in three cases, including prediction-orientated, native-orientated and full-list. Second, for the full-list metrics, the way of dealing with the last distance bin (i.e. distance  $> 20\text{\AA}$ ) is different. This bin was treated equally with others in the official assessment. In our opinion, as this bin covers a wider range of distance and contains less specific information than other bins, it should be treated in a different way. For multi-class metrics, the predicted label was taken as the bin with the highest probability in the official assessment. In contrast, we first compare the summed probability  $P(d_{ij} \leq 20\text{\AA})$  with  $P(d_{ij} > 20\text{\AA})$  to decide whether the prediction falls into the  $(0, 20\text{\AA}]$  interval. The predicted label may not be the last bin even if it has the highest probability (but smaller than 0.5). Third, the fluctuation of experimental structures is considered in our fuzzy metrics by assigning small rewards to the bins that are adjacent to the bin which the native distance belongs to. Finally, only long-range residue pairs (separation  $\geq 24$ ) are included in the official assessment, which are extended to the pairs with separation  $\geq 12$ .

In addition to the above technical differences, a few unique features exist in the proposed metrics. First, the metrics may find a broader application field as they are defined in three different cases (prediction-oriented, native-oriented and full-list). For example, when the native structure is unknown, the prediction-oriented metrics are especially useful, as already illustrated in Sections 3.5 and 3.6. Second, both regression-based and classification-based metrics are defined. These metrics are complementary to each other reflecting the accuracy of predicted distance from different aspects. Third, the objectiveness of the proposed metrics has been partly investigated by comparing with the quality of predicted structure model. This provides a guidance for selection of specific metrics and/or

defining more new metrics in future. Finally, open source codes are provided to enable easy implementation of the proposed metrics.

## 4 Conclusions

To guide further development of inter-residue distance prediction methods, we have performed a systematic study toward the assessment of predicted inter-residue distance. A total of 19 metrics were proposed to measure the accuracy of predicted distance. These metrics can be divided into three different cases based on the set of residue pairs assessed, i.e. *prediction-oriented* (9), *native-oriented* (6) and *full-list* (4). More specifically, the prediction-oriented metrics concentrates on the accuracy of the top predictions; the native-oriented case focuses on the performance for residue pairs with native distance  $\leq 20\text{\AA}$ ; the full-list metrics consider all residue pairs with separation  $\geq 12$ . In addition, based on the way of assessment, these metrics can be grouped into two categories: 6 regression-based and 13 classification-based.

The proposed metrics were discussed and compared on three benchmark datasets, with distance and structure models predicted by the trRosetta pipeline. The experiments show that a few metrics correlate well with the model accuracy measure TM-score. For example, a high Pearson's correlation coefficient ( $>0.7$ , on each of the three datasets) is observed between TM-score and the prediction-oriented regression measure, *distance precision*. Besides, we defined two variables *P20* and *mP20* to estimate the accuracy of the predicted distance. It turns out that these two variables have a high correlation with distance precision (PCC  $\geq 0.9$ ). In addition, the proposed metrics are used to rank the distance prediction groups in CASP14. The group ranking by our metrics coincides largely with the official ranking. These data suggest that the proposed metrics are objective for measuring distance prediction.

## Funding

The work was supported by the National Natural Science Foundation of China [NSFC 11871290 and 61873185].

*Conflict of Interest:* none declared.

## References

- Adhikari,B. (2020) A fully open-source framework for deep learning protein real-valued distances. *Sci. Rep.*, **10**, 13374.
- Baek,M. et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
- Ding,W. and Gong,H. (2020) Predicting the real-valued inter-residue distances for proteins. *Adv. Sci.*, **7**, 2001314.
- Greener,J.G. et al. (2019) Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.*, **10**, 3977.
- Haas,J. et al. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins*, **86**, 387–398.
- Hou,J. et al. (2019) Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins*, **87**, 1165–1178.
- Ji,S. et al. (2019) DeepCDpred: inter-residue distance and contact prediction for improved prediction of protein structure. *PLoS One*, **14**, e0205214.
- Jumper,J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
- Kryshtafovych,A. et al. (2019) Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins*, **87**, 1011–1020.
- Kucic,P. et al. (2014) Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. *BMC Bioinformatics*, **15**, 6.
- Mariani,V. et al. (2013) IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, **29**, 2722–2728.
- Pearce,R. and Zhang,Y. (2021) Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol.*, **68**, 194–207.
- Remmert,M. et al. (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

- Senior, A.W. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.
- Wu, T. *et al.* (2021) DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics*, **22**, 30.
- Wu, T. *et al.* (2020) Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics*, **36**, 1091–1098.
- Wuyun, Q. *et al.* (2018) A large-scale comparative assessment of methods for residue-residue contact prediction. *Brief. Bioinf.*, **19**, 219–230.
- Xu, J. (2019) Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. USA*, **116**, 16856–16865.
- Xu, J. and Wang, S. (2019) Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins*, **87**, 1069–1081.
- Xu, J. and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889–895.
- Yang, J. *et al.* (2020) Improved protein structure prediction using predicted inter-residue orientations. *Proc. Natl. Acad. Sci. USA*, **117**, 1496–1503.
- Zemla, A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
- Zhang, Y. and Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality. *Proteins*, **57**, 702–710.