Zheng Wei (Orcid ID: 0000-0002-2984-9003)
Li Yang (Orcid ID: 0000-0003-2480-1972)
Zhang Chengxin (Orcid ID: 0000-0001-7290-1324)
Zhang Yang (Orcid ID: 0000-0002-2739-1916)

Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14

(Short title: Structure prediction by D-I-TASSER/D-QUARK)

Wei Zheng[1, †], Yang Li[1, 2, †], Chengxin Zhang[1, †], Xiaogen Zhou[1], Robin Pearce[1], Eric W. Bell[1], Xiaoqiang Huang[1], Yang Zhang[1, 3, *]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA.

[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing 210094, China.

[3]Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA.

* To whom correspondence should be addressed. Tel: +1 734 647 1549; Fax: +1 734 615 6443; Email: zhng@umich.edu

† These authors contributed equally

ABSTRACT

In this article, we report 3D structure prediction results by two of our best server groups ("Zhang-Server" and "QUARK") in CASP14. These two servers were built based on the D-I-TASSER and D-QUARK algorithms, which integrated four newly developed components into the classical protein folding pipelines, I-TASSER and QUARK, respectively. The new components include: (i) a new multiple sequence alignment (MSA) collection tool, DeepMSA2, which is extended from the DeepMSA program; (ii) a contact-based domain boundary prediction algorithm, FUpred, to detect protein domain boundaries; (iii) a residual convolutional neural network-based method, DeepPotential, to predict multiple spatial restraints by co-evolutionary features derived from the MSA; and (iv) optimized spatial restraint energy potentials to guide the structure assembly simulations. For 37 FM targets, the average TM-scores of the first models produced by D-I-TASSER and D-QUARK were 96% and 112% higher than those constructed by I-TASSER and QUARK, respectively. The data analysis indicates noticeable improvements produced by each of the four new components, especially for the newly added spatial restraints from DeepPotential and the well-tuned force field that combines spatial restraints, threading templates, and generic knowledge-based potentials. However, challenges still exist in the current pipelines. These include difficulties in modeling multi-domain proteins due to low accuracy in inter-domain distance prediction and modeling protein domains from oligomer complexes, as the co-evolutionary analysis cannot distinguish inter-chain and intra-chain distances. Specifically tuning the deep learning-based predictors for multi-domain targets and protein complexes may be helpful to address these issues.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/prot.26193

KEYWORDS: *ab initio* folding, residue-residue distance prediction, deep learning, multiple sequence alignment, domain partition, protein structure prediction, CASP14

1. INTRODUCTION

Despite significant effort and achievement in the field of protein 3D structure prediction, it has remained a central unsolved problem in computational biology. Based on whether homologous template structures from the PDB library are used in the prediction, the computational methods that aim to solve this problem can be categorized as template-based modeling (TBM) or template-free modeling (FM). In previous Critical Assessment of Protein Structure Prediction (CASP) experiments[1-4], two major pipelines, "Zhang-Server" and "QUARK" were developed by our group. The first version of the "Zhang-Server" pipeline was based on the I-TASSER algorithm, which is a TBM method that constructs models by first identifying homologous template structures through alignment of the query sequence to the structures (i.e. "threading") and then assembling those template structures into the full-length model using Replica Exchange Monte Carlo (REMC) simulations[5-7]. The earlier version of the "QUARK" server was based on the QUARK algorithm, which is an FM method that builds models from scratch by assembling short structural fragments without using global template structures[8,9]. Since their development, the capabilities of these two pipelines have been consistently extended by introducing new features and components, most notably, the introduction of contact maps into the assembly simulations. A contact has been defined as a pair of residues where the distance between their Cα or Cβ atoms is ≤ 8Å, provided they are separated by at least five residues in the sequence. Inspired by rapid progress in residue-residue contact map prediction based on multiple sequence alignments (MSAs), the predicted contacts from several deep learning predictors[10-13] were integrated into "Zhang-Server" and "QUARK" through a newly developed contact energy potential[10,14,15], which significantly improved the modeling quality for both servers. However, the model quality quickly approached a ceiling, due to the limited information provided by binary contact prediction (i.e., it can only tell whether the distance between the Cβ atoms from two residues is below 8Å, without providing the actual distance). Thus, how to use deep learning predictors to predict more accurate spatial information, and how to utilize the information to efficiently guide the structural assembly simulations were two central issues that we wanted to address after CASP13.

In CASP14, four newly developed components were integrated into the Zhang-Group servers, which includes "Zhang-Server", "QUARK" and the other three newly developed frameworks ("Zhang-CEthreader", "Zhang-TBM" and "Zhang_Ab_Initio"). First, we extended our iterative MSA construction program, DeepMSA[16], to DeepMSA2 by adding two additional pipelines which search large whole-genome and metagenome databases and a contact-based scoring mechanism to rank MSAs generated by the three MSA pipelines. Second, a predicted contact-based domain partitioning program, FUpred[17], was incorporated into the servers in combination with the previously implemented threading-based domain predictor, ThreaDom[18] to create a more accurate composite domain partition prediction. Third, a deep residual convolutional neural network-based predictor, DeepPotential, was developed to predict residue-to-residue spatial restraints including contact maps, distance maps, inter-residue orientations and hydrogen-bond networks. Finally, new log-odds potentials were designed for implementing these spatial restraint predictions and were subsequently integrated with the existing I-TASSER/QUARK energy potentials, including threading-based energy terms and the inherent knowledge/physics-based potentials, in order to balance the performance for both TBM targets and FM targets. With these new developments, the modeling quality of "Zhang-Server" and "QUARK" has been significantly improved when compared with the pipelines utilized in previous CASPs. The new "Zhang-Server" and "QUARK" pipelines have

been named as D-I-TASSER and D-QUARK, which represent the new distance-guided aspects of I-TASSER and QUARK, respectively.

2. METHODS

2.1 Overview of the Zhang-Group pipelines used in CASP14

In CASP14, Zhang-Group had five servers that participated in the tertiary structure prediction category. These servers are "Zhang-Server" based on the D-I-TASSER algorithm, "QUARK" based on the D-QUARK algorithm, "Zhang-CEthreader" based on DEthreader, "Zhang-TBM" based on LOMETS3, and "Zhang_Ab_Initio" based on the D-QUARK-FM algorithm.

The overall procedures of the protein structure prediction methods used by the Zhang-Group servers during CASP14 are depicted in **Figure 1**. Starting from a query sequence, the domain boundaries were predicted by FUpred[17] and ThreaDom[18], which are based on the predicted contact maps (for FM targets) and the threading template coverage (for TBM targets), respectively. Then, the DeepMSA2 pipeline was used to construct multiple sequence alignments for the full-length protein and each domain sequence by iteratively searching genomics and metagenomics sequence databases. The MSA was then passed to DeepPotential to predict geometric restraints, including contact maps, distances, inter-residue orientations and hydrogen-bond networks, for the domain level and full-length query sequences. These predicted geometric restraints, along with the DeepMSA2 MSA, were used for template detection by LOMETS3 for all query sequences, where the domain-level targets were defined as "Trivial", "Easy", "Hard" or "Very Hard"[19] based on the quality and number of threading alignments detected by LOMETS3. For the "Zhang-Server", "Zhang-CEthreader" and "Zhang-TBM" servers, the decoy conformations were generated by D-I-TASSER folding simulations, which is based on a Replica Exchange Monte Carlo (REMC) simulation guided by the full set of C-I-TASSER force fields[10,15] plus the newly added negative logarithm style distance, orientation and hydrogen bond-network energy potentials. However, the threading templates and spatial restraints used by these three servers were different (See **section 2.8**). For the "QUARK" and "Zhang_Ab_Initio" servers, the decoys were built from D-QUARK folding simulations guided by the full set of C-QUARK[10,14] energy potentials plus the distance and orientation energy potentials, similar to D-I-TASSER. The major difference between "QUARK" and "Zhang_Ab_Initio" is whether threading-based templates/fragments are used in the folding simulations (see **section 2.8** and **Text S1**). Decoy structures from the REMC simulations are then clustered by SPICKER[20] and refined by FG-MD[21] or ModRefiner[22] to derive the full-atom structure models. Lastly, the side-chains of the models are repacked by FASPR[23] to remove steric clashes. For multi-domain proteins, both individual domain structures and a rough full-length structure are first generated by D-I-TASSER/D-QUARK. The individual domain structures are then assembled into the final full-length structure using a modified DEMO[24] protocol that uses the original DEMO energy function, but instead of using structure templates as a guide for domain assembly, the previously generated rough full-length model is used.

2.2 Multiple sequence alignment construction by DeepMSA2

Multiple sequence alignments (MSAs) for the query sequence are generated by DeepMSA2, which utilizes HHblits[25], Jackhmmer[26] and HMMsearch[26] to iteratively search two whole-genome sequence databases (Uniclust30[27] and UniRef90[28]) and four metagenome sequence databases (Metaclust[29], BFD[30], Mgnify[31] and IMG/M[32]) (see **Figure 2A**). DeepMSA2 contains three approaches, dMSA, qMSA, and mMSA), where the dMSA pipeline is short for the DeepMSA[16] pipeline which we used in CASP13.

In the dMSA pipeline, HHblits2, Jackhmmer and HMMsearch are used to search the query against Uniclust30 (version 2017_04), UniRef90 and Metaclust, respectively. In Stages 2 and 3 of dMSA, homologs identified by Jackhmmer and HMMsearch, respectively, are constructed into a custom HHblits formatted database, which is searched through by HHblits2 using the MSA input from the previous stage to generate new MSAs. As an extension of dMSA, qMSA (which stands for "quadruple MSA") is composed of 4 stages that perform HHblits2, Jackhmmer, HHblits3, and HMMsearch searches against Uniclust30 (version 2020_01), UniRef90, BFD, and Mgnify, respectively. Similar to dMSA Stages 2 and 3, the sequence hits from Jackhmmer, HHblits3 and HMMsearch in Stages 2, 3 and 4 of qMSA are converted into HHblits formatted databases, against which the HHblits2 search based on the MSA input from the previous stage is performed. In mMSA (or "multi-level MSA"), the qMSA Stage 3 alignment is used as a probe by HMMsearch to search through the IMG/M database and the resulting sequence hits are converted into a sequence database. This mMSA database is then used as the target database, which is searched by HHblits2 with three seed MSAs (MSAs from dMSA stage 2 and qMSA stages 2 and 3), to derive three new MSAs. These steps result in 10 MSAs in total (i.e., 3 from dMSA, 4 from qMSA, and 3 from mMSA), which are scored by TripletRes[12] contact prediction, where the MSA with the highest probabilities for the top 10*L* (*L* is the sequence length) all range contacts (Cβ-Cβ distances<8Å) is selected as the final MSA.

2.3 Spatial geometric restraint prediction by DeepPotential

The spatial geometric restraints used in the D-I-TASSER and D-QUARK folding simulation include contact maps, distances, inter-residue orientations and hydrogen-bond networks. These four kinds of restraints are predicted by DeepPotential, NeBcon[10,13], ResPRE[11], ResTriplet[33] and TripletRes[12]. Since the later four are previously established pipelines for contact map prediction, we only introduce the newly developed DeepPotential pipeline in this section.

In the DeepPotential pipeline (see **Figure 2B**), a set of co-evolutionary features are extracted from the MSA obtained by DeepMSA2. These co-evolutionary features, which are inherently two-dimensional, include the raw coupling parameters from the pseudo likelihood maximized (PLM) 22-state Potts model[34] and the raw mutual information (MI) matrix. The 22 states of the Potts model represent the 20 standard amino acids, the non-standard amino acid type and the gap state. The corresponding parameters for each residue pair in the PLM and MI matrices are also extracted as additional features that measure query-specific co-evolutionary information in an MSA. The field parameters and the self-mutual information are considered as the one-dimensional features, incorporated with Hidden Markov Model (HMM) features. The one-hot representation of the MSA and other descriptors, such as the number of sequences in the MSA, are also considered. These one-dimensional features and two-dimensional features are fed into deep convolutional neural networks separately, where each of them goes through a set of one-dimensional and two-dimensional residual blocks[35], respectively, and are then tiled together. The feature representations are considered as the inputs of another fully residual neural network which outputs several inter-residue interaction terms. The Cα-Cα contacts, Cβ-Cβ contacts, Cα-Cα distances, Cβ-Cβ distances, Cα-Cβ torsional angle terms, and Cα-based hydrogen-bond network geometry descriptors between residues are considered as prediction terms. The contact, distance, orientations, and hydrogen-bond geometry values are discretized into binary descriptors; using these binary values, the neural networks were trained using cross-entropy loss. For the Cα-Cα distances and Cβ-Cβ distances, four thresholds were selected as the upper range for the distance prediction, including 10Å, 13Å, 16Å, and 20Å, while for inter-residue orientations and hydrogen-bonds, the thresholds are 20Å and 10Å.

2.4 Template detection by LOMETS3

The templates for most of the Zhang-Group servers (except for "Zhang-CEthreader" and "Zhang_Ab_Initio", see **section 2.8**) were detected by LOMETS3, an updated meta-threading server based on LOMETS2[36] that currently contains six profile-based threading methods[37-42] and five contact-/distance-based methods[43-46]. The MSA generated by DeepMSA2 was used to produce sequence profiles (or profile HMMs) for the six profile-based threading methods and to predict contact maps by DeepPotential for the five contact-based threading methods in LOMETS3. The running speed of the contact-based threading methods are much slower than the profile-based threading methods, due to algorithm design limitations. Therefore, to speed up the contact-based threading approaches, we isolated the top 1000 templates identified by HHsearch[39], and then re-ranked only these templates by the five contact-based threading methods individually. For proteins that were defined as "Hard" or "Very Hard" targets by the original LOMETS3 threading methods, the predicted contacts were used to re-rank the templates identified by the profile-based threading methods using the contact map overlap score[44] (CMO). The final 110 templates (10 templates from each individual threading method) were used as the initial conformations for the D-I-TASSER folding simulations.

2.5 Distance and hydrogen-bond energy potential

Three kinds of energy potentials based on the predicted spatial restraints provided by the deep learning predictors, including the distance energy potential, orientation energy potential[47] and hydrogen-bond energy potential, were newly implemented with the full set of C-I-TASSER and C-QUARK force fields to guide the folding simulations in D-I-TASSER and D-QUARK, respectively.

For distances, four upper limit threshold distances were predicted by DeepPotential, including 10Å, 13Å, 16Å, and 20Å. Considering that DeepPotential tends to have a higher confidence for distance models with shorter distance cutoffs, four sets of distance (i.e. likelihoods that the true inter-residue distance fall within predefined distance bins) were generated with distance ranges from [2, 10], [2, 13], [2, 16], and [2, 20] Å, where the four ranges were divided into 18, 24, 30, and 38 distance bins, respectively; only the distance profiles from the lower distance cutoffs were selected, i.e., distances from [2-10) Å were selected from model Set-1, distances from [10-13) Å from Set-2, [13-16) Å from Set-3, and [16-20] Å from Set-4. The combined distances were then converted into a negative logarithm style function used as the distance potential as described by **Eq 1**:

$$E_{disance}(d_{ij}) = -\log(\frac{P_{ij}(d_{ij})+P_{ij}^{N}}{2P_{ij}^{N}}) \quad (1)$$

Here, for a residue pair (*i* and *j*), d_{ij} is the distance between *i* and *j*, which follows a predicted probability distribution P_{ij}, where $P_{ij}(d_{ij})$ is the probability that the distance d_{ij} is located at, and P_{ij}^{N} is the probability of the last distance bin below the upper threshold (i.e., 10Å, 13Å, 16Å, and 20Å).

The inter-residue orientation definitions are the same as defined by previous research[47], and the energy potential used by our servers can be found in **Text S1** and **Eq. S1**.

The hydrogen-bonds[6] used in D-I-TASSER are defined as the inner cross products of two local Cartesian coordinates system formed by a residue pair *i* and *j* (**Figure 2C**). For residue *i*, three unit direction vectors, A_i, B_i and C_i are used to define the local coordinate system that describes the hydrogen direction. Here B_i is the direction vector of the plane formed by three neighboring atoms $C\alpha_{i-1}$, $C\alpha_i$ and $C\alpha_{i+1}$, and A_i, C_i are mutually perpendicular vectors located in the plane. The DeepPotential pipeline predicts the angles between the corresponding unit vectors of residue *i* and residue *j* (i.e., A_i

and A_j, etc.) if the distance between *i* and *j* is below 10Å. The predicted probability distribution of angles is then converted into an energy potential with a similar form as the distance energy, where the potential is described in **Eq 2**:

$$E_{HB}(\theta_{ij}) = -\log\left(\frac{P_{ij}(\theta_{ij})+\varepsilon}{P_{ij}^{N}+\varepsilon}\right) \quad (2)$$

where θ_{ij} is the hydrogen angle between *i* and *j*, which follows a predicted probability distribution P_{ij}, where $P_{ij}(\theta_{ij})$ is the probability that the angle θ_{ij} is located at, and ε =1.0×E-04 is a pseudo count introduced to avoid the logarithm of zero.

2.6 D-I-TASSER folding pipeline

The "Zhang-Server" pipeline in CASP14 was based on the new protein folding algorithm, D-I-TASSER, which is an extension of I-TASSER and C-I-TASSER that integrates deep learning-based distance and hydrogen-bond networks with iterative threading assembly simulations (see **Figure 3A)**.

In the D-I-TASSER pipeline, starting from the query sequence of the domain-level or full-length protein, an MSA is constructed by DeepMSA2. The MSA is then passed to LOMETS3 and DeepPotential for template detection and geometric restraint prediction, respectively. Fragments are extracted from the aligned regions of the template structures and assembled into models using a modified REMC simulation procedure. A force field, which combines the spatial restraints obtained from the LOMETS3 templates and deep learning predictors with the inherent knowledge-based energy terms, is used to guide the D-I-TASSER structural assembly simulations.

Three types of REMC simulations (labeled as 'A', 'M' and 'F') are run depending on a target's category, i.e., 'A' keeps all Cα atoms on a 0.87Å lattice with the REMC simulations starting from random conformations; '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 implemented only for "Trivial" and "Easy" targets whose template alignments have a higher confidence. For each pipeline, five REMC simulations are performed, where the structural decoys from eight (or three for "Hard" and "Very Hard" targets) low-temperature replicas are submitted to SPICKER for structure clustering and model selection.

The SPICKER clusters are refined at the atomic level using fragment-guided molecular dynamics (FG-MD) simulations[21], and finally, the side-chain rotamer structures repacked by FASPR[23]. To select models generated from different pipelines, a set of six model quality assessment programs (MQAPs), including the D-I-TASSER C-score[15], the satisfaction rate of predicted contact maps, structural consensus measured by pair-wise TM-score[48,49], and three statistical potentials (RW, RWplus[50], and Rotas[51]), are implemented. The final model quality is determined by a meta-MQAP consensus score, calculated as the sum of the rank of the six MQAP scores. The top five models with the lowest consensus MQAP scores are selected for submission. The residue level quality of these models is estimated by ResQ[52], a Support Vector Regression-based predictor to predict the deviation of each residue position in the models from the native residue positions.

2.7 D-QUARK folding pipeline

The tertiary structure prediction of the "QUARK" server in CASP14 is based on D-QUARK, an extension of QUARK and C-QUARK, which integrates deep learning-based distance and orientation predictions with replica-exchange Monte Carlo fragment assembly simulations (**Figure 3B**).

In the D-QUARK algorithm, the query sequence is first passed to DeepMSA2 to construct an MSA, which is then used by DeepPotential to generate geometric restraints.

Meanwhile, continuous fragments ranging from 1 to 20 residues are generated by two different approaches. In the "QUARK" server, all fragments are generated by gapless threading of the query through the PDB structure library[8]. In addition to position-specific local fragments, for "Hard", "Easy" and "Trivial" targets, LOMETS3 is utilized for template detection, which are used to collect restraints and create initial conformations for the REMC folding simulations.

Three types of REMC simulations (labeled as 'QE', 'QN' and 'QT') are run depending on a target's category, i.e., 'QE' runs the *ab initio* D-QUARK protocol with initial conformations created from random fragment connection, without including the LOMETS-based restraints in the force field. 'QN' is similar to 'QE' but with the initial conformations created from the LOMETS templates. 'QT' is similar to 'QN' but with LOMETS-based restraints included in the force field. 'QE' is run for Very Hard and Hard targets, 'QN' for Hard and Easy targets, and 'QT' for Easy and Trivial targets, respectively. For each pipeline, five REMC simulations are performed, where the structural decoys from the 10 lowest-temperature replicas are submitted to SPICKER[20] for structure clustering and model selection. The atomic model generation step and model selection step of the D-QUARK algorithm are the same as the D-I-TASSER pipeline, but the models are first refined by ModRefiner[22] before refinement by FG-MD[21].

2.8 Other servers and human groups

In addition to "Zhang-Server" and "QUARK", there were three other servers from the Zhang-Group. Among them, the "Zhang-CEthreader" and "Zhang-TBM" servers used a similar pipeline to D-I-TASSER and "Zhang_Ab_Initio" used a similar workflow to D-QUARK. For the "Zhang-CEthreader" server, the threading algorithm was based on CEthreader[44] and DEthreader, instead of using LOMETS3. The DEthreader algorithm is extended from CEthreader, in which we added a distance map-based energy term to guide the template search through dynamic programming. Here, the predicted distance map for a query is estimated from the DeepPotential residue-residue distance distribution. For a residue pair (i, j), the central distance value of the bin with the largest predicted probability is used as the estimated distance between residue *i* and residue *j*. The "Zhang-TBM" server was based on the LOMETS3 threading algorithm with a major focus on the threading components. For example, the contact map-based threading methods were directly used to scan the whole database instead of a subset (see **LOMETS3** in the **METHODS** section). Because of this strong focus on threading, the folding simulation of the "Zhang-TBM" sever was run for a shorter time compared to the "Zhang-Server" pipeline, and the upper threshold of the predicted distances used for guiding the simulations was 20Å instead of combining multiple distance thresholds together. The "Zhang_Ab_Initio" server focused purely on free modeling, thus only "QE" simulations without any LOMETS3 information were used; in addition, the fragments library was only built from L-BFGS, without any fragments from the PDB database (see **Text S1**).

Two human groups from the Zhang-Group, "Zhang" and "DeepPotential", participated in CASP14. Since "DeepPotential" will be separately reported in another paper in this special issue[53], this section mainly introduces "Zhang". The "Zhang" human group used essentially the same pipeline as our "Zhang-Server" group, except that the whole set of structure models generated by the CASP servers, instead of the in-house LOMETS3 templates, were used as the starting models of the D-I-TASSER pipeline. In addition, a bug in the MSA generation pipeline, which affected the first 22 targets for "Zhang-Server"

(and "QUARK"), was identified and corrected in the Zhang human group.

3. RESULTS

In the results section, we will mainly focus on two of our best performing server groups, "Zhang-Server" (D-I-TASSER) and "QUARK" (D-QUARK), for the analysis since the other three servers and human groups used similar pipelines. 91 domains from 65 targets were assessed in this work. Based on the difficulty of modeling, these 91 domains were categorized as 26 "TBM-easy" targets, 28 "TBM-hard" targets, 14 "FM/TBM" targets and 23 "FM" targets by the official CASP definitions. In the following analysis, we treated "TBM-easy" and "TBM-hard" targets as TBM targets, while "FM" and "FM/TBM" targets were treated as FM targets.

3.1 Overall performance of Zhang-Group servers

Five automatic servers from the Zhang-Group participated in the tertiary structure prediction section of CASP14, the performance of the first models and the best models among the top five submitted models of those servers are listed in **Table S1**. Overall, "Zhang-Server" (D-I-TASSER) and "QUARK" (D-QUARK) were ranked as the best two servers, followed by the "Zhang-CEthreader", "Zhang-TBM" and "Zhang_Ab_Initio" servers based on the average TM-scores of the first models. TM-score is a measure used to assess the global similarity of a structural model relative to its native structure. The TM-score ranges between 0 and 1, with TM-scores $\geq$0.5 indicating that the structure models have correct global topologies. In CASP14, "Zhang-Server" utilized a new folding approach, namely, D-I-TASSER, which takes advantage of the strengths of both threading templates and sequence-based spatial restraints derived from our deep learning approach (DeepPotential). D-I-TASSER is an extended pipeline of the classic template-based I-TASSER algorithm and the contact-associated C-I-TASSER algorithm, where "Zhang-Server" was based on I-TASSER before CASP12, and on C-I-TASSER for CASP12 and CASP13. Similar to "Zhang-Server", the "QUARK" server utilized the D-QUARK pipeline in CASP14, which also combines the restraints from threading templates and the predicted spatial restraints from deep learning. The classic fragment-assembly version of QUARK was used before CASP12 for the "QUARK" server, while the version of C-QUARK based on predicted contacts from direct coupling analysis (DCA) was used in CASP12 and predicted contacts from deep learning approaches was used in CASP13 by the "QUARK" server.

To confirm the effect of implementing the new deep learning-based spatial restraints, especially for the distance maps, orientations and hydrogen-bond networks, into D-I-TASSER and D-QUARK, we ran C-I-TASSER (I-TASSER) and C-QUARK (QUARK) for each CASP14 target using the same domain partitions and the same set of templates used by D-I-TASSER and D-QUARK during CASP14. The results of the head-to-head comparison are shown in **Figure 4**. For the 54 TBM targets, the first models of 43 (46) targets obtained by D-I-TASSER (D-QUARK) were better than the corresponding C-I-TASSER (C-QUARK) models and 45 (45) targets were better than the corresponding I-TASSER (QUARK) models. The average TM-score of the D-I-TASSER (D-QUARK) first models was 0.7757 (0.7694), which was 9% (8%) better than that of the C-I-TASSER (C-QUARK) models with a P-value of 1.44E-07 (6.23E-08), and was 13% (11%) better than that of the I-TASSER (QUARK) models with a P-value of 5.41E-08 (1.55E-07). Interestingly, the improvement from C-I-TASSER to D-I-TASSER (9%) was larger than that from I-TASSER to C-I-TASSER (3%), indicating that adding distance and hydrogen-bond prediction information can provide more helpful restraints than contact information for

TBM targets. A larger improvement can be observed for the 37 FM targets when compared with the TBM targets. For instance, the average TM-score of the first models generated by D-I-TASSER was 0.6055, which was 32% and 96% higher than those of C-I-TASSER's and I-TASSER's first models (P-value = 8.00E-10 and 5.09E-11). In particular, 24 FM targets were foldable[48] (TM-score>0.5) by D-I-TASSER, which was 60% (300%) higher than the number (15 targets) of targets that were foldable by C-I-TASSER (6 targets by I-TASSER) (see **Table S2** for details). In addition, 24 of the 37 FM targets were successfully folded by D-QUARK, while C-QUARK (QUARK) could only fold 11 (2) of the targets. The average TM-score of D-QUARK's first models was 0.6084 for the 37 FM targets, which was 49% (112%) higher (with P-value=1.51E-09 and 8.95E-08) than that for C-QUARK (QUARK) (see **Table S3** for details). These data show that inclusion of the predicted spatial restraints (especially distance maps, orientations and hydrogen-bond networks) from deep learning can improve the modeling performance, in particular for the FM targets.

To more specifically examine the contribution from the deep learning-based distance maps, and hydrogen-bond networks (orientations) to FM targets in the D-I-TASSER (D-QUARK) pipeline, we compared the historical data of CASP FM targets based on the best models generated by either the "Zhang-Server" or "QUARK" server (see **Figure S1**). Here, pure fragment-assembly-based I-TASSER and QUARK could only fold 10% (3 of 30) of the FM targets with an average TM-score of 0.36 in CASP11. With the DCA-based contacts or deep learning-based contacts added to the C-I-TASSER and C-QUARK pipelines, 43% and 66% of the FM targets were foldable in CASP12 and CASP13, respectively. In CASP14, with the further inclusion of residue-residue distance maps, orientations and hydrogen-bond networks to D-I-TASSER and D-QUARK, 70% of the 37 FM targets were foldable, and the average TM-score of the CASP14 FM targets was 77.8%, 39.1% and 18.5% better than those from CASP11-13, respectively. It is remarkable that each time a new feature is added, the average TM-score can be improved by around 0.1 during the last four CASPs. **Figure S2** further summarizes the TM-scores of the best "Zhang-Server" or "QUARK" models vs the target lengths in CASP11-14. In CASP11, only two targets with lengths greater than 100 residues were correctly folded, while in CASP12 and CASP13, 10 and 25 targets with lengths greater than 100 residues were foldable through the introduction of residue-residue contact prediction. It is notable that both C-I-TASSER in CASP13 and D-I-TASSER in CASP14 could fold more than 20 FM targets. Furthermore, the quality of the models generated by D-I-TASSER for targets with lengths ranging from 100-300 residues was better than that for C-I-TASSER. For example, the average TM-score for this length range was 0.6359 for D-I-TASSER models, which was 12% better than that of C-I-TASSER models (0.5665). These data demonstrate that inclusion of deep learning predicted distance maps, orientations and hydrogen-bond networks is able to improve the modeling quality of "Zhang-Server" and "QUARK", especially for the large FM targets.

3.2 Deep learning-based distance prediction and interplaying predicted distance restraints and template information improve the modeling quality

The high accuracy modeling quality of "Zhang-Server" and "QUARK" may be attributed to the newly added deep learning-based spatial restraints predicted by DeepPotential, in particular the predicted residue-residue distance maps.

Figure 5A provides a closer look at the impact of residue-residue distances on the predicted structural models, where the x-axis represents the mean absolute error between distances derived from experimental structures and predicted distances (MAE_n, see definition in **Text S2**) for the long-range top 5L distances (L is the length of the protein) from DeepPotential, and the y-axis represents the quality of

the first models produced by D-I-TASSER (black points) and D-QUARK (grey points). Here, MAE_n represents the predicted distance error, where the lower the MAE_n value is, the better the distance prediction from DeepPotential is. **Figure 5A** shows a strong correlation between the quality of predicted structural models from D-I-TASSER (D-QUARK) and the MAE_n values, as the Pearson correlation coefficient (PCC) is -0.72 (-0.73). It is remarkable that 96% (96%) of the targets were foldable by D-I-TASSER (D-QUARK) when the predicted distance error MAE_n was less than 2Å. Another key factor for producing successful models is whether D-I-TASSER (D-QUARK) simulations can generate models that fit the DeepPotential predicted distances, i.e., whether the newly added distance energy potential can guide the models to fold towards the conformation that fits well with the DeepPotential predicted distances. **Figure 5B** summarizes the relationship between the model quality of D-I-TASSER (D-QUARK) and the model fitting error, MAE_m, which is defined as the mean absolute error between distances calculated from the models and the predicted distances for the long-range top 5*L* distances from DeepPotential (see **Text S2** for details). In general, a better fit (i.e., lower MAE_m values) resulted in better model quality, where the PCC was -0.68 (-0.67) between the model quality and MAE_m value, and similar with **Figure 5A**, when the fitting error was lower than 2Å, 88% (88%) of targets were foldable by D-I-TASSER (D-QUARK).

It is clear that the high accuracy of predicted distances from DeepPotential lead to better final model quality. For example, **T1094-D2**, is an FM target that contained 207 residues. It is an α+β protein with 11 α-helices and 7 β-strands (**Figure 5C**). The best template (PDB ID: 4bj1A) identified by LOMETS3 had a low TM-score of 0.18 (**Figure 5D**), indicating that this target is very difficult to model solely based on the information from threading templates. However, the residue-residue distance map prediction was very accurate for this target (**Figure 5E,** the predicted distance map is shown in the upper triangle matrix and the distance map derived from the experimental structure is shown in the lower triangle matrix), where the MAE_n by DeepPotential was 0.96Å. The D-I-TASSER (D-QUARK) model fit to the predicted distance map very well, as the MAE_m was 0.52Å (0.54Å). As a result, the D-I-TASSER (D-QUARK) model (**Figure 5F**) achieved a very high TM-score of 0.91 (0.91). To demonstrate that the predicted distances were critical for successfully folding this target, we also predicted the structure of this target using C-I-TASSER (C-QUARK) and I-TASSER (QUARK), with the same set of contact predictions and templates utilized in CASP14. As a result, the first models generated by C-I-TASSER (C-QUARK) only had a TM-score of 0.71 (0.70) (**Figure 5G**), while I-TASSER (QUARK) had a TM-score of 0.26 (0.16) (**Figure 5H**). This target demonstrates the importance of predicted distances from deep learning, especially when good templates are not detected.

Even though deep learning-based distance prediction is a critical feature for generating successful predicted models, template information from LOMETS3 is still a very important factor, especially for TBM targets. For example, **T1026-D1**, is a β-protein that contains 146 residues that form 8 β-strands (**Figure 5I**). D-I-TASSER (D-QUARK) folded this target where the final model had a TM-score of 0.78 (0.75) (**Figure 5J**). The successful models produced by D-I-TASSER and D-QUARK were not due to the predicted distances, since the MAE_n of this target was 3.2Å. As shown in **Figure 5K**, the predicted distance map generated by DeepPotential (upper triangle) was obviously different from the distance map derived from the experimental structure (lower triangle). The low accuracy of the predicted distance map for this target was mainly due to poor MSA quality, as the number of detected homologous sequences was 22 and the number of effective sequences (*Neff*, see definition in **Text S3**) was 0.8. Although it was a relatively hard target for deep learning methods to generate correct distance predictions, T1026-D1 was defined as an "Easy" target by LOMETS3. As a result, 20 good templates[48] (TM-score>0.5) were

detected and the best template (PDB ID: 6f2sH) had a TM-score of 0.71 (**Figure 5L**). The I-TASSER (QUARK) model, which depends mainly on the templates, had a TM-score of 0.761 (0.765). Thus, the success of D-I-TASSER (D-QUARK) for this target was mainly from the template contributions identified by LOMETS3. To investigate the influence of LOMETS3 template quality on T1026-D1 modeling, we ran two control tests for this target after CASP14. For the first test, we excluded all 20 good templates and utilized the same predicted distances from DeepPotential that were used in CASP14 by D-I-TASSER. For the second test, we ran the D-QUARK pure *ab initio* folding approach without using any LOMETS3 templates. The purpose of these two tests is to remove the influence of the high-quality templates, and primarily use the predicted distances to guide the simulations. As a result, D-I-TASSER and D-QUARK could only fold this target with TM-scores of 0.43 and 0.41 (**Figure 5M**), respectively. These results suggest that the identification of good templates is still an important component for protein structure prediction when the deep learning predicted distances are not accurate.

We noticed that for two TBM targets, **T1030-D1** and **T1030-D2**, the accuracies (MAE_n=16.3Å and MAE_n=9.6Å) of the predicted distances from DeepPotential were much worse than the rest of the TBM targets (MAE_n<3.3Å). This was because "Zhang-Server" (D-I-TASSER) and "QUARK" (D-QUARK) server had an issue when running DeepMSA2 to generate the MSA for target **T1030**. The failed MSA also affected LOMETS3 (see **METHODS** for LOMETS3), and thus all of the detected templates were incorrect. This made both "Zhang-Server" and "QUARK" fail for these two targets, as the TM-scores of the first models by D-I-TASSER and D-QUARK for T1030-D1 were 0.27 and 0.26, and for T1030-D2 they were 0.40 and 0.32. After CASP14, we re-ran D-I-TASSER (D-QUARK) for target T1030 using the correctly generated MSA by DeepMSA2 (**Figure S3**), where the models for T1030-D1 and T1030-D2 generated by D-I-TASSER (D-QUARK) achieved TM-scores of 0.66 (0.64) and 0.69 (0.62).

The data shown in **Figure 5** demonstrate that the predicted distances from deep learning are a key factor for generating correct models, and template information is also still a very important component of D-I-TASSER and D-QUARK. Furthermore, the two cases for T1094-D2 and T1026-D1, highlight the careful balance of the contributions from templates and predicted distances, which is a major reason why D-I-TASSER and D-QUARK achieved good results in both TBM and FM target modeling.

3.3 Deeper MSA construction improves the modeling quality

The high accuracy of predicted residue-residue distances used in "Zhang-Server" and "QUARK" may be attributed to the newly added deep learning-based spatial restraint prediction method, DeepPotential. The features of DeepPotential are derived from the co-evolutionary information obtained from multiple sequence alignments (MSA) of homologous proteins. Hence, a sufficient number of homologous sequences is critical to ensure the accuracy of predicted distances, and to further ensure the quality of 3D structure construction. In CASP14, we utilized a new MSA construction method, DeepMSA2, which is an extension of DeepMSA, to collect homologous sequences from several metagenomics databases. To examine the influence of DeepMSA2 on the distance prediction and tertiary structure prediction, we ran a control test after CASP14. In the control test, we re-ran D-I-TASSER and D-QUARK using MSAs from the DeepMSA pipeline as input and used the same version of the template library that was used during CASP14. To make a fair comparison, we corrected the modeling results for T1030-D1 and T1030-D2 by DeepMSA2, which was mentioned in **section 3.2**.

Table S4 summarizes the information from MSAs generated by DeepMSA2 and DeepMSA for the 91 CASP14 targets. In general, DeepMSA2 detected more homologs than DeepMSA for both TBM and FM targets. For TBM and FM targets, the number of effective sequences (*Neff*) increased 40% and 150%,

respectively, when utilizing DeepMSA2. These data indicate DeepMSA2 can generate "deeper" MSAs than the former MSA collection pipeline. To check the influence of DeepMSA2 on distance prediction, we made a head-to-head comparison of predicted distance errors (MAE_n) using MSAs from DeepMSA2 or DeepMSA as input, which is shown in **Figure 6A**. By utilizing DeepMSA2 MSAs, the MAE_n of the predicted distances for 62 targets was improved. Particularly for the 37 FM targets, the error of the predicted distances (average MAE_n=2.1Å) from DeepPotential starting from MSAs generated by DeepMSA2 was significantly lower than that (average MAE_n=3.2Å) obtained by starting from MSAs generated by DeepMSA, with a P-value of 4.88E-06 (see detailed data in **Table S5**). These results indicate the MSAs from DeepMSA2 can improve the accuracy of predicted distances from DeepPotential.

To further investigate the impact of DeepMSA2 on protein structure prediction, we show the comparison of the first models generated by D-I-TASSER (D-QUARK) which used the MSAs from DeepMSA2 and DeepMSA as input, respectively, in **Figure 6B** (**Figure 6C**). For D-I-TASSER (D-QUARK), the model quality of 68 (76) targets was improved after using DeepMSA2. Especially for the FM targets, the average TM-score of D-I-TASSER (D-QUARK) models increased from 0.4842 (0.4751) to 0.6055 (0.6084). Such significant (P-value=3.71E-06 and 1.86E-06) improvement for FM targets is understandable, because the accuracy of distance prediction from deep learning is more important for FM targets where threading templates are not reliable. Thus, the quality of the MSAs has a larger effect on the structure prediction for FM targets. The detailed data can be found in **Table S6** (for D-I-TASSER) and **Table S7** (for D-QUARK). **Figure 6D** lists 12 FM targets, including **T1031-D1**, **T1035-D1**, **T1037-D1**, **T1042-D1**, **T1053-D2**, **T1082-D1**, **T1090-D1**, **T1093-D1**, **T1093-D3**, **T1094-D2**, **T1096-D1** and **T1096-D2**, where the TM-score differences of D-I-TASSER (D-QUARK) models were over 0.05 when using different MSA pipelines.

In summary, the newly developed MSA generation pipeline, DeepMSA2, can generate "deeper" MSAs when compared with our CASP13 MSA collection algorithm (DeepMSA). The better quality of the MSAs can help DeepPotential produce more reliable distance prediction, which can further result in more accurate protein structure prediction, especially for FM targets.

3.4 Domain boundary prediction and domain assembly effect structure prediction quality

The domain partitioning and domain assembly procedures remain important factors that affect the structure modeling quality. In CASP14, 19 multi-domain targets were released, and we assessed the domain partition results for 17 of these targets whose experimental structures are available. In the Zhang-Group server pipelines, the domain partitions for multi-domain targets was based on ThreaDom and FUpred, where ThreaDom split domains depending on threading alignment coverage and FUpred predicted the domain boundaries based on contact maps from deep learning (**Figure 1**). FUpred is a newly added domain boundary predictor and achieved state-of-the-art performance on discontinuous domain boundary detection[17]. **Table S8** lists a comparison of the Zhang-Group domain boundary predictions and actual domain splits based on the experimental structures. Here, the normalized domain overlap score[54] (NDO-score) implemented in the former CASP assessment was utilized to assess the domain boundary prediction accuracy. The NDO-score evaluates the overlap between the predicted domain regions and the true domain regions. On average, the NDO-score was 0.865 for the 17 multi-domain targets. Compared with the value for CASP13 targets[10] generated by ThreaDom solely, the NDO-score increased 17%. In particular for the targets containing discontinuous domains, the NDO-score increased 34%. Using the new domain partition pipeline, the performance of boundary prediction for discontinuous domain targets was slightly lower than that for continuous domain targets, as the NDO-

score was 0.82 for discontinuous domains and 0.90 for continuous domains. In contrast, the value for discontinuous domains was much lower than that from continuous domains in CASP13[10] (0.61 vs 0.79).

Here we took **T1094** as an example to illustrate that the correct domain partition and assembly will lead to correct domain-level and full-length models. T1094 was a 496-residue protein that contained two domains. The first domain, **T1094-D1** was a discontinuous domain with two fragments 1-126 and 334-484, and the second domain, **T1094-D2** was a continuous domain starting from residue 127 to 333 (**Figure 7A**). FUpred correctly predicted T1094 as a two domain protein and gave a relatively accurate domain boundary (1-143,298-496 and 144-297) with an NDO-score of 0.76 (**Figure 7B**). As a result, D-I-TASSER (D-QUARK) generated models for these two targets with TM-scores of 0.64 (0.63) and 0.91 (0.91) even though these two targets were FM targets (**Figure 7C**). Furthermore, the relatively good inter-domain distance prediction from DeepPotential (**Figure 7D**) led to a high quality rough full-length reference model (TM-score=0.71) from D-I-TASSER. Finally, the two domain-level models were assembled by DEMO using the reference rough full-length model as the template, and the final assembled full-length model (**Figure 7E**) from D-I-TASSER (D-QUARK) for T1094 had a very good quality with a TM-score of 0.74 (0.73).

On the other hand, the wrong domain partition and assembly can result in poor structure modeling quality. For example, **T1061** contained three domains, T1061-D1, T1061-D2 and T1061-D3 (**Figure S4A**), where the first domain **T1061-D1** was a discontinuous domain with two fragments from residues 1-170 and 442-735. However, ThreaDom and FUpred predicted T1061 to be a five-domain protein with boundaries "1-170;160-455;445-580;570-735;725-838" (**Figure S4B**). That is, T1061-D1was divided into three fragments in our domain partition (here, we name these three fragments as "T1061-D1_f1", "T1061-D1_f1" and "T1061-D1_f3"). For these three fragments, D-I-TASSER generated all correct foldable models with TM-scores of 0.76, 0.67 and 0.74, respectively (**Figure S4C**). However, the distances among these three fragments from the full-length sequences were not accurate, which resulted in the wrong orientation in the final assembled model by DEMO. Thus, the D-I-TASSER model for T1061-D1 only had a TM-score of 0.45 (**Figure S4D**).

In summary, the domain partitioning and domain assembly largely affect the structure modeling quality, particularly for the full-length models. In **Table S8**, we also summarized the TM-scores of the full-length models for each target and the average TM-score of the component domains of each target. The D-I-TASSER and D-QUARK model quality for full-length targets (TM-score=0.583 or 0.586) was considerably worse than the domain-level model quality (TM-score=0.733 or 0.733). Here, full-length model quality depends on the accuracy of predicted inter-domain distances. However, the number of inter-domain distances (<20Å) was much less than that of intra-domain distances, resulting in the difficulty in the training procedure for deep learning. This is why the current results for full-length modeling were worse than domain-level modeling. Thus, designing a deep learning predictor (i.e., enlarging the weights for inter-domain distances in the loss function) that specifically optimizes the accuracy of inter-domain distances could be one solution to solve this issue.

3.5 Problems with modeling proteins from oligomers or complex structures

As we mentioned in **section 3.3**, the targets, especially FM 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 final D-I-TASSER (D-QUARK) models for the 37 FM targets. As shown in **Figure S5**, the final model TM-scores were positively correlated with the MSA *Neff* values. We noticed that five targets had low TM-scores (TM-score<0.5), although most of the targets with *Neff*>8 were foldable by D-I-

TASSER (D-QUARK). Among those five targets, **T1061-D1** had a low TM-score mainly due to the domain partition and assembly problem as we discussed in **section 3.4**, **T1029-D1** had good local fragment packing but wrong N-terminal and C-terminal orientations because of the sparse MSA in the terminal regions (see **Figure S6**), while the remaining three targets were all from protein complexes or oligomers. The failure of modeling these three targets is because the deep learning distance predictor cannot correctly deal with the inter-chain distances and intra-chain distances from complexes or oligomers. Here, we used **T1070-D1** as an example to explain the issue of the current pipeline in modeling proteins from oligomer complexes.

T1070-D1, was a β-protein that contained 79 residues that formed eight β-strands, where three copies (named here as chain A, B and C) of the same monomer protein formed a symmetric oligomer complex, similar to a triple helix (**Figure 8A**). In monomer T1070-D1, the last three β-strands (S6, S7 and S8) formed an anti-parallel β-sheet, while the other five strands (S1-S5) formed a parallel β-sheet with the other strands from the symmetric copies. For example, the strands S1 from chain B, S2 from chain A, and S3 from chain C formed one parallel β-sheet as shown in **Figure 8A**. The topology adopted by T1070-D1 is very different from the generally existing monomer structures, which often form local secondary structure segments by intra-chain residues. The D-I-TASSER model for this target had a very low TM-score of 0.34 (**Figure 8B**). Examining the local segments of the D-I-TASSER model and the T1070-D1 oligomer structure, it can be easily observed that the strand S5 of the D-I-TASSER model (structure in blue in **Figure 8C**) should not be close to strand S6 in the experimental structure (structure in red in **Figure 8C**), but appeared in the position of S5 from another copy (structure by yellow in **Figure 8C**), and formed an anti-parallel sheet with S6. The wrong position of S5 was mainly due to the incorrect distances predicted from DeepPotential. In **Figure 8D**, we show the predicted distance map by DeepPotential and the distance map calculated from the T1070-D1 oligomer complex. Here, the two upper triangle matrices are the predicted distance maps for two T1070-D1 monomer copies, chain A and chain B. The lower triangle matrix shows the distance maps derived from the experimental oligomer structures, containing two intra-chain distance maps for chains A and B, and the inter-chain distance map formed by chains A and B. From the DeepPotential distance map, a 5Å distance between intra-chain residue 39 and residue 54 was predicted. However, in the experimental distance map, the intra-chain distance of these two residues was around 11Å. When we re-checked the inter-chain distance map, the 5Å distance can be observed from residue 54 of chain A and residue 39 from chain B. The co-evolutionary relationship should exist between residues from S6 of copy A and residues from S5 of the neighboring copy B, but the co-evolution-based analysis cannot classify it as inter-chain co-evolution or intra-chain co-evolution when the complex is a homo-oligomer. Since the main input features of DeepPotential are co-evolutionary features, it is not unexpected for the distances to be incorrectly predicted. As a result, during the D-I-TASSER modeling, residue 39 of strand S5 and residue 54 of strand S6 were brought together to form a 5Å distance (**Figure 8E**) as DeepPotential predicted. The other stands (S1-S4) of the D-I-TASSER model had a similar issue as S5 and S6 during the folding simulations, and thus a very compact segment packing can be found in the N-terminal region, while that region of the experimental structure had an extended packing. Thus, the D-I-TASSER model had very poor quality in the N-terminal region.

This case demonstrates that incorrect distance predictions for oligomer structures may lead to incorrect local folding. Furthermore, the incorrect distance prediction is mainly because the co-evolution-based analysis cannot classify it as inter-chain co-evolution or intra-chain co-evolution, which is an

obvious methodology-level limitation in current deep learning-based distance prediction methods that use features from co-evolution-based analysis when dealing with homo-oligomer complexes.

4. CONCLUSIONS

In CASP14, five fully automated protein structure prediction servers from Zhang-Group, including "Zhang-Server", "QUARK", "Zhang-CEthreader", "Zhang-TBM" and "Zhang_Ab_Initio", were deployed. Here, we mainly report the results from two servers, "Zhang-Server" and "QUARK", which were built on the D-I-TASSER and D-QUARK algorithms, respectively. D-I-TASSER and D-QUARK are the extended versions of the previously established I-TASSER and QUARK pipelines with four major developments, including an extended deep MSA generation method, a new domain partition system, a deep learning-based predictor for spatial restraints (contact maps, distance maps, inter-residue orientations and hydrogen-bond networks), and a newly optimized folding force field including balanced deep learning spatial energy potentials, template-based energy potentials, and knowledge-based potentials.

The performance analysis demonstrated that the high model quality of D-I-TASSER and D-QUARK may mainly be attributed to the careful balance of template information from LOMETS3 and spatial restraints (especially the residue-residue distance maps) from the deep learning-based predictor, DeepPotential. Furthermore, the newly developed MSA generation method, DeepMSA2, can generate "deeper" MSAs with more effective sequences by searching more and larger metagenomics databases, and thus produce more accurate evolutionary coupling information for distance prediction. Hence, the MSA generation method is also a vital factor to help improve the accuracy of protein structure prediction, especially for FM targets. Moreover, the new domain partition and assembly system, which combined the threading template-based method ThreaDom and predicted contact-based method FUpred for domain boundary prediction, and DEMO for domain assembly, showed remarkably accurate performance on the domain boundary prediction and the full-length model assembly for multi-domain targets. With the help of these new developed components, the final models from D-I-TASSER and D-QUARK had significantly better accuracies than the models from the previous C-I-TASSER/I-TASSER and C-QUARK/QUARK pipelines, especially for FM targets.

However, significant challenges still remain in the current pipelines. First, although the new domain partition and assembly system works better than what we used in CASP13, multi-domain protein modeling performance was still not satisfactory; in particular the inter-domain distance prediction often had low accuracy, which impacted the full-length model assembly performance. The second problem of the D-I-TASSER and D-QUARK pipelines came from the modeling of protein domains from oligomer complexes, which was mainly because the co-evolution-based features used by DeepPotential could not distinguish inter-chain and intra-chain distances. Thus, how to utilize the specifically tuned deep learning-based predictors for multi-domain proteins or proteins from oligomer complexes should be considered in the future to help address these issues. The third problem is the limited computational resources. Currently, DeepPotential is trained with only a single GPU and 10GB of memory usage. Thus, we have discarded one group of useful features, a raw Precision matrix (PRE), which was used by TripletRes in CASP13 and exhibited excellent performance. Furthermore, we did not consider deeper/wider neural networks. Thus, more computational resources are needed in the future to help improve the accuracy and scalability of DeepPotential. Finally, with the rapid increase in the size of metagenomics databases, the MSA collection stage took longer and required more resources than previous CASPs. For instance, the MSA construction by DeepMSA from the MetaClust (~100GB)

database took around 1 hour using 1 CPU for a 150-residue protein, while it took around 4 hours using 50 CPUs for the same length protein by searching the 5TB IMG/M metagenome database in DeepMSA2. Most recently, we analyzed the metagenome assisted Pfam family structure modeling data and found that there is an inherent linkage between the microbiome niches and their homologous protein families[55]. When using MSAs constructed from an individual biome that is the most closely linked with the target protein family, the amount of memory requested and searching speed of MSA construction was significantly improved, compared to the more expensive MSA search from the whole set of the combined microbiome genome database. Meanwhile, the quality of the 3D structure modeling of the Pfam families was simultaneously improved compared to the latter. This result provides an interesting and promising avenue to improve both quality and speed of future MSA construction, which is especially important when the rapid accumulation of sequences in the metagenome databases makes a comprehensive sequence database search increasingly prohibitive.

ACKNOWLEDGMENT

The authors thank Qiqige Wuyun, Dr. Xiaoqiong Wei, Dr. Xi Zhang, and Zi Liu for discussion and assistance. The authors thank Jonathan Poisson from University of Michigan and Mahidhar Tatineni from XSEDE for the IT supporting. The D-I-TASSER and D-QUARK 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, S10OD026825 to Y.Z), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.), and the National Science Foundation (IIS1901191, DBI2030790, MTM2025426 to Y.Z.).

REFERENCES

1. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics.* 2019;87(12):1011-1020.
2. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics.* 2018;86(S1):7-15.
3. 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.
4. 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.
5. 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.
6. 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.
7. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Research.* 2015;43(W1):W174-W181.
8. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins: Structure, Function, and Bioinformatics.* 2013;81(2):229-239.
9. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics.* 2012;80(7):1715-1735.
10. 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.
11. 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.
12. 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.
13. He B, Mortuza SM, Wang Y, Shen H-B, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics.* 2017;33(15):2296-2306.
14. S. M. Mortuza WZ, Chengxin Zhang, Yang Li, Yang Zhang. C-QUARK: Template-free protein structure modeling using low-accuracy contact-map predictions. in preparation.
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.* 2021:100014.
16. 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.

17. 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.
18. Xue Z, Xu D, Wang Y, Zhang Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics.* 2013;29(13):i247-i256.
19. 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.
20. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry.* 2004;25(6):865-871.
21. Zhang J, Liang Y, Zhang Y. Atomic-Level Protein Structure Refinement Using Fragment-Guided Molecular Dynamics Conformation Sampling. *Structure.* 2011;19(12):1784-1795.
22. Xu D, Zhang Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. *Biophysical Journal.* 2011;101(10):2525-2534.
23. 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.
24. Zhou X, Hu J, Zhang C, Zhang G, Zhang Y. Assembling multidomain protein structures through analogous global structural alignments. *Proceedings of the National Academy of Sciences.* 2019;116(32):15930.
25. 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.
26. 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.
27. 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.
28. 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.
29. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nature Communications.* 2018;9(1):2542.
30. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods.* 2019;16(7):603-606.
31. Mitchell AL, Almeida A, Beracochea M, et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research.* 2020;48(D1):D570-D578.
32. 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.
33. 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.
34. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E.* 2013;87(1):012707.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016.
36. 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.

37. 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.
38. 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.
39. Söding J. Protein homology detection by HMM–HMM comparison. *Bioinformatics.* 2005;21(7):951-960.
40. 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.
41. 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.
42. Meier A, Söding J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLOS Computational Biology.* 2015;11(10):e1004343.
43. Bhattacharya S, Roche R, Bhattacharya D. DisCovER: distance-based covariational threading for weakly homologous proteins. *bioRxiv.* 2020:2020.2001.2031.923409.
44. 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.
45. Buchan DWA, Jones DT. EigenTHREADER: analogous protein fold recognition by efficient contact map threading. *Bioinformatics.* 2017;33(17):2684-2690.
46. Ovchinnikov S, Park H, Varghese N, et al. Protein structure determination using metagenome sequence data. *Science.* 2017;355(6322):294.
47. 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.
48. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics.* 2010;26(7):889-895.
49. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics.* 2004;57(4):702-710.
50. Zhang J, Zhang Y. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. *PLOS ONE.* 2010;5(10):e15386.
51. Park J, Saitou K. ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. *BMC Bioinformatics.* 2014;15(1):307.
52. Yang J, Wang Y, Zhang Y. ResQ: An Approach to Unified Estimation of B-Factor and Residue-Specific Error in Protein Structure Prediction. *Journal of Molecular Biology.* 2016;428(4):693-701.
53. Yang Li CZ, Wei Zheng, Xiaogen Zhou, Eirc W. Bell, Dong-Jun Yu, Yang Zhang. Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. in preparation.
54. Tai C-H, Lee W-J, Vincent JJ, Lee B. Evaluation of domain prediction in CASP6. *Proteins: Structure, Function, and Bioinformatics.* 2005;61(S7):183-192.
55. Yang P, Zheng W, Ning K, Zhang Y. Decoding microbiome and protein family linkage to

improve protein structure prediction. *bioRxiv.* 2021:2021.2004.2015.440088.

FIGURES

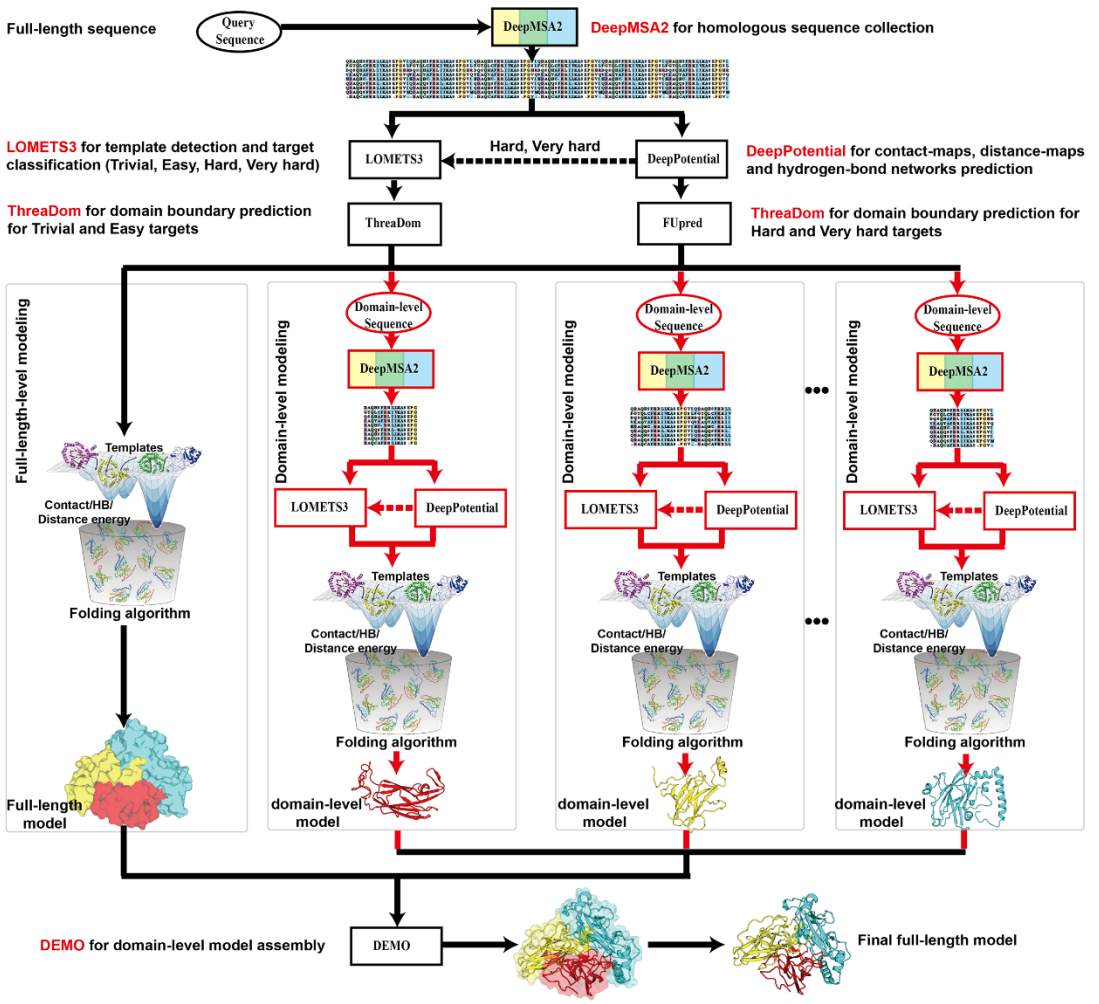


Figure 1. An overview of the common procedures shared by the five automated pipelines of Zhang-Group servers in CASP14 ("Zhang-Server",,"QUARK","Zhang-CEthreder", "Zhang-TBM" and "Zhang_Ab_Initio") on target classification, domain splitting and multi-domain structure assembly.

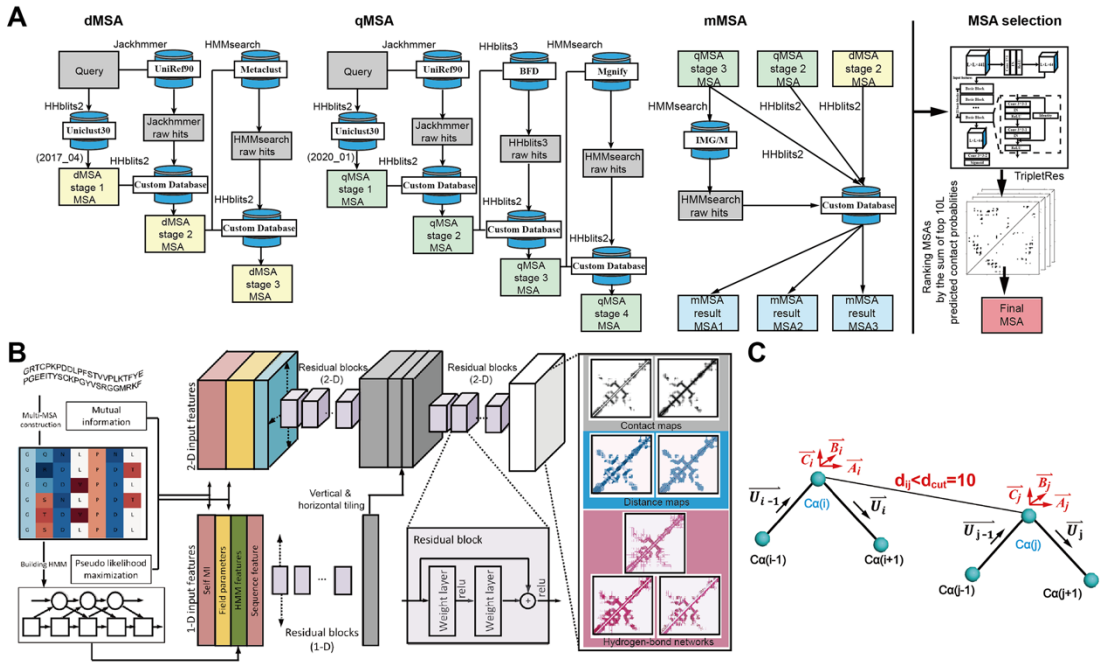


Figure 2. (A) DeepMSA2 pipeline for multiple sequence alignment (MSA) generation, which contains four approaches, dMSA, qMSA, mMSA and MSA selection. (B) DeepPotential pipeline for generating spatial geometric restraints, which include contact maps, distances, orientations and hydrogen-bond networks. (C) The illustration of the hydrogen-bond definition used in D-I-TASSER and D-QUARK.

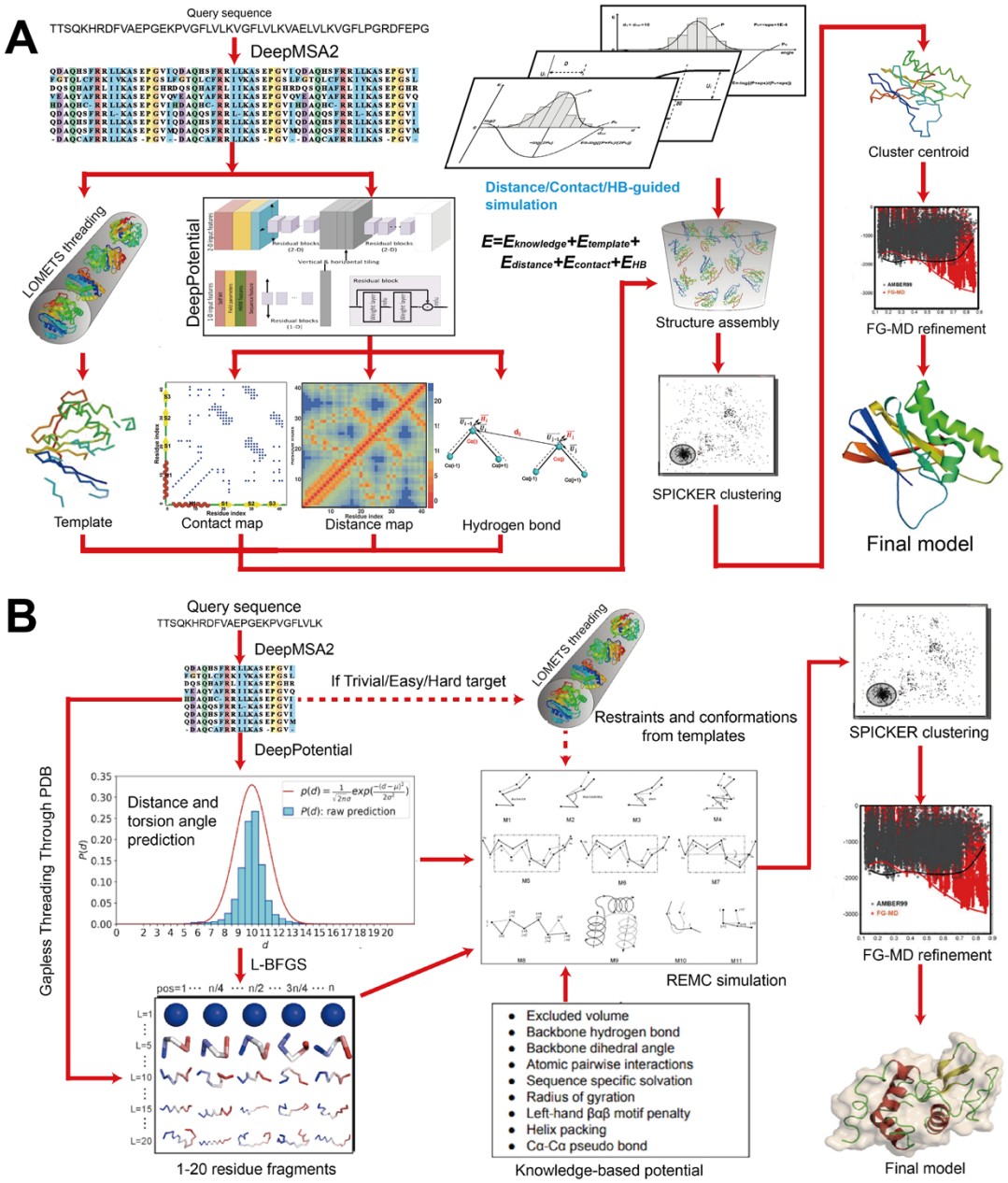


Figure 3. (A) D-I-TASSER pipeline, which is an extension of I-TASSER and C-I-TASSER that integrates deep learning-based distance and hydrogen-bond networks with iterative threading assembly simulations. (B) D-QUARK pipeline, which is an extension of QUARK and C-QUARK that integrates deep learning-based distance and orientation predictions with replica-exchange Monte Carlo fragment assembly simulations.

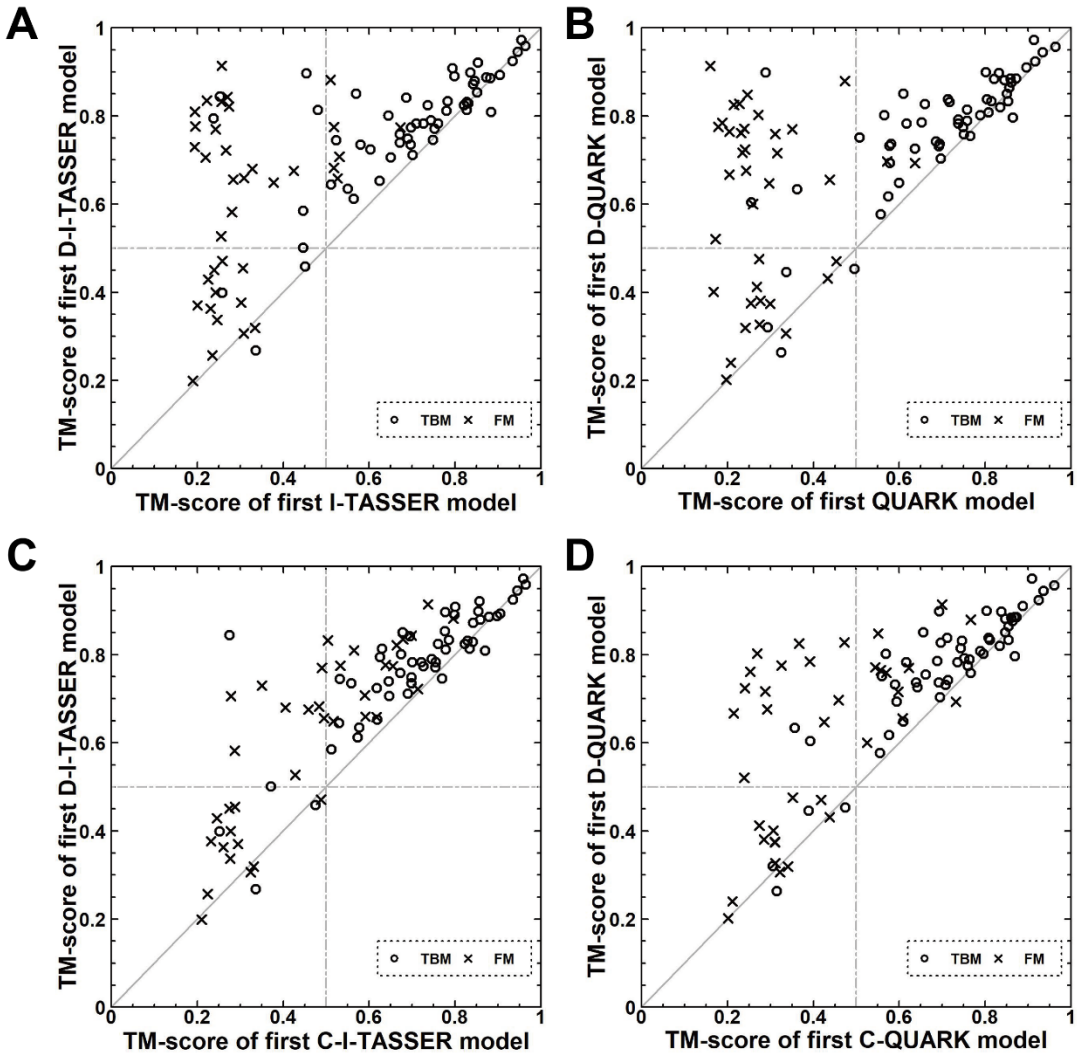


Figure 4. Head-to-head comparisons between (A) D-I-TASSER and I-TASSER, (B) D-QUARK and QUARK, (C) D-I-TASSER and C-I-TASSER, (D) D-QUARK and C-QUARK. C-I-TASSER, I-TASSER, C-QUARK, and QUARK were run using the same domain partitions and the same set of templates used by D-I-TASSER and D-QUARK during CASP14.

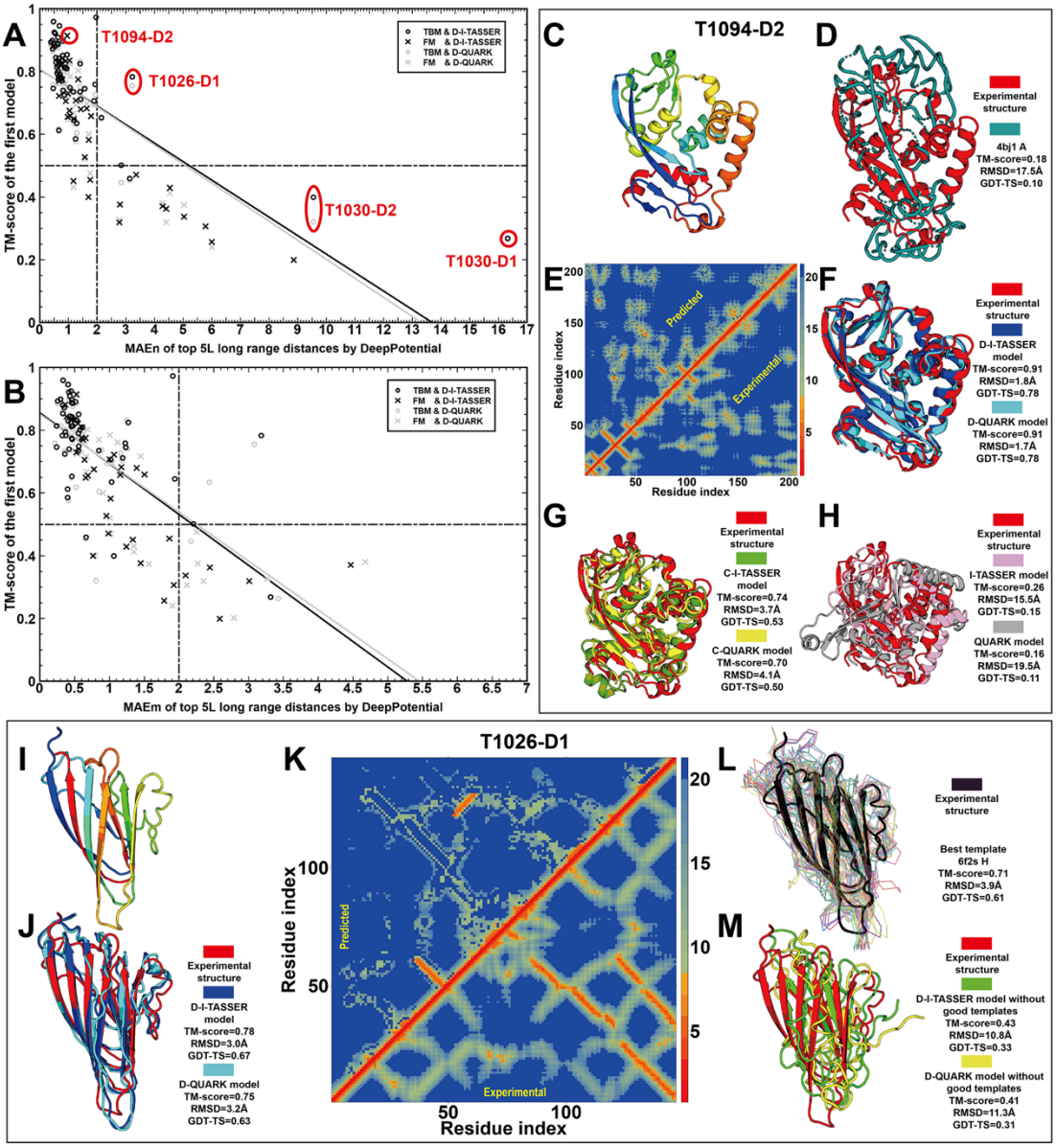


Figure 5. (A) The relationship between the model quality of D-I-TASSER/D-QUARK and MAE_n, which represents the mean absolute error between distances derived from the experimental structures and predicted distances for the long-range top 5*L* distances (*L* is the length of the protein) from DeepPotential. (B) The relationship between the model quality of D-I-TASSER/D-QUARK and MAE_m, which is defined as the mean absolute error between the distances calculated from the model and predicted distances for the long-range top 5*L* distances from DeepPotential. (C) The experimental structure of T1094-D2. (D) The superposition between the experimental structure and the best template (PDB ID: 4bj1A) identified by LOMETS3 for T1094-D2. (E) The residue-residue distance map prediction for T1094-D2, where the predicted distance map is shown in the upper triangle matrix and the distance map derived from the experimental structure is shown in the lower triangle matrix. The D-I-TASSER and D-QUARK models (F), the C-I-TASSER and C-QUARK models (G) and the I-TASSER and QUARK models (H) of T1094-D2 superposed with the experimental structure. (I) The experimental structure of T1026-D1. (J) The D-I-TASSER and D-QUARK models of T1026-D1 superposed with the experimental structure. (K) The residue-residue distance map for T1026-D1, where the predicted distance map is shown in the upper triangle matrix and distance map calculated from the experimental structure is shown in the lower triangle

matrix. (L) The superposition of the experimental structure and the high-quality templates identified by LOMETS3 for T1026-D1. (M) The D-I-TASSER and D-QUARK models for T1026-D1 after excluding good templates superposed with the experimental structure.

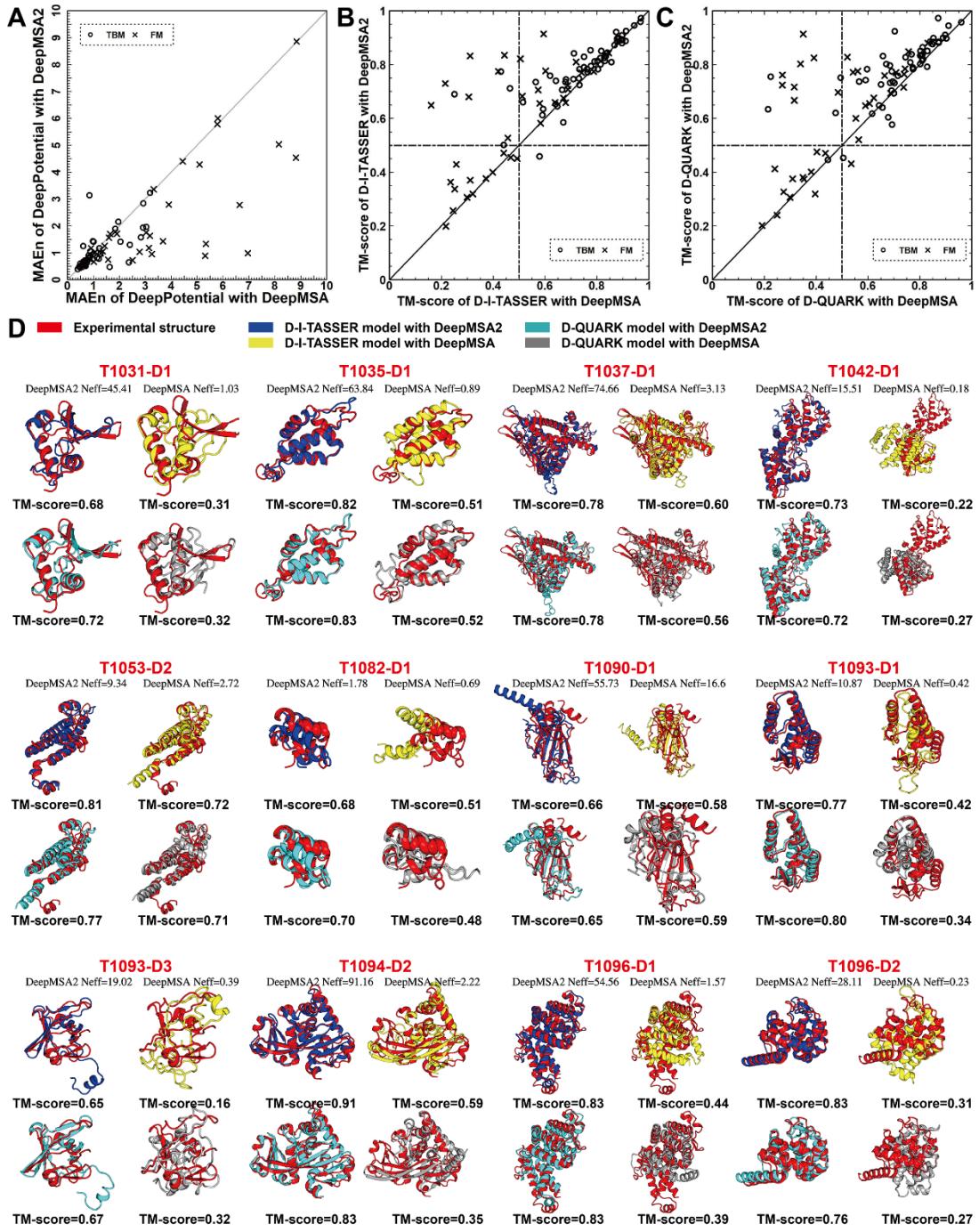


Figure 6. (A) The head-to-head comparison of predicted distance errors (MAE_n) between MSAs from DeepMSA2 and DeepMSA. (B) The head-to-head comparison of the model quality generated by D-I-TASSER between MSAs from DeepMSA2 and DeepMSA. (C) The head-to-head comparison of the model quality generated by D-QUARK between MSAs from DeepMSA2 and DeepMSA. (D) 12 FM targets, where the TM-score differences of the D-I-TASSER (D-QUARK) models were over 0.05 when using different MSA pipelines.

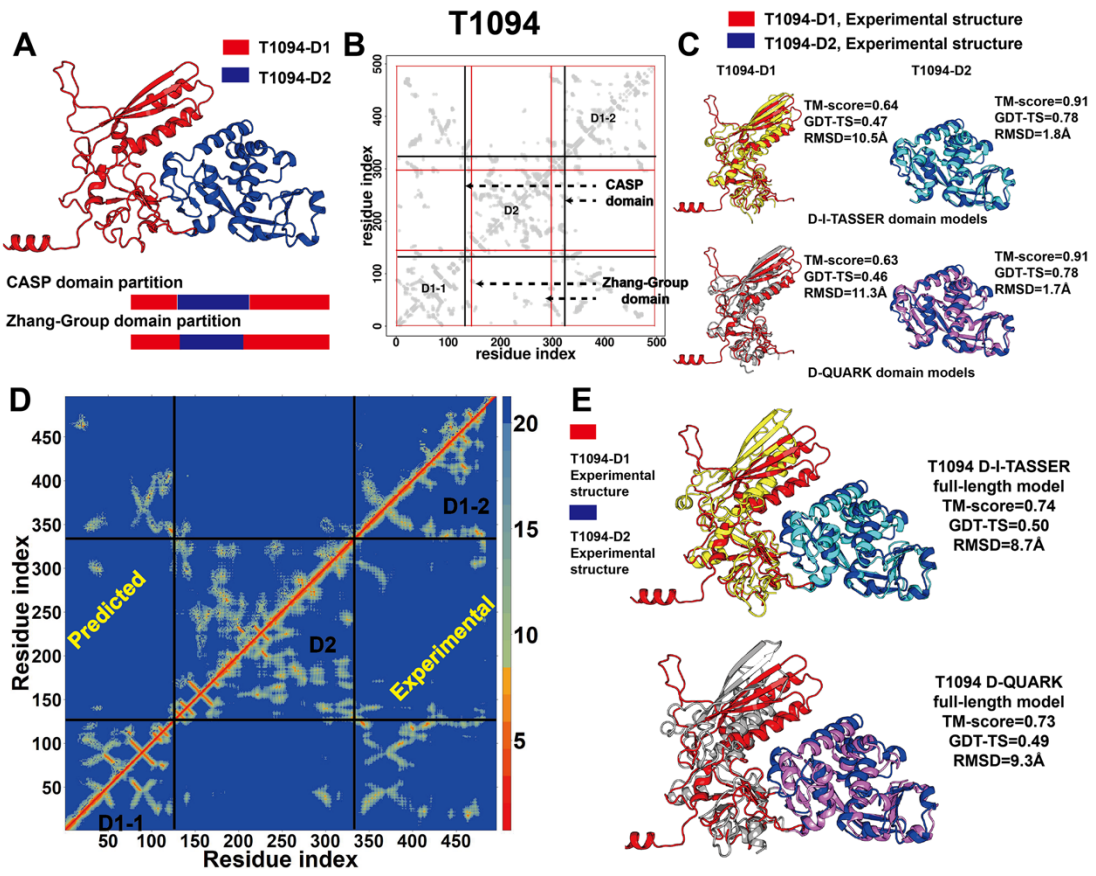


Figure 7. (A) The experimental structure and domain partition for T1094. (B) The illustration of predicted domain boundaries by FUpred based on the DeepPotential contact map. (C) The D-I-TASSER and D-QUARK models for two domains of T1094 superposed with the experimental structures. (D) The residue-residue distance map predicted from DeepPotential (upper triangle) and the distance map calculated from the experimental structure (lower triangle) for T1094. (E) The D-I-TASSER and D-QUARK full-length models of T1094 superposed with the experimental structures.

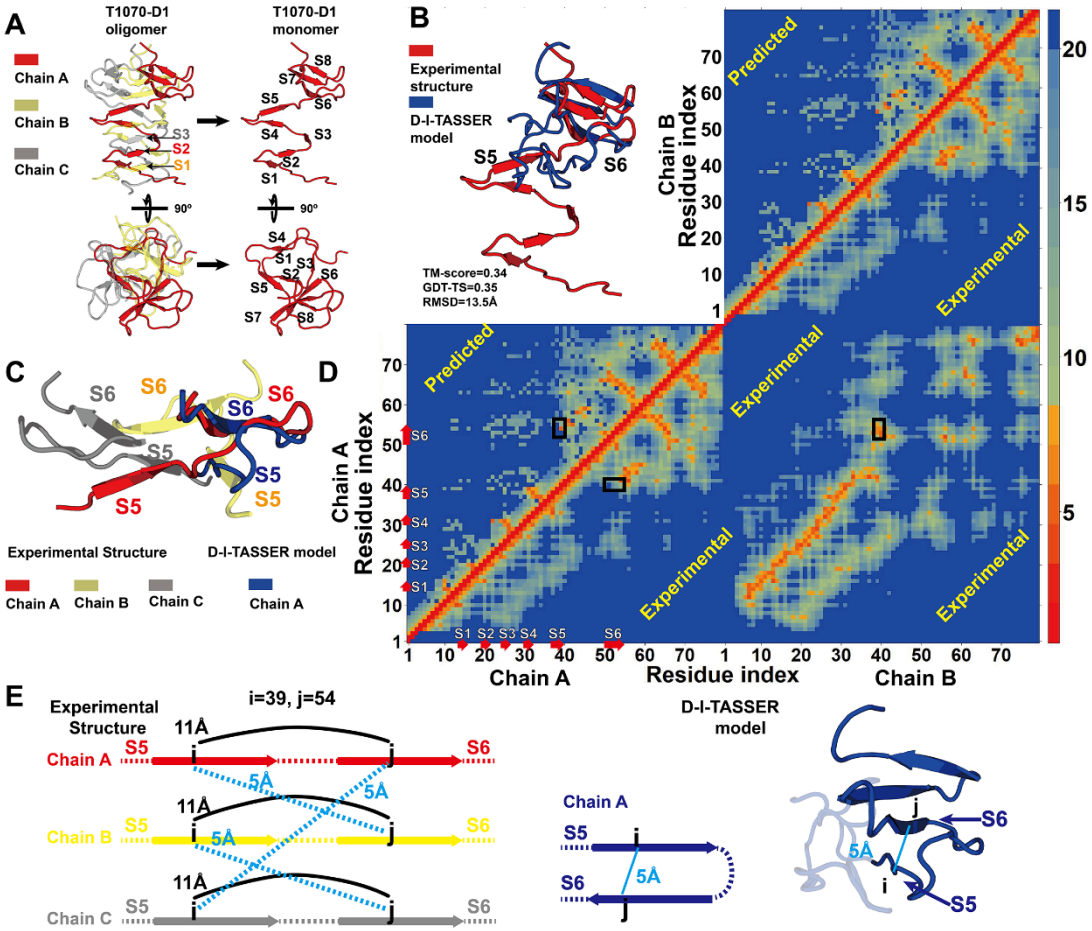
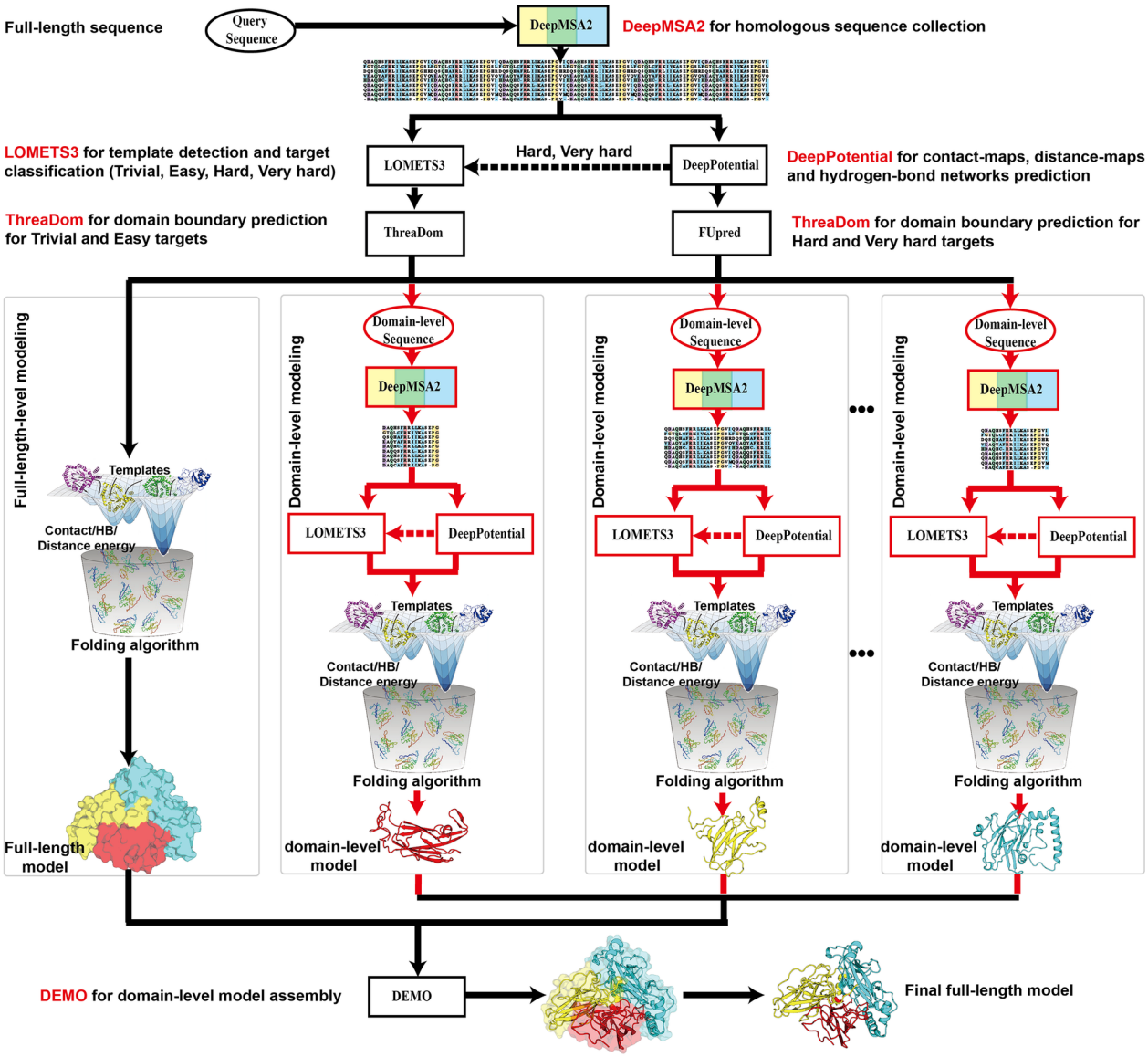
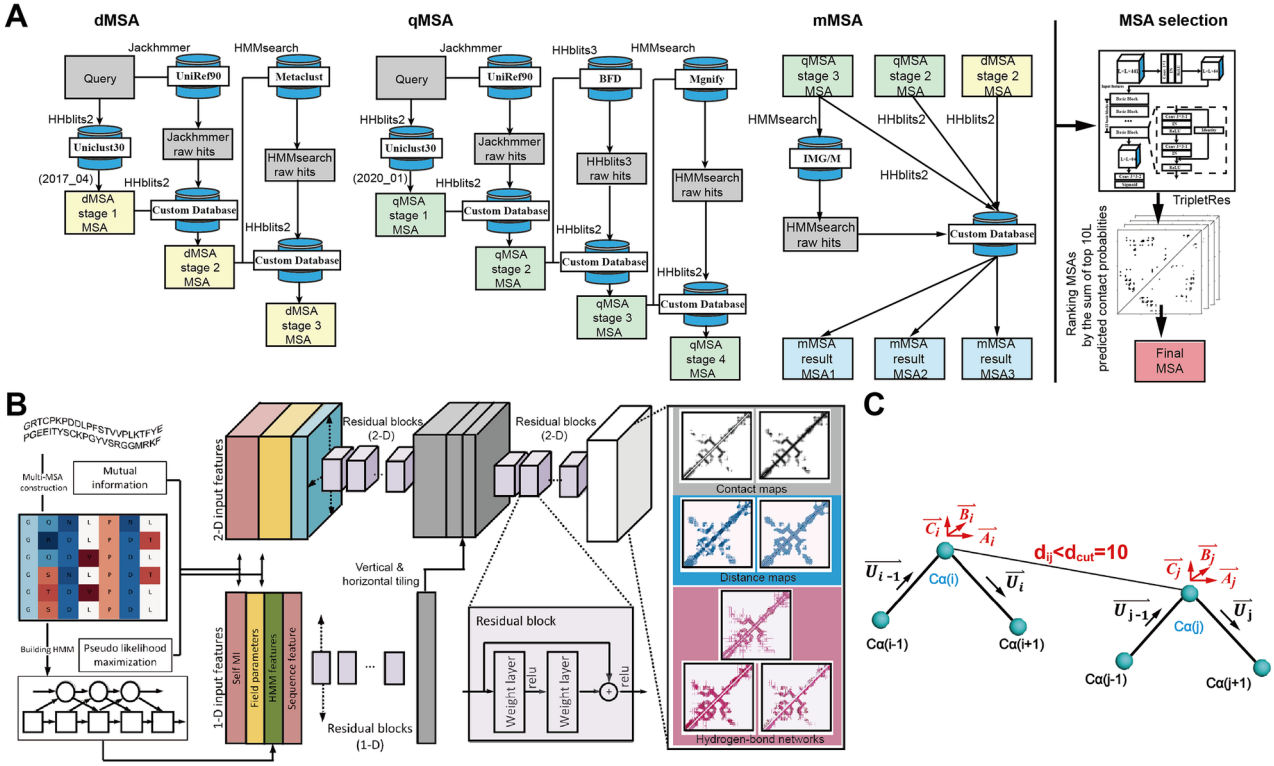


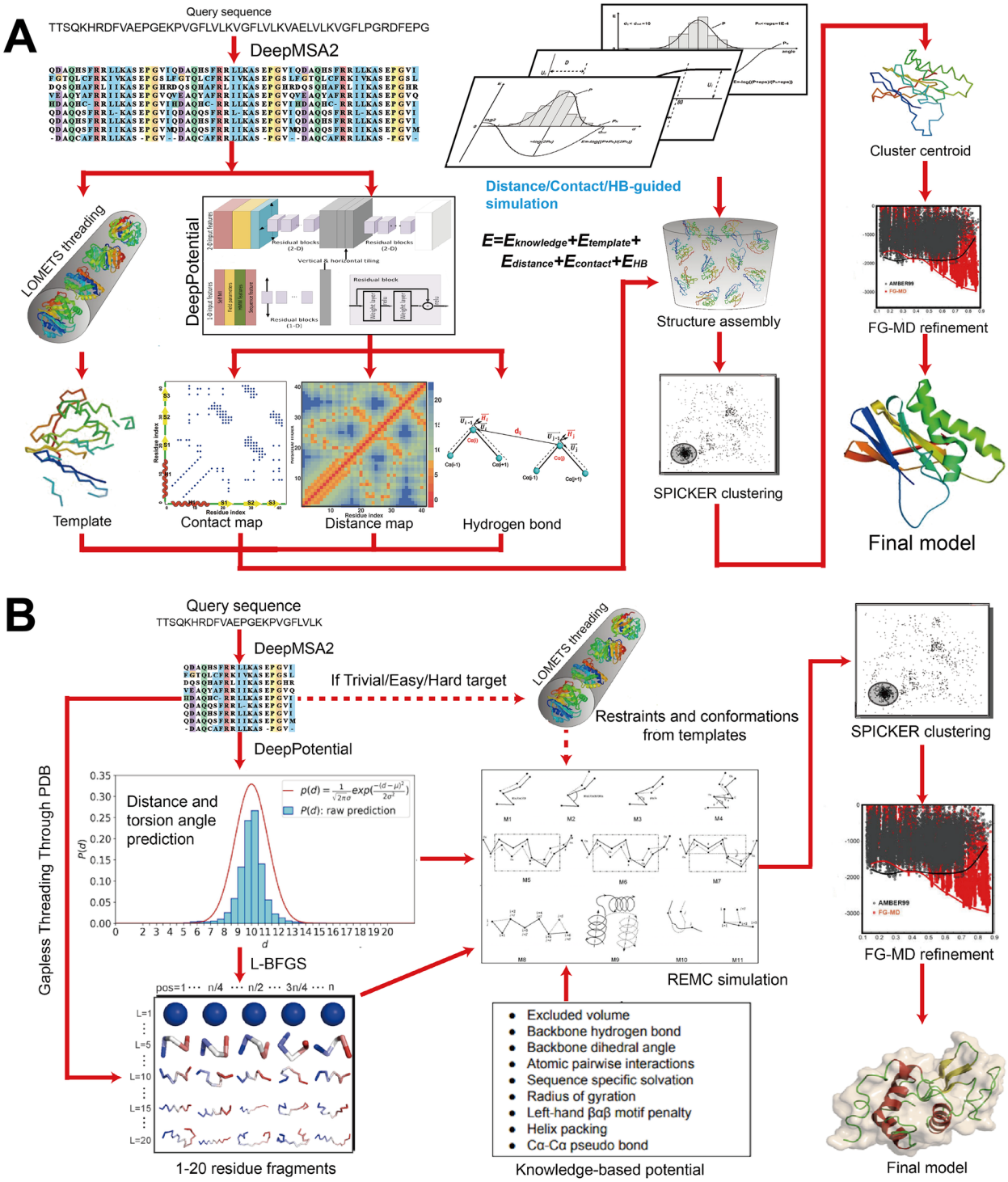
Figure 8. (A) Three copies (named here as chain A, B and C) of the same monomer protein of T1070-D1 form a symmetric oligomer complex. (B) The D-I-TASSER model of T1070-D1 superposed with the experimental structure. (C) The local segments of β-strands S5 and S6 from the D-I-TASSER model and the T1070-D1 oligomer structure. (D) The predicted distance map by DeepPotential and the distance map calculated from the T1070-D1 oligomer complex. The bottom left and upper right matrices are two intra-chain distance maps for two T1070-D1 monomer copies, chains A and B, respectively, where the two upper triangle matrices are the predicted distance maps and the lower triangle matrices are derived from the experimental oligomer structures. The bottom right matrix is the inter-chain distance map formed by chains A and B which was calculated from the T1070-D1 oligomer complex. (E) The illustration of the intra-chain and inter-chain distances between residue 39 and residue 54 in the experimental structure and the D-I-TASSER model.



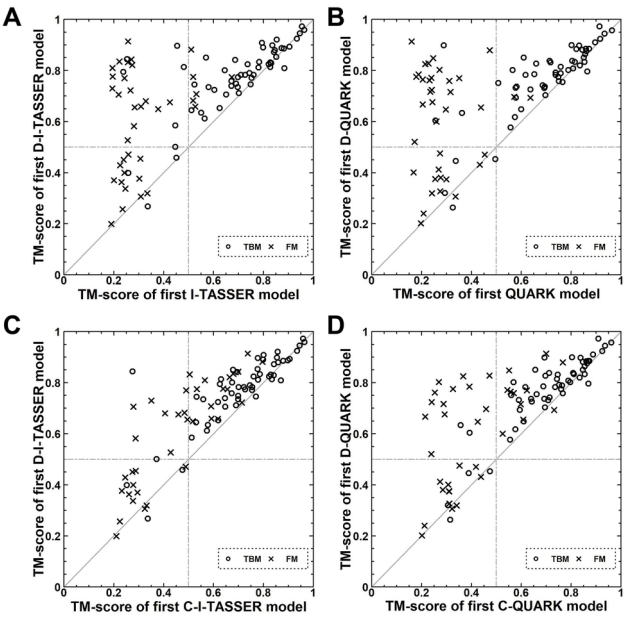
PROT_26193_Figure_1.tif



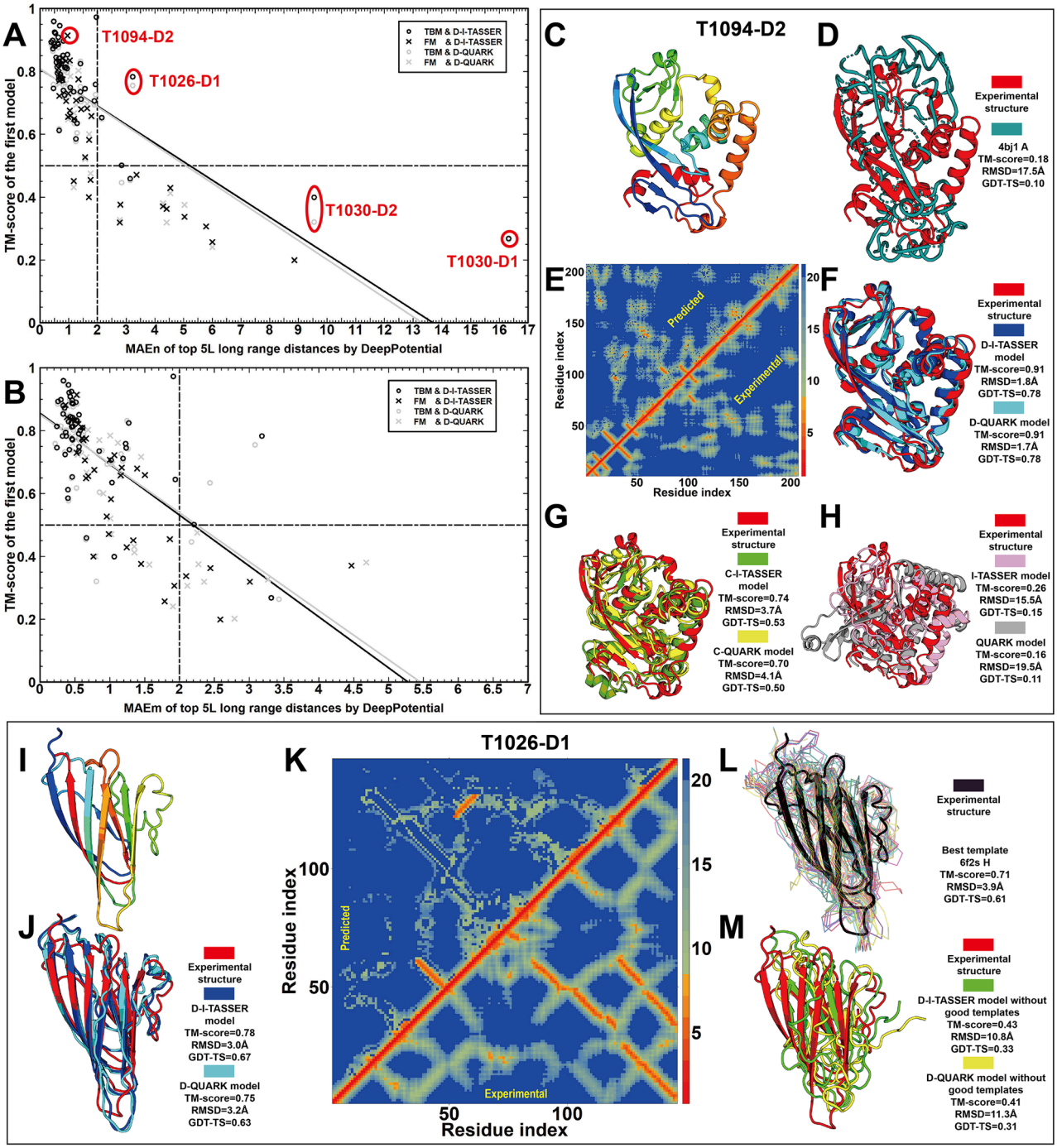
PROT_26193_Figure_2.tif



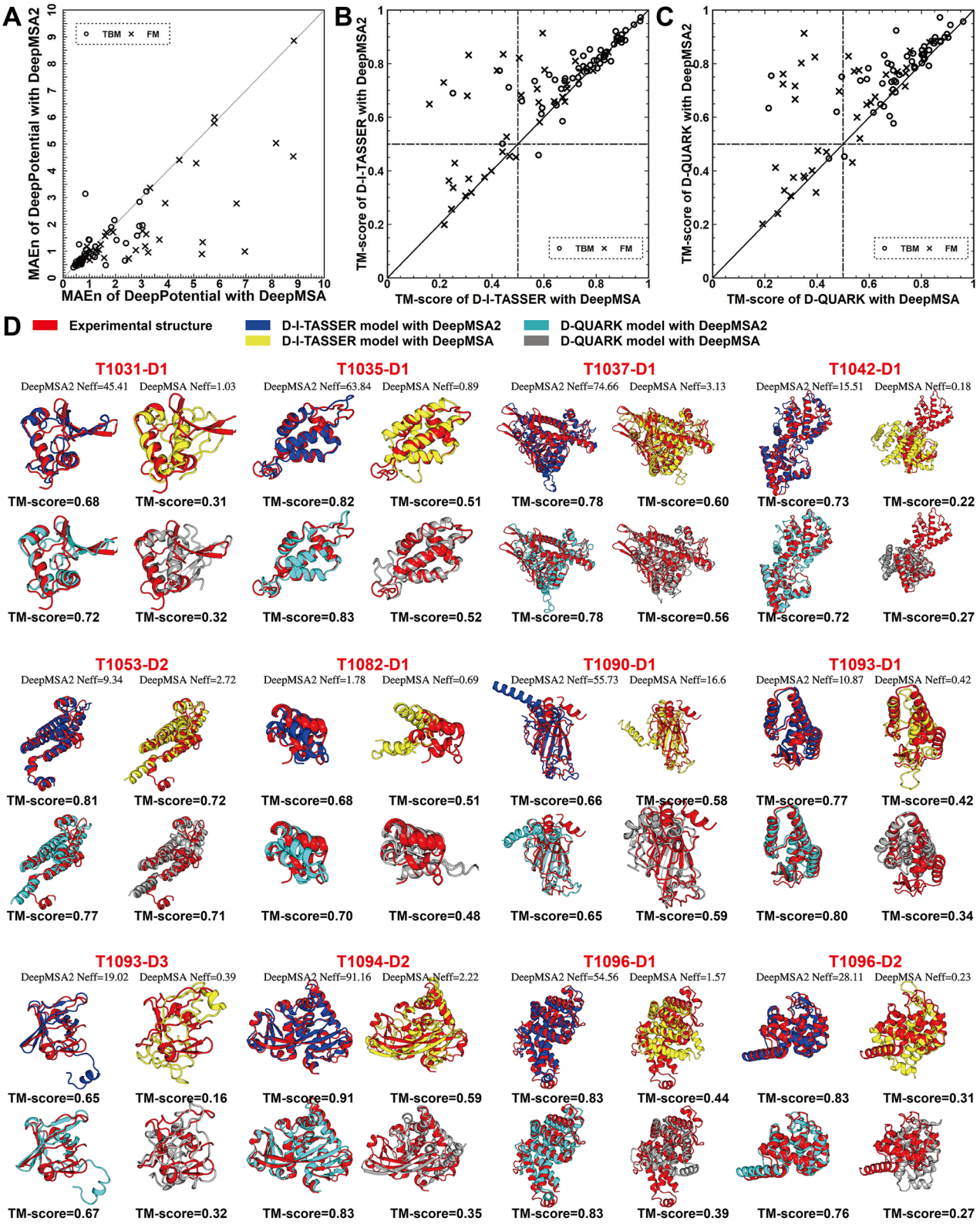
PROT_26193_Figure_3.tif



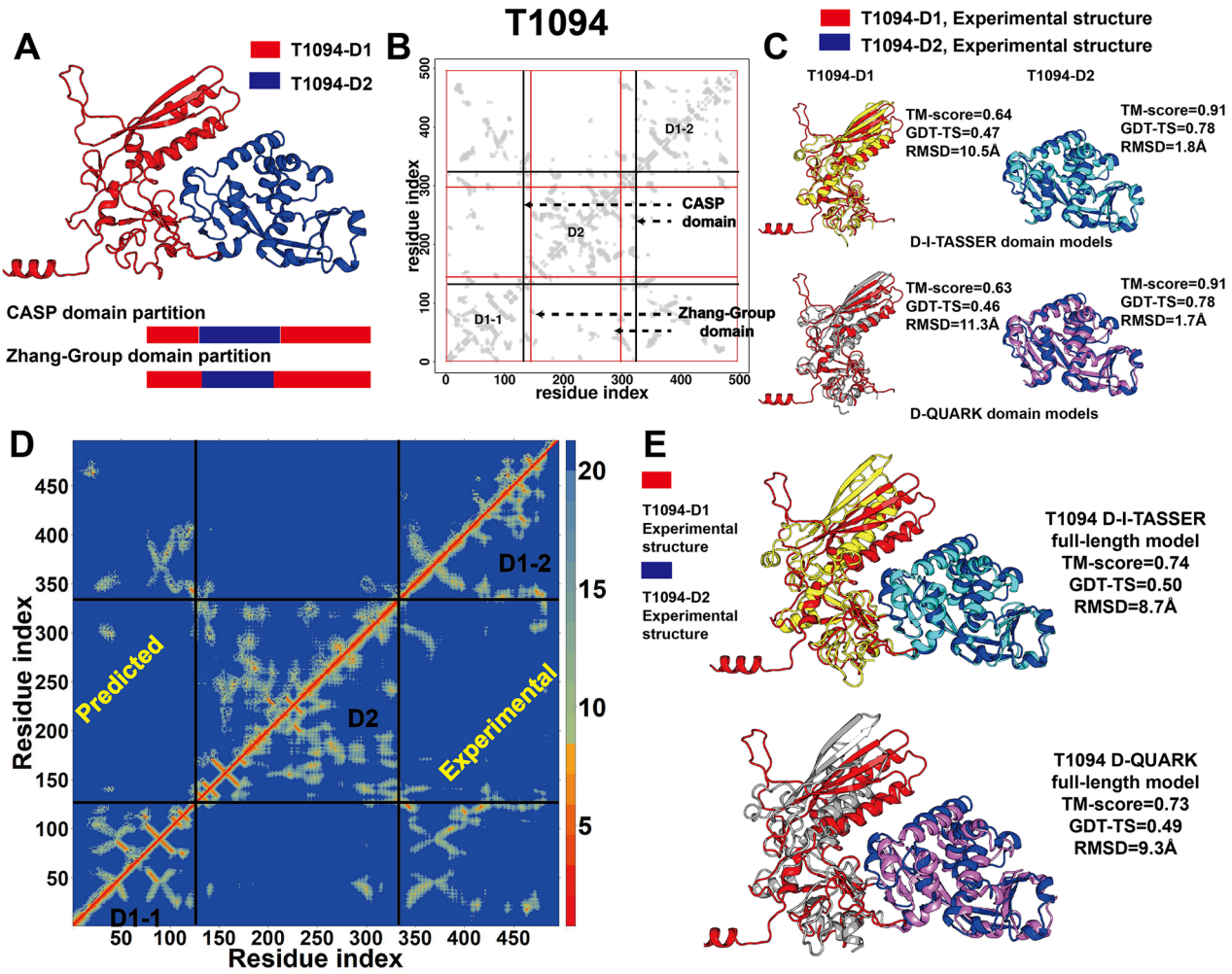
PROT_26193_Figure_4.tif



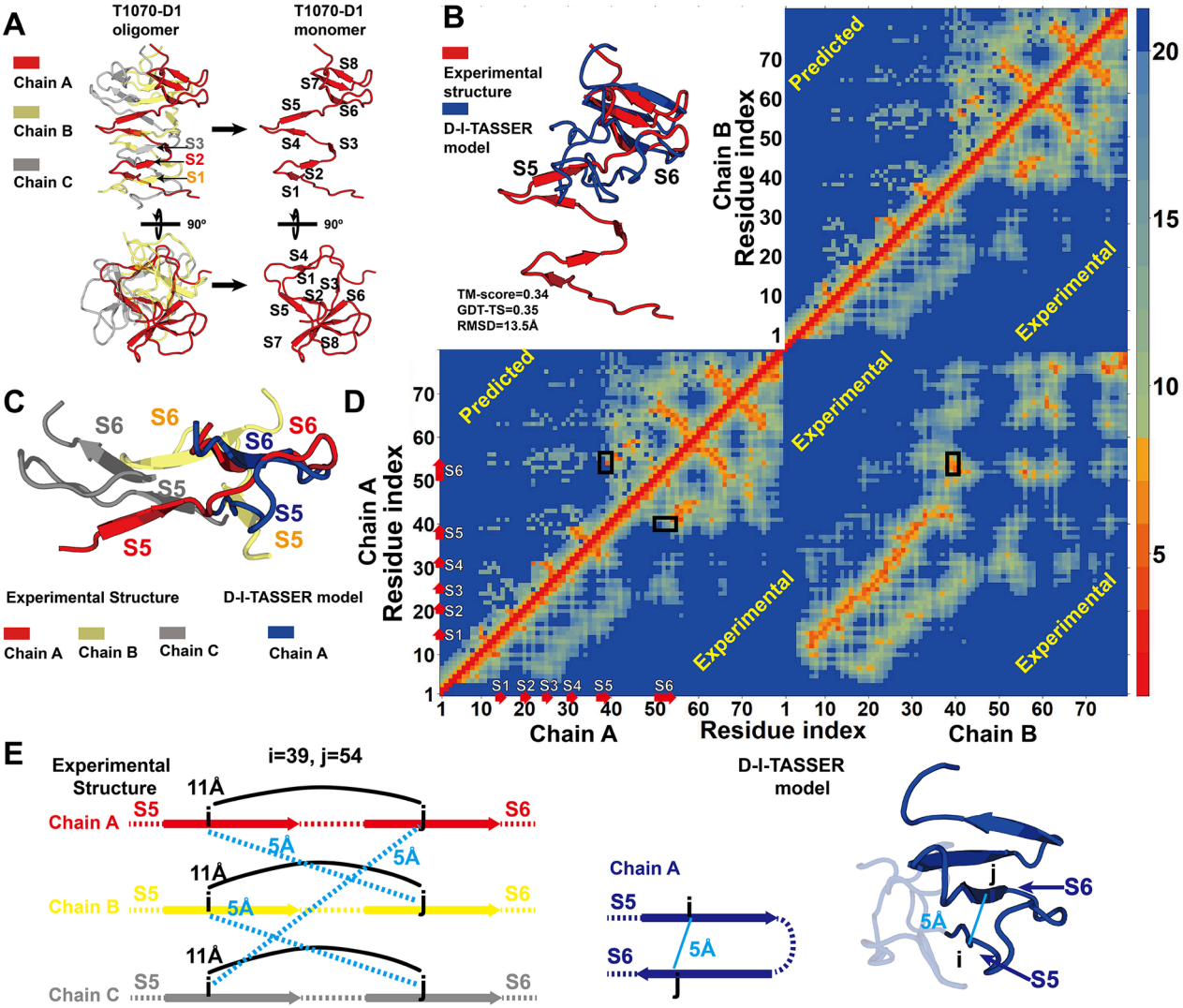
PROT_26193_Figure_5.tif



PROT_26193_Figure_6.tif



PROT_26193_Figure_7.tif



PROT_26193_Figure_8.tif