ORIGINAL ARTICLE
Journal Section

When Homologous Sequences Meet Structural Decoys: Accurate Contact Prediction by tFold in CASP14 - (tFold for CASP14 Contact Prediction)

Tao Shen*, Jiaxiang Wu*, Haidong Lan, Liangzhen Zheng, Jianguo Pei, Sheng Wang, Wei Liu, and Junzhou Huang[1]

[1]Tencent AI Lab

Correspondence
Junzhou Huang, Tencent AI Lab, Shenzhen 518057, China
Email: joehhuang@tencent.com

Funding information

In this paper, we report our *tFold* framework's performance on the inter-residue contact prediction task in the 14th Critical Assessment of protein Structure Prediction (CASP14). Our *tFold* framework seamlessly combines both homologous sequences and structural decoys under an ultra-deep network architecture. Squeeze-excitation and axial attention mechanisms are employed to effectively capture inter-residue interactions. In CASP14, our best predictor achieves 41.78% in the averaged top-L precision for long-range contacts for all the 22 free-modeling (FM) targets, and ranked 1st among all the 60 participating teams. The tFold web server is now freely available at: `https://drug.ai.tencent.com/console/en/tfold`.

KEYWORDS
CASP14, contact prediction, protein folding, deep convolutional residual neural network

1 | INTRODUCTION

Accurate inter-residue distance/contact prediction has become the focal research point in the realm of *de novo* protein tertiary structure prediction since its significant success in CASP13 [1]. Emerging predictors such as AlphaFold [2], CONFOLD2 [3], DMPfold [4] and trRosetta [5] put great emphasize on distance prediction accuracy, which instantly promotes precision of predicted 3D structures.

Early attempts for inter-residue contact prediction involve manually-designed co-evolution features from homologous sequences, *e.g.* GREMLIN [6], PSICOV [7], and CCMpred [8]. Specifically, GREMLIN [6] adopts a Markov random field model to capture the conservation and co-evolution in a multiple sequence alignment, while CCMpred [8] provides a highly-optimized pseudo-likelihood maximization (PLM) implementation for contact prediction. Inspired by deep network's overwhelming success in various domains, most of recent methods adopt convolutional neural networks (CNNs) for end-to-end learning from multiple-sequence alignment (MSA) to inter-residue contact/distance prediction [5, 9, 10, 11, 12]. In addition to MSA based features, RaptorX [13] first described how to integrate template information with deep learning to improve contact/distance prediction. The prediction task is formulated as a pixel-wise regression/classification problem, where residual connection [14] and dilated convolution [15] are often utilized to increase the model capacity for distance prediction. However, there are still much left to explore on designing more effective network architectures for inter-residue distance prediction.

In this work, we propose a novel framework, namely *tFold*, to seamlessly integrate diverse information from homologous sequences and structural decoys for inter-residue distance prediction. Specifically, we collect structural decoys via template search and decoy generation with various protein folding methods. Such structural decoys reveal possible inter-residue interactions directly in the 3D conformation space, which is complementary to the co-evolution information in homologous sequences. The *tFold* framework consists of two sub-networks, *tFold-DistNet* for MSA-based distance prediction and *tFold-RefineNet* for refining distance predictions with structural decoys. We construct

*Equally contributing authors.

these networks with squeeze-and-excitation residual units [16] as basic building blocks, and axial attention layers [17] to better capture the global information for long-range interactions. The squeeze-and-excitation module allows more efficient communication among all the feature map channels, which is critical in constructing an ultra-deep network. The axial attention mechanism, allowing each residue to adaptively attend to all the other residues, circumvents the local receptive field limitation of standard convolution layers. The network is trained under a progressive learning manner for efficient optimization. Furthermore, we present a "cluster-and-rank" approach to fuse inter-residue distance predictions derived from different MSA features, automatically filtering out low-quality ones for more effective ensemble. With a pre-trained fold classification network, all the candidate inter-residue distance predictions are firstly clustered into groups based on the pairwise similarity, and then ranked by a probability distribution based quality assessment metrics to determine the optimal one.

We summarize our main contributions in three folds:

- We propose a unified framework to combine both homologous sequences and structural decoys for inter-residue distance prediction.
- We design an ultra-deep network architecture with squeeze-and-excitation and axial attention mechanisms, and present a progressive learning method for efficient training.
- We present a "cluster-and-rank" strategy for adaptive fusion of inter-residue distance predictions from various sequence databases and sensitivity thresholds.

We entered the contact prediction track in CASP14[1] with six variants of our *tFold* framework (*tFold*, *tFold-IDT*, *tFold-CaT*, *tFold_human*, *tFold-IDT_human*, and *tFold-CaT_human*). Among them, *tFold-CaT_human* ranked 1st among 60 teams, with the averaged top-L long-range contact prediction precision of 0.41783 for 22 free-modeling (FM) domains. In this paper, we present technical details and extensive evaluation results of our *tFold* framework, and discuss what went well/wrong in CASP14.

2 | MATERIALS AND METHODS

2.1 | The tFold Distance Prediction Pipeline

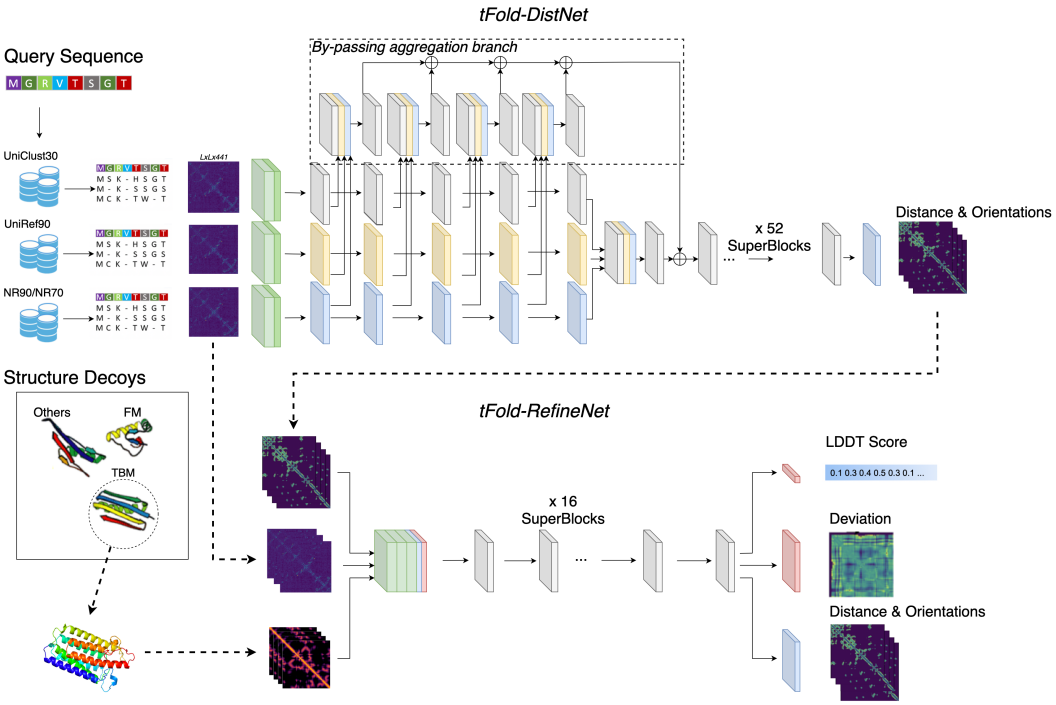

FIGURE 1 The overall architecture of tFold distance prediction pipeline.

We illustrate the overall architecture of our proposed *tFold* framework in Figure 1. The *tFold* framework consists of two sub-networks, where *tFold-DistNet* produces initial inter-residue distance and orientation predictions from MSA features, and *tFold-RefineNet* further refine such predictions assisted by structural decoys.

Specifically, in the first phase, *tFold-DistNet* accepts a set of MSAs searched against various sequence databases, and transforms them into following features: PSSM (Position Specific Scoring Matrix) of size $L \times 21$ and MRF (Markov Random Fields) [6] of size $L \times L \times 441$, where L is the input sequence length. The model produces four different histograms, one for distance and three for orientations. These histograms

[1]CASP14 contact prediction results: `https://predictioncenter.org/casp14/zscores_rrc.cgi`

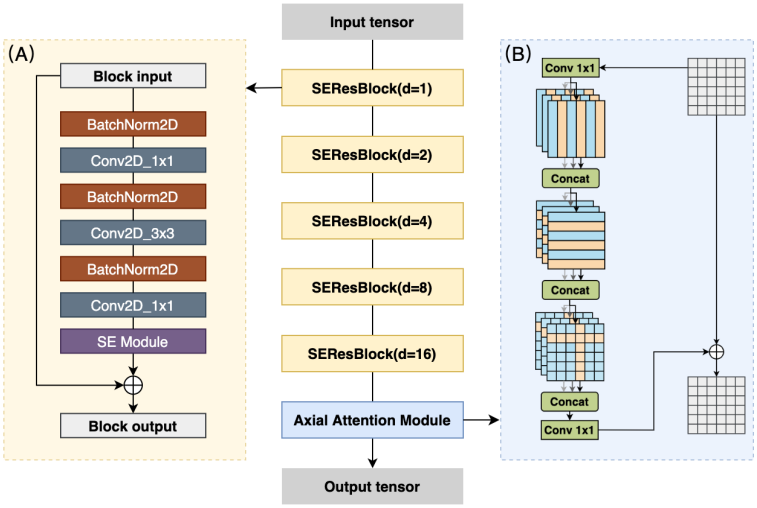

FIGURE 2 The *SuperBlock*'s architecture, which consists of five *SEResBlock* units with gradually increased dilation rates (in 3×3 convolutional layers) and one axial attention module.

.

are identical with those in trRosetta [5]:

- d_{ij}: distance defined by C_β^i-C_β^j atoms
- ω_{ij}: dihedral angle defined by C_α^i-C_β^i-C_β^j-C_α^j atoms
- θ_{ij}: dihedral angle defined by N^i-C_α^i-C_β^i-C_β^j atoms
- φ_{ij}: plane angle defined by C_α^i-C_β^i-C_β^j atoms

where N^i, C_α^i, and C_β^i denote the i-th residue's N, C_α, and C_β atoms, respectively. trRosetta[5] pointed it out that the introduction of three types of orientations are beneficial to the model training, as they provide auxiliary supervision from complementary aspects. During the second phase, *tFold-RefineNet* takes MSA features and structural decoys as additional inputs to further refine *tFold-DistNet*'s predictions. Similarly, we cooperate auxiliary supervision, *i.e.* per-residue lDDT scores [18] and per-decoy C_β^i-C_β^j distance deviations, to better guide the refinement of inter-residue distance and orientation predictions.

We highlight the *tFold* framework in two aspects: 1) *tFold-DistNet* and *tFold-RefineNet* share the same network building block, only distinguished by input features, output predictions, and training/inference protocols; and 2) *tFold-RefineNet* accepts structure decoys from a wide variety of sources, including partial structural templates from PDB and 3D structural decoys generated from protein folding algorithms. *tFold-RefineNet* is robust with some low-quality decoys (lDDT < 0.4), *i.e.* its refinement results are boosted even if input decoys contain misleading inter-residue interactions.

2.2 | Network Architecture

Both *tFold-DistNet* and *tFold-RefineNet* are constructed with the same network building block, namely *SuperBlock*, which consists of multiple *SEResBlock* units followed by one axial attention module. We visualize the *SuperBlock*'s architecture in Figure 2.

Each *SEResBlock* [16] involves three convolutional layers forming a bottleneck structure for computational efficiency and one squeeze-and-excitation module to capture inter-channel dependencies. Dilated convolution is adopted to enlarge the receptive field under the same computational complexity, which better captures long-range inter-residue interactions. Nevertheless, dilated convolution only allows square-shaped receptive fields, which may not be optimal for modeling one residue's interaction with all the other ones. Therefore, we employ axial attention [19, 20, 17] at the end of each *SuperBlock*. Each axial attention module consists of two orthogonal axial attention layers and one criss-cross attention layer, which is capable of formulating non-local dependencies for each residue.

The main branch of *tFold-DistNet* model is composed of 52 *SuperBlock*s, as we empirically find that such depth is indeed beneficial for fully exploitation of MSA features. We discover that different sequence databases, *e.g.* UniClust30 [21], UniRef90 [22], and NR90/70 [23], lead to highly diverse MSA features. Therefore, for each sequence database, we adopt 5 additional *SuperBlock*s to pre-process its corresponding PSSM and MRF features before concatenating all the resulting feature maps together, and add a by-passing aggregation branch for multi-layer fusion, as depicted in Figure 1.

We first developed *tFold-DistNet* and designed *SuperBlock* for distance prediction. After *tFold-DistNet* was completed, the *SuperBlock* was used as the building block of *tFold-RefineNet*. For computational efficiency, multi branch fusion was not used in *tFold-RefineNet* and it is a relatively shallow network (16 *SuperBlock*s). Multiple structural decoys are fed into *tFold-RefineNet*, after extracting their inter-residue

distance and orientation as input features. *tFold-RefineNet* is expected to refine *tFold-DistNet*'s distance and orientation predictions based on additional structural information embedded in these decoys. Furthermore, we use per-decoy local lDDT scores and distance deviations as auxiliary supervision to aid the model training. The distance deviation is defined as the difference in C_β^i-C_β^j distance, uniformly discretized into 20 bins (-10Å to 10Å). Thus, *tFold-RefineNet* is aware of the quality of each input decoy, and therefore is more robust with bad decoys.

2.3 | Training Protocol

It often takes huge amount of computational resources and careful engineering to train deep learning models, especially for ultra-deep ones. To overcome such difficulty, we adopt the following progressive training protocol:

1. Build an initial shallow and train it until convergence.
2. Extend the model with more *SuperBlock*s with identity mapping initialization [24].
3. Train the extended model with a smaller learning rate until convergence.
4. Repeat step 2 & 3 until model accuracy no longer improves.

The above protocol starts by training an initial model, and progressively improves the accuracy by adding more *SuperBlock*s. The extended model is relatively faster to converge, as only newly added *SuperBlock* are optimized. Compared with traditional approaches that train the entire model from scratch, our approach is more efficient and easier to train such ultra-deep models.

We trained all the models on a single Nvidia DGX-2 workstation equipped with 16 V100 GPUs. The initial *tFold-DistNet* consists of 16 *SuperBlock*s and is gradually extended to 52 *SuperBlock*s (16-24-32-40-48-52). The initial model is trained for 40 epochs with Adam optimizer (learning rate: 1e-4), and fine-tuned for 10 epochs (learning rate: 5e-5) whenever extra *SuperBlock*s are added. Similarly, the *tFold-RefineNet* model starts with 8 *SuperBlock*s and stops at 16 *SuperBlock*s (8-12-16). The initial model is trained for 20 epochs with Adam (learning rate: 1e-4), and further fine-tuned for 5 epochs (learning rate: 5e-5) whenever the network depth grows. The difference in *tFold-DistNet*'s and *tFold-RefineNet*'s network depth and number of training epochs are mainly due to their respective task difficulty and data availability. The networks were always trained on 128 × 128 regions of the distance matrix, which reduces the GPU memory consumption to allow deeper networks to be trained

For CASP14, the training subset is created from PDB40-20200301 database, containing 36,505 structures. We randomly select 1,000 structures from it to constitute the test subset, in addition to 31 CASP13 FM structures (domain-level) and 141 CAMEO structures (chain-level). We collect structural decoys for *tFold-RefineNet* from: 1) CNFpred's [25] top-10 decoys for the CATH database (version: 20180316); and 2) all the servers' submissions for 879 CAMEO [26] targets during 20190511-20200502. As different CAMEO servers produce highly diverse structural decoys for the same target, the *tFold-RefineNet* model is more robust with input decoys' quality.

2.4 | Inference Protocol

The inference process of our *tFold* framework mainly consists of following procedures:

1. Search through sequence databases to generate MSAs.
2. Predict inter-residue distance and orientations with *tFold-DistNet*.
3. Fuse distance and orientation predictions for further refinement.
4. Generate/collect structural decoys for refinement with *tFold-RefineNet*.
5. Refine distance and orientation predictions with *tFold-RefineNet*.
6. Fuse distance and orientation predictions for final submissions.

For each query sequence, we firstly search it against UniRef30 (2020-03), UniRef90 (2020-04), and NCBI-nr (2020-04) databases with HHSuite [27], HMMER [28], and BLAST [29], respectively. The resulting MSA may be further used to search against the MetaClust50 (2018-06) database with HHMER if its NEFF (**N**ormalized number of **EFF**ective sequences) index is lower than the threshold. By varying MSA search hyper-parameters (E-value and number of iterations) for different databases, we obtain a total of 12 MSAs or 83 MSAs. *tFold-DistNet* is employed to predict inter-residue distance and orientations for each MSA, following by a "cluster-and-rank" strategy for multi-MSA fusion (as detailed in Appendix A). Top-ranked predictions are then further refined by *tFold-RefineNet*, using structural decoys generated by our in-house protein folding pipeline and other servers' submissions (only available for human groups). The refined predictions are again fused with the "cluster-and-rank" strategy to obtain final submissions. We registered 6 contact prediction servers for CASP14, as listed in 1, with different MSA search strategies and structural decoys.

TABLE 1 Inference protocols of tFold predictors in CASP14.

Predictor	Model	Input MSAs	Input Decoys
tFold	*tFold-DistNet*	12 MSA	None
tFold-IDT	*tFold-DistNet + tFold-RefineNet*	12 MSA	tFold server decoys (TBM)
tFold-CaT	*tFold-DistNet*	83 MSA	None
tFold_human	*tFold-DistNet + tFold-RefineNet*	12 MSA	tFold server decoys (TBM + FM)
tFold-IDT_human	*tFold-DistNet + tFold-RefineNet*	12 MSA	All server decoys
tFold-CaT_human	*tFold-DistNet + tFold-RefineNet*	83 MSA	All server decoys

TABLE 2 tFold predictors' top-L/k precision for contact prediction on CASP14 FM targets.

	Short Range			Medium + Long Range			Long Range		
Method	L	L/2	L/5	L	L/2	L/5	L	L/2	L/5
tFold	24.56	39.84	61.79	43.41	55.87	66.99	35.13	47.74	57.96
tFold-IDT	25.02	41.05	64.15	45.63	58.33	68.08	36.52	49.11	60.93
tFold-CaT	24.90	41.50	65.23	46.23	59.36	69.49	37.11	50.35	61.22
tFold_human	25.40	42.31	64.73	47.57	59.68	70.49	38.06	51.21	61.66
tFold-IDT_human	**25.86**	**42.89**	66.29	49.93	61.81	73.03	40.50	52.87	64.75
tFold-CaT_human	25.77	42.84	**66.30**	**50.81**	**63.09**	**73.89**	**41.78**	**55.64**	**66.55**

3 | RESULTS

3.1 | Contact Prediction Accuracy in CASP14

CASP14 involves a total of 68 target proteins, officially divided into 107 structural domains. In CASP14's contact prediction track, 60 participanting teams (30 server groups and 30 human groups) submitted predictions for 15 TBM/FM and 22 FM domains, which were then evaluated over various metrics.

Here, we follow the widely-used evaluation protocol for quantitative comparison of various contact prediction methods. Two residues are "in-contact" if the distance between their C_β atoms is lower than 8Å. Based on the sequence separation, contacts are further categorized into short-range (6-11), medium-range (12-23), and long-range (over 24) ones. Previous studies [30, 31] have demonstrated that long-range contacts are most informative and critical for subsequent protein structure optimization procedures. In Table 2, we report the short/medium+long/long-range contact prediction accuracy of all the our *tFold* predictors on 22 FM domains, measured by the top-L/k precision.

As the baseline method, *tFold* only employs *tFold-DistNet* with 12 MSAs and no structural decoys, and achieves the top-L long-range precision of 35.13%. *tFold-IDT* is intended to integrate the template information into the MSA-based contact prediction process. Although only a limited number of templates can be found for FM domains, *tFold-IDT* still consistently improves the top-L long-range precision to 36.52%. On the other aspect, *tFold-CaT*'s 37.11% top-L long-range precision indicates that using more diverse MSA features (83 MSAs *vs.* 12 MSAs) is also critical in improving the contact prediction accuracy.

For three human group predictors, *tFold_human* improves over *tFold-IDT* by further cooperating FM-based folding results by the tFold server, although no human-group-only information is used in this predictor. The improvement of *tFold-IDT_human* over *tFold_human* (40.50% *vs.* 38.06%) further indicates that involving more structural decoys, regardless of their quality, is beneficial for improving the contact prediction accuracy. Our best performing predictor, *tFold-CaT_human*, integrates the above two effective strategies (more diverse MSAs and structural decoys). It achieves the top-L long-range precision of 41.78%, and ranks 1st among all the 60 predictors in CASP14.

It is worthy noting that many other top-ranked predictors in contact prediction and 3D structure prediction have utilized the BFD database [32] for MSA search, while we did not use during CASP14. Please refer to Section 3.3 for more detailed analysis on how the BFD database affects the contact prediction accuracy.

3.2 | Ablation Study on *tFold-DistNet*

We further conduct the ablation study on several variants of the *tFold-DistNet* model, to validate each component's effectiveness in the contact prediction accuracy. The architectural details of these models are as follows:

Baseline: Baseline model is a similar network architecture with trRosetta[5]. The model consists of 80 *SEResBlock*s with 128 channels,

cycling through dilations 1, 2, 4, 8, 16. The baseline model takes MRF features from single MSA as input. During inference time, predictions are perfromed using single MSA.

Baseline w/ multi branch: We add multi branch fusion module to the baseline model. For MSAs from multipule databases, MRF features are fed into the network at the same time.

Baseline w/ axial attention: We add axial attention module to the baseline model. After every cycle of dilation rate, an axial attention module is added after 2D convolution.

***tFold-DistNet* 16 blk**: We add both multi branch fusion module and axial attention module to the baseline model. It is the lightweight version of *tFold-DistNet* with 16 *SuperBlocks*.

***tFold-DistNet* 52 blk**: The full ultra deep version of *tFold-DistNet* with 52 *SuperBlocks* trained using the progressive training protocol described in Section 2.3.

In Figure 3, we report the top-L and top-L/5 precision for medium- and long-range contacts on CASP14 FM targets. We start with the baseline model described above, which has the same number of convolution layers as 16 *SuperBlocks*. By introducing the multi-branch architecture to fuse MSA features generated by different sequence databases, the top-L precision is improved from 32.42% to 34.44%. The axial attention module is able to effectively capture long range relations, which is very important for inter-residue geometry predictions. Adding axial attention after 2D convolution gave a clear improvement in contact accuracy (36.42% *vs.* 32.42%). By combining these two components together, "*tFold-DistNet* (16-blk)" achieves the top-L precision of 39.34%, which is significantly better than the baseline. Such improvement can be further enlarged by using more *SuperBlocks* (16 → 52), which leads to a boost in the top-L precision from 39.34% to 43.41%. This indicates that deeper networks are indeed more powerful and expressive for the inter-residue distance prediction problem.

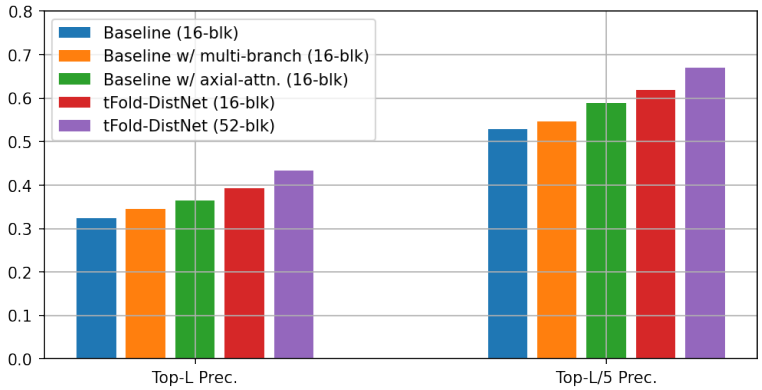

FIGURE 3 Comparison on the top-L/k precison for medium+long-range contacts on CASP14 FM targets.

3.3 | MSA Quality's Impact on Contact Predictions

As the fundamental input features for inter-residue contact/distance predictions, multiple sequence alignment's quality directly affects the prediction accuracy. Particularly, top-ranked structure prediction teams in CASP14, including AlphaFold2 [33], BAKER [34], and ZHANG [35], used large-scale metagenomics databases for MSA search, such as BFD [32], MGnify [36], and JGI [23]. However, during CASP14, the *tFold* framework only used MetaClust50 [37], which is a relatively small metagenomics database, in conjunction with UniClust30 [21], UniRef90 [22], and NR90/NR70 [23]. In the section, we replay our *tFold* framework with the BFD database enabled, to quantitatively measure how the MSA quality affects the contact prediction accuracy.

In Figure 4, we present the per-target top-L precision comparison for the *tFold* framework, with or without the BFD database. Although *tFold* is the simplest predictor as listed in Table 1, using BFD-generated MSA features leads to a huge boost in the contact prediction accuracy (44.77% *vs.* 35.13%), even outperforms the best predictors (*tFold-CaT_human*) in CASP14 by 2.99%. Further improvement can be observed by cooperating more structural decoys where *tFold-Cat_human-BFD* scores 51.40%, 9.62% higher than tFold-CaT_human. For targets with few homologous sequences (T1031-D1, T1039-D1, T1042-D1, T1074-D1, and T1096-D2), BFD dataset can help to find more sequences and distance prediction is significantly improved. However, for targets T1090-D1 and T1093-D3, high quality MSA can be aquired without BFD, adding metegenomic databases for MSA construction can even negatively impact precision.

3.4 | Decoy Quality's Impact on Contact Predictions

The *tFold-RefineNet* model takes as inputs structural decoys of unknown quality. A natural question arises: how the structural decoy's quality affects the contact prediction accuracy? In this section, we analyze the correlation between input structural decoy's quality and relative improvement in the contact prediction accuracy.

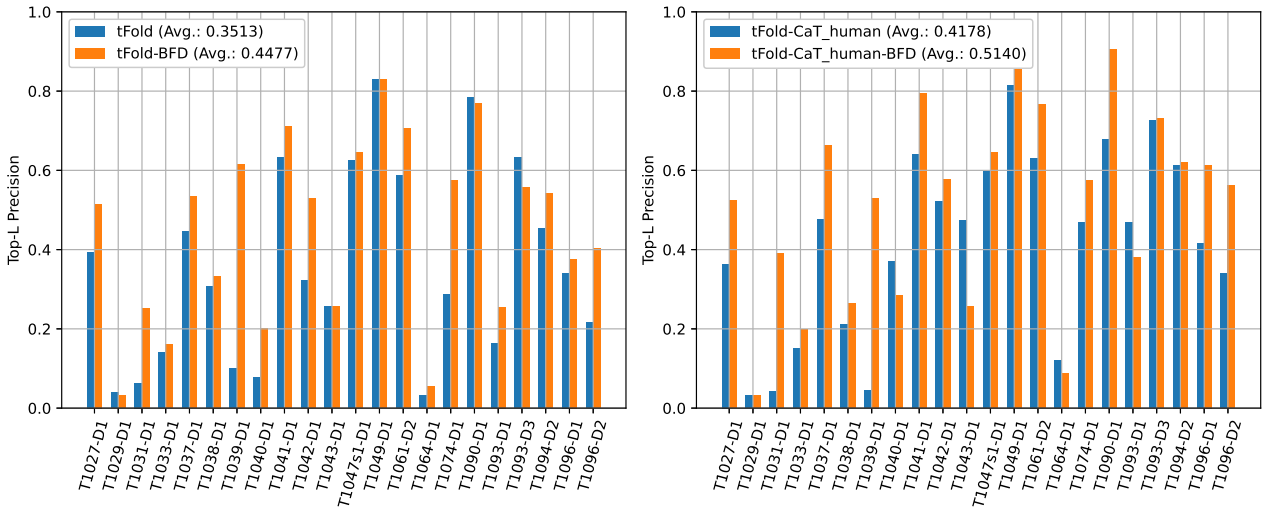

FIGURE 4 Comparison on the top-L precision for long-range contacts on CASP14 FM targets, with or without the BFD database being used.

For 20 chain-level FM targets in CASP14, we collect a total of 1,339 decoys from server submissions, whose lDDT scores range from 0.190 to 0.647. In Figure 5, we visualize the relative improvement in the top-L precision for medium- and long-range contacts for each structural decoy. Most of the dots are above the dashed gray line, indicating that refining contact predictions with the additional structural decoy is indeed beneficial. As is shown in Figure 5, *tFold-RefineNet* is robust with some low-quality input decoys (lDDT < 0.4). However, there are still some cases where the model failed to refine distance prediction. Performance of the model can be improved by discarding low quality decoys. In this case, an accurate accuracy estimation model is needed to filter the input decoys.

Additionally, we report the per-target Pearson correlation coefficient in Table S1. The overall Pearson correlation coefficient is 0.6718 (averaged across all targets), suggesting that better structural decoys leads to larger boost in the contact prediction accuracy. For FM targets (in the left column), we observe strong positive correlation for most targets (Pearson correlation coefficient over 0.5) with only a few exceptions (T1029, T1047s1, T1049, and T1090). For non-FM targets (in the right column), the performances are more stable (all non-FM targets have Pearson correlation coefficients over 0.4) because of higher quality MSAs.

Take T1029 as an example, its top-L precision shows negligible improvements when structural decoys are used. The main reason is that for T1029, we could not find sufficient homologous sequences during MSA search, while our *tFold* framework still heavily depends on MSA features.

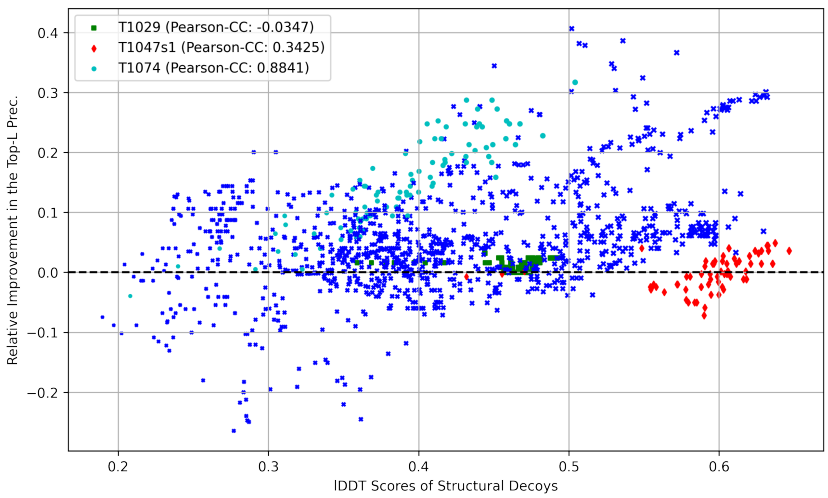

FIGURE 5 Visualization of decoy quality's impact on the top-L precision for medium- and long-range contacts.

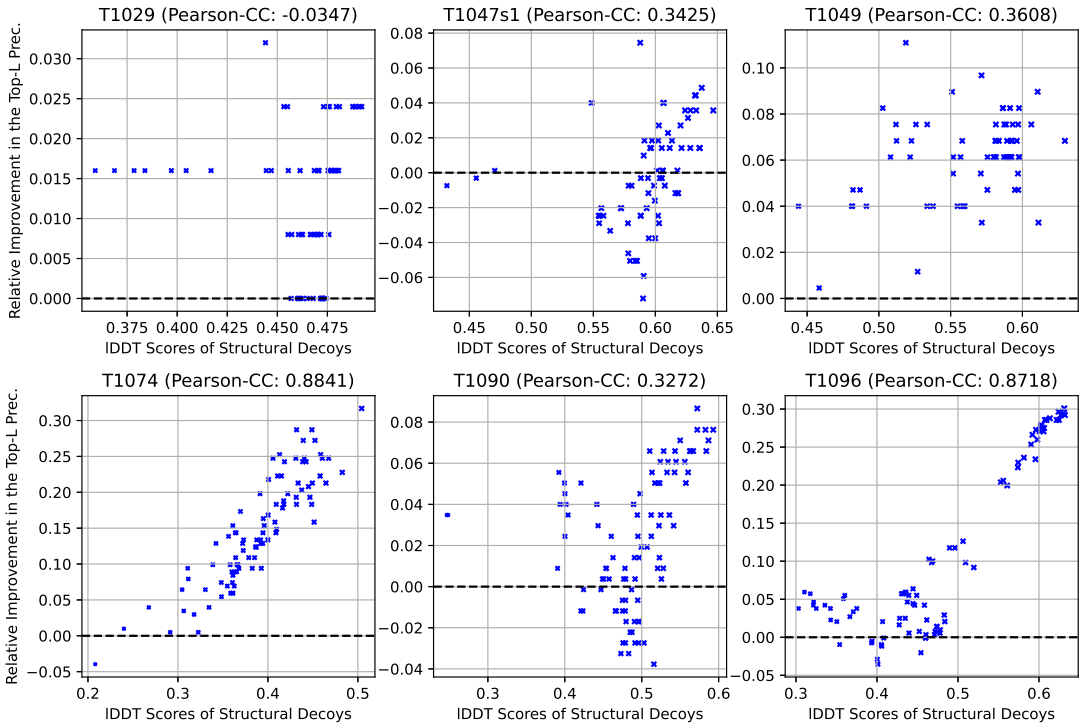

FIGURE 6 Per-target visualization of decoy quality's impact on the top-L precision for medium- and long-range contacts.

4 | DISCUSSION AND CONCLUSION

4.1 | Case Study: T1043

T1043, which is a part of H1044 (2,180 residues), is an extremely hard target in CASP14. No homologous sequences can be found with the default MSA search protocol in *tFold*. Therefore, we used the full sequence of H1044 for MSA search and then cropped out the corresponding 148 residues for T1043. As depicted in Figure S2, the resulting MSA's NEFF index improves significantly, among which the highest NEFF achieves 23.016. We feed these enhanced MSAs into our 6 *tFold* predictors, and the best-performing one, *tFold-CaT_human*, achieves 47.30% top-L precision for long-range contacts. At the same time, the best submission from other groups only reaches 9.4% top-L precision, which is notably lower than all of *tFold* predictors. However, when we apply the similar strategy on T1039 and T1040, as they are also a sub-sequence of H1044, no similar phenomenon is observed. This indicates that MSA search with an extended sequence could be sometimes effective in improving the MSA quality, but the inconsistent improvement for different targets may need further investigation. Lastly, we would like to note that due to lack of both homologous templates and high-quality structural decoys, the refinement with *tFold-RefineNet* did not work for T1043.

4.2 | Case Study: T1074

As mentioned in Section 3.3, the contact prediction accuracy of T1074 can be dramatically boosted once the BFD database is used. Since we failed to use the BFD database during CASP14, while most of top-ranked contact/structure prediction teams did, our best server-group submission for this target only achieves 35.61% (*tFold-CaT*), ranked 7th out of 60 teams. Meanwhile, *Yang-Server* and *TripletRes* teams scored 41.67% and 38.60% top-L precision for long-range contacts, respectively. Therefore, it is quite likely that other servers' structure prediction submissions reveal some BFD-exclusive inter-residue interactions that are not observed in our server-group submissions.

With this regard, *tFold-RefineNet* played a critical role in improving the contact prediction accuracy, as structural decoys submitted by other teams could be cooperated. We used 95 decoys from server submissions, including aforementioned teams' structure prediction counterparts. From Figure 5, we observe that the contact prediction accuracy can be improved for most of structural decoys. On the other hand, as visualized in Figure 7, *tFold-CaT_human* correctly suppressed false-positive contacts in the red bounding-box, and promoted false-negative contacts in the yellow bounding-box. As a result, *tFold-CaT_human* (46.97%) not only outperformed our MSA-only predictor (*tFold-CaT*: 35.61%), but also surpassed the best server-group submission (*Yang-Server*: 41.67%). This indicates that our *tFold-RefineNet* model

is indeed capable of improving initial contact predictions with additional information embedded in structure decoys.

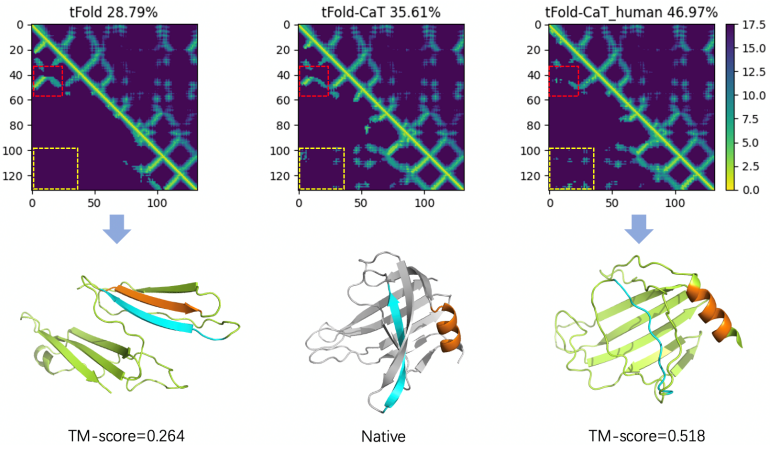

FIGURE 7 T1074-D1: *tFold*, *tFold-CaT*, and *tFold-CaT_human*'s contact predictions and folding results.

4.3 | Conclusion

In this paper, we present our *tFold* framework, which ranked 1st in the contact prediction track in CASP14. Both homologous sequences and structural decoys are exploited under the same ultra-deep network, which can be efficiently trained in a progressive learning manner. We adopt squeeze-excitation and axial attention mechanisms to better formulate inter-residue interactions, and propose a "cluster-and-rank" strategy for effective multi-MSA fusion.

The most important lesson we learnt from CASP14 is that the quality of multiple sequence alignment is critical to contact prediction. Modern metagenomics databases, *e.g.* BFD [32], are almost indispensable for those targets with limited number of homologous sequences. We failed to utilize the BFD database during the CASP14 competition, but later experimental results indicate that this could lead to 9-10% improvement in the top-L precision. On the other hand, this also raises a important question to future researches: how to alleviate the dependency on multiple sequence alignment for contact / structure prediction methods.

In summary, the main features of our proposed tFold framework include: 1) ultra-deep networks efficiently optimized via progressive training; 2) integration of homologous sequences and structural decoys with *tFold-RefineNet*; and 3) adaptive fusion of various candidate inter-residue distance predictions. The joint utilization of above three components leads to the SOTA performance in the CASP14 contact prediction.

references

[1] Cheng J, Choe MH, Elofsson A, Han KS, Hou J, Maghrabi AH, et al. Estimation of model accuracy in CASP13. Proteins: Structure, Function, and Bioinformatics 2019;87(12):1361–1377.

[2] AlQuraishi M. AlphaFold at CASP13. Bioinformatics 2019;35(22):4862–4865.

[3] Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC bioinformatics 2018;19(1):1–5.

[4] Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nature communications 2019;10(1):1–13.

[5] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences 2020;117(3):1496–1503.

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

[7] Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28(2):184–190.

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

[9] Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology 2017;13(1):e1005324.

[10] Zhu J, Wang S, Bu D, Xu J. Protein threading using residue co-variation and deep learning. Bioinformatics 2018;34(13):i263–i273.

[11] Xu J. Distance-based protein folding powered by deep learning. Proceedings of the National Academy of Sciences 2019;116(34):16856–16865.

[12] Li Y, Hu J, Zhang C, Yu DJ, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics 2019;35(22):4647–4655.

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

[14] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

[15] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: International conference on learning representations; 2016. .

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

[17] Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen LC. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision Springer; 2020. p. 108–126.

[18] Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29(21):2722–2728.

[19] Ho J, Kalchbrenner N, Weissenborn D, Salimans T. Axial attention in multidimensional transformers. arXiv preprint arXiv:191212180 2019;.

[20] Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 603–612.

[21] Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic acids research 2017;45(D1):D170–D176.

[22] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926–932.

[23] Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, et al. The genome portal of the department of energy joint genome institute. Nucleic acids research 2012;40(D1):D26–D32.

[24] Bachlechner T, Majumder BP, Mao HH, Cottrell GW, McAuley J. Rezero is all you need: Fast convergence at large depth. arXiv preprint arXiv:200304887 2020;.

[25] Ma J, Peng J, Wang S, Xu J. A conditional neural fields model for protein threading. Bioinformatics 2012;28(12):i59–i66.

[26] Haas J, Barbato A, Behringer D, Studer G, Roth S, Bertoni M, et al. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. Proteins: Structure, Function, and Bioinformatics 2018;86:387–398.

[27] Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC bioinformatics 2019;20(1):1–15.

[28] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research 2011;39(suppl_2):W29–W37.

[29] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 1997;25(17):3389–3402.

[30] Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin AM. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. Proteins: Structure, Function, and Bioinformatics 2018;86:51–66.

[31] Shrestha R, Fajardo E, Gil N, Fidelis K, Kryshtafovych A, Monastyrskyy B, et al. Assessing the accuracy of contact predictions in CASP13. Proteins: Structure, Function, and Bioinformatics 2019;87(12):1058–1068.

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

[33] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Tunyasuvunakool K, et al. High Accuracy Protein Structure Prediction Using Deep Learning. In: Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book); 2020. p. 22–24.

[34] Anishchenko I, Baek M, Park H, Dauparas J, Hiranuma N, Mansoor S, et al. Protein structure prediction guided by predicted inter-residue geometries. In: Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book); 2020. p. 30–31.

[35] Li Y, Zheng W, Zhang C, Bell E, Huang X, Pearce R, et al. Protein 3D Structure Prediction by Zhang Human Group in CASP14. In: Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book); 2020. p. 328–330.

[36] Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic acids research 2020;48(D1):D570–D578.

[37] Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nature communications 2018;9(1):1–8.

[38] Sillitoe I, Bordin N, Dawson N, Waman VP, Ashford P, Scholes HM, et al. CATH: increased structural coverage of functional space. Nucleic Acids Research 2021;49(D1):D266–D273.

Appendix A: Additional Figures and Tables

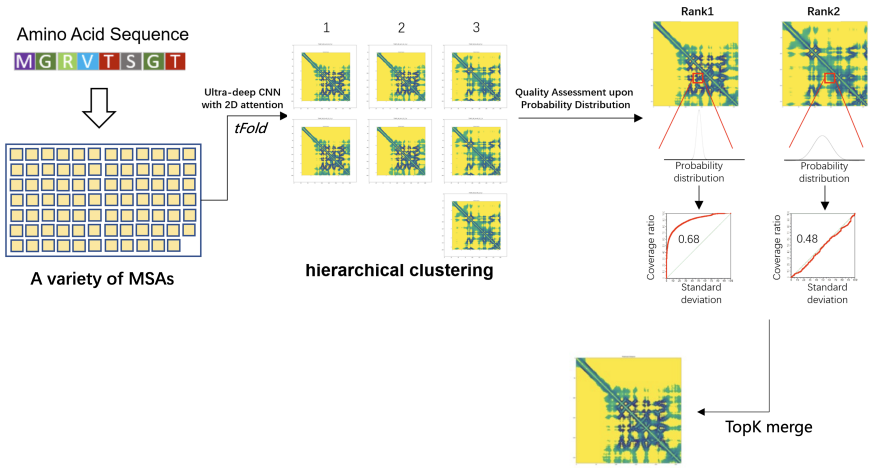

FIGURE S1 The "cluster-and-rank" strategy for multi-MSA fusion of distance predictions.

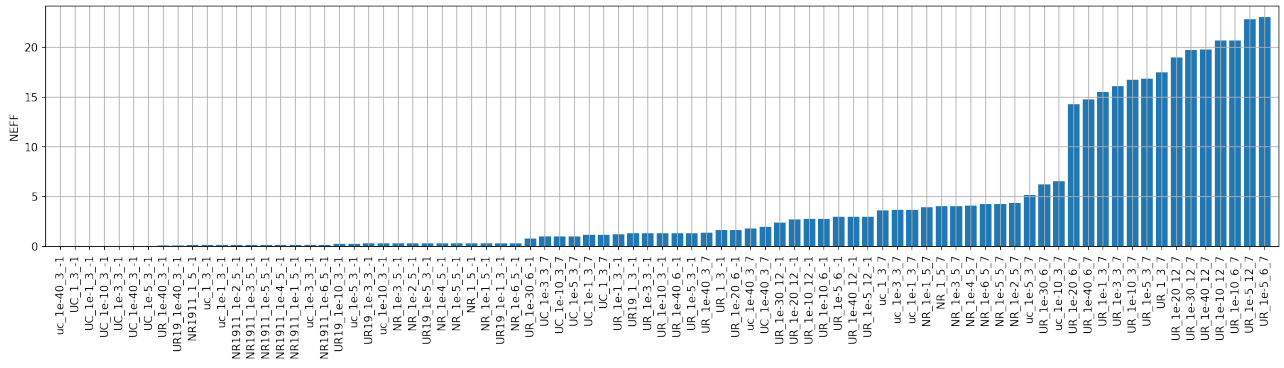

FIGURE S2 The NEFF index for all the 83 MSAs generated for T1043.

Appendix B: Neural network hyperparameters

| *SuperBlock*

- 5 *SEResBlock* blocks with 128 channels, cycling through dilations 1, 2, 4, 8, 16
- Nonlinearity: ELU
- The number of attention heads and head dims for two orthogonal axial attention layer: 8 and 64
- The number of attention heads and head dims for criss cross attention layer: 8 and 64
- 0.90 dropout keep probability

| *tFold-DistNet*

- Three branch fusion with 52 SuperBlocks
- Batch of 1 crops on each of 16 GPU workers
- Weight decay: 0.0001
- Progressive training with 16-24-32-40-48-52 *SuperBlock*s

| *tFold-RefineNet*

- Single branch with 16 SuperBlocks
- Batch of 1 crops on each of 16 GPU workers.
- Weight decay: 0.0001
- Auxiliary loss weights: deviation prediction = 1.0, lddt prediction = 10.0
- Progressive training with 8-12-16 *SuperBlock*s

Appendix C: The "Cluster-and-Rank" Strategy for Multi-MSA Fusion

The "cluster-and-rank" strategy works by firstly grouping distance predictions into clusters, and then select top-ranked ones from them. To define the pairwise similarity between distance predictions, another convolutional network is trained to predict the fold class from distance predictions, following the CATH fold definition [38]. Afterwards, we extract the final layer's activations as the embedding vector, and compute the cosine similarity between two embedding vectors for hierarchical clustering. Isolated clusters with only one member are filtered out. Subsequently, we calculate the averaged distance prediction of each cluster and then rank the results to identify the best one for the final submission.

We rank distance predictions according to a probability distribution based quality assessment metric, which is defined as the AUC score of distance distribution's standard deviation *vs.* coverage ratio of inter-residue pairs. This metric aims at measuring the concentration degree of predicted probabilistic distributions. The idea is based on the assumption that when MSA's co-evolution information strongly indicates that two residues are closely-related, then the predictor should be highly confident in its predictions. This corresponds to a single high peak in predicted probabilities. In contrast, if such co-evolution information does not exist in the MSA data, then the predictor tends to produce less concentrated probabilistic distributions since the confidence is lower.

For the distance prediction for each inter-residue pair, we compute its standard deviation as:

$$\begin{aligned} \mu_{d,i,j} &= \sum_{k=1}^{K_d} p_{d,i,j}^{(k)} \cdot d_k \\ \sigma_{d,i,j} &= \sqrt{\sum_{k=1}^{K_d} p_{d,i,j}^{(k)} \cdot \left(d_k - \mu_{d,i,j}\right)^2} \end{aligned} \tag{1}$$

where d_k is the average distance value for the k-th histogram bin, $\mu_{d,i,j}$ is the expected distance value, and $\sigma_{d,i,j}$ is the standard deviation of predicted distance distribution.

Furthermore, we design a normalization routine to ensure that this quantitative metric falls into range $[0, 1]$ for any distance predictions. We introduce a hyper-parameter, σ_{max}, as the maximal possible value of standard deviations. It is easy to prove that with each histogram bin's distance value given ($d_1 < d_2 < \cdots < d_{K_d}$), the maximal possible standard deviation is:

$$\sigma_{max} = \frac{d_{K_d} - d_1}{2} \tag{2}$$

With the maximal possible standard deviation given, we compute the AUC (area under curve) score for the (σ, r) curve, where r is the ratio of inter-residue pairs whose standard deviation $\sigma_{d,i,j}$ is smaller than the threshold $\sigma \in [0, \sigma_{max}]$. The AUC score is guaranteed to take values ranging from 0 and 1, and the prefect distance prediction (one-hot prediction for each inter-residue pair) indeed achieves the maximal AUC score, *i.e.* 1. In practice, instead of enumerating every possible threshold, we uniformly select a fixed number of thresholds (*e.g.* 1000) and then compute the approximated AUC score. This offers a good balance between approximation precision and computational efficiency.

Appendix D: Normalized Number of EFFective sequences (NeFF)

We take a commonly adopted approach to evaluate the MSA depth: normalized number of effective sequences (NeFF):

$$N_f = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m\neq n}^{N} \mathbb{1}\left(S_{m,n} \geq 0.8\right)} \tag{3}$$

where L is the length of query sequence, N is the number of sequences in the MSA, $S_{m,n}$ is the sequence identity between the m-th and n-th sequences, and $\mathbb{1}(\cdot)$ is the indicator function, *i.e.* $\mathbb{1}\left(S_{m,n} \geq 0.8\right)$ equals to 1 if $S_{m,n} \geq 0.8$, and zero otherwise.

TABLE S1 Comparison on the Pearson correlation coefficient for FM (left) and non-FM (right) targets between structural decoys' lDDT scores and relative improvement in the top-L precision for medium- and long-range contacts. The baseline top-L precision is also reported, which corresponds to contact predictions without any structural decoys.

Target	Top-L Prec.	# of Decoys	Pearson CC	Target	Top-L Prec.	# of Decoys	Pearson CC
T1027	0.2679	56	0.7896	T1026	0.5697	60	0.9332
T1029	0.0320	61	-0.0347	T1030	0.4322	63	0.8198
T1031	0.0211	61	0.8342	T1032	0.4049	68	0.8106
T1033	0.1600	58	0.5310	T1034	0.9743	56	0.4069
T1037	0.4455	60	0.9338	T1035	0.2941	62	0.9355
T1038	0.4221	61	0.7751	T1045s2	0.7861	66	0.4157
T1039	0.0683	65	0.8583	T1046s1	0.5945	68	0.6865
T1040	0.0692	68	0.8556	T1046s2	0.7676	70	0.8564
T1041	0.6281	70	0.8242	T1047s2	0.9095	73	0.6968
T1042	0.3183	64	0.7411	T1050	0.7471	55	0.8066
T1043	0.2500	60	0.8706	T1052	0.8413	57	0.9346
T1047s1	0.6078	71	0.3425	T1053	0.5775	70	0.7230
T1049	0.7801	68	0.3608	T1054	0.7631	68	0.6132
T1061	0.6280	20	0.9219	T1055	0.5067	71	0.8741
T1064	0.0283	72	0.5806	T1056	0.7096	65	0.6009
T1074	0.2525	95	0.8841	T1057	0.8710	70	0.5501
T1090	0.7772	93	0.3272	T1058	0.7356	72	0.8535
T1093	0.3867	86	0.5120	T1060s2	0.7953	95	0.5456
T1094	0.3810	60	0.6739	T1060s3	0.8714	100	0.4873
T1096	0.2694	90	0.8718	T1065s1	0.6299	94	0.4982
				T1065s2	0.7142	96	0.4616
				T1067	0.2765	99	0.7583
				T1068	0.3886	94	0.4558
				T1070	0.7015	94	0.5685
				T1076	0.9836	84	0.8101
				T1078	0.5579	95	0.5094
				T1079	0.8039	90	0.5340
				T1080	0.0195	53	0.8689
				T1082	0.3608	85	0.5309
				T1083	0.6122	93	0.5156
				T1084	0.5616	88	0.6393
				T1087	0.4731	90	0.8674
				T1089	0.7920	97	0.5366
				T1091	0.6581	95	0.5619
				T1092	0.6502	88	0.5023
				T1095	0.6165	78	0.7232
				T1099	0.2404	91	0.7228
				T1100	0.5325	99	0.6577
				T1101	0.9119	98	0.7111