Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15

High-quality protein monomer and complex structure prediction in CASP15 utilizing deep learning, threading templates, and multi-MSA strategy

Wei Zheng[1, 2,] *, Qiqige Wuyun[3], Peter L Freddolino[1, 2], Yang Zhang[1,2,4, 5]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA.
[2]Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA.
[3]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
[4]Department of Computer Science, School of Computing, National University of Singapore, 117417 Singapore.
[5]Cancer Science Institute of Singapore, National University of Singapore, 117599, Singapore.

* To whom correspondence should be addressed.
Email: zhengwei@umich.edu

ABSTRACT (<250 words, now=250?)

We report the results of the 'UM-TBM' and 'Zheng' groups in CASP15 for protein monomer and complex structure prediction. These prediction sets were obtained using the D-I-TASSER and DMFold-Multimer algorithms, respectively. For monomer structure prediction, D-I-TASSER introduced four new features during CASP15: (i) a multiple sequence alignment (MSA) generation protocol that combines multi-source MSA searching and a structural modeling-based MSA ranker; (ii) attention-network based spatial restraints; (iii) a multi-domain module containing domain partition and arrangement~~assembly~~ for domain-level templates; (iv) an optimized I-TASSER-based folding simulation system for full-length model creation guided by a combination of deep learning restraints, threading alignments, and knowledge-based potentials. For 47 FM targets and 47 TBM targets in CASP15, the first models predicted by D-I-TASSER showed average TM-scores 19% and 4% higher than the standard AlphaFold2 program, respectively. We thus showed that traditional Monte Carlo-based folding simulations, when appropriately coupled with deep learning algorithms, can generate models with improved accuracy over end-to-end deep learning methods. For protein complex structure prediction, DMFold-Multimer generated models by integrating a new MSA generation algorithm (DeepMSA2) with the end-to-end modeling module from AlphaFold2-Multimer. For the 38 complex targets, DMFold-Multimer generated models with an average TM-score of 0.83 and Interface Contact Score of 0.60, both significantly higher than those of competing complex prediction tools. Our analyses on complexes highlighted the critical role played by MSA generating, ranking, and pairing in protein complex structure prediction. We also discuss future room for improvement in the areas of viral protein modeling and complex model ranking.

1. INTRODUCTION

Protein structure prediction is a long-studied fundamental problem in structural biology. The template-based modeling method I-TASSER (**I**terative **T**hreading **ASSE**mbly **R**efinement)[1-4] was designed to address this problem, and has proven to provide a highly robust and versatile framework for ongoing advances in protein structure prediction. 'Zhang-Server', which was based on the I-TASSER algorithm, joined the Critical Assessment of Protein Structure Prediction (CASP) experiments[5-9] from CASP7 to CASP11[10-14]. During this period, I-TASSER predicted protein structure mainly based on threading template information and knowledge-based potentials. Since the introduction of co-evolution and deep learning techniques, two versions of contact-guided I-TASSER (C-I-TASSER)[15] were developed by introducing the direct coupling analysis (DCA)-based and deep learning-based residue-residue contact prediction[16-18] into I-TASSER folding. These two versions of C-I-TASSER were used in CASP12[19] and CASP13[20], respectively. Subsequently, a more comprehensive deep-learning guided I-TASSER (D-I-TASSER) was developed using predicted contacts, distances, and hydrogen-bonds[21,22] to guided I-TASSER Replica Exchange Monte Carlo (REMC) simulation, as applied in CASP14[23]. All these I-TASSER-based protein structure prediction methods followed a similar two-step workflow: first, collecting geometric restraints either from templates or deep learning predictors; second, converting those features into energy potentials that are combined with the inherent knowledge-based potentials to guide the REMC folding simulations. This workflow shows strong robustness and versatility, because new restraint features from new algorithms and techniques can be conveniently introduced to the pipeline and result in improved modeling quality. This is evident by the fact that the overall accuracy of the TASSER series algorithms has consistently increased from CASP11 to CASP14, as improved deep learning-based constraints were incorporated into the underlying I-TASSER modelling framework.

CASP14 saw a remarkable shift in the field, due to the fact that the end-to-end deep learning pipeline AlphaFold2[24] generated excellent models for most targets. Different from our two-step protein folding strategy, AlphaFold2 feeds the raw multiple sequence alignment (MSA) into a deep neural network and directly creates the structure models by the network learning. It must be noted that AlphaFold2 still has difficulties in modeling some proteins, such as those with multiple domains and those with no homologous sequences[25]. However, the success of AlphaFold2 marked a solution to the structure prediction problem through pure machine learning from the large pool of experimentally solved structures in the Protein Data Bank (PDB)[26]. Meanwhile, it also prompted us to ask several fundamental questions: (i) Is the end-to-end learning method the only way to solve the protein structure prediction problem? (ii) Are knowledge-based potentials useless in the era of AI and deep learning? (iii) How can multi-domain protein modeling best incorporate the power of deep learning methods? These were central questions in protein monomer structure prediction that we wanted to address since CASP14.

After CASP14, the AlphaFold2 framework was subsequently extended to AlphaFold2-Multimer[27] for predicting the structures of multi-chain protein complexes. The AlphaFold2-Multimer pipeline has been demonstrated to have the ability to produce high-quality models for numerous instances. The quality of input MSAs largely decides the modeling performance of AlphaFold2-Multimer[27-29]. However, the shallow MSAs produced by its default MSA pipeline, and the mechanism of combining MSAs without optimal diversity, had restricted the predictive power of AlphaFold2-Multimer in terms of complex structure predictions. Thus, addressing the

MSA generation and pairing problems in protein complex structure predictions was another central topic that we wanted to address.

In CASP15, we used two methods, D-I-TASSER (group name ‘UM-TBM’) and DMFold-Multimer (group name ‘Zheng’), to participate in the protein monomer and protein complex structure prediction categories, respectively.

Compared with the D-I-TASSER pipeline used in CASP14[23], four newly developed components were integrated into the D-I-TASSER version used in CASP15. First, a new MSA construction pipeline, DeepMSA2, has been created. This pipeline searches large-scale whole-genome and metagenome databases for generating multi-source MSAs, and utilizes deep learning structure modeling for scoring MSAs. Second, two new attention-based deep neural network predictors, AttentionPotential and AlphaFold2, have been developed or introduced. These two methods as well as our previous method, DeepPotential[22], are combined to predict residue-to-residue spatial restraints, including contact maps, distance maps, inter-residue orientations, and hydrogen-bond networks. Third, a domain partition and arrangement ~~assembly~~ module has been developed for handling multi-domain targets. For domain partition, the contact-based method FUpred[30] and threading-based method ThreaDom[31] were utilized for non-homologous and homologous targets, respectively. In the domain arrangement ~~assembly~~ stage, DEMO2[32] was employed to merge ~~assemble~~ domain-level templates and extract spatial restraints. The merged ~~assembled~~ features were subsequently used in the folding simulation. Fourth, the D-I-TASSER REMC folding system has been optimized to predict protein structures with the combined guidance of deep learning-based restraint potentials, template-based potentials, and knowledge-based potentials. With the developments outlined above, the qualities of models obtained using D-I-TASSER are significantly improved when compared with AlphaFold2. The success of D-I-TASSER in CASP15 demonstrates that an end-to-end deep learning pipeline may not be a unique solution to the protein structure prediction problem. A two-step modeling strategy can also achieve this goal (with higher performance) by appropriately combining highly accurate deep learning-based spatial restraints with knowledge-based potentials, and potentially carries the additional benefit of more easily and transparently incorporating ongoing advances in the field through its ability to combine information from several different deep learning pipelines.

The DMFold-Multimer pipeline used a three-stage process for multi-chain complex prediction. First, monomer MSAs were constructed by DeepMSA2. Second, the top-ranked monomer MSAs were combinatorically paired to generate a diverse set of joint MSAs. Third, the end-to-end modeling module from AlphaFold2-Multimer was used to generate models using the paired MSAs as input. The advancements in DMFold-Multimer have led to a substantial enhancement in the model qualities when compared with AlphaFold2-Multimer and other competing methods.

2. METHODS

2.1 Overview of ‘UM-TBM’ server and ‘Zheng’ human group in CASP15

The ‘UM-TBM’ server group, utilizing the D-I-TASSER algorithm (**Fig. 1**), participated in the protein monomer modeling category in CASP15. Meanwhile, the 'Zheng' human group, based on the DMFold-Multimer method (**Fig. 2**), participated in the protein complex modeling category. Although the ‘Zheng’ group pipeline is fully automated, the extensive running time required for large protein complexes prevented participation in categories other than the human group. Below we describe the components of the D-I-TASSER monomer pipeline in **sections 2.2-2.5**, with a final summary of how the methods connect in **section 2.6**; DMFold-Multimer's components are described in **sections 2.2** and **2.7**.

2.2 Multiple sequence alignment construction for protein monomers by DeepMSA2

We utilized DeepMSA2 (**Fig. 2A**) to generate the multiple sequence alignments required in subsequent stages of both of our pipelines. DeepMSA2 contains two stages: (i) MSA generation using three sub-pipelines, and (ii) MSA ranking based on the structure model-associated confidence score.

During the MSA construction step, three sub-methods (dMSA, qMSA, and mMSA) are employed to generate a maximum of ten potential multiple sequence alignments (MSAs). The first sub-method, dMSA, is a prior MSA construction program (DeepMSA[33]) created for CASP13. dMSA utilizes three stages (labeled stage 1-3) where HHblits[34], Jackhmmer[35], and HMMsearch[35] are used to query the input sequence against Uniclust30[36], Uniref90[37], and Metaclust[38] databases, respectively. qMSA is an extended version of dMSA with a new search added between stages 2 and 3, utilizing HHblits to explore the BFD[39] metagenomics database. Additionally, qMSA employs UniRef30[36] as the database used in stage 1, and adds a new iteration stage (stage 4) to search through the Mgnify[40] metagenomics database. The construction of both dMSA and qMSA will terminate at any searching stage where the number of effective sequences (*Neff*) value (**Eq. 1**) is greater than 128, yielding a maximum of seven distinct MSAs generated by stages 1-3 of dMSA and stages 1-4 of qMSA. Here, *Neff* is defined as:

$$Neff = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1, m \neq n}^{N} I\left[S_{m,n} \geq 0.8\right]} \quad (1)$$

where L is the length of the query sequence, N is the number of sequences contained in the MSA, $S_{m,n}$ is the sequence identity between the m-th and n-th sequences, and I[ ] represents the Iverson bracket, which takes the value $I\left[S_{m,n} \geq 0.8\right] = 1$ if $S_{m,n} \geq 0.8$, and 0 otherwise.

The mMSA pipeline subsequently uses the MSA obtained from stage 3 of qMSA, generated from the BFD database, as the starting point for HMMsearch to explore three in-house metagenome databases (IMG/M[41], TaraDB[42], and MetaSourceDB[43]), which contain more sequences compared to the Metaclust, BFD, and Mgnify databases. The resulting sequence hits are converted into a sequence database, which is used as the target database for stage 3 of dMSA, stage 3 of qMSA, or stage 4 of qMSA to generate additional three MSAs.

In the MSA ranking step, the ten MSAs generated by DeepMSA2 are utilized as inputs for separate AlphaFold2 runs, where the template detection module is turned off and the embedding parameter is set to one to enable rapid model generation. The MSA linked with the highest pLDDT score among the AlphaFold2 models is selected as the final output of DeepMSA2.

Based on benchmarking of 73 protein monomer targets, the time complexity for DeepMSA2 scales roughly linearly with sequence lengths; run times can be estimated by 0.003L+2.055 hours, where L is the length of the protein (**Fig. S1A**). The benchmarking is based on ten CPU cores and four GPU A40 cards.

2.3 Template detection by LOMETS3

The templates used for D-I-TASSER simulation are detected by the LOMETS3[44] pipeline (**Fig. 1**), which contains two steps: (i) template detection by individual threading programs, and (ii) template re-ranking by the LOMETS3 scoring function.

In the template detection step, two groups of threading algorithms, profile-based threading and contact-based threading methods, are employed to gather the initial templates. The MSA that was generated in the previous step is utilized to create sequence profiles or profile Hidden Markov Models (HMMs) for six profile-based threading methods including FFAS3D[45], HHpred[46], HHsearch[47], MRFsearch[48], MUSTER[49], and SparksX[50]. For these six profile-based threading methods, a template re-ranking algorithm is implemented based on a scoring function, $Zscore(i,j)$, which combines the original profile-based alignment score (*Prof*), contact map overlapping score (*CMO*), and mean absolute distance error of the template (*MAET*). The $Zscore(i,j)$, where *i* represents *i*-th template and *j* represents *j*-th threading program, is defined as following:

$$Zscore(i,j)=w_1 Zscore^{MAET}(i,j)+w_2 Zscore^{CMO}(i,j)$$

$$+w_3 Zscore^{Prof}(i,j)(2)$$

Here, $Zscore^{X}(i,j)$ could be calculated as:

$$Zscore^{X}(i,j)=\frac{X(i,j)-\langle X(j)\rangle}{\sigma(X(j))}(3)$$

where $\langle X(j)\rangle$ and $\sigma(X(j))$are the average and standard deviation of the scoring function X, and X represents for *CMO*, *MAET* and *Prof*. Here, *CMO* is defined as:

$$CMO=\frac{N(CM^{query},CM^{template})}{N(CM^{query})}(4)$$

where $N(CM^{query},CM^{template})$ is the number of overlapping contacts between the predicted query contact map and the contact map derived from the aligned template, and $N(CM^{query})$ is the number of predicted contacts. *MAET* is defined as:

$$MAET=\frac{\sum_{m}^{ali}\sum_{n>m}^{ali}\text{¿}\,d_{m,n}^{query}-d_{m,n}^{template}\vee\text{¿}}{\sum_{m}^{ali}\sum_{n>m}^{ali}1}(5)\text{¿}$$

where $d_{m,n}^{query}$ is the predicted distance between residue *m* and *n* in the query structure, $d_{m,n}^{template}$ is the corresponding distance between the residue in the template structure that aligned to position *m* and *n* of the query, and *ali* means the length of alignment. The contact map and distance map, obtained from the top-ranked AlphaFold2 model using the MSA generated by DeepMSA2, are utilized by five contact-based threading methods. These methods include CEthreader[51], DisCovER[52], Map_align[53], EigenThreader[54], and Hybird-CEthreader[51]. To increase the efficiency of the contact-based threading approaches, we choose the top 1,000 templates identified by HHsearch, and then re-rank these templates using each of the five contact-based threading methods.

For the final template re-ranking step, the top 20 ranked templates will be selected from each individual threading method, resulting in 220 templates. Those templates are re-ranked based on the following scoring function that integrates Zscore and sequence identity between the identified template and query sequence[55]:

$$score(i,j)=conf\frac{(j)*Zscore(i,j)}{Zscore_0(j)}+seqid(i,j)(6)$$

where $seqid(i,j)$ is the sequence identity between the query and the *i*-th template from the *j*-th program, $conf(j)$ is the confidence score for the *j*-th program, and $Zscore_0(j)$ is the Zscore cutoff for defining good/bad templates for the *j*-th program. The target will be defined as 'Easy' or 'Hard' based on the number of high-quality threading alignments ($Zscore(i,j) > Zscore_0(j)$) detected by LOMETS3, where the 'Easy' target is roughly corresponding to CASP 'TBM-easy' (and 'TBM-hard') target, and 'Hard' target is roughly corresponding to CASP 'FM/TBM' (and 'FM') target. For the different type of targets, the modeling strategy in later steps will be different.

2.4 Spatial restraint prediction by the deep learning module

Three deep learning algorithms (**Fig. 1**), AlphaFold2[24] (8-embedding), AttentionPotential, and DeepPotential[22], are applied to accurately predict residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bond networks for D-I-TASSER, utilizing the DeepMSA2 final MSAs.

DeepPotential[22] is a computational tool that we developed during CASP14 to predict residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bond networks for both Cα-Cα and Cβ-Cβ residues. The DeepPotential pipeline utilizes a combination of two-dimensional co-evolutionary features and one-dimensional sequence-based features as machine learning inputs. The co-evolutionary features consist of raw coupling parameters from the 22-state Potts model optimized through pseudo-likelihood maximization (PLM), and the raw mutual information (MI) matrix derived from the co-evolutionary information of the given multiple sequence alignment (MSA). Sequence features include Potts model field parameters, Hidden Markov Model (HMM) features, self-mutual information, and one-hot representation of the MSA. These features are then input separately into deep convolutional residual neural networks, where they are passed through sets of one-dimensional and two-dimensional residual blocks, respectively, before being tied together. The resulting tiled feature representations serve as inputs to another fully residual neural network that contains 40 2-D residual blocks, which is trained using cross-entropy loss and outputs several types of spatial restraints.

AttentionPotential is an advanced computational pipeline derived from DeepPotential that leverages an MSA transformer[56] and AlphaFold2 Evoformer[24]. Unlike DeepPotential, AttentionPotential extracts co-evolutionary information directly through an attention mechanism that can more accurately capture the interactions between residues. Starting from a MSA m_{si}^{init}, with *S* aligned sequences and *L* positions, the 'InputEmbedder' module is applied to get the embedded MSA representation m_{si} and the pairwise representation z_{ij}. Additionally, the MSA embeddings and attention maps from MSA transformer, i.e., m_{si}^{esm} and z_{ij}^{esm}, are linearly projected and added to m_{si} and z_{ij} respectively. The representations obtained are subsequently input into the Evoformer model, which consists of 48 Evoformer stacks used to predict residue-residue contact maps, distance distributions, inter-residue torsion angles, and hydrogen-bond networks.

In addition to DeepPotential and AttentionPotential, the Cβ-Cβ distance distribution derived from AlphaFold2 is also utilized to guide the D-I-TASSER simulation. The final MSA of DeepMSA2 is input into AlphaFold2, where the default templates are replaced by LOMETS3 templates and the embedding parameter is set to eight. Other parameters (e.g., modeling recycles, dropout rate, number of sampling decoys, etc.) of AlphaFold2 as utilized were left at their default values. Finally, AlphaFold2 generates five models, and the distance output from the model with the highest pLDDT score is selected as the final output.

To assess the accuracy of the distance predictions relative to experimental results, the mean absolute distance error (*MAE*) of the top 5*L* (*L* is the protein length, in amino acids) long-range (¿$i-j\lor\geq 24$) predicted distances is considered:

$$MAE=\frac{1}{5L}\sum_{(i,j)}^{5L}\left|d_{i,j}^{pred}-d_{i,j}^{\exp}\right|(7)$$

where $d_{i,j}^{\exp}$ is the distance between residue *i* and *j* in the experimental structure, and $d_{i,j}^{pred}$ is the predicted distance between residue *i* and *j* from prediction, the latter is estimated as the middle value of the bin with the highest probability.

2.5 Domain partition and arrangement ~~assembly~~ by the multi-domain handling module

A novel domain partition and arrangement ~~assembly~~ module (**Fig. 1**) has been incorporated into D-I-TASSER to tackle the complex issue of multi-domain protein modeling. Unlike our earlier domain handling module employed in CASP14, which attempted to merge ~~assemble~~ the final predicted domain-level models, the new module strives to ~~assemble~~ reconstruct a full-length model from the domain-level inputs, i.e., the templates and spatial restraints, for subsequent D-I-TASSER folding simulation.

The new domain partition module integrated into D-I-TASSER incorporates two domain boundary prediction algorithms, ThreaDom[31] and FUpred[30]. ThreaDom is a template-based approach utilized for 'Easy' targets, whereas FUpred is designed for 'Hard' targets and predicts domains based on deep learning predicted contact maps. ThreaDom predicts domain boundaries by relying on LOMETS3 threading alignment coverage, where a domain conservation score (DCS) is calculated for each residue by combining information from template domain structures, terminal and internal gaps, and insertions. The domain boundary information is then derived from the DCS profile distribution. On the other hand, FUpred uses a recursive strategy to detect domain boundaries based on predicted contact maps and secondary structure information. This algorithm retrieves domain boundary locations by maximizing the number of intra-domain contacts while minimizing the number of inter-domain contacts from the contact maps. For a full-length sequence, LOMETS3 and the deep learning module are first used to collect whole chain-based templates and predicted contact maps, respectively, which are subsequently utilized by ThreaDom and FUpred to predict domain boundaries. Each individual predicted domain is then input again to LOMETS3 for domain-level template detection and to the deep learning module for domain-level spatial restraint prediction.

The final templates for the domains are merged ~~assembled~~ into a 'full-length' template using DEMO2[32]. DEMO2 first identifies ten global templates that cover as many domains as possible from a non-redundant multi-domain protein library by aligning each domain model to the template using TM-align[57]. A Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization is then performed starting from each initial global template to detect each domain's optimal translation vectors and rotation angles for domain-level templates. The optimization is guided by a comprehensive energy function that includes the knowledge-based potential, template-based potential, and inter-domain spatial restraints from the deep learning module. The translation vectors and rotation angles with the lowest energy are selected to construct the final 'full-length' template. In addition to merging ~~assembling~~ the full-length templates, the domain-level predicted spatial restraints are also merged ~~assembled~~ into full-length spatial restraints. Finally, the merged ~~assembled~~ full-length templates, the merged ~~assembled~~ full-length spatial restraints, and the whole chain-based full-length spatial restraints are all used as inputs for the D-

I-TASSER simulation (see below), with the whole chain-based full-length spatial restraints primarily used for providing inter-chain restraints during the modeling process.

2.6 D-I-TASSER folding pipeline for protein monomers

The full D-I-TASSER pipeline (**Fig. 1**) includes five steps: (i) MSA construction by DeepMSA2, (ii) template detection by LOMETS3, (iii) spatial restraints prediction by the deep learning module, (iv) domain partition and arrangement ~~assembly~~ by the multi-domain handling module, and (v) full-length atomic model generation by D-I-TASSER folding simulation.

Steps (i)-(iv) have been introduced in **sections 2.2-2.5** above, while step (v) involves the D-I-TASSER folding simulation, which includes three sub-stages. Firstly, initial conformations are generated based on LOMETS3 templates and deep learning-based models. Next, full-length Cα models are arranged ~~assembled~~ using D-I-TASSER Replica Exchange Monte Carlo (REMC) simulation, which is guided by template-based restraints, deep learning spatial restraints, and knowledge-based potentials. Finally, a full-length atomic model is generated and refined.

In the initial conformation generation step, a total of 15 full-length models are created by AlphaFold2 or DeepFold[58] L-BFGS folding system utilizing spatial restraints collected from LOMETS3 templates (see **section 2.3**) and predicted by the deep learning module (see **section 2.4**). To provide further details on the DeepFold system, it should be noted that the probabilities of distance terms for each pair of residues are converted into smooth potentials for the gradient-descent based protein folding system. The negative log of raw probability histogram is then interpolated by cubic spline as potentials. For distance probability histogram of residue pair *i* and *j*, the probability, $P(i,j)_{dis}$, is a fusion probability combining the raw probability $P(i,j)_{dis}^{dp}$ predicted from DeepPotential (or AttentionPotential) and statistical probability $P(i,j)_{dis}^{tem}$ derived from LOMETS3 top *N* ranked templates with alignment coverages > 0.5 for 'Easy' targets (alignment coverages >0.6 for 'Hard' targets). Here, *N* is 50 for an 'Easy' target, while 30 for a 'Hard' target. The fusion probability $P(i,j)_{dis}$ can be calculated as

$$P(i,j)_{dis} = \wp(i,j)_{dis}^{dp} + (1-w)P(i,j)_{dis}^{template} (8)$$

where *w* is a weight and equals to 0.8. Five models were generated using DeepFold, with varying random seeds, utilizing restraints from either DeepPotential or AttentionPotential combined with LOMETS3 templates. Along with five models from AlphaFold2, a total of 15 models are collected from the deep learning module. These 15 models are ranked as five AlphaFold2 models, five AttentionPotential-based models, and five DeepPotential-based models. They will be then employed as initial conformations, together with 220 LOMETS3 templates, for D-I-TASSER REMC folding simulations.

During the D-I-TASSER REMC folding simulation stage, three different types of REMC simulations (labeled as 'A', 'M' and 'F') are carried out based on the target's category. The 'A' simulation retains all Cα atoms on a 0.87Å lattice and the Cα atoms move along the lattice, with REMC simulation conformations initiated from LOMETS3 templates and 15 deep learning models and gaps filled from random conformations. On the other hand, 'M' freely rotates and translates fragments excised from the threading alignments; and 'F' keeps the threading-aligned fragments frozen with changes only to the unaligned regions. 'M' and 'F' are conducted only for 'Easy' targets whose template alignments have a higher confidence. For each pipeline, five REMC simulations are performed, and the decoy structures from eight (or three for 'Hard' targets) low-temperature replicas are subjected to structural clustering. The REMC simulation is guided by knowledge-based potentials, template-based restraints, and deep learning-based spatial restraints potentials. The deep learning-based spatial restraints potentials consist of residue-

residue contact maps[15,59] (described in **Eq. 9**), distance distributions[23] (described in **Eq. 10**), inter-residue torsion angles, and hydrogen-bond networks potentials, with the first two terms being the primary energy terms during the folding simulation.

$$E_{contact}(d_{ij}) = \begin{cases} -U_{ij}, d_{ij} < 8\,Å \\ \frac{-1}{2} U_{ij}\left[1 - \sin\left(\frac{d_{ij} - \left(\frac{8+D}{2}\right)}{d_b}\pi\right)\right], 8\,Å \le d_{ij} < D \\ \frac{1}{2} U_{ij}\left[1 + \sin\left(\frac{d_{ij} - (\frac{D+80}{2})}{(80-D)}\pi\right)\right], D \le d_{ij} \le 80\,Å \\ U_{ij}, d_{ij} > 80\,Å \end{cases} \quad (9)$$

$$E_{con}(d_{ij}) = \begin{cases} -U_{ij}, d_{ij} < 8\,Å \\ \frac{-1}{2} U_{ij}\left[1 - \sin\left(\frac{d_{ij} - \left(\frac{8+D}{2}\right)}{d_b}\pi\right)\right], 8\,Å \le d_{ij} < D \\ \frac{1}{2} U_{ij}\left[1 + \sin\left(\frac{d_{ij} - (\frac{D+80}{2})}{(80-D)}\pi\right)\right], D \le d_{ij} \le 80\,Å \\ U_{ij}, d_{ij} > 80\,Å \end{cases}$$

$$E_{disance}(d_{ij}) = -\log\left(\frac{P_{ij}(d_{ij}) + P_{ij}^{N}}{2P_{ij}^{N}}\right) (10)$$

where, *i* and *j* are a residue pair, $U_{ij}U_{ij}$ is the depth of the potential, d_{ij} is the Cβ-Cβ or Cα-Cα distance between residue *i* and *j* in the simulation decoys, and *D* is a hyper parameter to control the well width of the contact potential function term. In distance energy potential, $P_{ij}(d_{ij})$ is the probability of the distance d_{ij}, and P_{ij}^{N} is the probability of the last distance bin.

In the structure refinement stage, the 10,000 decoy conformations obtained from the REMC simulation are subjected to clustering using SPICKER[60], which yields five clusters. These five cluster centers are then subjected to fragment-guided molecular dynamics (FG-MD[61]) simulations, leading to the generation of five full atomic models. Subsequently, FASPR[62] is employed for repacking the side-chain rotamer structures of these models, while locPREFMD[63] and Amber[64] refinement techniques, which were used in AlphaFold2, are applied sequentially to refine the models further. The models are ranked by a confidence score that is calculated based on the significance of threading template alignments, contact map satisfaction rate, mean absolute error between model distances and predicted distances, and convergence of D-I-TASSER simulations.

Based on benchmarking of 85 protein monomer targets, the time complexity of D-I-TASSER is linear with protein length, and can be estimated by 0.172*L*-3.822 hours, where *L* is the length of the protein (**Fig. S1B**). The benchmarking is based on ten CPU cores and four GPU A40 cards.

2.7 DMFold-Multimer folding pipeline for protein complexes

The DMFold-Multimer (**D**eep**M**SA-based **Fold**ing for protein **Multimer**) pipeline (**Fig. 2B**) is designed for modeling protein complexes using the DeepMSA2 method for multiple sequence alignment (MSA) generation and the AlphaFold2-Multimer algorithm for protein complex modeling. The DeepMSA2 pipeline generates ten ranked MSAs for each individual protein. As a result, for a protein complex, each constituent protein sequence is linked with ten ranked MSAs. In the case of homomer (homo-oligomer) complexes, all ten MSAs are utilized in DMFold-Multimer for the purpose of generating paired MSAs, as the same MSA can be used for all monomers (simply tiled the correct number of times). However, in the instance of heteromer (hetero-oligomer) complexes, an additional selection process is implemented to generate an ideal set of paired MSAs based on the combinations of the individual constituent MSAs. For each constituent protein, we select the top N ranked MSAs based on monomeric pLDDT scores. These MSAs are then used to generate potential paired MSAs, where each selected MSA for one constituent protein can be paired with the MSA of another constituent. Thus, for a heteromeric complex containing M different constituent proteins, N^M distinct paired MSAs are generated and evaluated. To guarantee that modeling with N^M set of paired MSAs could be completed within a reasonable time, N is selected as the maximal value to satisfy $N^M \leq 100$. For example, if a complex contains three unique protein components (A2B2C1), then N would be set to 4 ($64=4^3 \leq 100$). In other word, for each protein components in this complex, we will select best top 4 MSAs, and build a set of MSAs for the complex with 64 different combinations, using every possible combinatorial pairing of those four MSAs for each of the constituents. After the MSAs for M different constituent proteins are paired, for example for paired MSAs (MSA-P1$_{i1}$, MSA-P2$_{i2}$, MSA-P3$_{i3}$, … , MSA-PM$_{iM}$) ($1 \leq i_k \leq N$; $1 \leq k \leq M$), the sequences within each MSA are concatenated using the AlphaFold2-Multimer default sequence connection pipeline[27]. This pipeline initially groups the sequences by the UniProt annotated species and subsequently connects the sequences in each group based on their order of sequence identity to the query sequence. In the final step of complex model generation, the selected N^M (or 10 for homomer) sets of MSAs are used as input to a modified AlphaFold2-Multimer pipeline. The major difference between DMFold-Multimer and AlphaFold2-Multimer is the MSA pipeline, so other parameters of AlphaFold2-Multimer folding engine utilized by DMFold-Multimer (i.e. modeling recycles, dropout rate, number of sampling decoys, etc.) were left at their default values during modeling. For each set of MSAs, 25 models are generated. Finally, the resulting $25N^M$ (or 250 for homomer) complex models are ranked by the predicted TM-scores[65] (a linear combination of protein monomer TM-score and protein interface TM-score with weight 0.2 and 0.8, respectively), and the top five complex models are selected as the final set of models. Based on the benchmarking of 32 protein complex targets, the time complexity of DMFold-Multimer is linear with sequence length and could be estimated by $0.059L-1.179$ hours (**Fig. S1C**), where L is the sum length of the component proteins in the protein complex. The time complexity benchmarking is based on ten CPU cores and four GPU A40 cards.

3. RESULTS

Using official CASP15 definition, 94 domains (**Table S1**) from 68 full-length monomer targets were assessed for protein regular modeling category by D-I-TASSER, and 38 multimer targets

(**Table S2**) were assessed for protein complex modeling category by DMFold-Multimer. Based on the difficulty of modeling, these 94 domains were categorized as 36 'TBM-easy' targets, 11 'TBM-hard' targets, 8 'FM/TBM' targets, and 39 'FM' targets by the official CASP definitions. To simplify the terminology in the following analysis, 'TBM-easy' and 'TBM-hard' targets were labeled as TBM targets, while 'FM' and 'FM/TBM' targets were labeled as FM targets. Of the 38 multimer targets, based on chemical stoichiometry information, the protein complexes include 19 hetero-oligomer (heteromer) targets and 19 homo-oligomer (homomer) targets. In the following discussions, the analyses were made on the datasets mentioned above, if there is no specific explanation.

3.1 The evolution of a series of I-TASSER algorithms after introducing more accurate deep learning-based spatial restraints

The protein monomer modeling pipeline in CASP15 was based on the D-I-TASSER algorithm, which is an extended method from our classic template-based I-TASSER algorithm and the contact-associated C-I-TASSER algorithm. **Fig. 3A** and **3B** show the protein monomer modeling results of a series of I-TASSER algorithms from CASP11 to CASP15. Since some experimental structures in CASP11 to CASP14 are still not released, we directly downloaded the 'Zhang-Server' results from the CASP official website, and all analyses in **Fig. 3A** and **3B** were based on TM-scores from those official results. I-TASSER, which is a pure fragment assembly-based method, took part in the CASP11, and only folded two FM targets with an average TM-score of 0.335 (**Table S3**). In CASP12 and CASP13, two versions of C-I-TASSER algorithms, using DCA-based contacts or deep learning-based contacts to guide the folding simulation, showed better folding abilities, where C-I-TASSER folded (TM-score>0.5[66]) over 45% and 54% of the FM targets in CASP12 and CASP13, with average TM-scores of 0.470 and 0.486, respectively. After deep learning-predicted distances were introduced, D-I-TASSER in CASP14 generated models with an average TM-score of 0.610 for FM targets. In CASP15, the new version of D-I-TASSER, which hybridizes distance predictions from AlphaFold2, AttentionPotential, and DeepPotential, generated substantially better models with an average TM-score of 0.833 for FM targets, increasing around 36% compared with CASP14. For TBM targets, since the modeling performance highly depends on the template quality, the tendency was slightly different from FM targets.

Since the target difficulties in the past five CASP experiments were slightly different, to make a more fair comparison between these I-TASSER-based algorithms and highlight the advantages of the new version of the D-I-TASSER algorithm, we re-ran I-TASSER, C-I-TASSER, and CASP14 version of D-I-TASSER algorithm on 65 CASP15 full-length targets with the same domain boundary predictions we used in CASP15, and all templates released after May 1 2022 excluded. **Fig. 3C-E** shows the resulting head-to-head comparisons between the D-I-TASSER algorithm and these three previous I-TASSER methods. Overall, D-I-TASSER generated models for FM and TBM targets with average TM-scores of 0.840 and 0.925, which were 152% (121%) and 23% (23%) better than the models generated by I-TASSER (C-I-TASSER) method, with *p*-values of 2.13E-14 (7.11E-15) and 1.42E-14 (7.11E-15), respectively (**Table S4**). Most notably, when compared with previous D-I-TASSER pipeline (CASP14 version) that solely utilized the DeepPotential spatial restraints, the updated D-I-TASSER pipeline generated 98% (=46/47) and 98% (=46/47) models with better TM-scores for TBM and FM targets on the CASP15 dataset, respectively.

One reason why the new D-I-TASSER pipeline outperformed the previous version used in CASP14 was the introduction of AlphaFold2-derived distance restraints. To test whether the improvement of D-I-TASSER arises only from the advantage of AlphaFold2, we made a direct comparison between D-I-TASSER and AlphaFold2 on 94 CASP15 domains (**Fig. 3F**). The AlphaFold2 models were taken from the CASP standard AlphaFold2 server (the 'NBIS-AF2-standard' group in CASP15, which used the public release AlphaFold2 at that time with default parameters run by the Elofsson Lab). Overall, D-I-TASSER generated 39 (31) models with better TM-scores than AlphaFold2 models for 47 FM (47 TBM) targets. Especially for FM targets, the average TM-score of D-I-TASSER model was 19% better than AlphaFold2 with a *p*-value of 4.02E-06 (**Table S5**). It was notable that D-I-TASSER constructed correct folds (TM-score>0.5[66]) for 12 targets (10 FM targets and 2 TBM targets) on which AlphaFold2 failed in generating correct models. **Fig. 3G** lists five such FM targets (**T1125-D1**, **T1125-D2**, **T1125-D5**, **T1130-D1**, and **T1169-D1**), for which D-I-TASSER predicted correct models, while the other five of these FM targets formed a protein complex (H1137) by a simple helix-strand fold, and thus are not listed here. For these five targets, D-I-TASSER predicted models with TM-scores that were all above 0.7, while AlphaFold2 models had TM-scores that were all below 0.45. The better performance of D-I-TASSER showed its folding ability already went beyond the premier end-to-end deep learning method, AlphaFold2.

In summary, D-I-TASSER outperformed the previous I-TASSER-based algorithms after incorporating AlphaFold2 (and other deep learning-derived) distances. The result comparison of a series of I-TASSER algorithms, both in historical CASP data and re-analysis of CASP15 data, showed the improvement of D-I-TASSER folding system by including more state-of-the-art deep learning methods within the I-TASSER simulation framework.

3.2 Contributions of DeepMSA2 MSA and threading template information to D-I-TASSER

To investigate why D-I-TASSER performed better than AlphaFold2 ('NBIS-AF2-standard' group), we performed a comparative analysis on modeling results from the standard version AlphaFold2, AlphaFold2 with LOMETS3 templates (shortened as 'AlphaFold2-L'), AlphaFold2 with LOMETS3 templates and DeepMSA2 MSA (shortened as 'AlphaFold2-LD'), and the full D-I-TASSER method (**Fig. 4A**). Overall, AlphaFold2-L performed slightly better than standard AlphaFold2 in both FM and TBM targets. Especially for FM targets, the average TM-score was 0.725 against 0.707 (**Table S5**), demonstrating the usefulness of LOMETS3-detected templates. Furthermore, after giving Alphafold2-L the improved DeepMSA2-derived MSA as input, AlphaFold2-LD could generate much better models, with the average TM-scores increasing 7% and 1% for FM and TBM targets, respectively. These results showed the power of integrating deeper MSA to AlphaFold2 pipeline, especially for FM targets. As we mentioned in **section 2.4**, in the full D-I-TASSER pipeline the models from AlphaFold2-LD would be used as the initial conformations for D-I-TASSER folding simulation and the derived distances from those models would be used to guide the folding simulation. Thus, the LOMETS3 threading templates and DeepMSA2 MSA contributed to the final D-I-TASSER modeling performance. Finally, there was a substantial jump in quality between the models from AlphaFold2-LD and D-I-TASSER, i.e., TM-scores were 0.775 vs. 0.840 for FM targets and 0.913 vs. 0.925 for TBM targets, arising from the contributions of other components in the D-I-TASSER pipeline, such as multi-domain handling module and folding simulation with the comprehensive force field.

Based on the above analysis, it is clear that deep learning-based spatial prediction was a critical feature for generating successful predicted models. However, the usefulness of the template information still could not be ignored. For example, **T1110** (**T1110-D1**) is an isocyanide hydratase with 227 residues. It is a single-domain TBM target with 8 α-helices and 8 β-strands forming a α/β fold (**Fig. 4B**). **T1109** (**T1109-D1**) is the same protein as T1110 with a single site mutation D183A (highlighted in red color in **Fig. 4B**). Both T1109 and T1110 form a homo-dimer complex (targets **T1109o** and **T1110o** in complex modeling category) in their crystal structure. The structures of the main region (residues 1-205) of these two targets are almost identical, while the C-terminus loop regions (residues 206-227) have different orientations. In T1110, the C-terminus loop forms intra-chain contacts to the N-terminus main region, while for T1109, the C-terminus loop shifts to the opposite direction almost without any contacts to N-terminus main body region. For the wild type protein (T1110), all component threading methods of LOMETS3 detected a good template, 3nooA, with an average TM-score of 0.80 (**Fig. 4C**). All residues from the template that aligned to the key residue D183 of the query sequence were all observed as aspartate (**Fig. S2**). Thus, the C-terminus loops of the templates also showed the same orientation pattern as the experimental structure of T1110. In addition, the predicted distances indicated the C-terminus loop had contact with the main region, which was consistent with the threading templates (**Fig. S3**). The conserved and high-quality templates from LOMETS3 helped D-I-TASSER construct a high-quality model with a TM-score of 0.97 that was nearly identical to the experimental structure (**Fig. 4E**). In contrast, for the mutated target T1109, LOMETS3 detected a template, 3b38A (**Fig. 4C**), which has a similar main region to the experimental structure with a TM-score of 0.74, but without the C-terminal region. Thus, when D-I-TASSER built models for T1109, the C-terminus loop was constructed *ab initio* guided by spatial restraints from the deep learning pipelines. D-I-TASSER generated a very accurate predicted distance map (**Fig. 4D**) with the mean absolute distance error (*MAE*, see **Eq. 7**) of 0.28Å. It clearly showed that the distances between the C-terminus loop and the main region were greater than 20Å in the predicted distance map. As a result, the final D-I-TASSER model (**Fig. 4E**) achieved a very high TM-score of 0.96. In contrast, for the wild-type T1110, The success of modeling both the wild-type T1110 and the mutated T1109, and particularly in identifying different templates and distance restraints on the basis of the mutation, showed the benefits of correct template information to protein structure prediction, especially for predicting some single site mutation targets.

Current deep learning-based protein structure prediction protocols are highly dependent on the quality of MSAs. To further investigate the impact of DeepMSA2 on D-I-TASSER modeling, we highlight one single-domain FM target, **T1179 (T1179-D1)**. T1179-D1 is an all β protein with 261 residues (**Fig. 4G**). AlphaFold2 generated a relatively good model for this target with a TM-score of 0.76, as its MSA pipeline detected 136 homologous sequences with a *Neff* of 6.84. However, the D-I-TASSER model had a much better quality with a TM-score of 0.93, since DeepMSA2's MSA had 324 homologous sequences with a slightly higher *Neff* of 8.84. To provide further insight into the relationship between MSA depth and model quality on T1179, we separately considered the models generated by D-I-TASSER using each of the component MSAs created by the DeepMSA2 pipeline. As expected, we saw a strong correlation between deeper MSAs (higher *Neff*) and increasing quality of D-I-TASSER. For example, when only genome sequence databases were used (dMSA and qMSA stage 1&2), numbers of sequences in MSAs were approximately 140, resulting in D-I-TASSER models with TM-scores lower than 0.7. After giving DeepMSA2 the third-party metagenomics sequence databases (Metaclust, BFD and Mgnify), models from D-I-TASSER achieved slightly better TM-scores due to more homologous sequences detected. Finally, if in-house metagenomics sequence database was utilized, two of three D-I-TASSER models achieved TM-scores greater than 0.9, since the MSAs contained around 300 homologous sequences with *Neffs* greater than 8.0. The case of T1179-D1 where D-I-TASSER with different MSAs generated different quality of models showed again MSA is an important feature of the success of D-I-TASSER.

Overall, the newly developed MSA generation pipeline, DeepMSA2, provides deeper MSAs which helped D-I-TASSER produce more reliable protein structure prediction results, especially for FM targets. Although the modeling performance improvement obtained by introducing threading templates was relatively small, the template information still showed its usefulness to correctly pick up the correct folds for some TBM targets.

3.3 A case study to highlight the advantage of optimized D-I-TASSER folding system to model domains from large multi-domain targets

As noted above, both threading templates and MSAs were important features for the D-I-TASSER modeling pipeline. However, it is difficult to determine whether solely using one feature could lead to the success of D-I-TASSER on any particular target. The optimized D-I-TASSER folding system which integrates threading template information, high quality MSAs, an efficient multi-domain handling module, and REMC folding simulation with the comprehensive force field, results in high overall prediction performance. Especially when modeling multi-domain targets, this optimized folding system showed remarkable superiority over other methods, demonstrating the importance of the improved domain-level handling in the most recent iteration of D-I-TASSER. In the 94 CASP15 domain targets, 48 came from the single-domain targets, while 46 domains came from 20 multi-domain targets. **Fig. S4** compared the modeling performance of D-I-TASSER for domains from single-domain targets vs. multi-domain targets. It was interesting to see that D-I-TASSER modeling quality for these two groups of domains were comparable for both FM and TBM targets, where average TM-scores of domains from single-domain targets were 0.840 and 0.926 for FM and TBM targets, and average TM-scores of domains from multi-domain targets were 0.840 and 0.924 for FM and TBM targets, respectively (**Table S6**).

Here we use the case of T1125-D2 to highlight the advance of D-I-TASSER in modeling large multi-domain targets, by combining multi-domain handling module and combining multi-source distances for the folding simulation. **T1125** is a large multi-domain target with 1,200 residues (**Fig. 5A**). The solved experimental structure covers only residues 327-1162, whereas the N-terminus and C-terminus are disordered regions. The solved region could be split into six domains, T1125-D1 to T1125-D6, with domain boundaries defined by the CASP organizers as '327-460; 461-608; 609-797; 798-946; 947-1096; 1097-1162' (with each pair of numbers denoting the residue range for a single domain). Since the entire T1125 was defined as a 'Hard' target by LOMETS3, FUpred (see **section 2.5**) was utilized to predict the domain boundaries. The entire T1125 was predicted as a seven domain targets where the last five domains covered the equivalent regions in the experimental structures of T1125-D1 to T1125-D6, with the predicted domain boundaries as '331-460; 461-610; 611-810; 811-925; 926-1200'. **T1125-D2** is an all β protein with 148 residues. FUpred almost perfectly predicted the domain boundary (461-610 vs. 461-608). The standard AlphaFold2 predicted a model for the T1125-D2 domain with a very low TM-score of 0.38 if modeling T1125 as a whole target without any domain partitions (**Fig. 3G**). Interestingly, with the MSAs from DeepMSA2, even the same modeling strategy utilized in AlphaFold2, it could generate a much better model for the T1125-D2 domain, with a TM-score of 0.75, associated with a derived distance map with an *MAE* of 0.53Å (**Fig. 5B** and

5C). However, due to the excellent domain boundary prediction for T1125-D2 and the deeper domain-level MSA with more homologous sequences, the domain-level distance map provided by domain-level AlphaFold2 modeling, had a lower error with an *MAE* of 0.49Å (**Fig. 5D)**, associated with model with a better TM-score of 0.79 (**Fig. 5E)**. Furthermore, as we mentioned in multi-domain handling module (**section 2.5**), the D-I-TASSER folding system combined the whole chain-level distance map and domain-level distance map, resulting in an even better distance map for T1125-D2 with an *MAE* of 0.48Å (**Fig. 5F**). Guided by this combined distance map, and initial conformation from the high-quality AlphaFold2 model with DeepMSA2 MSA (**Fig. 5E**), the D-I-TASSER folding simulation generated a more accurate final model with a TM-score of 0.83 (**Fig. 5G**). The modeling results of T1125-D2 again demonstrated the advantage of D-I-TASSER folding system in modeling domains from large-size multi-domain proteins by combining deep learning, threading template, MSA information, multi-domain handling module, and REMC folding simulation with the comprehensive force field, which taken together yields both improved intra-domain conformations and inter-domain orientations.

3.4 Overall performance of DMFold-Multimer for protein complex structure prediction

The 'Zheng' group protein complex modeling pipeline in CASP15 was based on the DMFold-Multimer method, which is an algorithm that combines DeepMSA2 multi-MSAs strategy and AlphaFold2-Multimer structure modeling module. In CASP15, four types of measures were used to assess the complex modeling quality, including TM-score, LDDT score, Interface Contact Score (ICS), and Interface Patch Score (IPS). The first two measures were used for assessing the global fold modeling quality, while ICS and IPS were used for quantifying the interface modeling performance. **Fig. 6A** summarizes the TM-scores of the DMFold-Multimer models vs. the target lengths for 38 CASP15 protein complex targets. Overall, DMFold-Multimer generated models for 36 complex targets with TM-scores greater than 0.50, and models for 30 targets with TM-scores greater than 0.70. In particular, for 63% of the complex targets, DMFold-Multimer models had a comparable quality with the experimental structures (TM-score>0.90). For all 38 complex targets, the DMFold-Multimer models achieved an average TM-score of 0.83, where for heteromer and homomer targets, average TM-scores were 0.869 and 0.792, respectively (**Table S7**). The reason why the average TM-score of homomer targets was slightly lower than heteromer targets was that for a homomer complex, given a residue *i*, the inter-chain distance to residue *j* in another chain or intra-chain distance to residue *j* in the same chain was more difficult to distinguish[23]. In contrast, the intra-chain vs. inter-chain distance problem was relatively rare in heteromer complexes. It was notable that DMFold-Multimer modeling quality was independent of the size of protein complex. As a proof, the average TM-score of the targets with residues greater than 1,500 was 0.881, which was even higher than the average TM-score (0.814) for targets with residues less than 1,500. For complex interface modeling, DMFold-Multimer generated models for 29 complex targets with ICS greater than 0.50, and for 17 targets with ICS greater than 0.70. For all 38 complex targets, the DMFold-Multimer models achieved an average ICS of 0.60, where for heteromer and homomer targets, average ICS were 0.61 and 0.59, respectively (**Table S7**). In **Fig 6B**, we presented the DMFold-Multimer models associated with the experimentally solved structures for 7 large-size complex targets (>1,500 residues) for which the predictions had a TM-score >0.8. These include H1111, H1114, H1137, T1170o, H1171, H1172, and T1181o, the sequences of which contain 8,460, 7,988, 4,592, 1,908, 1,956, 2,004, and 2,064 residues. For these 7 targets, DMFold-Multimer constructed impressive complex models with TM-scores (interface contact scores) of 0.98 (0.48), 0.91 (0.82), 0.94 (0.79), 0.93

(0.58), 0.93 (0.51), 0.91 (0.55), and 0.85 (0.60), respectively. Notably, the three largest targets are all heteromeric complexes with stoichiometry variable of ‘A9B9C9’, ‘A4B8C8’, and ‘A1B1C1D1E1F1G2H1I1’, respectively; DMFold-Multimer constructed high-accuracy models with average TM-score (ICS) of 0.94 (0.70) for them. These results demonstrated the ability of DMFold-Multimer to model large-size protein complexes.

Since the DMFold-Multimer method integrated the AlphaFold2-Mulitmer structure modeling module as its model generator, it was important to examine if DMFold-Multimer provided an improvement over the standard version of AlphaFold2-Multimer with default parameters. **Fig. 6C** shows a head-to-head comparison of the modeling quality between DMFold-Multimer and the standard version AlphaFold2-Multimer on CASP15 targets, where AlphaFold2-Multimer models came from ‘NBIS-AF2-multimer’ as operated by the Elofsson lab (using the public release of AlphaFold2-Multimer at that time with default parameters). Overall, DMFold-Multimer models outperformed AlphaFold2-Multimer for most targets both in terms of global quality and interface quality. Taking the TM-score and ICS score as examples, DMFold-Multimer models performed 15.6% and 29.1% better than AlphaFold2-Multimer models (both significant improvements, with *p*-values of 1.6E-02 and 3.4E-04, respectively; see **Table S8**). Furthermore, for heteromer and homomer targets, the DMFold-Multimer generated models with TM-scores (ICSs) of 0.869 (0.61) and 0.792 (0.59), which were 20.2% (41.9%) and 10.9% (18.0%) higher than those of AlphaFold2-Multimer models (0.723 (0.43) and 0.714 (0.50)), respectively. These results showed that the combination of improved MSAs and enhanced MSA pairing allowed DMFold-Multimer to substantially improve upon AlphaFold2-Multimer in predicting protein complex structures.

3.5 Contributions of MSA combination strategy and large-scale metagenomics database to DMFold-Multimer

Since DMFold-Multimer mainly focused on optimizing the input MSAs given to AlphaFold2-Multimer method (all other parameters during the modeling stage of DMFold-Multimer and AlphaFold2-Multimer are set same), it was important to investigate what MSA strategy led to the success of DMFold-Multimer. Compared to the default MSA pipeline in AlphaFold2-Multimer, two factors may contribute to the quality improvement: One is the integrated MSA creation and pairing mechanism, and the second is the inclusion of the additional huge in-house metagenomics databases used in DeepMSA2. To assess the relative contributions of these factors, in **Fig. 7A** and **Table S9**, we compared the complex modeling performance of AlphaFold2-Multimer and DMFold-Multimer using different sequence databases. Here H1111, H1114 and H1137 were excluded from the analysis, since the standard AlphaFold2-Multimer server (‘NBIS-AF2-multimer’ group) did not generate the full-length models for those three targets, which may affect the contribution analysis of MSA strategy. It was observed that even with the same sequence databases (from genomic sequences, BFD and Mgnify), DMFold-Multimer still outperformed AlphaFold2-Multimer with the average TM-score (ICS) increasing from 0.753 (0.48) to 0.769 (0.53), indicating the usefulness of MSA generating and pairing methods. After using the full version DMFold-Multimer including our expanded metagenome databases, the global fold modeling quality (TM-score) could be further increased by 6.6% from 0.769 to 0.820, and the interface modeling quality (ICS) could be further increased by 11.3% from 0.53 to 0.59, showing that the large metagenome databases were also beneficial for protein complex modeling.

To further check the importance of MSAs to protein complex structure prediction, three Nano-body antigen complexes, **H1140**, **H1141** and **H1144** are shown in **Fig. 7B**. For all three complex targets, AlphaFold2-Multimer could only generate models with TM-scores lower than 0.7 and ICS nearly 0. However, DMFold-Multimer could create models that have very high quality global folds with all TM-scores greater than 0.9, and good interfaces with all ICS greater than 0.5. Taking H1144 as an example to analyze why DMFold-Multimer produced high-quality models, we found that DMFold-Multimer could generate 2,500 decoys with TM-scores ranging from 0.62 to 0.99, and ICS ranging from 0.00 to 0.74 for this target (**Fig. 7C**), and thus most of the decoys had better quality than AlphaFold2-Multimer models. Furthermore, the predicted TM-score correctly selected one of those high-quality models as the first model (i.e., the model ranked as best by the predictor) for DMFold-Multimer, resulting in a very good final model with a TM-score of 0.99 and an ICS of 0.74 (**Fig. 7D**). It was notable that we observed that for H1144, all decoys with predicted TM-scores greater than 0.8 had very high-quality global fold with TM-score greater than 0.98, and good interface with ICS greater than 0.50. The quality of the DeepMSA2 paired MSA largely affected the accuracy of the protein complex modeling. For H1144, the number of sequences in DeepMSA2 paired MSAs ranged from 38 to 414, and the corresponding *Neff*s for those paired MSAs ranged from 1.8 to 16.3. **Fig. 7E-F** show the relationship between TM-score/ICS and *Neff*s of DeepMSA2 paired MSAs for H1144. It clearly indicated the first model of DMFold-Multimer came from the paired MSA with the highest *Neff*. It is understandable because the paired MSA with a high *Neff* could provide more co-evolutionary information, and thus lead to a better interface modeling quality. Interestingly, when giving DMFold-Multimer the same databases as AlphaFold2, it could still generate models with TM-scores greater than 0.9, and ICS roughly close to or greater than 0.5 for all three Nano-body antigen complex targets. Those results indicated that the enhanced MSA generating, ranking, and pairing used in DMFold-Multimer was critical for building correct models of those targets.

3.6 What went wrong in protein monomer and complex modeling using D-I-TASSER and DMFold-Multimer?

Although the D-I-TASSER and DMFold-Multimer have received excellent results in protein monomer and complex modeling, there are still some problems needed to be improved.

For D-I-TASSER protein monomer modeling pipeline, the prediction performance is highly reliant on the MSA quality and the targets with better MSA quality usually result in better structure models. Hence, we analyzed the relationship between the MSA *Neff* values and TM-scores of the D-I-TASSER models for different taxonomic categories on the 94 domains from 68 full-length monomer targets. As shown in **Fig. 8A**, DeepMSA2 is able to generate high-quality MSAs for most bacterial and eukaryotic targets, but relative low-quality MSAs for archaea and viruses. In particular, the virus-derived targets had low-quality MSAs with the lowest *Neff* of 13, and thus resulted in relatively low-performance models with an average TM-score of 0.801 (0.916 and 0.883 for bacterial and eukaryotic targets, respectively; see **Table S10**). This is mainly because there are fewer virus sequences in databases. Therefore, in the future, we plan to collect more virus data into our DeepMSA2 databases based on some virus specific databases such as Virus-Host DB[67] and NCBI Virus[68]. Although the overall modeling performance of eukaryotic targets was good (TM-score=0.883), for some targets, such as **T1130-D1** (**Fig. S5**), we observed relatively poor performance with a TM-score=0.754 compared with the best model from other groups with a TM-score=0.971. It appears that this failure was again due to the shallow MSA (*Neff*=0.16) created by DeepMSA2.

For our DMFold-Multimer protein complex modeling method, the predicted TM-score can distinguish high-quality models from low-quality models for most of the cases. However, we still found that the predicted TM-score sometimes was not sufficiently sensitive to rank high-quality models for some targets. For example, target **H1172** contains six copies of the 'A' protein and two copies of the 'B' proteins (A6B2 heterooctamer), where the two copies of 'B' proteins are in the neighboring positions in the experimental structure (**Fig. 8B**). DMFold-Multimer actually predicted all three possible states of this A6B2 complex (ortho, meta, or para orientations of the 'B' proteins). However, the first ranked model has the wrong relative positions of the 'B' proteins because it has a slightly higher predicted TM-score of 0.746 compared with Model3 with the correct positions (0.733) (**Fig. 8C**). Another example is the target **H1129**, which contains one copy of the 'A' protein and one copy of the 'B' protein (A1B1 heterodimer) (**Fig. 8D**). The models predicted by DMFold-Multimer all had predicted TM-scores lower than 0.4. Thus, the correct models cannot be picked out by the predicted TM-scores. In fact, DMFold-Multimer is able to predict a model (Model542) with a TM-score as high as 0.91. However, the best model has a predicted TM-score of only 0.335, which is 14% lower than the highest predicted TM-score of 0.388, and thus the correct structure could not be properly identified (**Fig. 8E**). Similarly, DMFold-Multimer predicted a model (Model79) with the highest ICS of 0.33. However, this model has a predicted TM-score of 0.371, which is still lower than the best predicted TM-score of 0.388, and thus the model showing the best ICS could not be picked out either (**Fig. 8E**). One reason for the failure of H1129 modeling is the poor predicted interface (ICS=0.01) between component protein 'A' and 'B'. This is due to the very shallow paired MSAs generated by DeepMSA2, which only contains two paired sequences (including the query sequences) with a *Neff* of 0.05. We noticed several groups built high-quality models for H1129 during CASP15, including the 'Wallner' group that utilized massive sampling with AlphaFold2-Multimer as their prediction strategy. To investigate whether the performance of DMFold-Multimer could be improved by combing the massive sampling with the DeepMSA2 multi-MSA strategy, we tested use of the same parameter settings as the Wallner group during the complex model generation stage for H1129. In detail, the settings include: using templates or not, increasing the modeling recycles, and turning on/off the dropout rate, resulting in a total of xx candidate structures as opposed to the yy generated by our standard pipeline. Overall, after applying massive sampling in DMFold-Multimer, the TM-score of the first model had been significantly improved to 0.96, and the ICS had also been largely improved to 0.72 (**Table S11**). Furthermore, we also tried this massive sampling strategy using DMFold-Multimer on other five dimer targets (H1142, T1121o, T1160o, T1161o, and T1187o) where we built relatively poor models with TM-scores <0.7 in the CASP15 evaluation. For all six targets, DMFold-Multimer produced better first models with average TM-score and ICS increased 26.4% and 120.0%, respectively, indicating that by combining the massive sampling with DeepMSA2 multi-MSAs strategy, the capability of DMFold-Multimer could be largely extended. However, we also noticed that the best models produced by DMFold-Multimer for those six targets were still substantially better than the first models according to our scoring, especially for T1121o and T1161o. Again, the ranking issue of the predicted TM-score prevents successfully selected those high-quality models. These findings suggest that improvements in massive sampling and model ranking have the potential to make major contributions to future performance enhancement for DMFold-Multimer, and that further improvements can be made through improved ranking to identify the best structures from an expanded ensemble of candidates.

4. CONCLUSIONS

We report two of our algorithms, D-I-TASSER and DMFold-Multimer, that participated in CASP15 as 'UM-TBM' server group for protein monomer structure prediction and 'Zheng' human group for protein complex structure prediction, respectively. The CASP15 version of D-I-TASSER includes four major developments compared with the previous version used in CASP14, including: a deep learning structure modeling ranking-based MSA generation method; new attention deep neural network-based spatial restraints predictors; a new domain partition and recombination ~~assembly~~ module; and a newly optimized folding system including balanced deep learning spatial energy potentials, template-based energy potentials, and knowledge-based potentials. The DMFold-Multimer pipeline combines the newly developed MSA constructor, DeepMSA2, for searching homologous sequences from large-scale genomic and metagenomics databases, with the structure model generator used in AlphaFold2-Multimer.

Based on analysis of the CASP15 targets, one important reason why D-I-TASSER and DMFold-Multimer generated high quality protein monomer and complex models was that the deeper and diverse MSAs generated by DeepMSA2. In particular, the large-scale metagenomics databases and the MSA pairing mechanism were key factors that helped improve the accuracy of protein complex structure predictions by DMFold-Multimer. On the other hand, by introducing more accurate geometric spatial restraints from new deep learning predictors, the D-I-TASSER method also showed its excellent performance compared with a series of previous I-TASSER methods. In addition, the newly designed domain handling module, associated with an optimized folding system that balanced deep learning spatial energy potentials, template-based energy potentials, and knowledge-based potentials, were major contributors for D-I-TASSER to produce high-quality models for multi-domain proteins. Finally, direct comparison between D-I-TASSER and standard AlphaFold2 control method in modeling protein monomer structures demonstrates that end-to-end deep learning is not the unique solution to achieve the goal of solving the protein folding problem.

Despite the success of D-I-TASSER and DMFold-Multimer, there are still significant challenges for the current pipelines. One was that shallow MSAs were still encountered for a few targets even though the DeepMSA2 and large-scale metagenomics databases had been utilized. Especially for some viral targets, due to rapid speed of evolution and the massive taxonomic spread of viruses, the number of homologous sequences was much fewer than other taxonomic groups, leading to a relatively lower performance of structure predictions. The other major challenge that we encountered was that the current model ranking score, predicted TM-score, produced by the AlphaFold2-Multimer structure module was not sensitive enough to distinguish models in some cases, especially when predicted TM-scores were very close, while model conformations were quite different. Those two issues probably could be solved by constructing a virus-specific sequence database[67,68], and developing a new model quality assessment tool that combines predicted TM-scores and decoys structure consensus, which will be the subject of our future efforts.

ACKNOWLEDGMENT

The authors thank Dr. Yang Li for discussion and assistance. We are grateful to Jonathan Poisson, Brock Palen, and the staff of the University of Michigan Advanced Research Computing team for the IT supporting. The D-I-TASSER and DMFold-Multimer servers use the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (ACI-1548562).

FUNDING

This work is supported in part by the National Institute of General Medical Sciences (GM136422 and S10OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to P.L.F. and Y.Z.), and the National Science Foundation (IIS1901191 and DBI2030790 to Y.Z., MTM2025426 to P.L.F. and Y.Z.).

REFERENCES

1. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols.* 2010;5(4):725-738.
2. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nature Methods.* 2015;12(1):7-8.
3. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Research.* 2015;43(W1):W174-W181.
4. Zheng W, Zhang C, Bell EW, Zhang Y. I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. *Future Generation Computer Systems.* 2019;99:73-85.
5. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction—Round VII. *Proteins: Structure, Function, and Bioinformatics.* 2007;69(S8):3-9.
6. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction—Round VIII. *Proteins: Structure, Function, and Bioinformatics.* 2009;77(S9):1-4.
7. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins: Structure, Function, and Bioinformatics.* 2011;79(S10):1-5.
8. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) — round x. *Proteins: Structure, Function, and Bioinformatics.* 2014;82(S2):1-6.
9. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function, and Bioinformatics.* 2016;84(S1):4-14.
10. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins: Structure, Function, and Bioinformatics.* 2007;69(S8):108-117.
11. Zhang Y. I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins: Structure, Function, and Bioinformatics.* 2009;77(S9):100-113.
12. Xu D, Zhang J, Roy A, Zhang Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins: Structure, Function, and Bioinformatics.* 2011;79(S10):147-160.
13. Zhang Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins: Structure, Function, and Bioinformatics.* 2014;82(S2):175-187.
14. Zhang W, Yang J, He B, et al. Integration of QUARK and I-TASSER for Ab Initio Protein Structure Prediction in CASP11. *Proteins: Structure, Function, and Bioinformatics.* 2016;84(S1):76-86.
15. Zheng W, Zhang C, Li Y, Pearce R, Bell EW, Zhang Y. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods.* 2021;1(3):100014.
16. Li Y, Hu J, Zhang C, Yu D-J, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics.* 2019;35(22):4647-4655.
17. Li Y, Zhang C, Bell EW, et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLOS Computational Biology.* 2021;17(3):e1008865.
18. Li Y, Zhang C, Bell EW, Yu D-J, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(12):1082-1091.
19. Zhang C, Mortuza SM, He B, Wang Y, Zhang Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins: Structure, Function, and Bioinformatics.* 2018;86(S1):136-151.

20. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(12):1149-1164.
21. Li Y, Zhang C, Zheng W, et al. Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins: Structure, Function, and Bioinformatics.* 2021;89(12):1911-1921.
22. Li Y, Zhang C, Yu D-J, Zhang Y. Deep learning geometrical potential for high-accuracy ab initio protein structure prediction. *iScience.* 2022;25(6):104425.
23. Zheng W, Li Y, Zhang C, et al. Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins: Structure, Function, and Bioinformatics.* 2021;89(12):1734-1751.
24. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596(7873):583-589.
25. Pearce R, Zhang Y. Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry.* 2021;297(1):100870.
26. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Research.* 2000;28(1):235-242.
27. Richard E, Michael ON, Alexander P, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv.* 2022:2021.2010.2004.463034.
28. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications.* 2022;13(1):1265.
29. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nature Methods.* 2022;19(6):679-682.
30. Zheng W, Zhou X, Wuyun Q, Pearce R, Li Y, Zhang Y. FUpred: detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics.* 2020;36(12):3749-3757.
31. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics.* 2013;29(13):i247-i256.
32. Zhou X, Peng C, Zheng W, Li Y, Zhang G, Zhang Y. DEMO2: Assemble multi-domain protein structures by coupling analogous template alignments with deep-learning inter-domain restraint prediction. *Nucleic Acids Research.* 2022;50(W1):W235-W245.
33. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics.* 2020;36(7):2105-2112.
34. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods.* 2012;9(2):173-175.
35. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Research.* 2018;46(W1):W200-W204.
36. 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.
37. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England).* 2015;31(6):926-932.
38. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nature Communications.* 2018;9(1):2542.
39. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods.* 2019;16(7):603-606.
40. Mitchell AL, Almeida A, Beracochea M, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research.* 2020;48(D1):D570-D578.

41. Chen IMA, Chu K, Palaniappan K, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research.* 2019;47(D1):D666-D677.
42. Wang Y, Shi Q, Yang P, et al. Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. *Genome Biology.* 2019;20(1):229.
43. Yang P, Zheng W, Ning K, Zhang Y. Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. *Proceedings of the National Academy of Sciences.* 2021;118(49):e2110828118.
44. Zheng W, Wuyun Q, Zhou X, Li Y, Freddolino PL, Zhang Y. LOMETS3: integrating deep learning and profile alignment for advanced protein template recognition and function annotation. *Nucleic Acids Research.* 2022;50(W1):W454-W464.
45. Xu D, Jaroszewski L, Li Z, Godzik A. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics.* 2014;30(5):660-667.
46. Meier A, Söding J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLOS Computational Biology.* 2015;11(10):e1004343.
47. Söding J. Protein homology detection by HMM–HMM comparison. *Bioinformatics.* 2005;21(7):951-960.
48. Ma J, Wang S, Wang Z, Xu J. MRFalign: Protein Homology Detection through Alignment of Markov Random Fields. *PLOS Computational Biology.* 2014;10(3):e1003500.
49. Wu S, Zhang Y. MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics.* 2008;72(2):547-556.
50. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Structure, Function, and Bioinformatics.* 2005;58(2):321-328.
51. Zheng W, Wuyun Q, Li Y, et al. Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLOS Computational Biology.* 2019;15(10):e1007411.
52. Bhattacharya S, Roche R, Moussad B, Bhattacharya D. DisCovER: distance- and orientation-based covariational threading for weakly homologous proteins. *Proteins: Structure, Function, and Bioinformatics.* 2022;90(2):579-588.
53. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science.* 2017;355(6322):294.
54. Buchan DWA, Jones DT. EigenTHREADER: analogous protein fold recognition by efficient contact map threading. *Bioinformatics.* 2017;33(17):2684-2690.
55. Zheng W, Zhang C, Wuyun Q, Pearce R, Li Y, Zhang Y. LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Research.* 2019;47(W1):W429-W436.
56. Roshan R, Jason L, Robert V, et al. MSA Transformer. *bioRxiv.* 2021:2021.2002.2012.430858.
57. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research.* 2005;33(7):2302-2309.
58. Pearce R, Li Y, Omenn GS, Zhang Y. Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. *PLOS Computational Biology.* 2022;18(9):e1010539.
59. Mortuza SM, Zheng W, Zhang C, Li Y, Pearce R, Zhang Y. Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nature Communications.* 2021;12(1):5011.
60. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry.* 2004;25(6):865-871.
61. Zhang J, Liang Y, Zhang Y. Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure.* 2011;19(12):1784-1795.
62. Huang X, Pearce R, Zhang Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics.* 2020;36(12):3758-3765.

63. Feig M. Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. *Journal of Chemical Information and Modeling.* 2016;56(7):1304-1312.

64. Case DA, Cheatham Iii TE, Darden T, et al. The Amber biomolecular simulation programs. *Journal of Computational Chemistry.* 2005;26(16):1668-1688.

65. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics.* 2004;57(4):702-710.

66. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics.* 2010;26(7):889-895.

67. Mihara T, Nishimura Y, Shimizu Y, et al. Linking Virus Genomes with Host Taxonomy. *Viruses.* 2016;8(3).

68. Hatcher EL, Zhdanov SA, Bao Y, et al. Virus Variation Resource – improved response to emergent viral outbreaks. *Nucleic Acids Research.* 2017;45(D1):D482-D490.

FIGURES

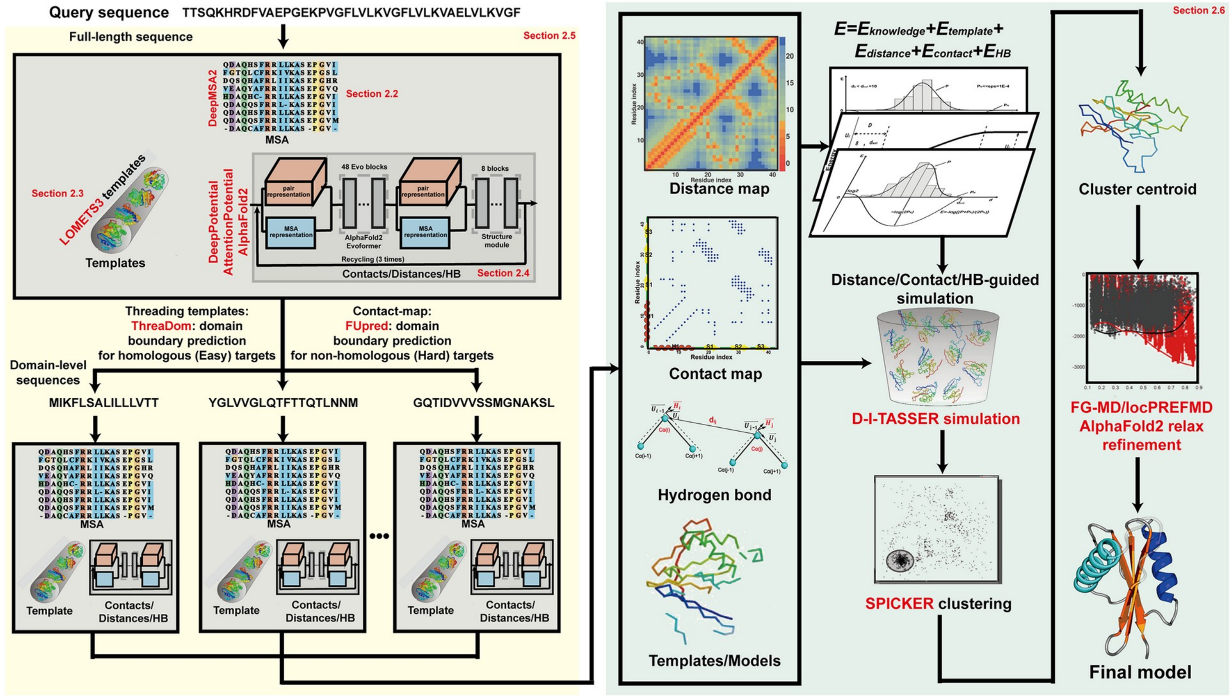

Figure 1. The pipeline of the 'UM-TBM' server (D-I-TASSER). The full D-I-TASSER pipeline is designed for modeling protein monomers through five steps: (i) MSA construction by DeepMSA2, (ii) template detection by LOMETS3, (iii) spatial restraints prediction by the deep learning module, (iv) domain partition and arrangement ~~assembly~~ by the multi-domain handling module (yellow box), and (v) full-length atomic model generation via D-I-TASSER folding simulation (green box).

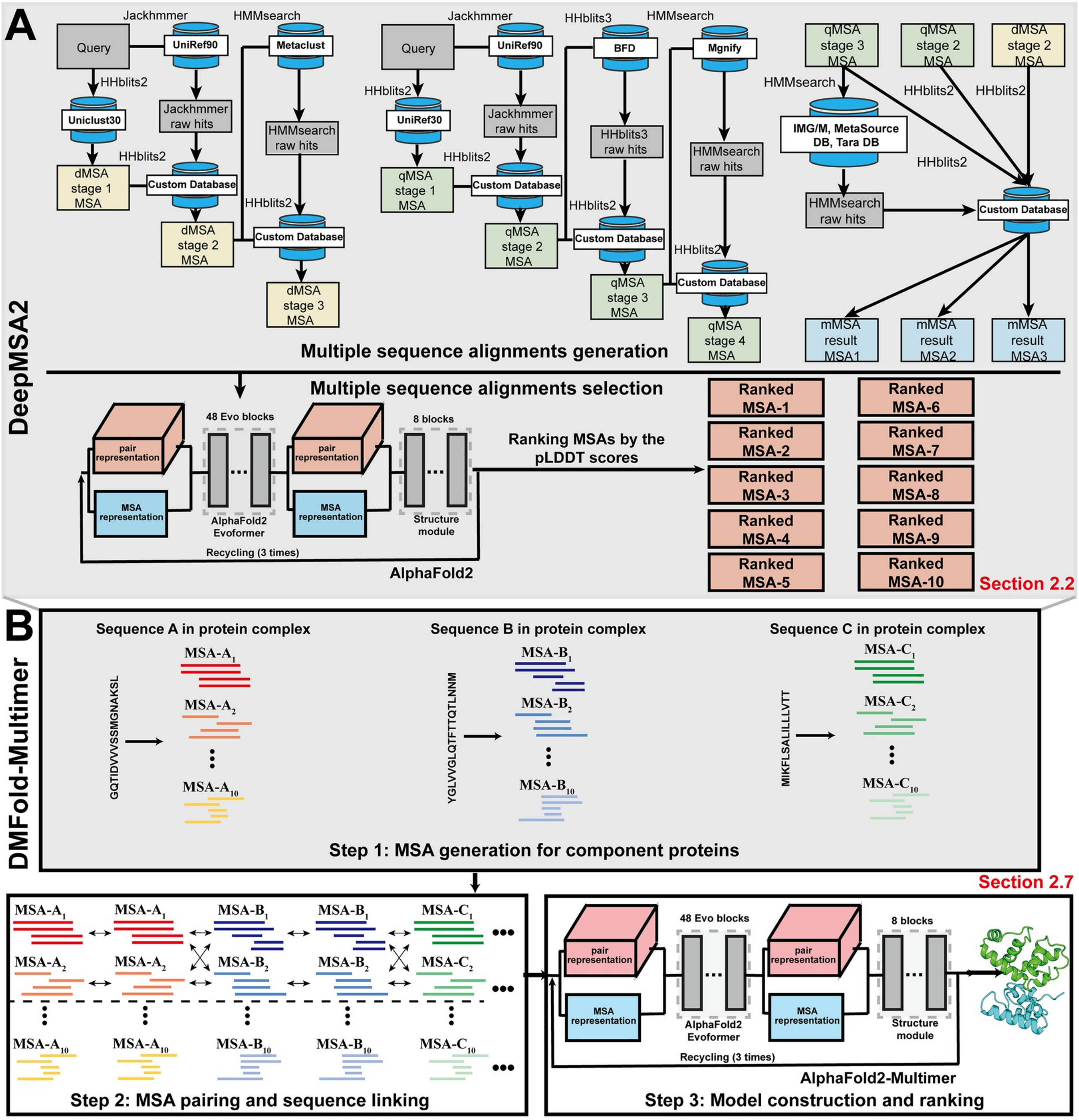

Figure 2. The pipelines of DeepMSA2 and 'Zheng' human group (DMFold-Multimer). (A) DeepMSA2 for generating the multiple sequence alignments, which contains two stages: (i) MSA generation using three sub-pipelines, and (ii) MSA ranking based on the structure model-associated confidence score. (B) The DMFold-Multimer pipeline is designed for modeling protein complexes by combining the DeepMSA2 MSA generation method with the AlphaFold2-Multimer complex modeling algorithm.

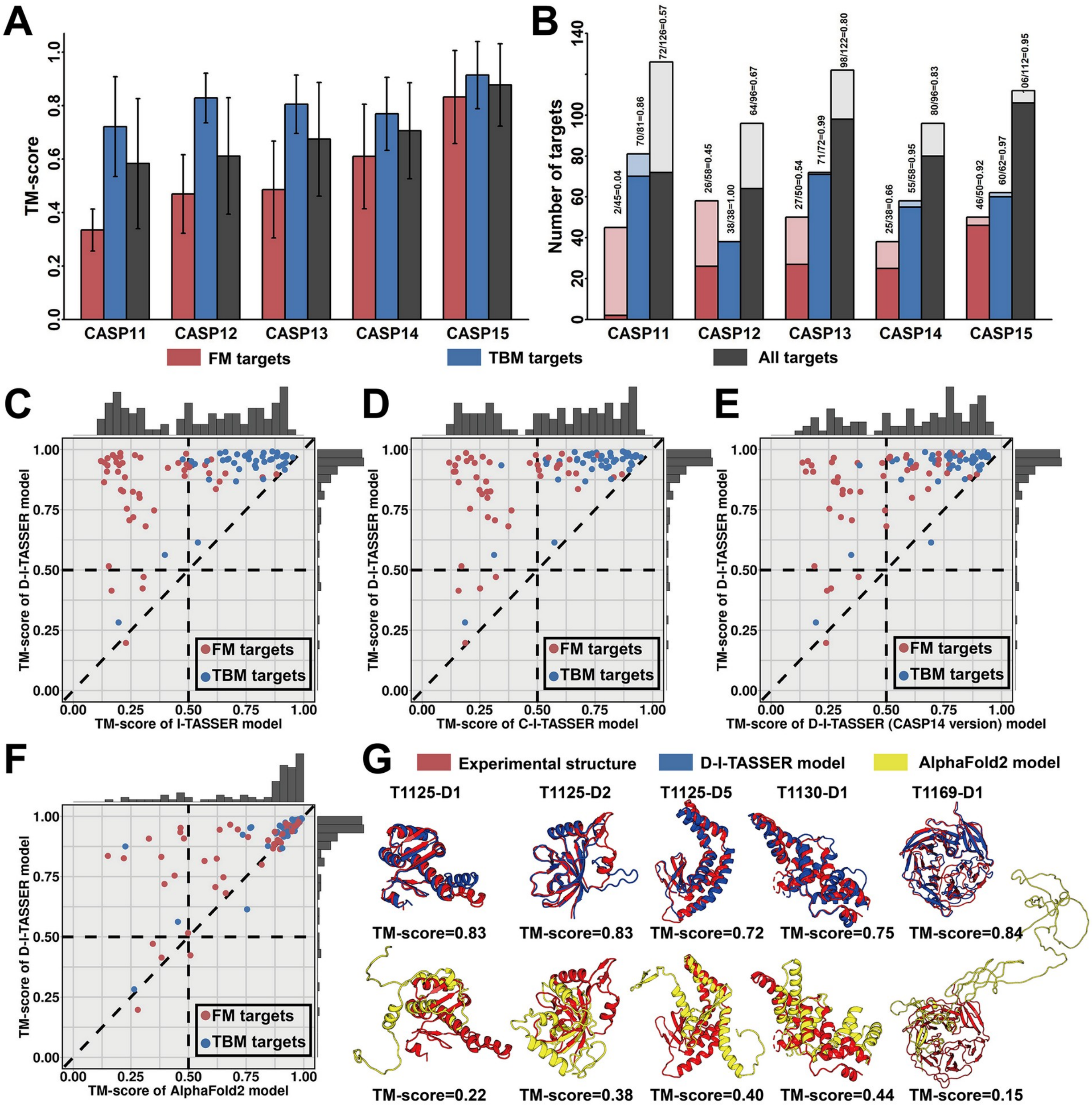

Figure 3. The series of I-TASSER methods for protein monomer predictions in CASP experiments 11-15 (years 2014 – 2013). (A) The TM-scores of a series of I-TASSER algorithms from CASP11 to CASP15, where the error bar means the standard deviation of the TM-scores. (B) The number of foldable targets (dark color) versus non-foldable target (light color) for the I-TASSER algorithms participating in CASP11 to CASP15. (C-F) Head-to-head comparisons between the CASP15 D-I-TASSER algorithm and four methods: (C) I-TASSER, (D) C-I-TASSER, (E) CASP14 version of D-I-TASSER, and (F) AlphaFold2 default method. (G) Five FM targets (T1125-D1, T1125-D2, T1125-D5, T1130-D1, and T1169-D1) where D-I-TASSER constructed correct folds (TM-score>0.5) while AlphaFold2 failed in generating correct models.

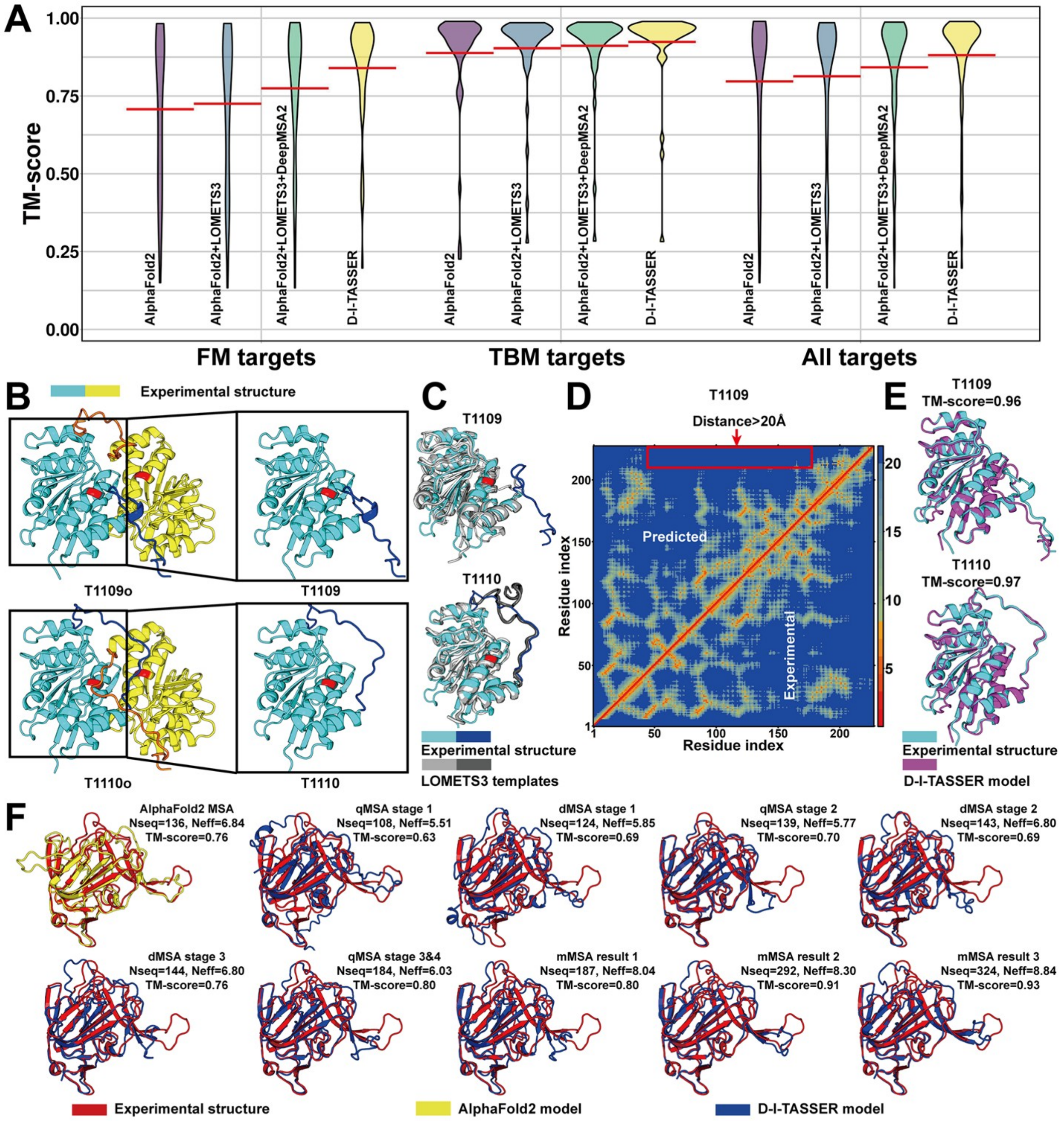

Figure 4. The impact of DeepMSA2 MSA and threading template information on D-I-TASSER. (A) The TM-scores of the first models built by the standard version of AlphaFold2, AlphaFold2 with LOMETS3 templates, AlphaFold2 with LOMETS3 templates and DeepMSA2 MSA, and the D-I-TASSER method. (B) The experimental structures for cases of T1109/T1109o and T1110/T1110o. (C) The LOMETS3 threading templates – PDB 3b38A superposed to T1109 experimental structure, and PDB 3nooA superposed to T1110 experimental structure. (D) The predicted (upper-triangle) and experimental (lower-triangle) distance maps for T1109. (E) The D-I-TASSER models for T1109 and T1110. (F) Comparative performance of the AlphaFold2 model and D-I-TASSER models from different stages of the DeepMSA2 pipeline for T1179.

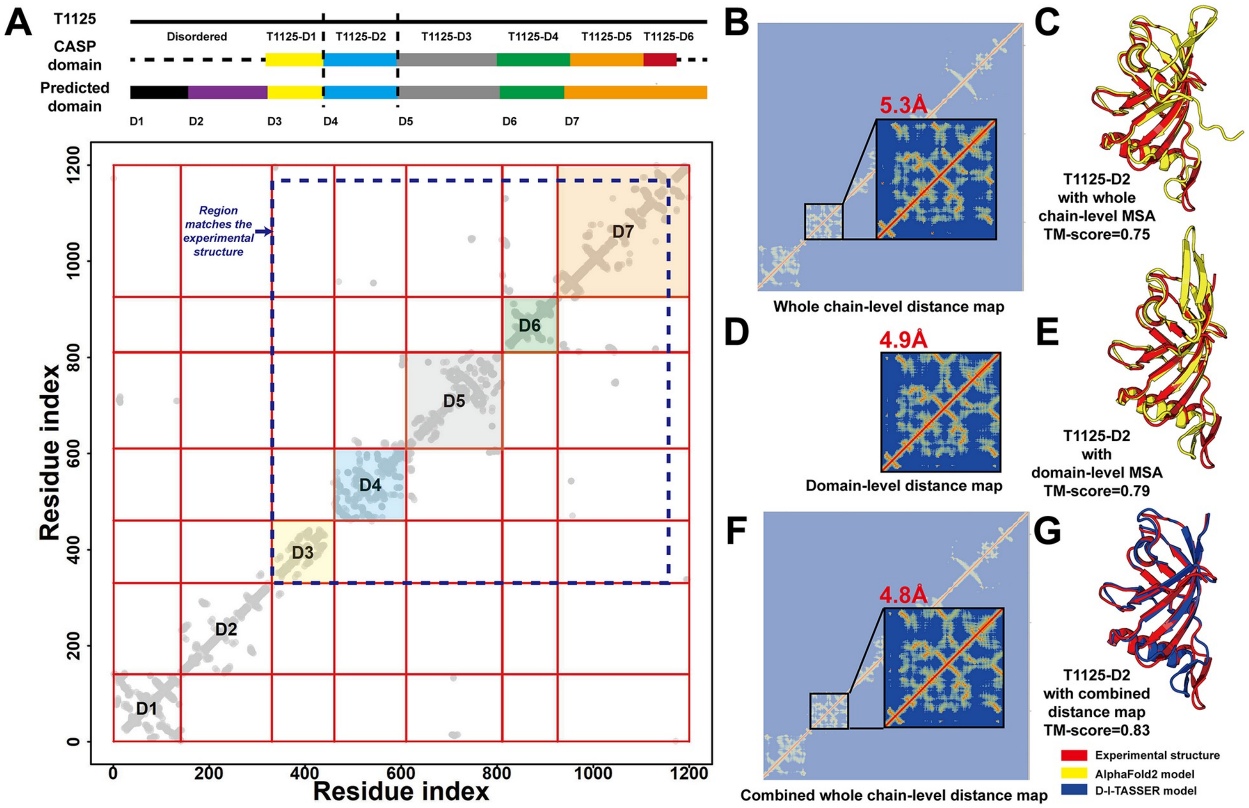

Figure 5. The impact of the multi-domain handling module and the folding simulation on D-I-TASSER in modeling domains from large multi-domain targets. (A) The official (CASP15) and predicted domain boundaries for T1125. (B-C) The whole chain-level distance map and structural model predicted by AlphaFold2 without any domain partitions for T1125-D2. (D-E) The domain-level distance map and structural model predicted by AlphaFold2 with domain partitions for T1125-D2. (F-G) The distance map and structural model predicted by D-I-TASSER that combined the whole chain-level distance map and domain-level distance map for T1125-D2.

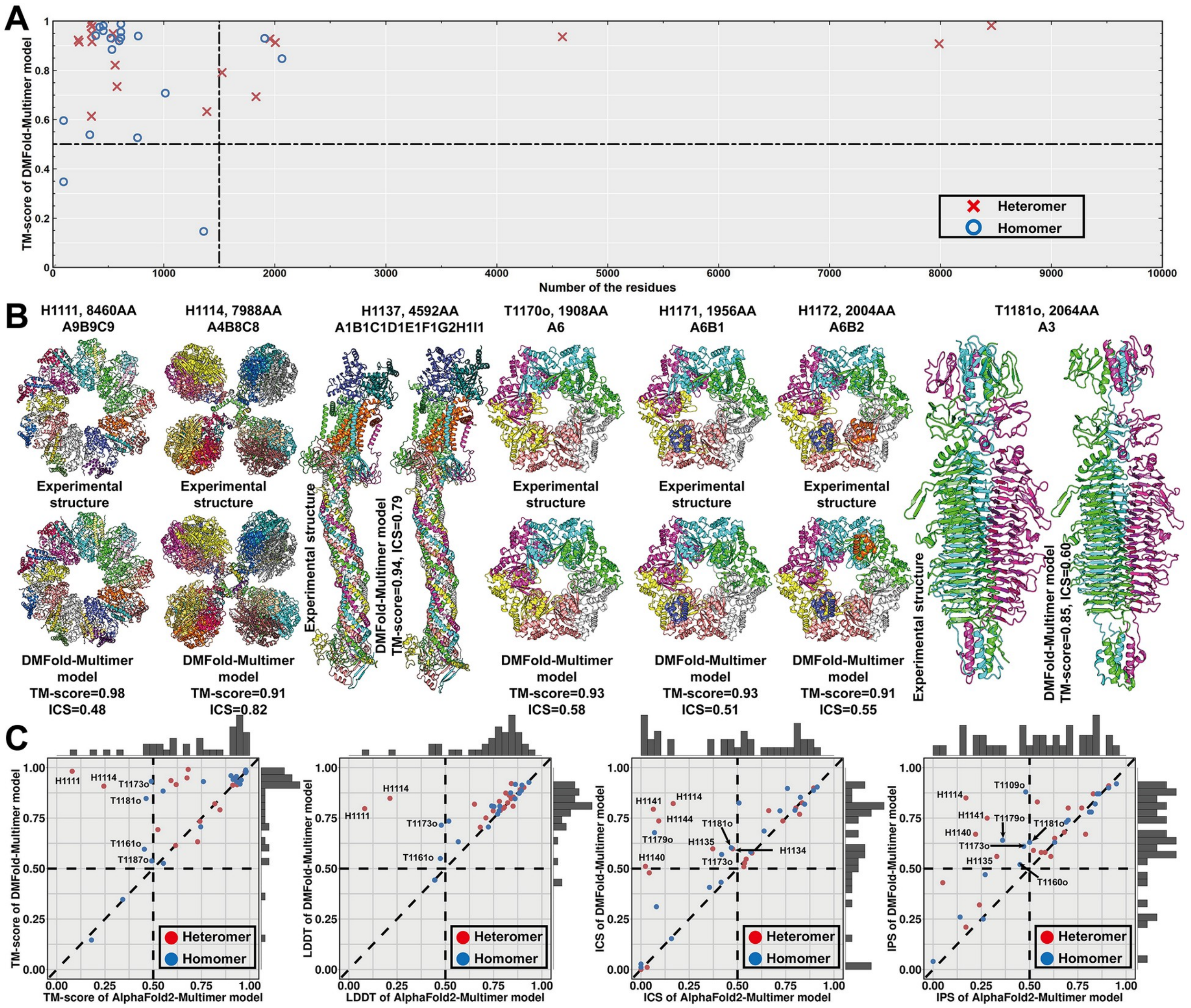

Figure 6. The performance of DMFold-Multimer for protein complex structure prediction in CASP15. (A) The TM-scores of the DMFold-Multimer models vs the target lengths for 38 CASP15 protein complex targets. (B) DMFold-Multimer models associated with the experimentally solved structures for 7 large-size complex targets (>1,500 residues) for which the predictions had a TM-score >0.8. (C) Head-to-head comparison of the modeling quality including TM-score, LDDT score, Interface Contact Score (ICS), and Interface Patch Score (IPS) between DMFold-Multimer and the standard AlphaFold2-Multimer.

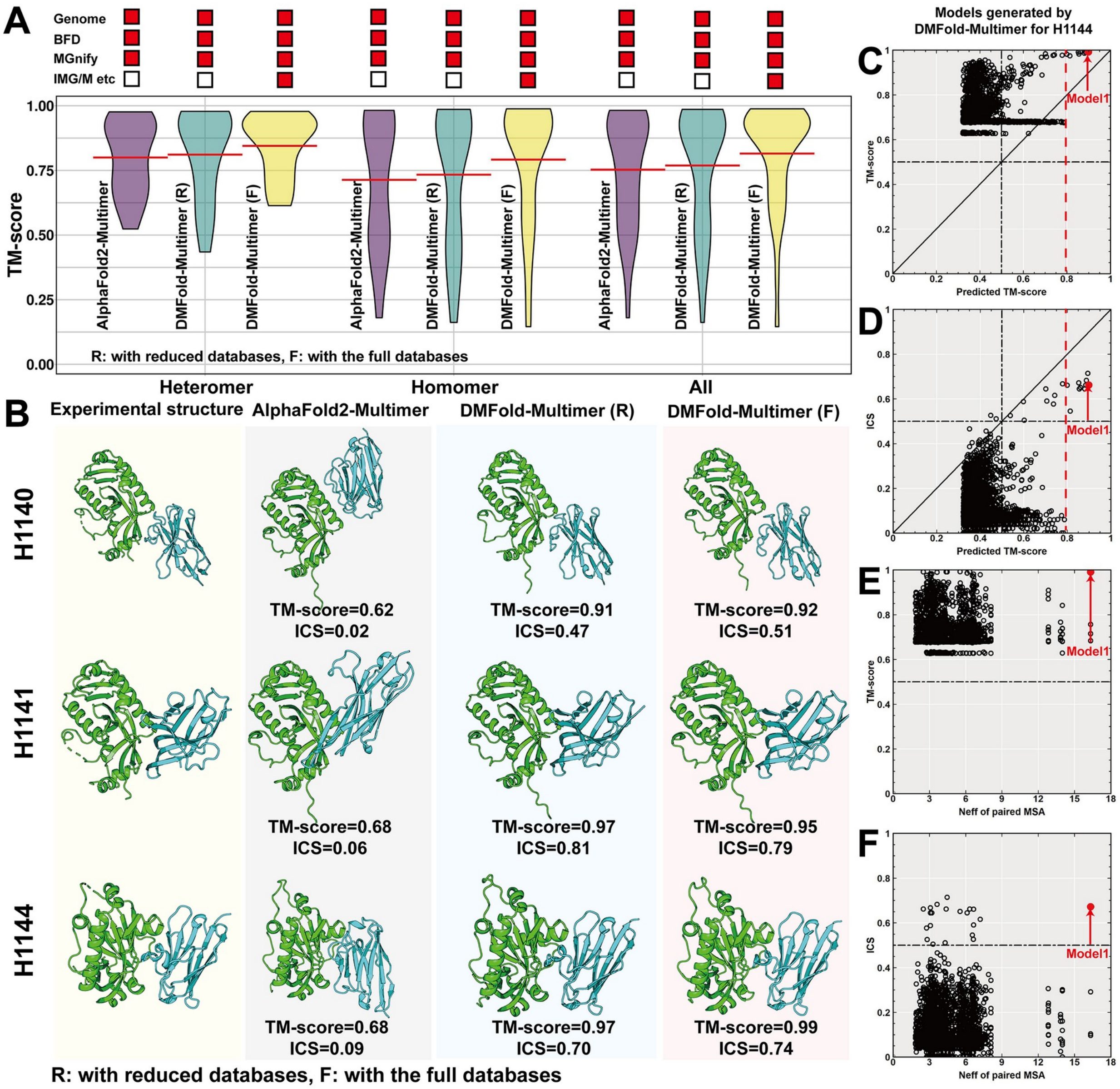

Figure 7. The impact of MSA combination strategy and large-scale metagenomics database on DMFold-Multimer. (A) The complex modeling performance of AlphaFold2-Multimer and DMFold-Multimer using different sequence databases; *n.b.* DMFold-Multimer (R) uses a new MSA selection/pairing strategy but the same databases as AlphaFold2-Multimer. (B) The experimental structures and models predicted by AlphaFold2-Multimer, DMFold-Multimer, and DMFold-Multimer without the in-house metagenomic database for three nanobody-antigen complexes, H1140, H1141 and H1144. (C) The relationship between TM-score and predicted TM-score for H1144. (D) The relationship between ICS and predicted TM-score for H1144. (E) The relationship between TM-score and *Neff* of the paired MSA for H1144. (F) The relationship between ICS and *Neff* of paired MSA for H1144.

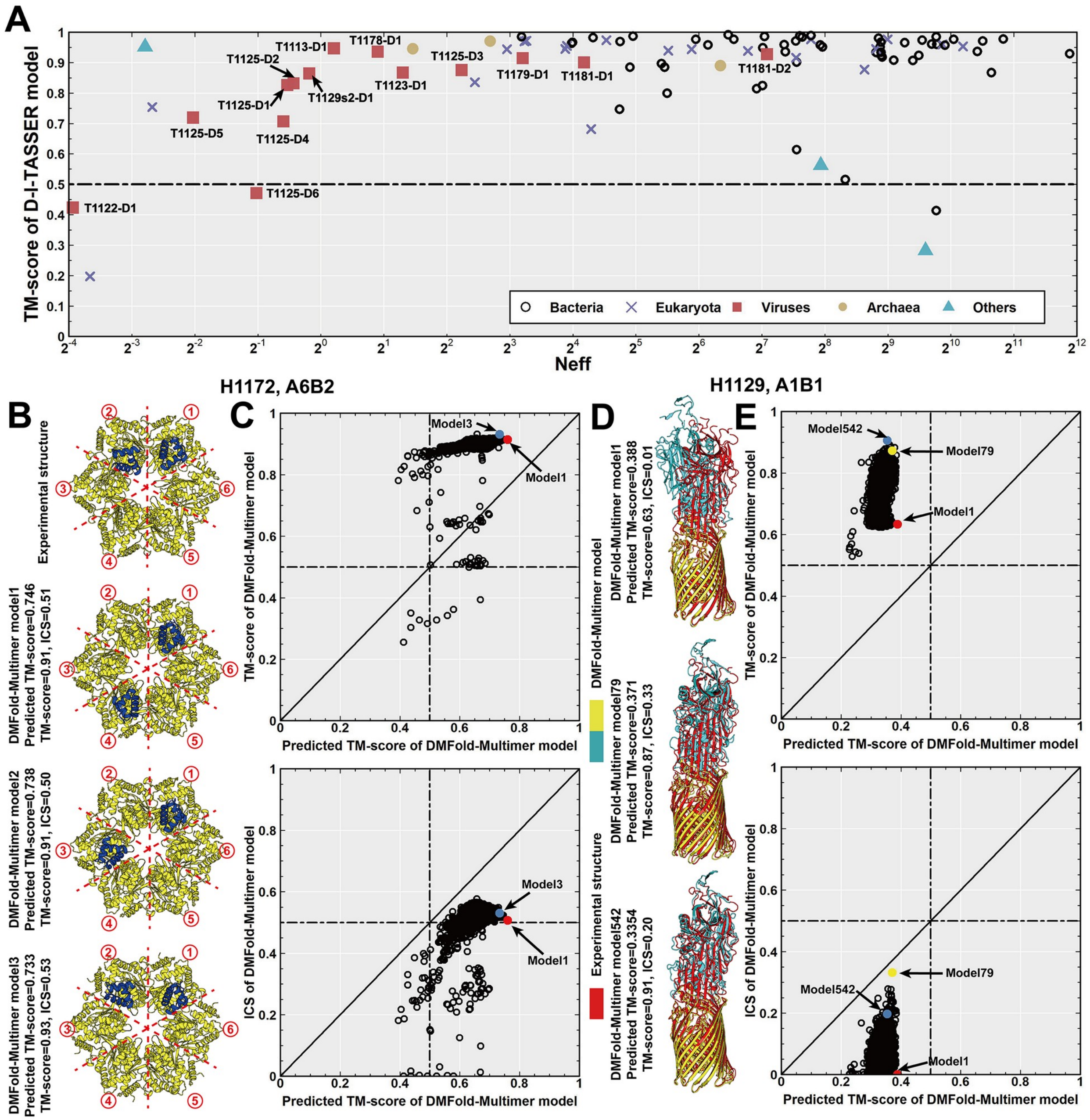

Figure 8. Problems in protein monomer and complex modeling. (A) The relationship between the MSA *Neff* values and TM-scores of the D-I-TASSER models for different taxonomic categories on the 94 domains from 68 full-length monomer targets, where three targets in the 'Others' group are one designed protein and two reconstructed ancient proteins. (B) The experimental structures and models predicted by DMFold-Multimer for H1172**.** (C) The relationship between TM-score (top)/ICS (bottom) and predicted TM-score for H1172. (D) The experimental structures and models predicted by DMFold-Multimer for H1129**.** (E) The relationship between TM-score (top)/ICS (bottom) and predicted TM-score for H1129.