nature biotechnology

Supplementary information

https://doi.org/10.1038/s41587-025-02654-4

Deep-learning-based single-domain and multidomain protein structure prediction with D-I-TASSER

In the format provided by the authors and unedited

Supplementary Information

Table of Contents

References

Supplementary Text

Text S1. Databases used in DeepMSA2

Sequence databases used in DeepMSA2 are categorized into two groups: genome and metagenome databases. For genome sequence databases, both **Uniclust30** and **UniRef30** contain HHblits-style Hidden Markov Model (HMM) profiles, where protein sequences in UniProtKB[1] are clustered at a threshold of 30% pairwise sequence identity employing MMseqs2[2]. Uniclust30 is the version of the database generated before 2019, while UniRef30 was created after 2019. **Uniref90** offers sequences sourced from UniProtKB, meticulously clustered at a threshold of 90% pairwise sequence identity utilizing MMseqs2. Within each cluster, the representative sequence is exclusively retained in the database, ensuring optimal representation. In total, those three genomics sequence databases contain 464 million sequences.

For metagenome databases, **Metaclust** was devised through the clustering and amalgamation of approximately 1.59 billion protein sequence fragments, which are predicted by Prodigal[3], sourced from around 2,200 metagenomics and meta-transcriptomic datasets acquired from JGI[4]. The clustering was carried out with a 50% sequence identity threshold, while ensuring a coverage of 90%. **Mgnify** was collected by the EBI Metagenomics project[5] and was clustered by MMseqs2 using coverage and sequence identity threshold at 90%. **BFD** is an HHblits-style HMM database that was created by clustering 2.5 billion protein sequences from UniProtKB30, Metaclust, soil reference catalog, and marine eukaryotic reference catalog assembled by Plass[6] using MMseqs2 with 30% pairwise sequence identity. Those three third-party metagenomics sequence databases contain ~3.2 billion sequences.

In addition, three additional metagenomics sequence databases, TaraDB, MetaSourceDB, and JGIclust were newly created for DeepMSA2. The three in-house databases, which were built using data collected from EBI Metagenomics project and the Joint Genome Institute (JGI), contain in total 35.6 billion sequences, which are approximately 11 times as large as the above-mentioned three third-party metagenomics databases (~3.2 billion). Among them, **TaraDB** was created from the 'Tara Oceans' project hosted on EBI Metagenomics with 245 metagenomics sequencing runs (https://www.ebi.ac.uk/metagenomics/studies/ERP001736). The raw read sequences were assembled by MEGAHIT v1.0 to contigs and only the contigs with >500 nucleotides were selected. Next, Prodigal (v2.6) was used with parameters '-c –m p meta' to identify ORFs from metagenome data and translate the gene to protein productions. Finally, CD-HIT (v4.6)[7] was utilized to cluster protein sequences in each sample, and the sequence identity threshold was set to 95% to remove the identical sequence. Next, **MetaSourceDB** collected metagenome data from four large environmental biomes from the EBI. Those four biomes, including 'Fermentor', 'Soil', 'Lake', and 'Gut', cover all typical biomes of the EBI database. In total, 1,705 high-quality samples were selected, assembled, and clustered by the similar pipeline used in Tara DB. In addition to Prodigal, FragGeneScan (v1.20)[8] was also used to predict ORFs from assembled contigs to avoid missing the short sequences. Finally, **JGIclust** was created from Joint Genome Institute (JGI), containing ~25,000 metagenomics and meta-transcriptomic samples. For each project, the assembled protein sequences ('*.assembled.faa') were downloaded and clustered with 90% sequence identity at 90% coverage by MMseqs2. For each cluster of one project, only the representative sequence was kept in the in-house JGIclust database. To further remove the redundancy, MetaSourceDB, TaraDB, and JGIclust were iteratively clustered to 50% identity using MMseqs2's linear cluster pipeline. Coverage was set at 0.8, using '—cov-mode 1'. Due to the storage and memory limitation, the entire sequence database was split to difference small chunks (<100GB) and clustered using the iterative greedy strategy. These chunks were merged into larger chunks, ensuring the merged databases did not exceed 200GB and the merged chunks were then re-clustered to 50% identity. The final large chunks that cannot further be merged were pairwise clustered. Redundant sequences were removed after each clustering round before proceeding to the next pairwise clustering. The process culminated in the final database clustered at 50% identity.

Text S2. dMSA, qMSA, and mMSA pipelines used in DeepMSA2

dMSA (which a short name of the original DeepMSA pipeline[9]) is comprised of three stages. In Stage 1, HHblits[10] from the HH-suite package[11] is used to search the query sequence against the Uniclust30 database[12] to generate the first-level MSA. If there are not enough homologous sequences in the first-level MSA, i.e., the number of effective sequences (*Neff*) of the first-level MSA generated by Stage 1 is <128, Stage 2 will be performed. In Stage 2, Jackhmmer from the HMMER package[13] is used to search the query sequence against the UniRef90 database[14] to generate homologous sequences (hits) for the construction of a custom HHblits-formatted database. Using the first-level MSA as input, HHblits is again applied to search against the custom database to generate the second-level MSA. If the *Neff* of the second-level MSA is still <128, Stage 3 will be performed. In Stage 3, similar to Stage 2, the second-level MSA is used to jump-start an HHblits search against a new custom HHblits-formatted database to get the third-

level MSA. The new custom database in Stage 3 is built by HMMsearch from HMMER to search a profile Hidden Markov Model (HMM) built by HMMbuild from the HMMER package against the Metaclust[15] metagenome sequence database.

qMSA (which stands for "quadruple MSA") contains four stages to perform Hhblits2, Jackhmmer, Hhblits3, and HMMsearch searches against UniRef30 (version 2020_01), UniRef90, BFD, and Mgnify, respectively. The sequence hits from Jackhmmer, HHblits3 and HMMsearch in Stage 2, 3 and 4 of qMSA are converted into an HHblits-formatted database, against which the HHblits2 search is performed using the MSA input from the previous stage.

mMSA (which stands for "multi-level MSA") utilizes the alignment in Stage 3 of qMSA as a probe by HMMsearch to search through the in-house metagenomics sequence databases (TaraDB, MetaSourceDB and JGIclust), and the resulting sequence hits are converted into a new sequence database. This database is then used as the target database, which is searched by HHblits2 with three seed MSAs (MSAs from stage 2 of dMSA, and stages 2 and 3 of qMSA), to derive three new MSAs.

Text S3. The definition of Zscore used in LOMETS3 pipeline

The $Zscore(i,j)$ in the above scoring function includes three score terms from contacts, distances, and hydrogen bond geometries predicted by AttentionPotential and DeepPotential, and one sequence profile score term from the original profile-based threading methods as follows:

$$Zscore(i,j) = w_1\, Zscore^{MAE}(i,j) + w_2\, Zscore^{CMO}(i,j) + w_3\, Zscore^{HB}(i,j) + w_4\, Zscore^{prof}(i,j) \quad (S1)$$

where $Zscore^{MAE}(i,j)$ is the Zscore of the mean absolute error (*MAE*) based on the predicted distance map, $Zscore^{CMO}(i,j)$ is the Z-score of the number of overlapping contacts based on the predicted contact map (CMO), $Zscore^{HB}(i,j)$ is the Z-score based on the predicted hydrogen bond geometry (HB), and $Zscore^{prof}(i,j)$ is a score which is based on the original profile threading scores. The formulas of these four Z-scores are as follows:

$$Zscore^{MAE}(i,j) = \frac{-MAE(i,j) - \langle -MAE(j)\rangle}{\sigma(-MAE(j))} \quad (S2)$$

$$MAE(i,j) = \frac{\sum_{m,n}^{ali}[\delta(m,n)|d_{m,n}^{query} - d_{m,n}^{template}| + (1-\delta(m,n))GapPenalty]}{\sum_{m,n}^{ali}\delta(m,n)} \quad (S3)$$

where $d_{m,n}^{query}$ is the predicted distance between residue *m* and *n* in the query structure, $d_{m,n}^{template}$ is the predicted distance between residue *m* and *n* in the template structure, GapPenalty = 1, ali means the length of alignment, and $\delta(m,n) = \begin{cases} 1, m\ and\ n\ are\ not\ gap \\ 0, else \end{cases}$. $\langle -MAE(j)\rangle$ and $\sigma(-MAE(j))$ are the average and standard deviation of the MAE scores across all templates for the *j*-th program, respectively.

$$Zscore^{CMO}(i,j) = \frac{CMO(i,j) - \langle CMO(j)\rangle}{\sigma(CMO(j))} \quad (S4)$$

$$CMO(i,j) = \frac{N(Overlap(CM^{query}, CM^{template}))}{N(CM^{query})} \quad (S5)$$

where $N(Overlap(CM^{query}, CM^{template}))$ is the number of overlapping contacts between the predicted contact map and the contact map derived from the aligned template, and $N(CM^{query})$ is the number of predicted contacts. $\langle CMO(j)\rangle$ and $\sigma(CMO(j))$ are the mean and standard deviation of the contact overlap scores across all templates for the *j*-th program, respectively.

$$Zscore^{HB}(i,j) = \frac{HBscore(i,j) - \langle HBscore(j)\rangle}{\sigma(HBscore(j))} \quad (S6)$$

$$HBscore(i,j) = \sum_{m,n}^{ali} \frac{1}{1 + (\frac{|min(|\theta_{m,n}^{query} - \theta_{m,n}^{template}|, \pi - |\theta_{m,n}^{query} - \theta_{m,n}^{template}|)|}{\theta})^2} \quad (S7)$$

where $\theta_{m,n}^{query}$ is the predicted hydrogen bond angle between residue *m* and *n* in the query structure, $\theta_{m,n}^{template}$ is the predicted hydrogen bond angle between residue *m* and *n* in the template structure, and $\theta = 15$. $\langle HBscore(j)\rangle$ and $\sigma(HBscore(j))$ are the average and standard deviation of the alignment scores across all templates for the *j*-th program, respectively.

$$Zscore^{prof}(i,j) = \frac{S(i,j) - \langle S(j)\rangle}{\sigma(S(j))} \quad (S8)$$

where $S(i,j)$ is the alignment score of the *i*-th template for the *j*-th program, and $\langle S(j)\rangle$ and $\sigma(S(j))$ are the average and standard deviation of the alignment scores across all templates for the *j*-th program, respectively.

Text S4. Five contact predictors used in D-I-TASSER

In addition to contact predictions from AttentionPotential and DeepPotential, D-I-TASSER also utilizes contact map information from TripletRes[16], ResTriplet[17], ResPRE[18], ResPLM[17], and NeBcon[19], the methods of which are outlined below.

TripletRes (https://zhanggroup.org/TripletRes)[16] is a recently developed contact map predictor, which we used in CASP13. It is noteworthy that the TripletRes method was ranked as the top contact predictor in the CASP13 experiment[20]. Starting from multiple sequence alignments created by DeepMSA2 (see "**Text S2**"), three co-evolutionary features are extracted and then ensembled directly by residual neural networks. Each input feature is fed into a set of residual blocks and transformed into the output feature with 64 channels. The three output features are concatenated along the channel dimension as the input of the last layers. The last set of layers try to learn patterns from the three transformed features using another 12 residual blocks. All residual blocks have a channel size of 64, and the kernel size of the convolutional layers is set to 3×3 with a padding size equal to one. Such a padding parameter set-up can keep the spatial information fixed through different layers. Here, we use a convolutional layer with a 1×1 kernel size to transform each co-evolutionary input feature and the concatenated features into 64 channels. The final contact map prediction is obtained by a sigmoid activation function.

ResTriplet[17] is another recent contact map predictor, which we used in CASP13. ResTriplet is a two-stage ensemble model that uses a stacking strategy. In Stage I, three individual base models are trained separately based on the three different sets of co-evolutionary features, PRE, PLM and COV, respectively as described above. The base models have the same training data and the same neural network structure consisting of 22 residual basic blocks. In Stage II, we use a shallow neural network structure to combine the predictions of the base models from Stage I. Thus, the predicted contact maps from the base models are considered as the input features in Stage II. To reduce the risk of over-fitting, predicted contact maps produced by each base model are generated by 10-fold cross-validation as the input features of Stage II. The predicted secondary structures, denoted as PSS, obtained using PSIPRED[21] are also adopted as an extra feature for the neural network model in Stage II. For shallow convolutional neural networks, the size of the receptive fields is usually limited. Hence, a dilated convolutional neural network structure with dilation equal to 2 is employed in order to enlarge the size of the receptive fields.

ResPRE (https://zhanggroup.org/ResPRE)[18] is a novel in-house contact map predictor, which consists of two consecutive steps of precision matrix-based feature generation and deep residual neural network-based contact inference. ResPRE is the average ensemble of ten base models trained by different subsets of the whole training data.

ResPLM[17] is another contact map predictor similar to ResPRE. The only difference is that ResPLM was trained using the PLM feature.

NeBcon (https://zhanggroup.org/NeBcon)[19] is a meta-approach designed for contact map prediction. In this study, we retrained NeBcon to improve its long-range contact prediction precision by using the a naïve Bayes classifier (NBC) to integrate eight state-of-the-art contact prediction methods, including four deep learning-based methods: DeepPLM[17], DeepCov[22], Deepcontact[23], and DNCON2[24], three co-evolution-based methods: GREMLIN[25], CCMpred[26], and FreeContact[27], and one meta-server-based methods MetaPSICOV2[28]. NeBcon has two variants, NeBconA and NeBconB, designed for C_α and C_β atoms, respectively.

Text S5. D-I-TASSER force field E-groups2-7

E-Group2: Template-based restraints

Four types of restraints have been derived from the LOMETS3 templates and used to guide the D-I-TASSER simulations.

Template-based short-range distance restraints. This energy term considers only the short-range interactions which occur for $|i\text{-}j| \leqslant 6$ for the i-th and j-th residues of the model.

$$E_{dist}^{Short} = \sum_{i=1}^{L-1} \sum_{j>i}^{i+6} E_{dist}^{Short}(d_{ij}) \tag{S9}$$

$$E_{dist}^{Short}(d_{ij}) = \begin{cases} 1, & if\ |d_{ij} - d_{ij}^{T}| > \sigma_{ij}^{T} \\ 0, & otherwise \end{cases} \tag{S10}$$

where d_{ij} is the C_α distance between the i-th and j-th residues of the model. d_{ij}^{T} and σ_{ij}^{T} are the average and the mean square deviation of the C_α distances, respectively, between the i-th and j-th residues that are collected from the threading templates.

Template-based long-range distance restraints. This energy term considers only the long-range interactions for $|i\text{-}j|>6$ for the i-th and j-th residues of the model.

$$E_{dist}^{Long} = \sum_{i=1}^{L-7} \sum_{j>i+6}^{L} E_{dist}^{Long}(d_{ij}) \tag{S11}$$

$$E_{dist}^{Long}(d_{ij}) = -\frac{1}{\max(1, |d_{ij} - d_{ij}^{T}|)} \tag{S12}$$

where d_{ij} is the Cα distance between the *i*-th and *j*-th residues of the model. d_{ij}^{T} is the average of the Cα distances between the *i*-th and *j*-th residues collected from the threading templates.

Template-based contact restraints for Cα. This energy term considers the contact information corresponding to Cα atoms, which is extracted from the templates.

$$E_{Tcon}^{C\alpha} = \sum_{i=1}^{L-1} \sum_{j>i}^{L} E_{Tcon}^{C\alpha}(d_{ij}) \tag{S13}$$

$$E_{Tcon}^{C\alpha}(d_{ij}) = \begin{cases} -U_{ij}, & if\ d_{ij} < 6.5\text{Å} \\ 0, & otherwise \end{cases} \tag{S14}$$

$$U_{ij} = \begin{cases} 1 + 4 * \left| conf_{ij}^{C\alpha} - conf_{cut}^{C\alpha} \right|, & if\ conf_{ij}^{C\alpha} > conf_{cut}^{C\alpha} \\ 1 - 2 * \left| conf_{ij}^{C\alpha} - conf_{cut}^{C\alpha} \right|, & otherwise \end{cases} \tag{S15}$$

where d_{ij} is the Cα distance between the *i*-th and *j*-th residues of the model; $conf_{ij}^{C\alpha}$ is the contact confidence score for the *i*-th and *j*-th Cα atoms of the model, where the confidence scores are based on the threading results; $conf_{cut}^{C\alpha}$ is the pre-tuned cut-off value for the contact confidence score for Cα atoms, which is query type-dependent.

Template-based contact restraints for the center of side-group heavy atoms (SG). This energy term considers the contact information corresponding to the center of side-group heavy atoms, which is extracted from the templates.

$$E_{Tcon}^{SG} = \sum_{i=1}^{L-1} \sum_{j>i}^{L} E_{Tcon}^{SG}(d_{ij}^{SG}) \tag{S16}$$

$$E_{Tcon}^{SG}(d_{ij}^{SG}) = \begin{cases} -U_{ij}^{SG}, & d_{ij}^{SG} < d_{cut}^{SG}(AA_i, AA_j) \\ -\frac{1}{2} U_{ij}^{SG} \left[1 - sin\left(\frac{d_{ij}^{SG} - \left(\frac{d_{cut}^{SG}(AA_i, AA_j) + D}{2} \right)}{D - d_{cut}^{SG}(AA_i, AA_j)} \pi \right) \right], & d_{cut}^{SG}(AA_i, AA_j) \le d_{ij}^{SG} < D \\ \frac{1}{2} U_{ij}^{SG} \left[1 + sin\left(\frac{d_{ij}^{SG} - \left(\frac{D + 80}{2} \right)}{(80 - D)} \pi \right) \right], & D \le d_{ij}^{SG} < 80\text{Å} \\ U_{ij}^{SG}, & d_{ij}^{SG} \ge 80\text{Å} \end{cases} \tag{S17}$$

$$U_{ij}^{SG} = \begin{cases} 1 + 4 * \left| conf_{ij}^{SG} - conf_{cut}^{SG} \right|, & if\ conf_{ij}^{SG} > conf_{cut}^{SG} \\ 1 - 2 * \left| conf_{ij}^{SG} - conf_{cut}^{SG} \right|, & otherwise \end{cases} \tag{S18}$$

where d_{ij}^{SG} is the distance between the *i*-th and *j*-th centers of the side-group heavy atoms in the model; $conf_{ij}^{SG}$ is the contact confidence score for the *i*-th and *j*-th pseudo side-group heavy atoms in the model, where the confidence scores are based on the threading results; $conf_{cut}^{C\alpha}$ is the pre-tuned cut-off value for the contact confidence score for the centers of the side-group heavy atoms, which is query type-dependent. $D = 2 + d_{cut}^{SG}(AA_i, AA_j)$, where $d_{cut}^{SG}(AA_i, AA_j)$ is an amino acid type-dependent cut-off value for the center of side-group heavy atoms.

E-Group3: Burial interaction restraints

This potential represents the general propensity of amino acids to be buried or exposed to the solvent.

$$E_{burial}^{SG} = -\sum_{i=1}^{L} E(x_i, y_i, z_i) * P(ASA_i) \tag{S19}$$

$$E(x_i, y_i, z_i) = \min(0, \max(-1, \frac{(x_i - x_c)^2}{x_0^2} + \frac{(y_i - y_c)^2}{y_0^2} + \frac{(z_i - z_c)^2}{z_0^2} - 2.5)) \tag{S20}$$

where $P(ASA_i)$ is the accessible surface (ASA) of the *i*-th residue predicted through PSSpred[29]. If the *i*-th residue is predicted as buried, the value of $P(ASA_i)$ is made negative. (x_i, y_i, z_i) is the coordinate for the center of the side-

group heavy atoms (SG) for the *i*-th residue. (x_0, y_0, z_0) is the length of the principal axes of the protein ellipsoid, and (x_c, y_c, z_c) is the center of the protein ellipsoid[30].

E-Group4: Secondary structure-based restraints

Secondary structure restraints for C_α. These three potential terms try to encourage local structures to form local secondary structures, where the secondary structure information for the query protein is predicted by PSSpred [29].

$$E_{sec}^{C\alpha} = w_{sec1}\sum_{i=1}^{L-4} E_{\text{sec}}^{\text{C}\alpha}(d_{i,i+4}) + w_{sec2}\sum_{i=1}^{L-4} E_{\text{sec}}^{\text{C}\alpha}(\overrightarrow{B_i}, \overrightarrow{B_{i+4}}) + w_{sec3}\sum_{i=1}^{L-2} E_{\text{sec}}^{\text{C}\alpha}(\overrightarrow{C_i}, \overrightarrow{C_{i+2}}) \tag{S21}$$

$$E_{\text{sec}}^{\text{C}\alpha}(d_{i,i+4}) = \begin{cases} -2 - \dfrac{DF_i * DF_{i+1} + DF_{i+3} * DF_{i+4}}{2}, & if\ \alpha - helix \\ -2 - (DF_i * DF_{i+1} + DF_{i+3} * DF_{i+4}), & if\ \beta - sheet \\ 0, & otherwise \end{cases} \tag{S22}$$

$$E_{sec}(\overrightarrow{B_i}, \overrightarrow{B_{i+4}}) = \begin{cases} -\dfrac{DF_i * DF_{i+1} + DF_{i+3} * DF_{i+4}}{2}, & if\ S_{i,i+4}\ is\ helix\ and\ \overrightarrow{B_i} * \overrightarrow{B_{i+4}} > 0.9 \\ -(DF_i * DF_{i+1} + DF_{i+3} * DF_{i+4}), & if\ \overrightarrow{B_i} * \overrightarrow{B_{i+4}} < -0.3\ or\ \overrightarrow{B_i} * \overrightarrow{B_{i+4}} > 0.5 \\ 0, & otherwise \end{cases} \tag{S23}$$

$$E_{sec}(\overrightarrow{C_i}, \overrightarrow{C_{i+2}}) = -\frac{DF_i + DF_{i+1} + DF_{i+2}}{2} * \frac{\min(0.71, \overrightarrow{C_i} * \overrightarrow{C_{i+2}})}{0.71} \tag{S24}$$

$$DF_i = \min\left(max\left(\frac{2.2 * L^{0.38}}{(x_i - x_c)^2 + (x_i - x_c)^2 + (x_i - x_c)^2}, 0.5\right), 1\right) \tag{S25}$$

where (x_i, y_i, z_i) is the coordinate for the Cα atom of the *i*-th residue. (x_0, y_0, z_0) is the length of the principal axes of the protein ellipsoid, and (x_c, y_c, z_c) is the center of the protein ellipsoid. $2.2 * L^{0.38}$ is the estimated radius of gyration for a protein with length *L*.

For the first term, the conditions for forming an $\alpha - helix$ include: $d_{i,i+4} < 7.53$Å$, 4$Å$ < d_{i,i+3} < 8$Å$, \overrightarrow{U_i} * \overrightarrow{U_{i+2}} < 0$, $\overrightarrow{U_{i+1}} * \overrightarrow{U_{i+3}} < 0$, $\overrightarrow{U_i} * \overrightarrow{U_{i+3}} > 0$, and the local segment $S_{i+1,i+3}$ is not predicted to be a sheet. Here, $\overrightarrow{U_i}$ is the unit vector starting from the *i*-th Cα atom and pointing to the (*i*+1)-th Cα atom. The conditions for forming β-sheets include: $d_{i,i+4} > 11$Å$,\ arccos\frac{\overrightarrow{B_{i+1}} * \overrightarrow{B_{i+3}}}{|\overrightarrow{B_{i+1}}| * |\overrightarrow{B_{i+3}}|} < 45°,\ arccos\frac{\overrightarrow{B_{i+1}} * \overrightarrow{B_{i+2}}}{|\overrightarrow{B_{i+1}}| * |\overrightarrow{B_{i+2}}|} > 135°$, and the local segment $S_{i+1,i+3}$ is not predicted to be a helix. $\overrightarrow{B_{i+1}}$ is the hydrogen bond direction of the (*i*+1)-th residue, which is equal to $\frac{\overrightarrow{U_i} \times \overrightarrow{U_{i+1}}}{|\overrightarrow{U_i} \times \overrightarrow{U_{i+1}}|}$. The second term focuses on the direction of the hydrogen bond $\overrightarrow{B_i}$, while the third term concerns $\overrightarrow{C_i}$, which is equal to $\frac{\overrightarrow{U_{i-1}} - \overrightarrow{U_i}}{|\overrightarrow{U_{i-1}} - \overrightarrow{U_i}|}$. $w_{sec1}, w_{sec2}, w_{sec3}$ are the weights used to balance each energy term.

Penalty for crumpling structures. This potential term imposes a penalty to the irregular crumpled structures.

$$E_{crumpling} = \sum_{i=1}^{L-8} E_{crumpling}(i) \tag{S26}$$

$$E_{crumpling}(i) = \begin{cases} 1, & if\ \overrightarrow{U_{i,i+4}} \cdot \overrightarrow{U_{i+4,i+8}} < 0,\ \overrightarrow{U_{i+4,i+8}} \cdot \overrightarrow{U_{i+8,i+12}} < 0\ and\ \overrightarrow{U_{i,i+4}} \cdot \overrightarrow{U_{i+8,i+12}} > 0 \\ 0, & otherwise \end{cases} \tag{S27}$$

where $\overrightarrow{U_{i,j}}$ is the unit vector starting from the *i*-th Cα atom and pointing to the *j*-th Cα atom.

Alpha/beta fragment restraints. This potential encourages the continuous alpha/beta fragments for secondary structures.

$$E_{sec}^{frag} = \sum_{i=1}^{L} E_{sec}^{frag}(i) \tag{S28}$$

$$E_{sec}^{frag}(i) = \begin{cases} \left|d_{i,i+7} - 10.5\right|, & if\ S_{i,i+7}\ is\ helix \\ \left|d_{i,i+6} - 19.1\right| * 2, & if S_{i,i+6}\ is\ sheet \\ 0, & otherwise \end{cases} \tag{S29}$$

E-Group5: Statistical pairwise potentials

C_α-SG pairwise potential. This potential is used for atomic packing and solvation between Cα atom and side-group heavy atoms.

$$E_{pair}^{C\alpha-SG} = \sum_{i}^{L}\sum_{j\neq i}^{L} E_{pair}^{C\alpha-SG}\left(d_{ij}^{C\alpha-SG}\right) \tag{S30}$$

$$E_{pair}^{C\alpha-SG} = \begin{cases} \left(\frac{r_1}{d_{ij}^{C\alpha-SG}}\right)^2 - \frac{1}{2}, & if\ r_1 \le d_{ij}^{C\alpha-SG} < r_2 \\ \frac{1}{2}, & if\ d_{ij}^{C\alpha-SG} < r_1 \\ 0, & otherwise \end{cases} \tag{S31}$$

where $d_{ij}^{C\alpha-SG}$ is the distance between the Cα atom of the *i*-th residue and the center of the side-group heavy atoms for the *j*-th residue. r_1=3.14Å and r_2=5.22Å.

SG-SG pairwise potential. This potential is used for atomic packing and solvation between side-group heavy atoms.

$$E_{pair}^{SG} = \sum_{i}^{L}\sum_{j\neq i}^{L} E_{pair}^{SG}(d_{ij}^{SG}) \tag{S32}$$

$$E_{pair}^{SG}\left(d_{ij}^{SG}\right) = \begin{cases} U_{i,j}^{ori}, & if\ d_{ij}^{SG} < d_{cut}^{SG}\left(AA_i, AA_j\right) \\ 0, & otherwise \end{cases} \tag{S33}$$

where d_{ij}^{SG} is the distance between the *i*-th and *j*-th centers of the side-group heavy atoms in the model; $d_{cut}^{SG}\left(AA_i, AA_j\right)$ is an amino acid type-dependent cut-off value for d_{ij}^{SG}. $U_{i,j}^{ori}$ is the generic orientation-dependent contact potential derived from 6,500 non-redundant high-resolution PDB structures [31], and the contacts are weighted by the sum of the BLOSUM [32] mutation score between the residue pairs of the query and the PDB structures over a window of ± 5 neighboring residues. This potential is query sequence specific but an alignment between the query and the PDB structure is not needed since we count all the contact pairs in the PDB structures that have the same amino acid identity $\left(A_i,\ A_j\right)$ to the query, where A_i and A_j are the amino acid identities of the residues.

Parallel Cα-Cα pairwise potential. This potential is used for atomic packing and solvation between parallel Cα atoms.

$$E_P^{C\alpha} = \sum_{i}^{L-i}\sum_{j>i}^{L} E_P^{C\alpha}\left(d_{ij}\right) \tag{S34}$$

$$E_P^{C\alpha}\left(d_{ij}\right) = \begin{cases} \min\left(0, -\frac{r_1^2}{max\left(r_1^2,\ d_{ij}^2\right)} + \frac{1}{2}\right), & if\ \vec{C} * \overrightarrow{C_J} > 0.5 \\ 0, & otherwise \end{cases} \tag{S35}$$

Here, r_1=4.77Å. $\overrightarrow{C_\iota} * \overrightarrow{C_J} > 0.5$ indicates that the *i*-th Cα vector, $\overrightarrow{U_\iota}$, and the *j*-th Cα vector, $\overrightarrow{U_J}$, are parallel, where $\overrightarrow{U_\iota}$ is the unit vector starting from the *i*-th Cα atom and pointing to the (*i*+1)-th Cα atom, and $\overrightarrow{C_\iota} = \frac{\overrightarrow{U_{\iota-1}} - \overrightarrow{U_\iota}}{\left|\overrightarrow{U_{\iota-1}} - \overrightarrow{U_\iota}\right|}$ as shown in **Eq. 18**.

Non-parallel Cα-Cα pairwise potential. This potential is used for atomic packing and solvation between non-parallel Cα atoms.

$$E_{NP}^{C\alpha} = \sum_{i}^{L-i}\sum_{j>i}^{L} E_{NP}^{C\alpha}\left(d_{ij}\right) \tag{S36}$$

$$E_{NP}^{C\alpha}\left(d_{ij}\right) = \begin{cases} \frac{r_1^2}{d_{ij}^2} - \frac{1}{2}, & if\ \overrightarrow{C_\iota} * \overrightarrow{C_J} \le 0.5,\ d_{ij} < 5\text{Å} \\ 0, & otherwise \end{cases} \tag{S37}$$

Here, r_1=3.48Å. $\overrightarrow{C_\iota} * \overrightarrow{C_J} \le 0.5$ indicates that the *i*-th Cα vector, $\overrightarrow{U_\iota}$, and the *j*-th Cα vector, $\overrightarrow{U_J}$, are not parallel.

E-Group6: Hydrogen bond restraints

The hydrogen bonds in D-I-TASSER are specified by the backbone geometry following the STRIDE secondary structure assignments.

$$E_{HB} = \sum_{i=1}^{L-1} \sum_{j>i}^{L} E_{HB}\left(d_{ij}\right) \tag{S38}$$

$$E_{HB}\left(d_{ij}\right) = \begin{cases} -w_{HB}(1-|CC-CC_0|)(1-|BB-BB_0|)\left[\frac{1}{(1+|bri-br_0|)}+\frac{1}{(1+|brj-br_0|)}\right], \\ \qquad if\ \ helix\ and\ |i-j|=3 \\ -w_{HB}(|BB|*CC)\left[\frac{1}{1+bri/2}+\frac{1}{1+brj/2}\right], \\ \qquad if\ \ sheet \text{ and } |i-j|<4\ for\ parallel\ or\ |i-j|>20\ for\ antiparallel \end{cases} \tag{S39}$$

where $\mathrm{CC} = \overrightarrow{C_i} * \overrightarrow{C_j}$, $BB = \overrightarrow{B_i} * \overrightarrow{B_j}$, $bri = \left|\varepsilon\overrightarrow{H_i} - \vec{r}\right|$ and $brj = \left|\varepsilon\overrightarrow{H_j} - \vec{r}\right|$. Here, ε=5.0Å or 4.6Å if both donor and receptor residues are predicted as $\alpha - helices$ or $\beta - sheets$. Similarly, $w_{HB} = 1$ if both donor and receptor residues are predicted as $\alpha - helices$ and $\beta - sheets$; otherwise $w_{HB} = 0.5$. The cutoff parameters for standard hydrogen bonds (CC_0, BB_0, br_0) were calculated from an average of 500 high resolution PDB structures with their secondary structure elements assigned by STRIDE [33].

E-Group7: Statistical restraints from the PDB library

Short-range correlation restraints. This type of potential considers the short-range Cα distance correlation between residues. It includes three energy terms as follows.

$$\begin{aligned} E_{corr}^{C\alpha} = w_{corr1} &\sum_{i=1}^{L-2} corr\left(AA_i, AA_{i+2}, bin\left(d_{i,i+2}\right)\right) \\ +w_{corr2} &\sum_{i=1}^{L-3} corr\left(AA_{i+1}, AA_{i+2}, bin\left(d_{i,i+3}\right), \varepsilon_i,\ S_{i+1,i+3}\right) \\ +w_{corr3} &\sum_{i=1}^{L-4} corr\left(AA_{i+1}, AA_{i+2}, bin\left(d_{i,i+4}\right), S_{i+1,i+3}\right) \end{aligned} \tag{S40}$$

The first term $corr\left(AA_i, AA_{i+2}, bin(d_{i,i+2})\right)$ is the short-range Cα distance correlation between the *i*-th and the (*i*+2)-th residues, which comes from a look-up table. $d_{i,i+2}$ is the Cα distance between the *i*-th and (*i*+2)-th residues of the model. $bin(d_{i,i+2})$ indicates that $d_{i,i+2} < 6.03$ or that $d_{i,i+2} \geq 6.03$. The second term $corr\left(AA_{i+1}, AA_{i+2}, bin\left(d_{i,i+3}\right), \varepsilon_i,\ S_{i+1,i+3}\right)$ is from a look-up table for short-range Cα distance correlation between the *i*-th and the (*i*+3)-th residues. $d_{i,i+3}$ is the Cα distance between *i*-th and (*i*+3)-th residues of the model. $bin(d_{i,i+3})$ indicates that $d_{i,i+3} \in$ (0, 1Å], (1Å, 2Å], ⋯, or (11Å, ∞]. ε_i denotes the local structure chirality of three consecutive Cα-Cα vectors from the *i*-th to (*i*+3)-th residue. $S_{i+1,i+3}$ denotes that the local segment from the *i*-th to (*i*+3)-th residue is an alpha-helix, beta-sheet or coil. The third term $corr\left(AA_{i+1}, AA_{i+3}, bin\left(d_{i,i+4}\right), S_{i+1,i+3}\right)$ also comes from a look-up table for correlation between the *i*-th and the (*i*+4)-th residues. $d_{i,i+4}$ is the Cα distance between the *i*-th and (*i*+4)-th residues of the model. $bin(d_{i,i+4})$ indicates that $d_{i,i+4} \in$ (0, 1Å], (1Å, 2Å], ⋯, or (15Å, ∞]. $w_{corr1}, w_{corr2}, w_{corr3}$ are the weights used to balance each energy term.

Binary excluded volume restraints. This potential considers the general excluded volume interactions, which are represented by a smaller hard-sphere potential plus a 1/r type of soft-core potential with a slightly larger range. This mimics the minimal observed cutoff distance in real proteins, and allows a few atoms to approach closer than is normally observed with an accompanying penalty, thereby partly remedying the coarseness of the discrete lattice model.

$$E_{vol}^{SG} = \sum_{i}^{L-i} \sum_{j>i}^{L} E_{vol}^{SG}\left(d_{ij}^{SG}\right) \tag{S41}$$

$$E_{vol}^{SG}\left(d_{ij}^{SG}\right) = \begin{cases} 1,\ if \begin{cases} \overrightarrow{C_i} * \overrightarrow{C_j} > 0.5\ and\ d_{ij}^{SG} \in \left(d_{min}^{pa}\left(AA_i, AA_j\right), d_{max}^{pa}\left(AA_i, AA_j\right)\right) \\ or\ \ \overrightarrow{C_i} * \overrightarrow{C_j} < -0.5\ and\ d_{ij}^{SG} \in \left(d_{min}^{an}\left(AA_i, AA_j\right), d_{max}^{an}\left(AA_i, AA_j\right)\right) \\ or\ -0.5 \leq \overrightarrow{C_i} * \overrightarrow{C_j} \leq 0.5\ and\ d_{ij}^{SG} \in \left(d_{min}^{pe}\left(AA_i, AA_j\right), d_{max}^{pe}\left(AA_i, AA_j\right)\right) \end{cases} \\ 0,\ otherwise \end{cases} \tag{S42}$$

where d_{ij}^{SG} is the distance between the *i*-th and *j*-th centers of the side-group heavy atoms in the model. $\overrightarrow{C_i} * \overrightarrow{C_j} >$ 0.5 and $d_{ij}^{SG} \in \left(d_{min}^{pa}\left(AA_i, AA_j\right), d_{max}^{pa}\left(AA_i, AA_j\right)\right)$ indicate that the *i*-th Cα vector, $\overrightarrow{U_i}$, and the *j*-th Cα vector, $\overrightarrow{U_j}$, are

parallel. $\overrightarrow{C_i} * \overrightarrow{C_j} < -0.5$ and $d_{ij}^{SG} \in \left(d_{min}^{an}(AA_i, AA_j), d_{max}^{an}(AA_i, AA_j)\right)$ indicate that the *i*-th Cα vector, $\overrightarrow{U_i}$, and the *j*-th Cα vector, $\overrightarrow{U_j}$, are antiparallel. $\overrightarrow{C_i} * \overrightarrow{C_j} < -0.5$ and $d_{ij}^{SG} \in \left(d_{min}^{pe}(AA_i, AA_j), d_{max}^{pe}(AA_i, AA_j)\right)$ indicate that the *i*-th Cα vector, $\overrightarrow{U_i}$, and the *j*-th Cα vector, $\overrightarrow{U_j}$, are perpendicular. $\left(d_{min}^{pa}(AA_i, AA_j), d_{max}^{pa}(AA_i, AA_j)\right)$, $\left(d_{min}^{an}(AA_i, AA_j), d_{max}^{an}(AA_i, AA_j)\right)$ and $\left(d_{min}^{pe}(AA_i, AA_j), d_{max}^{pe}(AA_i, AA_j)\right)$, which correspond to parallel/antiparallel/perpendicular Cα vectors, are amino acid type-dependent statistical values that were extracted from the PDB.

Statistical excluded volume restraints. This potential is the upgrade version of excluded volume restraints.

$$E_{mvol}^{SG} = \sum_{i}^{L-i} \sum_{j>i}^{L} E_{mvol}^{SG}(d_{ij}^{SG}) \tag{S43}$$

$$E_{mvol}^{SG}(d_{ij}^{SG}) = \begin{cases} U^{pa}(AA_i, AA_j), if\ \overrightarrow{C_i} * \overrightarrow{C_j} > 0.5\ and\ d_{ij}^{SG} \in \left(d_{min}^{pa}(AA_i, AA_j), d_{max}^{pa}(AA_i, AA_j)\right) \\ U^{an}(AA_i, AA_j), if\ \overrightarrow{C_i} * \overrightarrow{C_j} < -0.5\ and\ d_{ij}^{SG} \in \left(d_{min}^{an}(AA_i, AA_j), d_{max}^{an}(AA_i, AA_j)\right) \\ U^{pe}(AA_i, AA_j), if\ -0.5 \le \overrightarrow{C_i} * \overrightarrow{C_j} \le 0.5\ and\ d_{ij}^{SG} \in \left(d_{min}^{pe}(AA_i, AA_j), d_{max}^{pe}(AA_i, AA_j)\right) \\ 0, \quad otherwise \end{cases} \tag{S44}$$

where $U^{pa}(AA_i, AA_j)$, $U^{pa}(AA_i, AA_j)$, and $U^{pe}(AA_i, AA_j)$, which correspond to parallel/antiparallel/perpendicular, are amino acid type-dependent statistical values that were extracted from the PDB.

Separated C_α-C_α pairwise potential. This potential considers the Cα distance between separated residues.

$$E_{Spair1-5}^{C\alpha} = \sum_{i=3}^{L-3} \sum_{j>i}^{L-1} E_{Spair1-5}^{C\alpha}(d_{ij}) \tag{S45}$$

$$E_{Spair1-5}^{C\alpha}(d_{ij}) = \begin{cases} -corr(AA_{i-1}, AA_{i+1}, bin(d_{i-2,i+2}), S_{i-1,i+1}) \\ * corr(AA_{j-1}, AA_{\mathrm{j}+1}, bin(d_{j-2,j+2}), S_{j-1,j+1}), \\ \quad if \begin{cases} \overrightarrow{C_i} * \overrightarrow{C_j} > 0.5\ and\ d_{ij}^{SG} \in \left(0, d_{max}^{pa}(AA_i, AA_j)\right) \\ or\ \overrightarrow{C_i} * \overrightarrow{C_j} < -0.5\ and\ d_{ij}^{SG} \in \left(0, d_{max}^{an}(AA_i, AA_j)\right) \\ or\ -0.5 \le \overrightarrow{C_i} * \overrightarrow{C_j} \le 0.5\ and\ d_{ij}^{\square\square G} \in \left(0, d_{max}^{pe}(AA_i, AA_j)\right) \end{cases} \\ 0, \quad otherwise \end{cases} \tag{S46}$$

where d_{ij} is the Cα distance between the *i*-th and *j*-th residues of the model; d_{ij}^{SG} is the distance between the *i*-th and *j*-th centers of the side-group heavy atoms in the model. $corr(AA_{i-1}, AA_{i+1}, bin(d_{i-2,i+2}), S_{i-1,i+1})$ is similar to the description in **Eq. S40**.

Contact profile constraints. The potential describes the contact environment.

$$E_{cprof} = \sum_{i=1}^{L} E_{cprof}(N_i^{pa}, N_i^{an}, N_i^{pe}, AA_i) \tag{S47}$$

where N_i^{pa}, N_i^{an}, N_i^{pe} are the number of residues that are in parallel/antiparallel/perpendicular contact with the *i*-th residue. $E_{cprof}(N_i^{pa}, N_i^{an}, N_i^{pe}, AA_i)$ is the statistic value from the PDB and calculated using the negative logarithm of the relative frequency histogram.

Contact number constraints. This potential accounts for the biases to the expected contact order and contact number.

$$E_{Ncon} = |N^{Con} - N_0^{Con}| + \left|\overline{S^{Con}} - S_0^{Con}\right| \tag{S48}$$

where N^{Con} is the number of contacts in a decoy structure and $\overline{S^{Con}}$ is the average sequence separation of the contacts. N_0^{Con} and S_0^{Con} are statistical values extracted from the PDB, which are a linear function of $\alpha * L$, where L is the protein length and α is 1.5.

Supplementary Figures

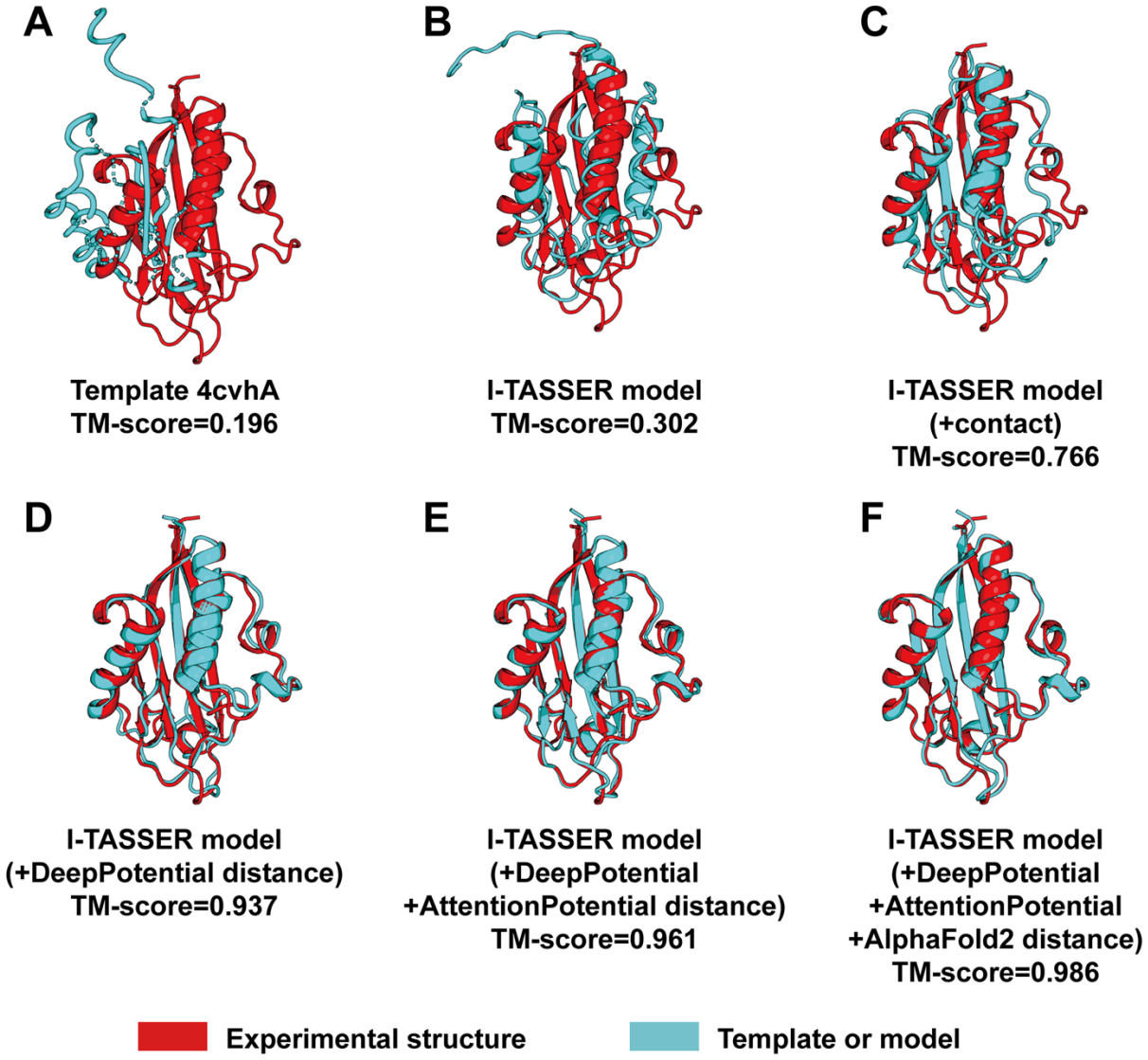


Fig. S1. Structural modeling of 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (PDB ID: 3fpiA) using various I-TASSER workflows. The images are shown for the superposition of the experimental structure (red) with predicted models by (A) the best LOMETS template (PDB ID: 4cvhA); (B) I-TASSER without using deep-learning restraints; (C) I-TASSER with contact-map prediction (C-I-TASSER); (D) I-TASSER with distance map by DeepPotential; (E) I-TASSER with distance maps by DeepPotential and AttentionPotential; (F) I-TASSER with distance maps by DeepPotential, AttentionPotential and AlphaFold2.

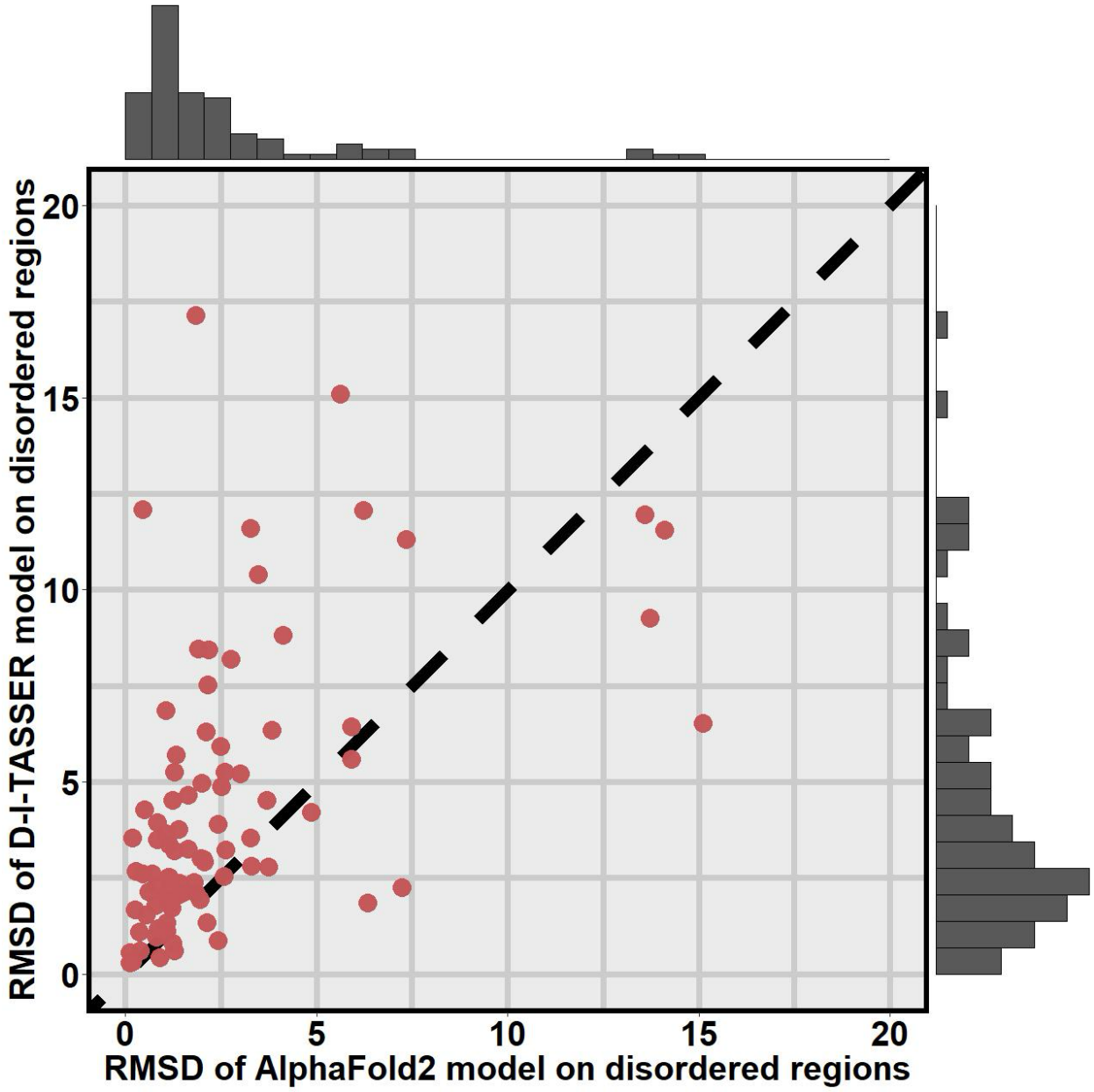


Fig. S2. The average RMSDs between the top five models generated by D-I-TASSER and those by AlphaFold2 for 91 disordered regions lacking experimentally determined structures on the *Benchmark-I* dataset of 1,262 proteins.

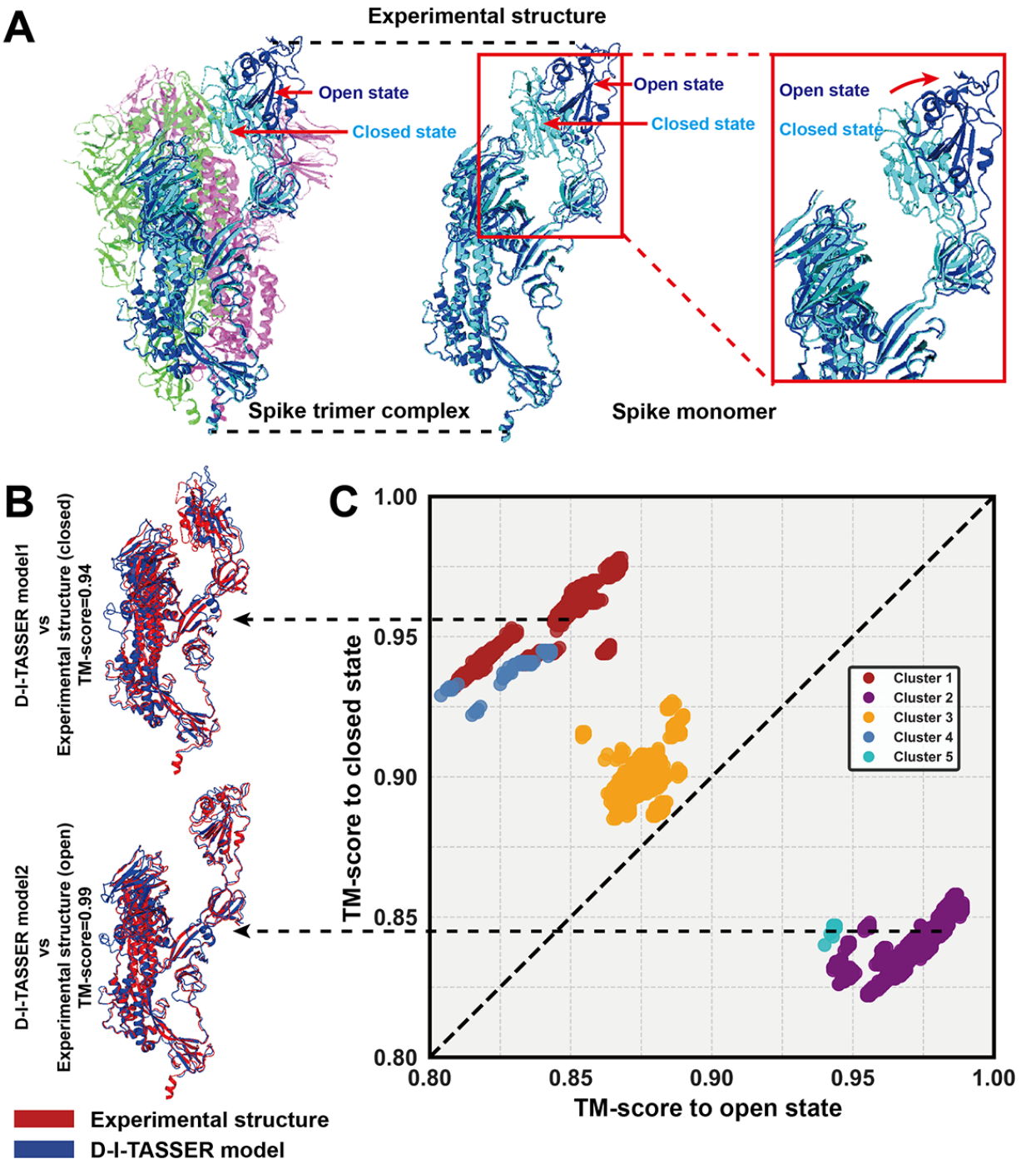


Fig. S3. Application of D-I-TASSER to multi-state modeling of the SARS-CoV-2 Spike protein. (A) Open and closed states of the experimental structure for the SARS-CoV-2 Spike protein. (B) Open and closed states of the D-I-TASSER models superposed with experimental structures for the SARS-CoV-2 Spike protein. (C) Head-to-head comparison between TM-scores of open and closed states of the 4,362 D-I-TASSER models for the SARS-CoV-2 Spike protein. Notably, the structure members of cluster1 and cluster2 are more similar, resulting in a higher degree of point overlap, which makes cluster1 and cluster2 appear relatively "smaller" than cluster3.

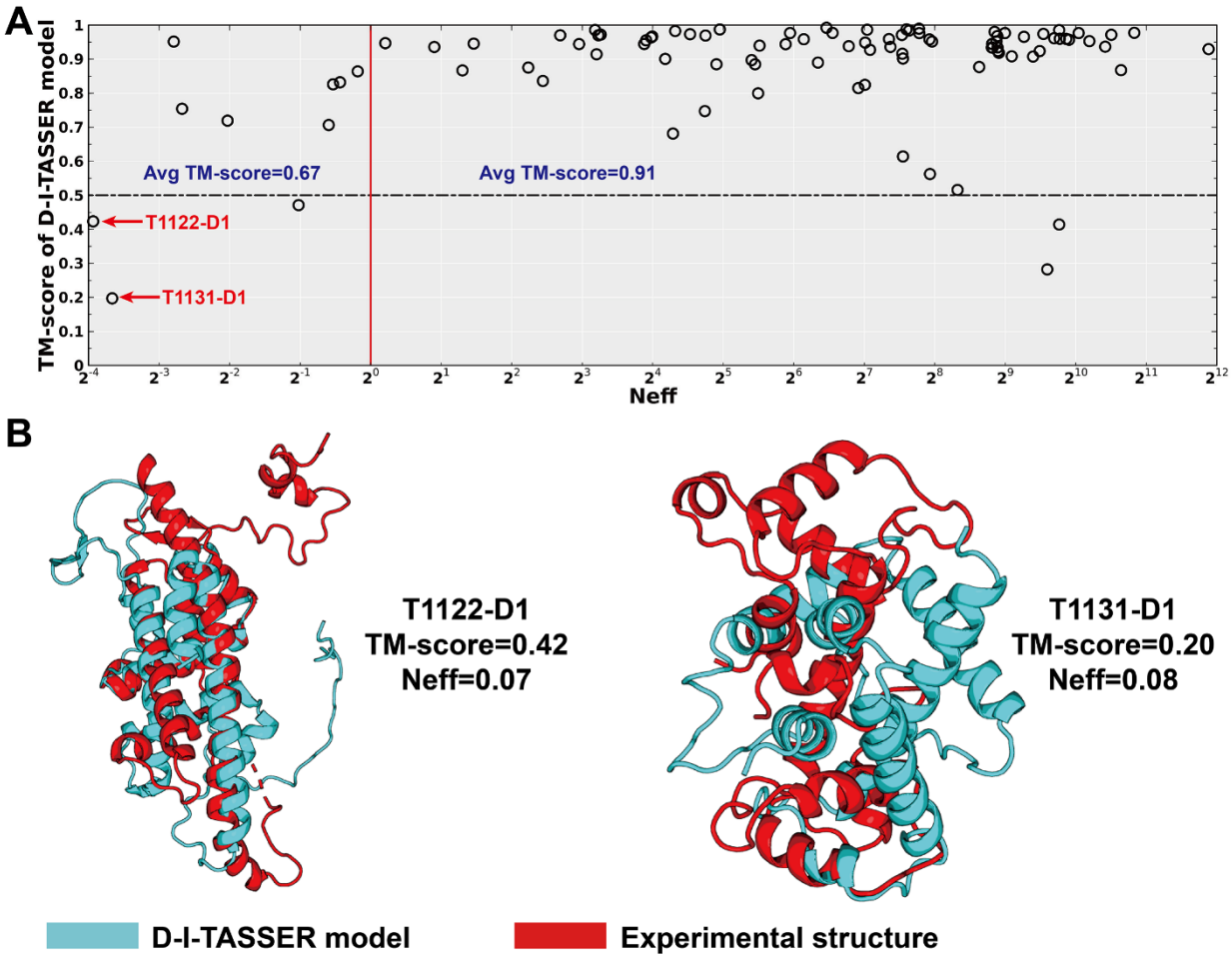


Fig. S4. The relationship between *Neff* and TM-score of D-I-TASSER models on CASP15 targets. (A) *Neff* versus TM-score of D-I-TASSER models on 94 CASP15 targets. (B) Two examples of orphan proteins for targets T1122-D1 and T1131-D1 for which poor modeling performance was observed due to low-information MSAs.

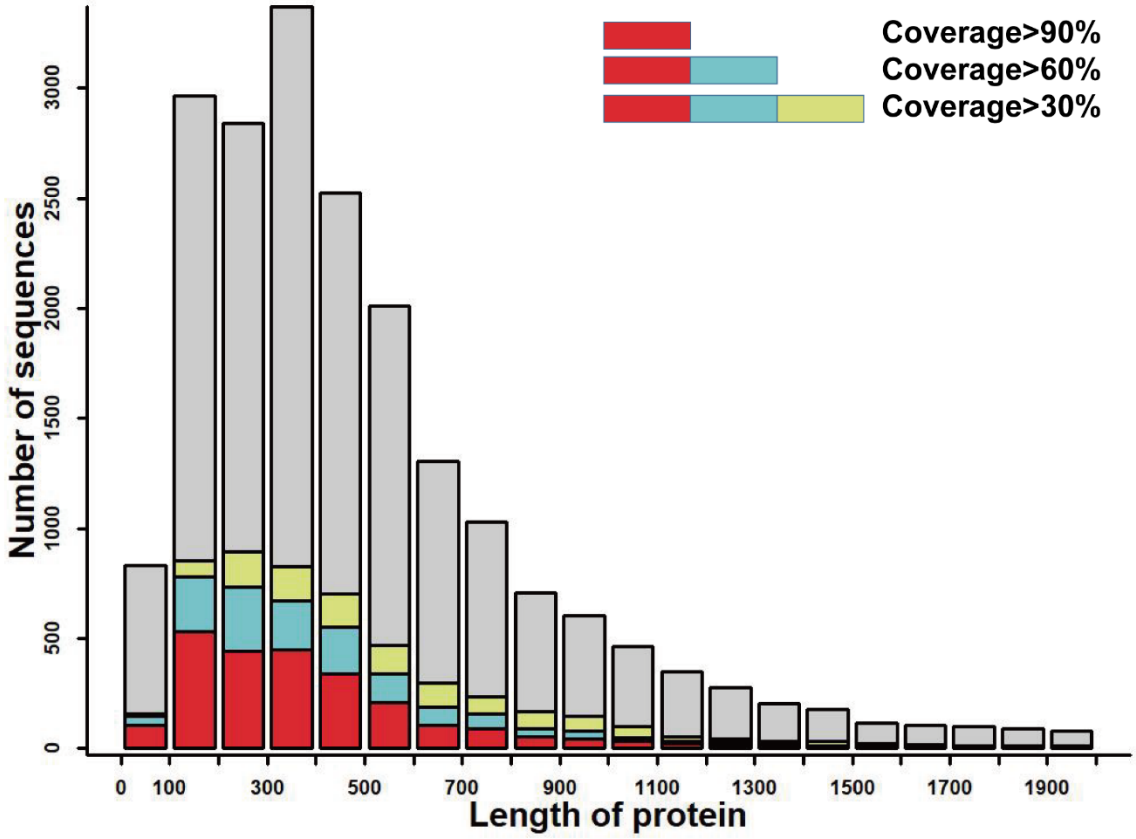


Fig. S5. Summary of the protein lengths and experimental structure coverage for the human proteome dataset of 20,596 proteins. The red bars represent the number of sequences with >90% coverage by known structures; the cyan bars correspond to the >60% and ⩽90% coverage; the yellow bars are for >30% and ⩽60% coverage.

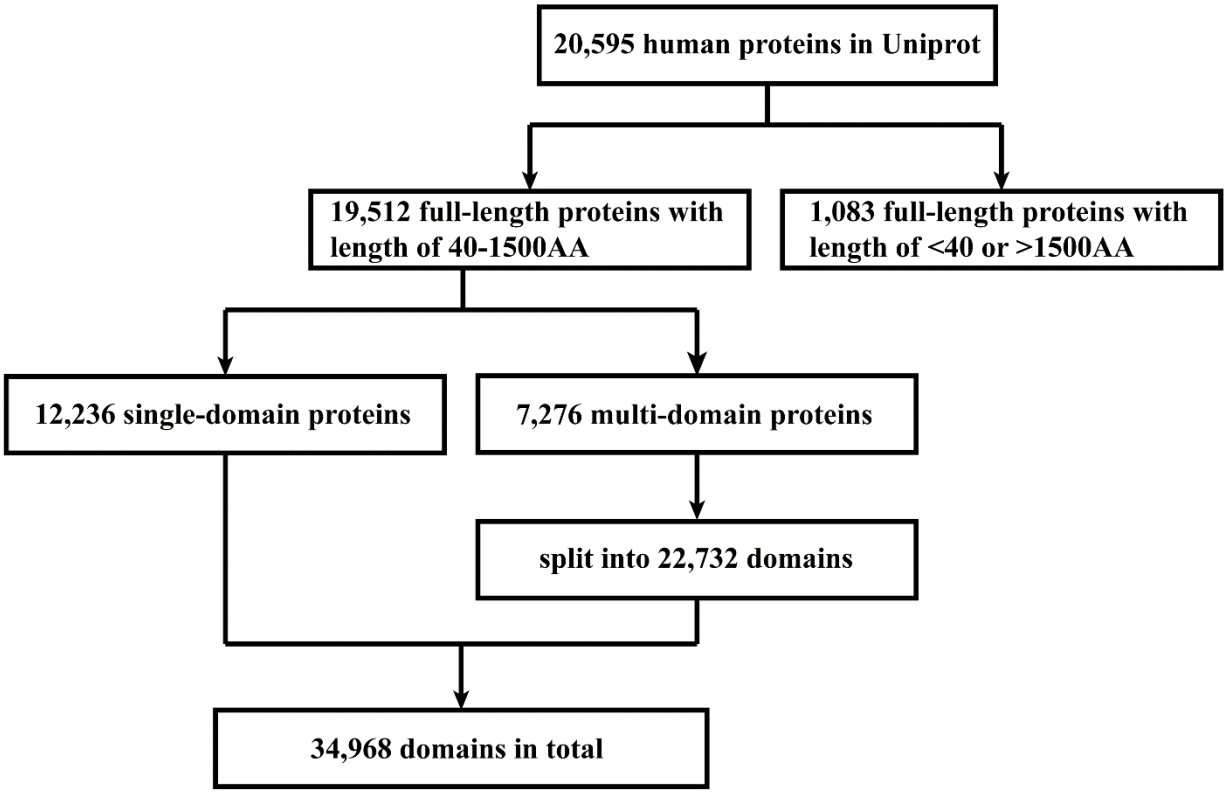


Fig. S6. Number of human proteins at each stage of the analysis, where each set is a subset of the previous set.

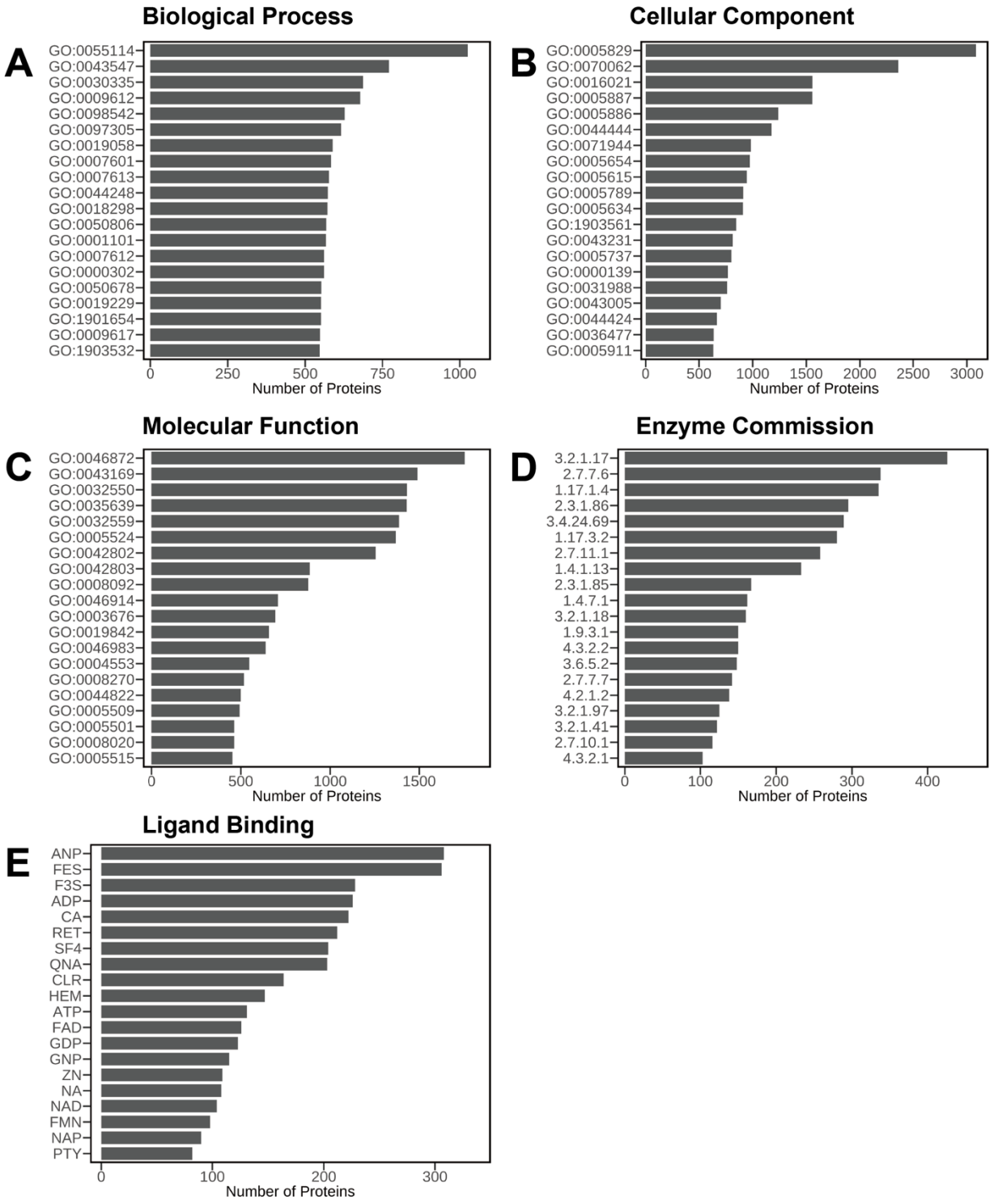


Fig. S7. Frequency analysis of the most commonly predicted functions for 19,512 proteins in the human proteome arising from our pipeline. The number of proteins on top 20 BP GO terms (A), CC GO terms (B), MF GO terms (C), EC terms (D) and non-peptide ligands (E).

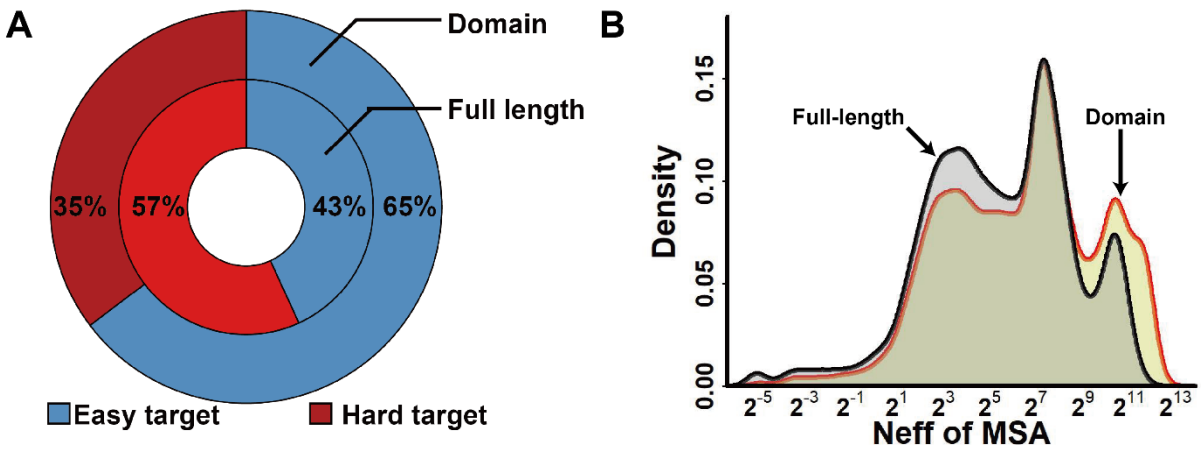


Fig. S8. Statistics on human proteome dataset of 19,512 proteins. (A) The ratio of Easy and Hard targets for the domain-level and full-chain human proteins. (B) MSA *Neff* value distribution for domain-level and full-chain human proteins.

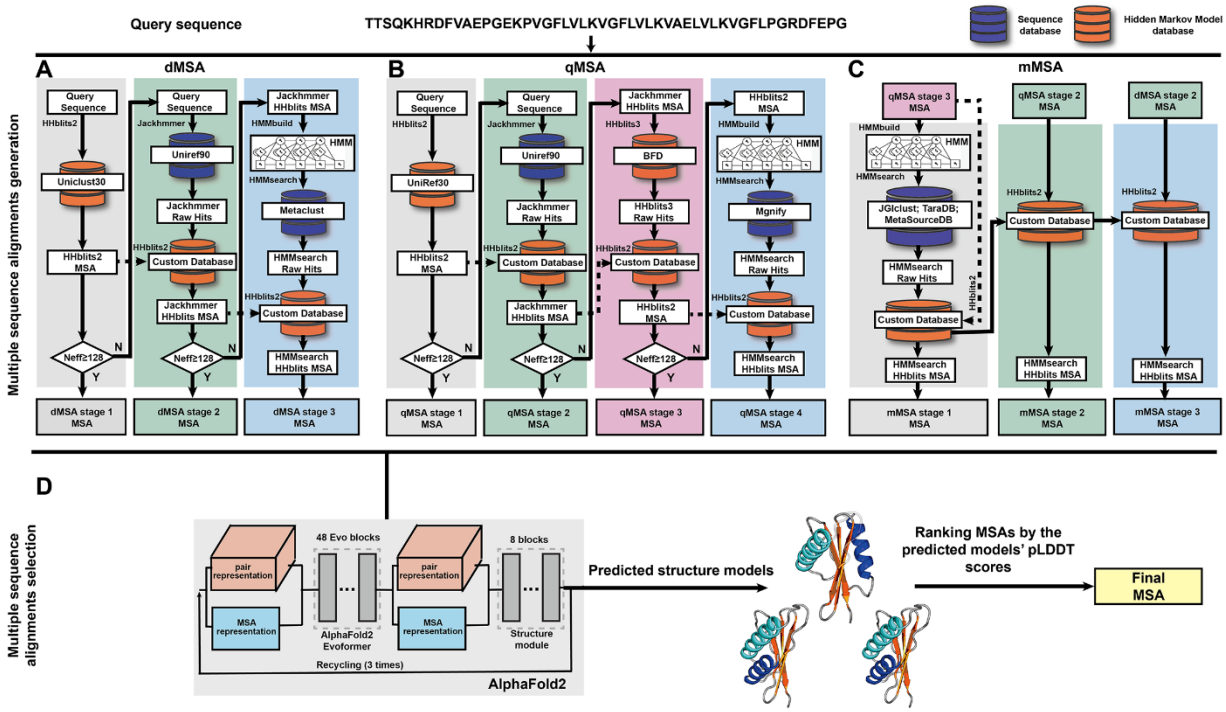
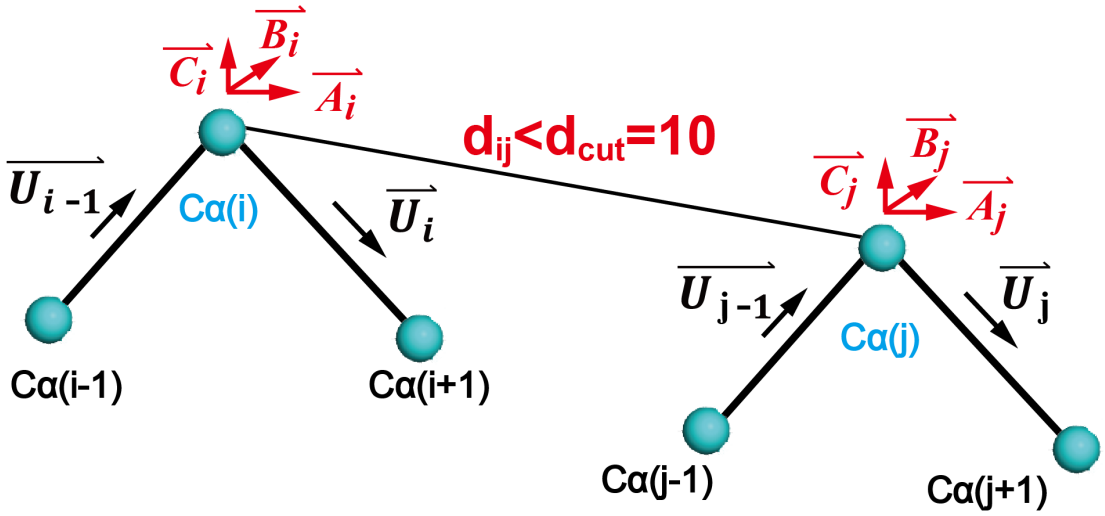


Fig. S9. Schematic of the DeepMSA2 pipeline, which contains four approaches, (A) dMSA, (B) qMSA, (C) mMSA and (D) MSA selection.



$$\overrightarrow{A_i} = \frac{\overrightarrow{U_{i-1}} + \overrightarrow{U_i}}{\left|\overrightarrow{U_{i-1}} + \overrightarrow{U_i}\right|}$$

$$\overrightarrow{B_i} = \frac{\overrightarrow{U_{i-1}} \times \overrightarrow{U_i}}{\left|\overrightarrow{U_{i-1}} \times \overrightarrow{U_i}\right|}$$

$$\overrightarrow{C_i} = \frac{\overrightarrow{U_{i-1}} - \overrightarrow{U_i}}{\left|\overrightarrow{U_{i-1}} - \overrightarrow{U_i}\right|}$$

$$AA = \overrightarrow{A_i} * \overrightarrow{A_j}$$

$$BB = \overrightarrow{B_i} * \overrightarrow{B_j}$$

$$CC = \overrightarrow{C_i} * \overrightarrow{C_j}$$

$$\theta_{AA} = arccos(\overrightarrow{A_i} * \overrightarrow{A_j})$$

$$\theta_{BB} = arccos(\overrightarrow{B_i} * \overrightarrow{B_j})$$

$$\theta_{CC} = arccos(\overrightarrow{C_i} * \overrightarrow{C_j})$$

Fig. S10. Definition of hydrogen bonds used by D-I-TASSER.

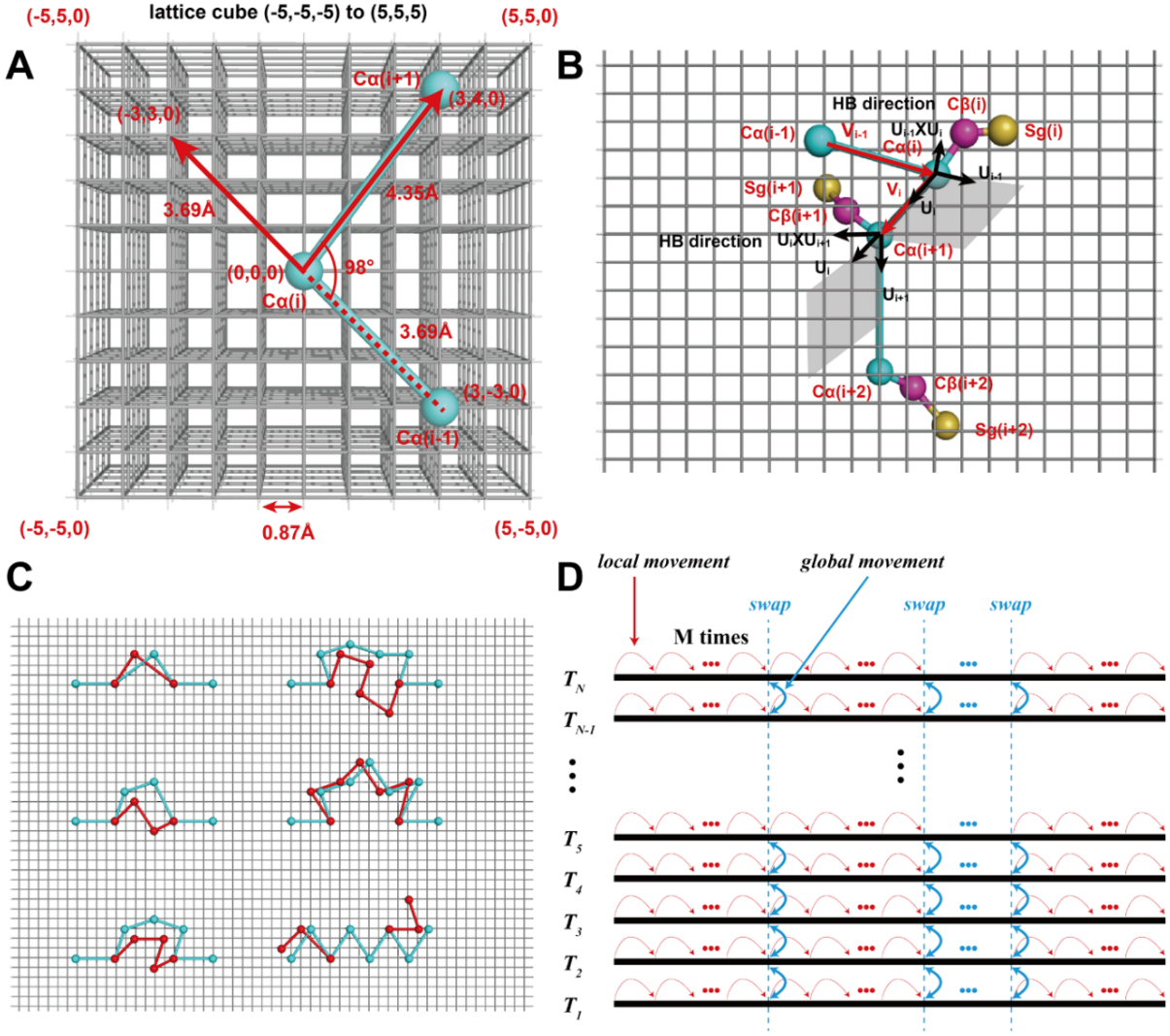


Fig. S11. Schematics of the modeling and simulation settings in D-I-TASSER. **(A)** Reduced representation of an amino acid using a three-dimensional underlying cubic lattice system with a lattice grid of 0.87 Å. Only the alpha carbon (Cα) atom of each residue is treated explicitly. Considering the Cα of the *i*-th residue, Cα(*i*), the lattice cube is from (-5,-5,-5) to (5,5,5). Cα(*i*) is located at (0,0,0). The Cα of the previous (*i*-1)-th residue, Cα(*i*-1) is located at (3,-3,0) and the Cα-Cα bond length between Cα(*i*-1) and Cα(*i*) is 3.69 Å. The Cα of the next (*i*+1)-th residue, Cα(*i*+1), is located at (3,4,0) and the Cα-Cα bond length between Cα(*i*+1) and Cα(*i*) is 4.35 Å. Additionally, the Cα-Cα bond angle is 98º. **(B)** Determination of the positions for the Cβ atom and the center of the side-group heavy atoms. The positions of three consecutive Cα atoms are used to define a local coordinate system for the determination of the beta carbon (Cβ) (except glycine), and the center of the side-group heavy atoms (SG) (except glycine and alanine). $\overrightarrow{\boldsymbol{V_{i-1}}}$ is the vector from Cα(*i*-1) to Cα(*i*), and $\overrightarrow{\boldsymbol{U_{i-1}}}$ is the unit vector for $\overrightarrow{\boldsymbol{V_{i-1}}}$. The cross product of $\overrightarrow{\boldsymbol{U_{i-1}}}$ and $\overrightarrow{\boldsymbol{U_i}}$, $\overrightarrow{\boldsymbol{U_{i-1}}} \times \overrightarrow{\boldsymbol{U_i}}$, is the direction of the hydrogen bond (HB). **(C)** Conformational movements in the D-I-TASSER Monte Carlo simulations. The cyan and red lines are the Cα traces before and after the movements, respectively. There are 6 types of conformational movements in the D-I-TASSER simulations: (1) 2-bond vector walk; (2) 3-bond vector walk; (3) 4-bond vector walk; (4) 5-bond vector walk; (5) 6-bond vector walk; (6) N- or C-terminal random walk. **(D)** Illustration of the local and global movements used during the REMC simulations. There are *N* replicas, which are implemented in parallel. After every 200**L* local conformational movements, where *L* is the protein length, a global swap movement between each pair of neighboring replicas is attempted following the standard Metropolis criterion.

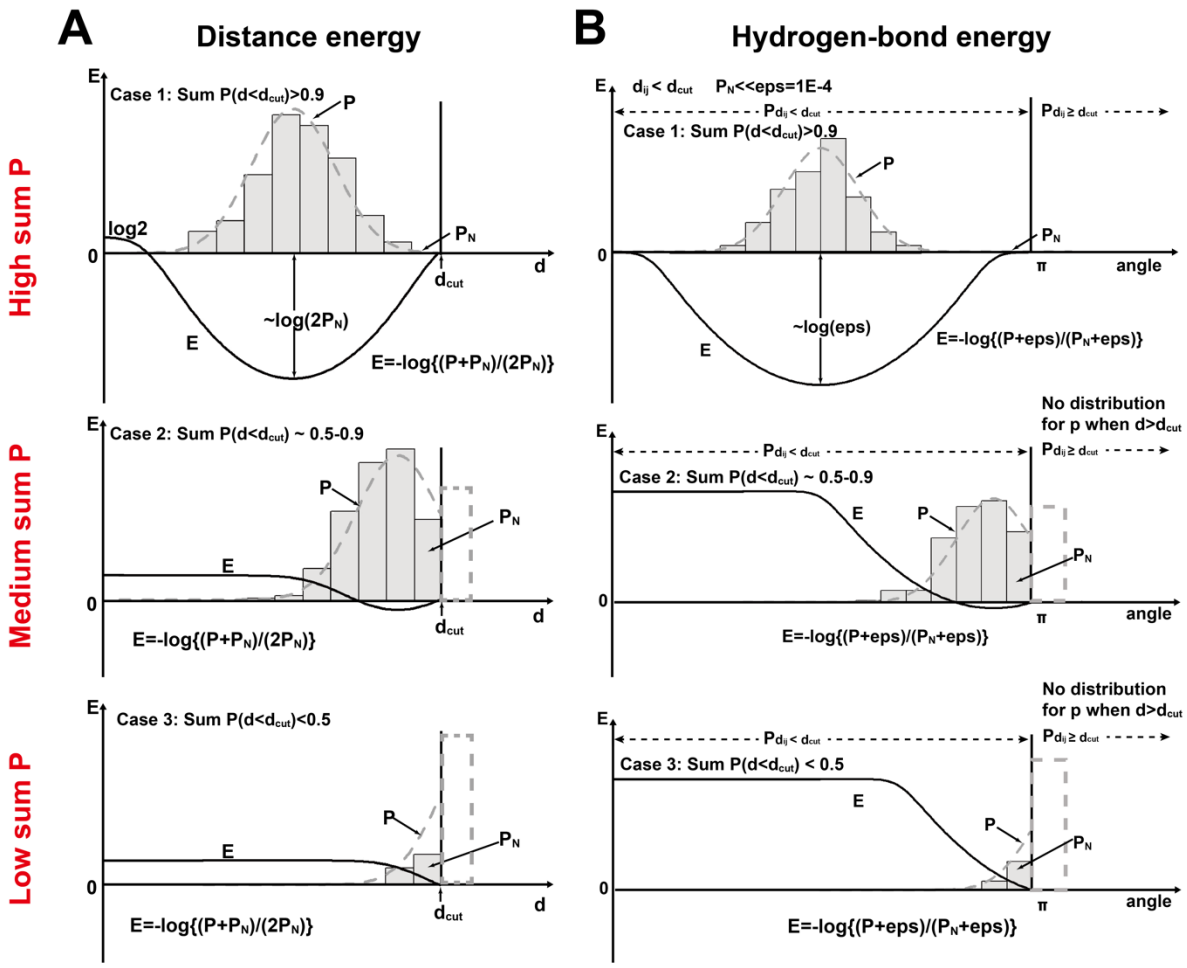


Fig. S12. Illustrations of (A) distance and (B) hydrogen bond potentials for three different situations.

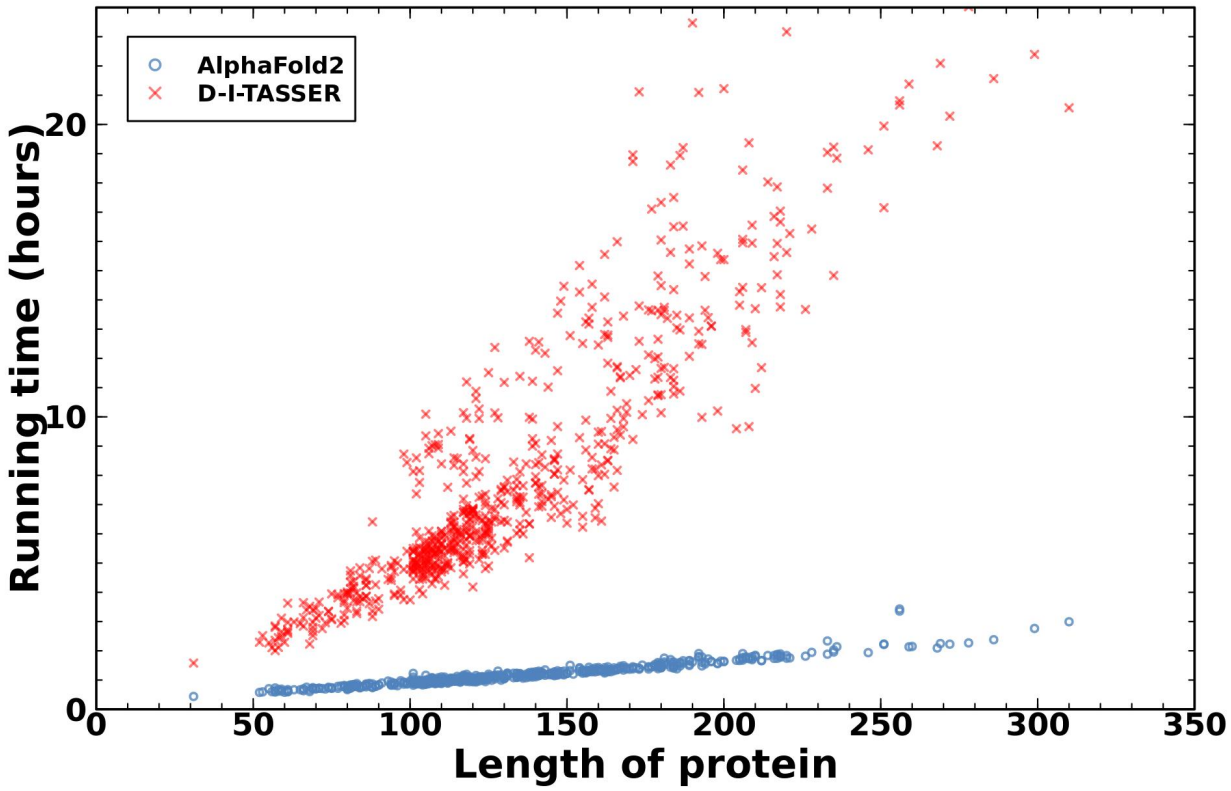


Fig. S13. Comparison of time requirements for D-I-TASSER and AlphaFold2 on different size proteins on a dataset of 645 proteins. Both programs were run using 10 CPUs with parallel processing, generating 5 models each. The AlphaFold2 program was executed with default settings, including 1 ensemble, full_dbs and monomer pipeline as implemented in AlphaFold version 2.2.0. The running time reported excludes the DeepMSA2 search time, as the speed of large database searches is largely influenced by I/O performance. For instance, storing databases on SSD or NVMe drives can significantly reduce search time.

Supplementary Tables

Table S1. Comparison of modeling results by D-I-TASSER with other methods for different target types on the 1,262 benchmark dataset (Benchmark-I). *P*-values were calculated between TM-scores by D-I-TASSER and others using paired one-sided Student's t-tests. #{TM-score >0.5} is the number of targets with a TM-score >0.5. Here, AlphaFold2 refers to version both 2.2 and 2.3.

Method	Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	All (1,262)	0.9097	-	1239
	Easy (762)	0.9359	-	759
	Hard (500)	0.8698	-	480
I-TASSER	All (1,262)	0.6062	1.73E-206	858
	Easy (762)	0.7290	6.87E-125	713
	Hard (500)	0.4191	9.66E-84	145
C-I-TASSER	All (1,262)	0.6852	9.07E-207	1066
	Easy (762)	0.7615	3.34E-125	737
	Hard (500)	0.5688	9.83E-84	329
AlphaFold2 (version 2.2)	All (1,262)	0.8814	1.52E-137	1213
	Easy (762)	0.9227	9.79E-78	757
	Hard (500)	0.8185	1.11E-61	456
AlphaFold2 (version 2.3)	All (1,262)	0.8869	1.15E-117	1218
	Easy (762)	0.9252	9.01E-76	760
	Hard (500)	0.8286	9.25E-46	458
AlphaFold2 +DeepMSA2	All (1,262)	0.8937	2.12E-121	1228
	Easy (762)	0.9281	2.94E-66	759
	Hard (500)	0.8413	2.89E-56	469

Table S2. The contributions of different spatial restraints used in I-TASSER folding simulations to the final modeling results, compared with different versions of AlphaFold (including AlphaFold3, AlphaFold2.3, AlphaFold2.2, AlphaFold2.1, and AlphaFold2.0) for all 500 Hard targets in our benchmark dataset (Benchmark-I). *P*-values were calculated between TM-scores by D-I-TASSER and others using paired one-sided Student's t-tests. #{TM-score >0.5} is the number of targets with a TM-score >0.5. Here, "I-TASSER+contact" indicates the standard I-TASSER method with contact potential used in folding simulation; "I-TASSER+DeepPotential distance+DeepMSA2" means standard I-TASSER method with DeepPotential distance restraints used in folding simulation in combination with DeepMSA2 for MSA generation; "I-TASSER+DeepPotential+AttentionPotential distance+DeepMSA2" means standard I-TASSER method with DeepPotential and AttentionPotential distance restraints used in folding simulation in combination with DeepMSA2 for MSA generation; "I-TASSER+ AlphaFold2 distance+DeepMSA2" means standard I-TASSER method with AlphaFold2 distance restraints used in folding simulation in combination with DeepMSA2 for MSA generation; "D-I-TASSER - DeepMSA2" means default D-I-TASSER method without using DeepMSA2 for MSA generation; "D-I-TASSER - pLDDT MSA ranking" means default D-I-TASSER method without pLDDT MSA ranking step.

Method	**TM-score**	***P*-value**	**#{TM-score>0.5}**
D-I-TASSER	0.8698	-	480
I-TASSER	0.4191	9.66E-84	145
I-TASSER +contact	0.5688	9.83E-84	329
I-TASSER +DeepPotential distance+DeepMSA2	0.6731	4.91E-82	393
I-TASSER +DeepPotential+AttentionPotential distance+DeepMSA2	0.7494	7.97E-76	428
I-TASSER +AlphaFold2 distance+DeepMSA2	0.8571	4.47E-16	472
D-I-TASSER -DeepMSA2	0.8362	3.63E-69	471
D-I-TASSER -pLDDT MSA ranking	0.8536	2.99E-38	476
AlphaFold3	0.8488	1.79E-07	466
AlphaFold2.3	0.8286	9.25E-46	458
AlphaFold2.2	0.8185	1.11E-61	456
AlphaFold2.1	0.8179	2.24E-62	453
AlphaFold2.0	0.8173	4.49E-63	452

Table S3. The comparison of D-I-TASSER with different versions of AlphaFold on 176 nun-redundant Hard targets whose structures were released after May 1, 2022. *P*-values were calculated between TM-scores by D-I-TASSER and each AlphaFold program using paired one-sided Student's t-tests. #{TM-score >0.5} is the number of targets with a TM-score >0.5.

Method	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	0.8101	-	164
AlphaFold3	0.7657	1.61E-12	157
AlphaFold2.3	0.7390	2.42E-23	148
AlphaFold2.2	0.7269	5.45E-28	150
AlphaFold2.1	0.7275	4.88E-27	150
AlphaFold2.0	0.7336	1.49E-26	151

Table S4. Comparison of full-chain-level modeling results by D-I-TASSER, AlphaFold2, and AlphaFold2+DeepMSA2 on the 230 multi-domain targets with different number of domains. *P*-values were calculated between TM-scores by D-I-TASSER and AlphaFold2 using paired one-sided Student's t-tests. #{TM-score >0.5} is the number of targets with a TM-score >0.5.

Method	Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	All (230)	0.7196	-	208
	2-domain (167)	0.7142	-	149
	3-domain (37)	0.7468	-	34
	≥4 domain (26)	0.7154	-	25
AlphaFold2 (version 2.2)	All (230)	0.6374	6.52E-28	193
	2-domain (167)	0.6393	2.59E-19	139
	3-domain (37)	0.6272	2.04E-06	30
	≥4 domain (26)	0.6400	5.96E-05	24
AlphaFold2 (version 2.3)	All (230)	0.6379	1.59E-31	194
	2-domain (167)	0.6401	5.34E-22	140
	3-domain (37)	0.6273	1.90E-06	30
	≥4 domain (26)	0.6386	2.41E-05	24
AlphaFold2 +DeepMSA2	All (230)	0.6723	7.86E-34	198
	2-domain (167)	0.6709	6.98E-24	142
	3-domain (37)	0.6842	1.43E-04	33
	≥4 domain (26)	0.6644	6.54E-06	23

Table S5. Comparison of domain-level modeling results between D-I-TASSER, AlphaFold2, and AlphaFold2+DeepMSA2 on the 557 domains that came from 230 multi-domain targets. *P*-values were calculated between TM-scores by D-I-TASSER and AlphaFold2 using paired one-sided Student's t-tests. #{TM-score >0.5} is the number of targets with a TM-score >0.5.

Method	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	0.8577	-	536
AlphaFold2.2	0.8341	1.45E-10	529
AlphaFold2.3	0.8345	2.31E-16	530
AlphaFold2+DeepMSA2	0.8504	1.61E-06	534

Table S6. Comparison of the structure prediction abilities of D-I-TASSER, NBIS-AF2-standard (AlphaFold2), and Wallner group predictions on 62 Template-based modeling (TBM) and 50 Free Modeling (FM) domains from the CASP15 experiment. *P*-values were calculated between TM-scores of D-I-TASSER and AlphaFold2 models using paired one-sided Student's t-tests. #{TM-score>0.5} is the number of predicted domains with a TM-score >0.5.

Method	Domain Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	All (112)	0.878	-	106
	TBM (62)	0.915	-	60
	FM (50)	0.833	-	46
NBIS-AF2-standard (AlphaFold2)	All (112)	0.801	9.35E-09	97
	TBM (62)	0.881	3.89E-04	59
	FM (50)	0.701	3.41E-06	38
Wallner	All (112)	0.809	1.30E-05	97
	TBM (62)	0.875	4.87E-04	58
	FM (50)	0.726	3.16E-03	39

Table S7. Comparison of structure predictions by D-I-TASSER, NBIS-AF2-standard (AlphaFold2), and Wallner group predictions on 55 single-domain and 22 multi-domain targets from the CASP15 experiment. *P*-values were calculated between TM-scores of D-I-TASSER and AlphaFold2 models using paired one-sided Student's t-tests. #{TM-score>0.5} is the number of predicted proteins with a TM-score >0.5.

Method	Target Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	All (77)	0.851	-	72
	Single-domain (55)	0.893	-	52
	Multi-domain (22)	0.747	-	20
NBIS-AF2-standard (AlphaFold2)	All (77)	0.787	3.67E-05	64
	Single-domain (55)	0.870	5.30E-03	51
	Multi-domain (22)	0.578	1.18E-03	13
Wallner	All (77)	0.795	1.11E-03	62
	Single-domain (55)	0.872	4.77E-02	49
	Multi-domain (22)	0.602	4.22E-03	13

Table S8. Results of all 132 groups (server and human) on 'Single-domain Structure Prediction' in CASP15. Data were copied from the CASP15 webpage at https://predictioncenter.org/casp15/zscores_final.cgi?formula=assessors, in which the Group rankings are based on Assessors' formulae, i.e., Assessor Score=1/6*(Z-score$_{GDT_HA}$ + Z-score$_{reLLG_lddt}$ + Z-score$_{ASE}$) + 1/16*(Z-score$_{LDDT}$ + Z-score$_{CAD_aa}$ + Z-score$_{SG}$ + Z-score$_{SC_error}$) + 1/12*(Z-score$_{MolProbity}$ + Z-score$_{BB_error}$ + Z-score$_{DipDiff}$), and two Z-score thresholds (-2.0 or -0.0) were used to excluded models. The D-I-TASSER server was registered as 'UM-TBM' (highlighted in bold) in the Table.

Groups	Rank (Z>-2.0)	Sum Z-score (>-2.0)	Avg Z-score (>-2.0)	Rank (Z>-0.0)	Sum Z-score (>-0.0)	Avg Z-score (>-0.0)	Groups	Rank (Z>-2.0)	Sum Z-score (>-2.0)	Avg Z-score (>-2.0)	Rank (Z>-0.0)	Sum Z-score (>-0.0)	Avg Z-score (>-0.0)
PEZYFoldings	2	28.2448	0.3014	1	70.8310	0.6620	Bhattacharya	67	-49.9268	-0.4580	67	23.1218	0.2121
UM-TBM	**1**	**33.4626**	**0.3070**	**2**	**68.5623**	**0.6290**	Zheng	81	-121.1419	0.1524	68	22.9098	0.5091
Yang-Server	4	20.0022	0.2037	3	61.2772	0.5674	FTBiot0119	69	-76.4316	-0.7012	69	22.8312	0.2095
DFolding	3	25.8625	0.2373	4	61.0695	0.5603	Graphen_Medical	72	-82.3454	-0.1417	70	22.4165	0.3071
Yang	5	16.2255	0.1688	5	59.0136	0.5464	GinobiFold	66	-43.2247	-0.3512	71	22.2223	0.2096
McGuffin	19	2.6214	0.0240	6	49.7262	0.4562	GinobiFold-SER	65	-36.4549	-0.2710	72	21.8518	0.2081
MULTICOM	21	2.3021	0.0211	7	48.8284	0.4480	Seder2022easy	73	-83.5517	-0.7316	73	21.8495	0.2061
MULTICOM_refine	6	13.6109	0.1249	8	48.8258	0.4479	NBIS-AF2-multimer	80	-114.7135	0.0657	74	21.2860	0.4257
BAKER	11	4.3748	0.0401	9	47.9903	0.4403	Yang-Multimer	82	-122.2089	0.1287	75	21.2429	0.4721
MULTICOM_human	20	2.4662	0.0226	10	47.9002	0.4395	ESM-single-sequence	70	-77.8559	-0.4931	76	20.0903	0.2160
MULTICOM_deep	8	9.8068	0.0900	11	46.2544	0.4244	RaptorX-Multimer	85	-128.2922	-0.0065	77	19.6842	0.4374
MULTICOM_qa	9	9.4572	0.0868	12	46.2344	0.4242	Takeda-Shitaka_Lab	84	-124.3999	0.0800	78	19.1241	0.4250
MULTICOM_egnn	7	11.9136	0.1093	13	45.7611	0.4198	Seder2022hard	77	-96.8854	-0.7115	79	19.0334	0.2025
Manifold-E	34	-5.7186	-0.0525	14	45.1978	0.4147	SHT	74	-86.0486	-0.6255	80	18.2032	0.1896
Kiharalab	25	-0.9270	-0.0085	15	45.1273	0.4140	Grudinin	87	-130.8786	-0.0200	81	16.7652	0.3810
MUFold_H	12	4.2297	0.0388	16	44.2540	0.4060	Agemo	76	-90.3782	-0.8292	82	16.1139	0.1478
ColabFold	10	4.9637	0.0455	17	44.0177	0.4038	DFolding-refine	75	-89.0381	-0.7834	83	14.5185	0.1370
colabfold_human	18	2.9425	0.0270	18	43.2644	0.3969	EMBER3D	88	-132.7949	-1.0739	84	13.8847	0.1509
Wallner	14	4.0416	0.0371	19	43.1163	0.3956	CoDock	89	-135.5668	-0.2461	85	12.3283	0.2623
Asclepius	26	-1.5361	0.0043	20	43.0518	0.3986	QUIC	79	-110.2350	-0.9537	86	11.2555	0.1093
bench	16	3.7085	0.0529	21	42.8291	0.3966	PICNIC	86	-129.3268	-1.1307	87	10.7159	0.1051
Manifold	39	-6.4946	-0.0596	22	42.8201	0.3928	Pierce	98	-169.8130	-0.2153	88	10.5512	0.3908
DFolding-server	13	4.0592	0.0372	23	41.9520	0.3849	RostlabUeFOFold	92	-155.9174	-1.2429	89	9.2273	0.1125
Elofsson	59	-23.6888	0.1121	24	41.7392	0.4537	Shen-CAPRI	94	-160.2618	-0.3018	90	8.4582	0.2488
MUFold	15	3.8458	0.0353	25	40.9513	0.3757	UNRES	78	-105.8942	-0.9116	91	8.1109	0.0787
RaptorX	17	2.9560	0.0271	26	40.5804	0.3723	Zou	95	-160.8293	-0.6056	92	8.0372	0.1960
Agemo_mix	32	-4.2573	-0.0391	27	40.4984	0.3715	WL_team	83	-123.6216	-0.9852	93	8.0090	0.0861
ShanghaiTech	47	-11.8306	-0.1085	28	40.1253	0.3681	wuqi	91	-143.2171	-1.1404	94	7.7618	0.0892
UltraFold_Server	22	1.0805	0.0099	29	40.0715	0.3676	ClusPro	100	-174.4625	-0.9634	95	5.6107	0.1336
UltraFold	30	-3.9657	0.0003	30	39.7243	0.3713	Fernandez-Recio	102	-181.0940	-0.9145	96	5.1969	0.1528
B11L	44	-10.2095	0.0372	31	38.2788	0.3753	TB_model_prediction	109	-191.2786	0.0555	97	5.1589	0.3968
GuijunLab-DeepDA	23	-0.4924	-0.0045	32	38.2400	0.3508	AIchemy_LIG2	113	-194.8586	-0.2199	98	4.5994	0.3538
BeijingAIProtein	49	-12.5932	-0.0058	33	38.0237	0.3692	AIchemy_LIG	113	-194.8586	-0.2199	98	4.5994	0.3538
ChaePred	33	-5.5144	-0.0142	34	37.9319	0.3545	AIchemy_LIG3	112	-194.8543	-0.2196	100	4.5960	0.3535
Shennong	28	-2.8649	0.0489	35	37.3845	0.3560	Panlab	90	-143.0865	-1.3127	101	4.3993	0.0404
MultiFOLD	48	-12.1034	-0.1110	36	36.6712	0.3364	Manifold-X	107	-189.7355	-0.4298	102	4.2898	0.2383
GuijunLab-Assembly	27	-2.6529	-0.0243	37	36.4567	0.3345	DELCLAB	93	-157.8925	-1.3803	103	3.8920	0.0401
GuijunLab-Human	36	-6.2358	-0.0209	38	36.2609	0.3389	Kozakov-Vajda	106	-188.8308	-0.9582	104	3.8780	0.1385
Kiharalab_Server	43	-10.1914	-0.0935	39	36.0746	0.3310	ACOMPMOD	111	-193.4552	-1.6853	105	3.7334	0.0479
server_124	40	-6.7143	-0.0616	40	35.8061	0.3285	SHORTLE	101	-174.8588	-1.1541	106	3.6861	0.0723
GuijunLab-Threader	31	-4.1330	-0.0379	41	35.6270	0.3269	TensorLab	116	-198.8770	-0.2615	107	3.5494	0.3227
hFold_human	24	-0.8240	-0.0076	42	35.1927	0.3229	Pan_Server	96	-164.4871	-1.4855	108	2.9232	0.0281
BAKER-SERVER	52	-14.2912	-0.1311	43	35.1349	0.3223	Manifold-LC-E	115	-196.1421	-0.5428	109	2.8952	0.1930
hFold	35	-5.7866	0.0020	44	35.0783	0.3309	Convex-PL	120	-206.2240	-0.0373	110	2.6877	0.4479
NBIS-AF2-standard	29	-2.8881	-0.0265	45	34.6335	0.3177	UTMB	119	-205.8203	0.0299	111	2.5274	0.4212
IntFOLD7	57	-20.4751	-0.1878	46	34.4688	0.3162	FALCON2	104	-182.3897	-1.6672	112	2.4164	0.0226
hks1988	38	-6.4477	-0.0592	47	34.4345	0.3159	FALCON0	103	-182.1541	-1.6650	113	2.4086	0.0225
DMP	64	-36.1319	-0.1055	48	34.0681	0.3549	noxelis	123	-207.4929	0.1014	114	2.4028	0.4806
FoldEver	42	-9.9724	-0.0915	49	33.8486	0.3105	KORP-PL	118	-204.0169	-0.2521	115	2.3884	0.2985
GuijunLab-Meta	37	-6.2794	-0.0213	50	33.6273	0.3143	MESHI_server	99	-172.4795	-1.4010	116	2.3765	0.0313
AP_1	51	-14.1961	-0.1302	51	33.5073	0.3074	MESHI	97	-165.2689	-1.2467	117	2.1919	0.0313
server_122	45	-10.4840	-0.0962	52	33.4822	0.3072	ddquest	122	-207.4390	0.1122	118	2.1296	0.4259
OpenFold	55	-19.6548	-0.1635	53	33.2518	0.3079	Convex-PL-R	121	-207.0127	-0.1688	119	1.9614	0.3269
server_125	46	-10.9894	-0.1008	54	33.0675	0.3034	zax	124	-209.0193	-0.8774	120	1.4544	0.1818
OpenFold-SingleSeq	56	-19.8857	-0.1656	55	32.9778	0.3054	Gonglab-THU	108	-190.9123	-1.7515	121	1.4451	0.0133
server_123	50	-13.1209	-0.1204	56	32.9353	0.3022	bio3d	129	-214.0009	-0.0005	122	1.2791	0.6396
FoldEver-Hybrid	58	-22.2976	-0.0624	57	32.9350	0.3261	MeilerLab	125	-211.5698	0.1434	123	1.2702	0.4234
server_126	41	-8.4517	-0.0775	58	32.8951	0.3018	Cerebra	110	-192.3969	-1.7651	124	1.0310	0.0095
Venclovas	71	-78.8584	-0.1197	59	32.2406	0.4357	Spider	117	-200.0590	-1.4563	125	0.9318	0.0282
ManiFold-serv	53	-14.4926	-0.1330	60	30.3083	0.2781	FEIGLAB	126	-212.2364	-0.0788	126	0.7872	0.2624
TRFold	60	-27.0165	-0.2316	61	29.4315	0.2725	BhageerathH-Pro	105	-185.0512	-1.6801	127	0.6974	0.0068
GuijunLab-RocketX	54	-15.7410	-0.1272	62	29.3099	0.2714	Sun_Tsinghua	127	-213.0850	-1.7766	128	0.6866	0.0312
trComplex	61	-28.1287	-0.2419	63	29.2205	0.2706	PerezLab_Gators	128	-213.5777	-0.5259	129	0.2630	0.0877
XRC_VU	68	-61.0768	-0.0385	64	26.1843	0.3273	CSRC_ICM	132	-217.1089	-1.1089	130	0.2540	0.2540
ShanghaiTech-TS-SER	62	-32.4376	-0.2327	65	26.0202	0.2478	coco	130	-215.3389	-0.6695	131	0.2100	0.1050
Coqualia	63	-35.7053	-0.2472	66	24.0265	0.2310	GatorsML	131	-216.4563	-1.4854	132	0.1091	0.0364

Table S9. Results of all 98 groups (server and human) on 'Inter-domain Structure Prediction' in CASP15. Data are copied from the official CASP15 webpage at https://predictioncenter.org/casp15/zscores_interdomain.cgi, in which the ranking of the groups is based on the linear combination Z-score (F1) + Z-score (Jaccard score) + Z-score (QS_best), with models having a Z-score below the tolerance threshold (-0.0) excluded. The D-I-TASSER server was registered as 'UM-TBM' (highlighted in bold) in the table.

#	Groups	Sum Z-score (>-0.0)	Avg Z-score (>-0.0)	#	Groups	Sum Z-score (>-0.0)	Avg Z-score (>-0.0)
1	**UM-TBM**	**35.5277**	**1.7764**	50	Kiharalab	5.4940	0.2747
2	Yang-Server	24.9602	1.2480	51	MULTICOM_deep	5.4897	0.2889
3	Yang	19.7115	0.9856	52	Seder2022easy	5.3198	0.2800
4	PEZYFoldings	18.0578	1.2039	53	GuijunLab-DeepDA	4.8945	0.2576
5	Manifold	14.9308	0.7858	54	XRC_VU	4.8243	0.6892
6	Venclovas	14.5386	0.7652	55	ColabFold	4.7985	0.2399
7	server_124	14.0810	0.7040	56	colabfold_human	4.7985	0.2399
8	DFolding	13.1098	0.6555	57	GuijunLab-Assembly	4.3734	0.2302
9	bench	12.0811	0.6041	58	Wallner	4.1617	0.2312
10	BAKER-SERVER	12.0030	0.6002	59	FoldEver	3.9173	0.2062
11	Manifold-E	11.6732	0.6144	60	MULTICOM	3.9011	0.2167
12	DFolding-server	11.5870	0.6098	61	MULTICOM_human	3.6862	0.2048
13	server_126	10.9291	0.5465	62	GuijunLab-Meta	3.6577	0.1925
14	Shennong	10.1011	0.5051	63	MULTICOM_qa	3.5934	0.1797
15	RaptorX	9.3845	0.4692	64	GuijunLab-Human	3.4694	0.1826
16	IntFOLD7	9.0429	0.4759	65	FoldEver-Hybrid	3.3126	0.2366
17	BAKER	8.8620	0.4431	66	MULTICOM_egnn	3.1465	0.1573
18	server_123	8.5973	0.4299	67	GinobiFold	2.7459	0.1615
19	Asclepius	8.5891	0.4521	68	Coqualia	2.7459	0.1615
20	MultiFOLD	8.2488	0.4583	69	Cerebra	2.5507	0.1500
21	DFolding-refine	8.1694	0.4300	70	MUFold	2.4377	0.1219
22	B11L	8.0841	0.4491	71	GuijunLab-RocketX	2.4240	0.1276
23	DMP	7.7417	0.5530	72	GuijunLab-Threader	2.4147	0.1342
24	MUFold_H	7.5603	0.3780	73	Bhattacharya	2.3720	0.1248
25	hFold	7.1212	0.4451	74	SHT	2.2983	0.1149
26	OpenFold-SingleSeq	7.0733	0.3723	75	BhageerathH-Pro	2.2781	0.1627
27	OpenFold	7.0733	0.3723	76	GinobiFold-SER	2.2577	0.1411
28	ShanghaiTech	7.0584	0.3529	77	FALCON2	2.2265	0.1113
29	ManiFold-serv	6.9819	0.3675	78	FALCON0	2.2265	0.1113
30	Graphen_Medical	6.8295	0.4878	79	hks1988	2.1535	0.1077
31	AP_1	6.8095	0.3405	80	**NBIS-AF2-standard**	**2.1141**	**0.1057**
32	Elofsson	6.6876	0.3520	81	Pan_Server	2.0335	0.1070
33	Agemo_mix	6.6477	0.3499	82	Gonglab-THU	1.8164	0.1068
34	Panlab	6.5871	0.3294	83	DELCLAB	1.1885	0.0660
35	McGuffin	6.5509	0.3448	84	ESM-single-sequence	1.1811	0.1312
36	TRFold	6.4502	0.3794	85	UNRES	1.1657	0.0833
37	MULTICOM_refine	6.4210	0.3379	86	QUIC	1.1187	0.0559
38	server_122	6.1989	0.3099	87	PICNIC	1.1002	0.0550
39	BeijingAIProtein	6.1858	0.3639	88	ShanghaiTech-TS-SER	0.8613	0.0538
40	UltraFold	6.1858	0.3639	89	Seder2022hard	0.5910	0.0591
41	UltraFold_Server	6.1858	0.3437	90	SHORTLE	0.5900	0.5900
42	server_125	6.0672	0.3034	91	wuqi	0.4176	0.0464
43	Agemo	5.9827	0.3519	92	MESHI_server	0.1894	0.0947
44	FTBiot0119	5.9501	0.2975	93	EMBER3D	0.1591	0.0159
45	ChaePred	5.7616	0.2881	94	Manifold-LC-E	0.0809	0.0809
46	WL_team	5.7417	0.3022	95	Manifold-X	0.0809	0.0809
47	Kiharalab_Server	5.7358	0.2868	96	RostlabUeFOFold	0.0346	0.0087
48	trComplex	5.6135	0.3302	97	MESHI	0.0000	0.0000
49	hFold_human	5.5688	0.3094	98	ACOMPMOD	0.0000	0.0000

Table S10. The comparison of D-I-TASSER with different versions of AlphaFold (including AlphaFold3, AlphaFold2.3, AlphaFold2.2, AlphaFold2.1, and AlphaFold2.0) on 50 Free Modeling (FM) domains and 22 multi-domain targets from the CASP15 experiment. *P*-values were calculated between TM-scores by D-I-TASSER and others using paired one-sided Student's t-tests. #{TM-score >0.5} is the number of targets with a TM-score >0.5.

Method	Target Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	FM (50)	0.8326	-	46
	Multi-domain (20)	0.7419	-	18
AlphaFold2.0	FM (50)	0.7149	1.04E-05	37
	Multi-domain (20)	0.5988	8.59E-03	13
AlphaFold2.1	FM (50)	0.7230	8.34E-06	38
	Multi-domain (20)	0.5980	6.81E-03	11
AlphaFold2.2	FM (50)	0.7212	6.10E-05	37
	Multi-domain (20)	0.5947	5.34E-03	12
AlphaFold2.3	FM (50)	0.7262	2.55E-04	38
	Multi-domain (20)	0.5920	8.59E-03	13
AlphaFold3	FM (50)	0.7265	4.65E-04	39
	Multi-domain (20)	0.6088	2.00E-02	12

Table S11. The structure prediction accuracy of D-I-TASSER and AlphaFold2 on 1,907 full-chain sequences from the human genome that have experimentally solved structures. These sequences contain 1,147 cases with single-domain and 760 cases with multi-domain structures. *P*-values were calculated between TM-scores of D-I-TASSER and AlphaFold2 models using paired one-sided Student's t-tests. #{TM-score>0.5} is the number of predicted proteins with a TM-score >0.5.

Method	Target Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	All (1,907)	0.931	-	1,872
	Single-domain (1,147)	0.918	-	1,119
	Multi-domain (760)	0.951	-	753
AlphaFold2	All (1,907)	0.916	3.17E-130	1,865
	Single-domain (1,147)	0.903	5.69E-84	1,113
	Multi-domain (760)	0.935	1.07E-47	752

Table S12. The results are the same as shown in Table S9, but the 1,907 proteins are categorized into two categories of 'Easy-zone' and 'Hard-zone' based on the D-I-TASSER and AlpahFold2 results. The 'Easy-zone' targets refer to those for which both D-I-TASSER and AlphaFold2 can achieve a TM-score >0.8, while the 'Hard-zone' targets are those for which at least one method performs poorly with a TM-score <0.8. *P*-values were calculated between TM-scores of D-I-TASSER and AlphaFold2 models using paired one-sided Student's t-tests. #{TM-score>0.5} is the number of predicted proteins with a TM-score >0.5.

Method	Target Type	TM-score	*P*-value	#{TM-score>0.5}
D-I-TASSER	All (1,907)	0.931	-	1,872
	Easy-zone (1,659)	0.966	-	1,659
	Hard-zone (248)	0.699	-	213
AlphaFold2	All (1,907)	0.916	3.17E-130	1,865
	Easy-zone (1,659)	0.958	2.47E-97	1,659
	Hard-zone (248)	0.633	1.17E-26	206

Table S13. Statistical summary of the top 20 most abundant prediction results for ligand-binding interactions, EC terms, and GO terms (BP, CC, and MF) for foldable full-chain human proteins. #{protein} is the number of proteins with the corresponding labels.

Type	ID	Name	#{protein}
Ligand-binding	ANP	ADENYLYL IMIDODIPHOSPHATE	308
	FES	DI-MU-SULFIDO-DIIRON	306
	F3S	TRI-MU-SULFIDO-MU3-SULFIDO-TRIIRON	228
	ADP	ADENOSINE 5'-DIPHOSPHATE	226
	CA	CALCIUM	222
	RET	RETINAL	212
	SF4	TETRA-MU3-SULFIDO-TETRAIRON	204
	QNA	1~{A}~{R},7~{B}~{S})-5-FLUORANYL-2,2-BIS(OXIDANYL)-1~{A},7~{B}-DIHYDRO-1~{H}-CYCLOPROPA[C][1,2]BENZOXABORININE-4-CARBOXYLIC ACID	203
	CLR	CHOLESTEROL	164
	HEM	PROTOHEME	147
	ATP	ADENOSINE-5'-TRIPHOSPHATE	131
	FAD	FLAVIN ADENINE DINUCLEOTIDE	126
	GDP	GUANOSINE-5'-DIPHOSPHATE	123
	GNP	PHOSPHOAMINOPHOSPHONIC ACID-GUANYLATE ESTER	115
	ZN	ZINC ION	109
	NA	SODIUM ION	108
	NAD	NICOTINAMIDE-ADENINE-DINUCLEOTIDE	104
	FMN	FLAVIN MONONUCLEOTIDE	98
	NAP	NICOTINAMIDE-ADENINE-DINUCLEOTIDE PHOSPHATE	90
	PTY	PHOSPHATIDYLETHANOLAMINE	82
EC	3.2.1.17	Lysozyme	426
	2.7.7.6	DNA-directed RNA polymerase	338
	1.17.1.4	Xanthine dehydrogenase	335
	2.3.1.86	Fatty-acyl-CoA synthase	295
	3.4.24.69	Bontoxilysin	289
	1.17.3.2	Xanthine oxidase	280
	2.7.11.1	Non-specific serine/threonine protein kinase	258
	1.4.1.13	Glutamate synthase (NADPH)	233
	2.3.1.85	Fatty-acid synthase	167
	1.4.7.1	Glutamate synthase (ferredoxin)	162
	3.2.1.18	Exo-alpha-sialidase	160
	1.9.3.1	Cytochrome-c oxidase	150
	4.3.2.2	Adenylosuccinate lyase	150
	3.6.5.2	Small monomeric GTPase	148
	2.7.7.7	DNA-directed DNA polymerase	142
	4.2.1.2	Fumarate hydratase	138
	3.2.1.97	Endo-alpha-N-acetylgalactosaminidase	125
	3.2.1.41	Pullulanase	122
	2.7.10.1	Receptor protein-tyrosine kinase	116
	4.3.2.1	Argininosuccinate lyase	103
BP	GO:0055114	oxidation-reduction process	1,026
	GO:0043547	positive regulation of GTPase activity	772
	GO:0030335	positive regulation of cell migration	687
	GO:0009612	response to mechanical stimulus	678
	GO:0098542	defense response to other organism	628
	GO:0097305	response to alcohol	616
	GO:0019058	viral life cycle	589
	GO:0007601	visual perception	584

	GO:0007613	memory	577
	GO:0044248	cellular catabolic process	573
	GO:0018298	protein-chromophore linkage	572
	GO:0050806	positive regulation of synaptic transmission	568
	GO:0001101	response to acid chemical	567
	GO:0007612	learning	561
	GO:0000302	response to reactive oxygen species	561
	GO:0050678	regulation of epithelial cell proliferation	552
	GO:0019229	regulation of vasoconstriction	551
	GO:1901654	response to ketone	551
	GO:0009617	response to bacterium	548
	GO:1903532	positive regulation of secretion by cell	547
CC	GO:0005829	cytosol	3,085
	GO:0070062	extracellular exosome	2,362
	GO:0016021	integral component of membrane	1,556
	GO:0005887	integral component of plasma membrane	1,555
	GO:0005886	plasma membrane	1,239
	GO:0044444	cytoplasmic part	1,175
	GO:0071944	cell periphery	983
	GO:0005654	nucleoplasm	973
	GO:0005615	extracellular space	945
	GO:0005789	endoplasmic reticulum membrane	912
	GO:0005634	nucleus	910
	GO:1903561	extracellular vesicle	846
	GO:0043231	intracellular membrane-bounded organelle	814
	GO:0005737	cytoplasm	801
	GO:0000139	Golgi membrane	769
	GO:0031988	membrane-bounded vesicle	762
	GO:0043005	neuron projection	703
	GO:0044424	intracellular part	666
	GO:0036477	somatodendritic compartment	637
	GO:0005911	cell-cell junction	632
MF	GO:0046872	metal ion binding	1,754
	GO:0043169	cation binding	1,490
	GO:0032550	purine ribonucleoside binding	1,432
	GO:0035639	purine ribonucleoside triphosphate binding	1,430
	GO:0032559	adenyl ribonucleotide binding	1,387
	GO:0005524	ATP binding	1,369
	GO:0042802	identical protein binding	1,255
	GO:0042803	protein homodimerization activity	886
	GO:0008092	cytoskeletal protein binding	878
	GO:0046914	transition metal ion binding	708
	GO:0003676	nucleic acid binding	693
	GO:0019842	vitamin binding	658
	GO:0046983	protein dimerization activity	638
	GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	547
	GO:0008270	zinc ion binding	517
	GO:0044822	poly(A) RNA binding	500
	GO:0005509	calcium ion binding	493
	GO:0005501	retinoid binding	463
	GO:0008020	G-protein coupled photoreceptor activity	463
	GO:0005515	protein binding	453

REFERENCES

1. UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523-D531 (2023).
2. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
3. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
4. Chen, I.M.A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research* **47**, D666-D677 (2019).
5. Hunter, S. et al. EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res* **42**, D600-606 (2014).
6. Steinegger, M., Mirdita, M. & Soding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* **16**, 603-606 (2019).
7. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
8. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research* **38**, e191-e191 (2010).
9. Zhang, C., Zheng, W., Mortuza, S.M., Li, Y. & Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **36**, 2105-2112 (2019).
10. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173-175 (2012).
11. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).
12. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research* **45**, D170-D176 (2016).
13. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).
14. Suzek, B.E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2014).
15. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature Communications* **9**, 2542 (2018).
16. Li, Y. et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLOS Computational Biology* **17**, e1008865 (2021).
17. Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics* **87**, 1149-1164 (2019).
18. Zheng, W. et al. Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLOS Computational Biology* **15**, e1007411 (2019).
19. He, B., Mortuza, S.M., Wang, Y., Shen, H.-B. & Zhang, Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics* **33**, 2296-2306 (2017).
20. Shrestha, R. et al. Assessing the accuracy of contact predictions in CASP13. *Proteins: Structure, Function, and Bioinformatics* **87**, 1058-1068 (2019).
21. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices11Edited by G. Von Heijne. *Journal of Molecular Biology* **292**, 195-202 (1999).

22. Jones, D.T. & Kandathil, S.M. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* **34**, 3308-3315 (2018).
23. Liu, Y., Palmedo, P., Ye, Q., Berger, B. & Peng, J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. *Cell Systems* **6**, 65-74.e63 (2018).
24. Adhikari, B., Hou, J. & Cheng, J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* **34**, 1466-1472 (2017).
25. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences* **110**, 15674 (2013).
26. Seemayer, S., Gruber, M. & Söding, J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics* **30**, 3128-3130 (2014).
27. Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. & Rost, B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15**, 85 (2014).
28. Buchan, D.W.A. & Jones, D.T. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. *Proteins: Structure, Function, and Bioinformatics* **86**, 78-83 (2018).
29. Yan, R., Xu, D., Yang, J., Walker, S. & Zhang, Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific Reports* **3**, 2619 (2013).
30. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nature Methods* **12**, 7-8 (2015).
31. Burley, S.K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research* **47**, D464-D474 (2018).
32. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915 (1992).
33. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics* **23**, 566-579 (1995).