nature portfolio

https://doi.org/10.1038/s41596-022-00728-0

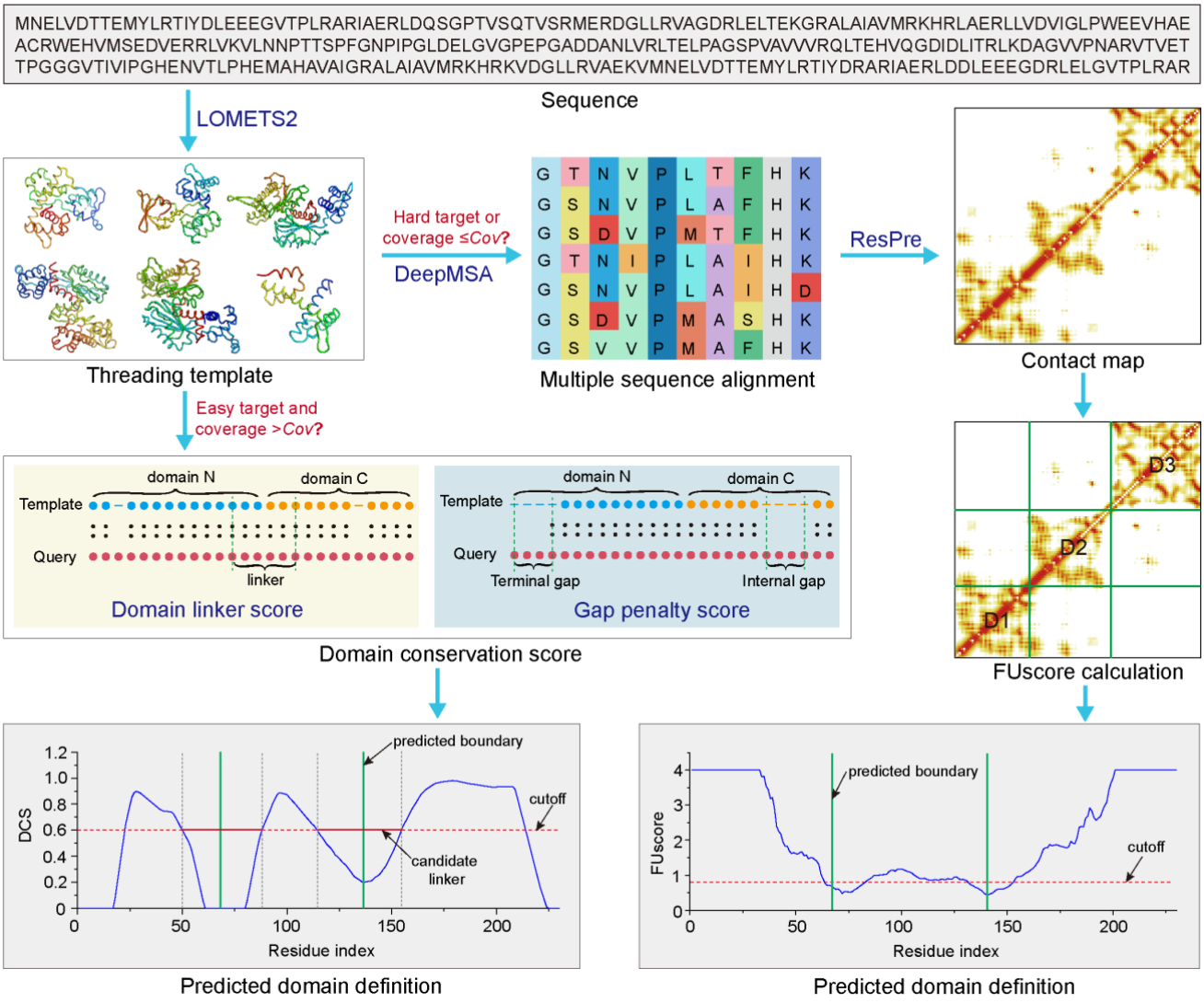
Supplementary information

I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction

In the format provided by the authors and unedited

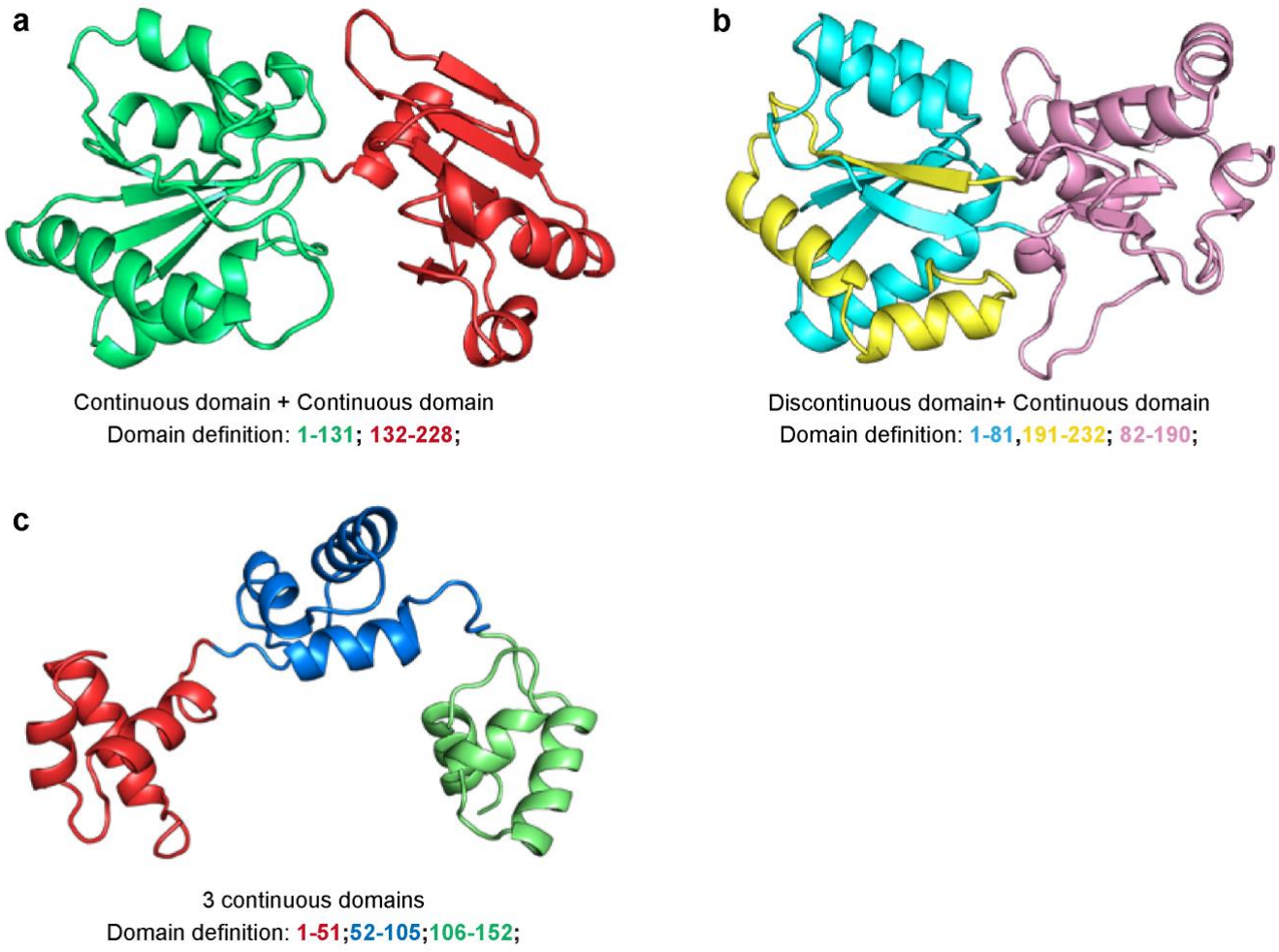
Supplementary Information

Supplementary Figures



Supplementary Figure 1

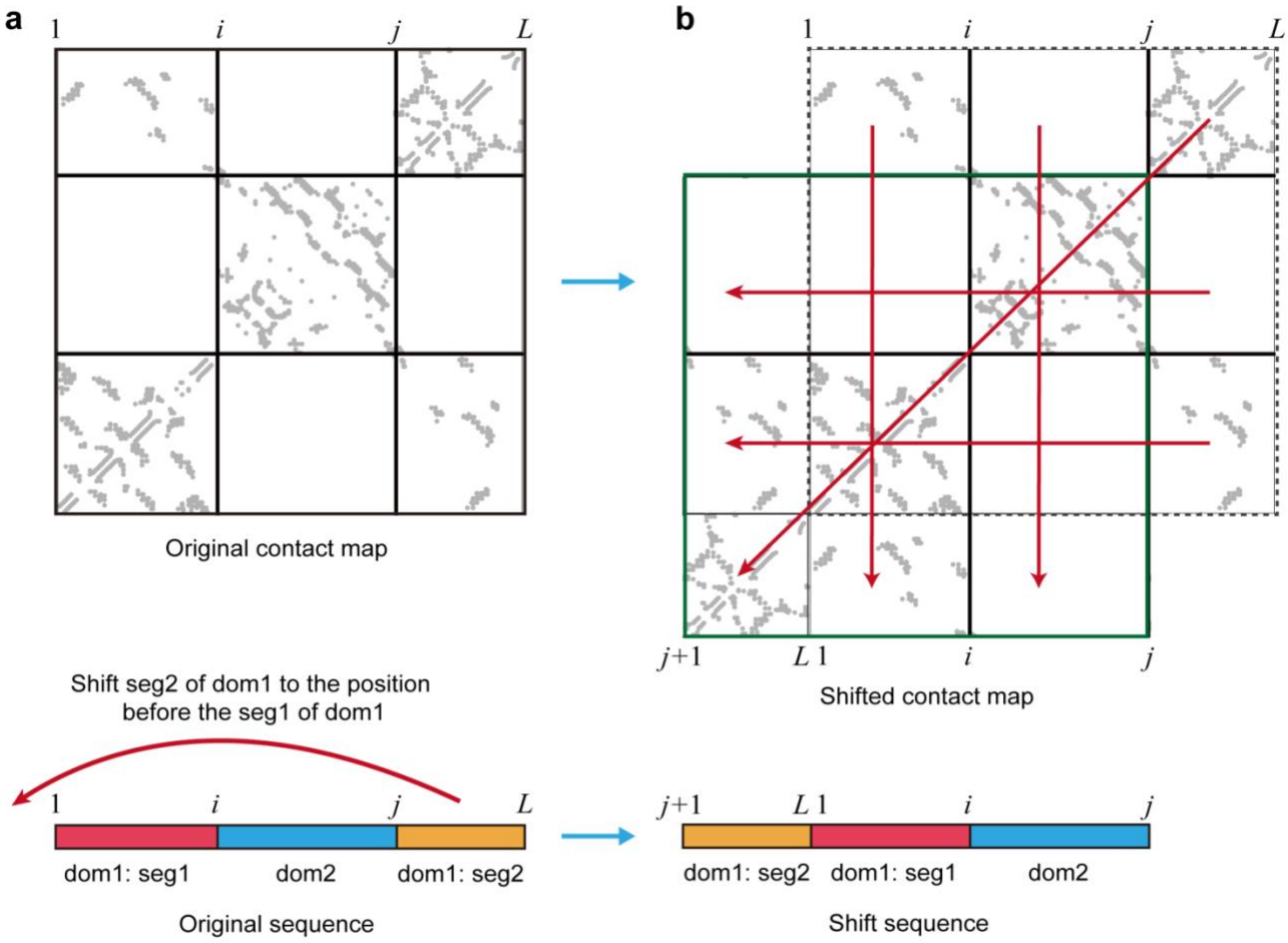
Pipeline of the domain boundary prediction based on ThreaDom and FUpred. Starting from the query amino acid sequence, LOMETS2 is first used to create multiple template alignments from the PDB. If the protein is defined as an Easy target by LOMETS2 and the alignment coverage is larger than a cutoff (*Cov*=95%), ThreaDom is employed to predict the domain boundaries according to the domain conservation score. Otherwise, the domain boundary will be predicted through FUpred by maximizing the number of intra-domain contacts and minimizing the number of inter-domain contacts on the contact-map predicted by a deep-learning based neural network program, ResPRE.



Supplementary Figure 2

Example of continuous and discontinuous domain.

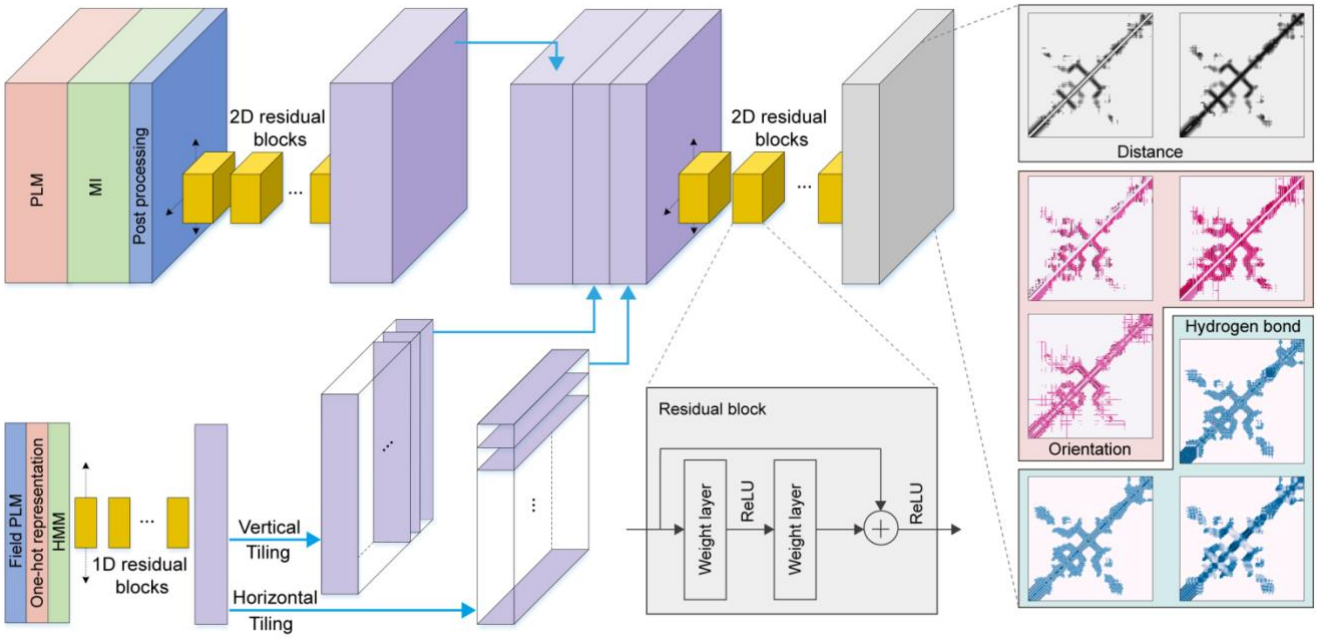
(**a**) A protein (PDBID: 2qbuA) contains two continuous domains, where the first domain (green) ranges from residue 1 to residue 131, and the second domain (red) covers residues from 132 to 228. (**b**) A protein (PDBID: 1atgA) consists of a discontinuous domain and a continuous domain. The first domain is a discontinuous domain which contains two separate segments at the sequence level, where the first segment (cyan) ranges from residue 1 to residue 81, and the second segment (yellow) ranges from residue 191 to residue 232. The second domain is a continuous domain inserted between the two segments of the discontinuous domain, and it covers the residues from 82 to 190. (**c**) A protein (PDBID: 1h88C) contains three continuous domains, where the first domain (red) ranges from residue 1 to residue 51, the second domain (blue) covers residues from 52 to 105, and the third domain (green) includes residues from 106 to 152.



Supplementary Figure 3

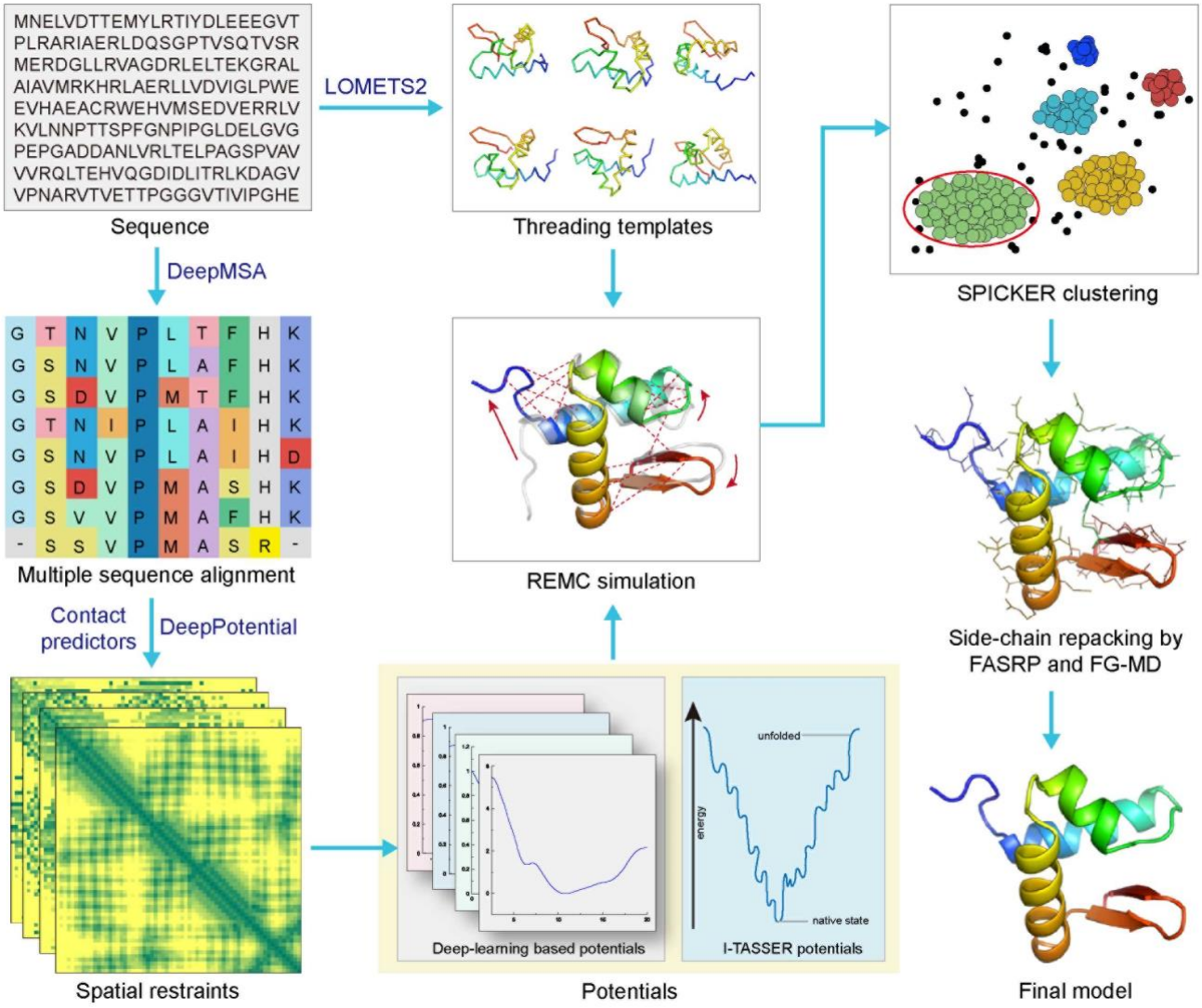
Example (PDBID:1bvp1) to show the contact map shift for the FUscore calculation of the discontinuous domain.

(a) The contact map predicted by ResPre, where residues 1 to *i* belong to the first segment of the first discontinuous domain (dom1: seg1, red), and the second segment of the first domain (dom1: seg2, orange) includes residues *j*+1 to *L*. The second domain (blue) is a continuous domain which covers residues from *i*+1 to *j*. The illustration of the sequence is shown below the contact map, where different colors indicate different domains or segments. (**b**) The contact map and the corresponding sequence illustration after shifting the second segment of the first domain to the position before the first segment of the first domain.



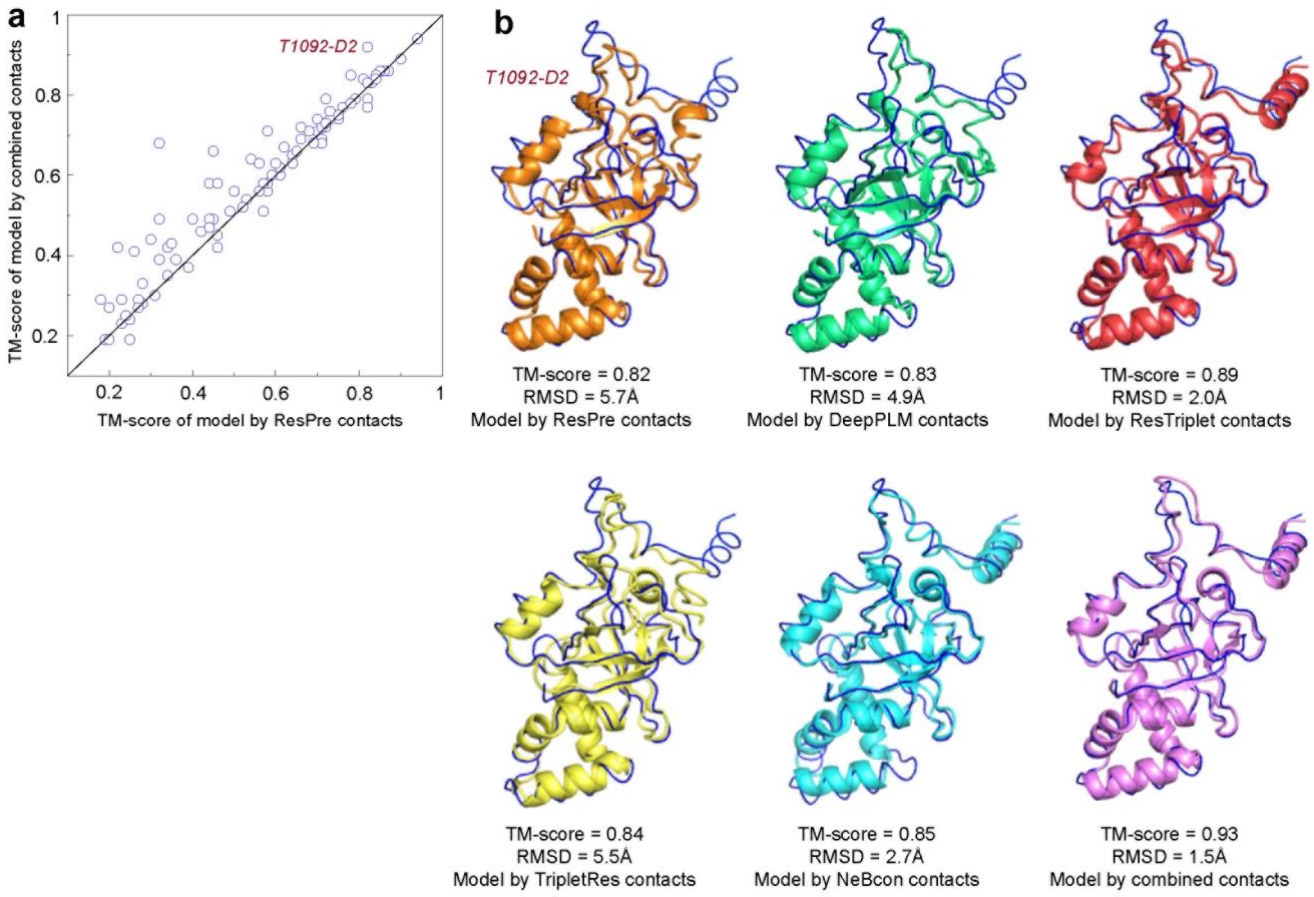
Supplementary Figure 4

Outline of DeepPotential for distance, orientation and hydrogen bond network prediction, where PLM, MI, and HMM represent the PseudoLikelihhod Maximized Potts model, Mutual Information, and Hidden Markov Model, respectively.



Supplementary Figure 5

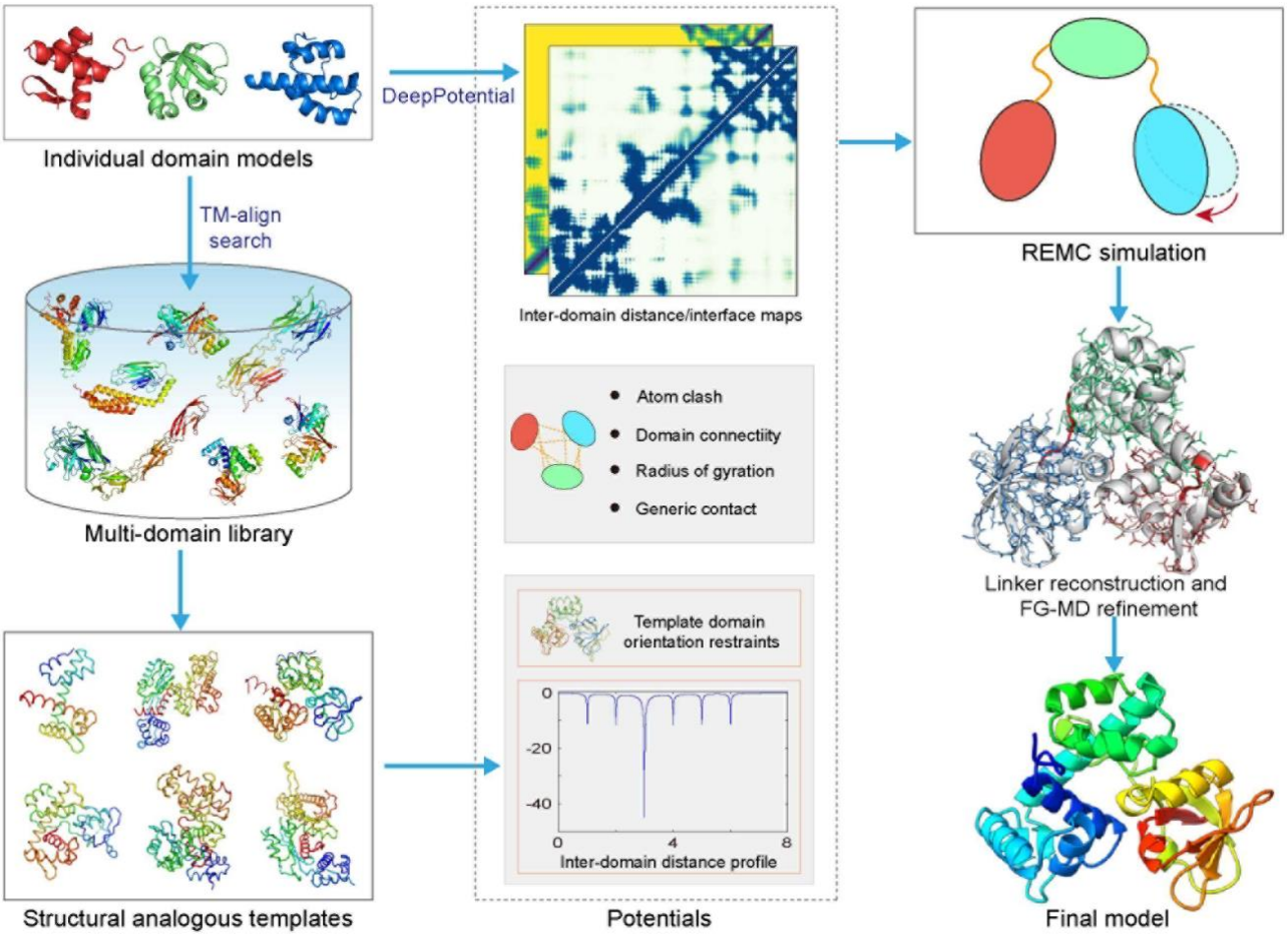
Outline of D-I-TASSER for individual domain model prediction, where contact predictors include ResPre[1], DeepPLM[2], ResTriplet[3], TripletRes[4], and NeBcon[5].



Supplementary Figure 6

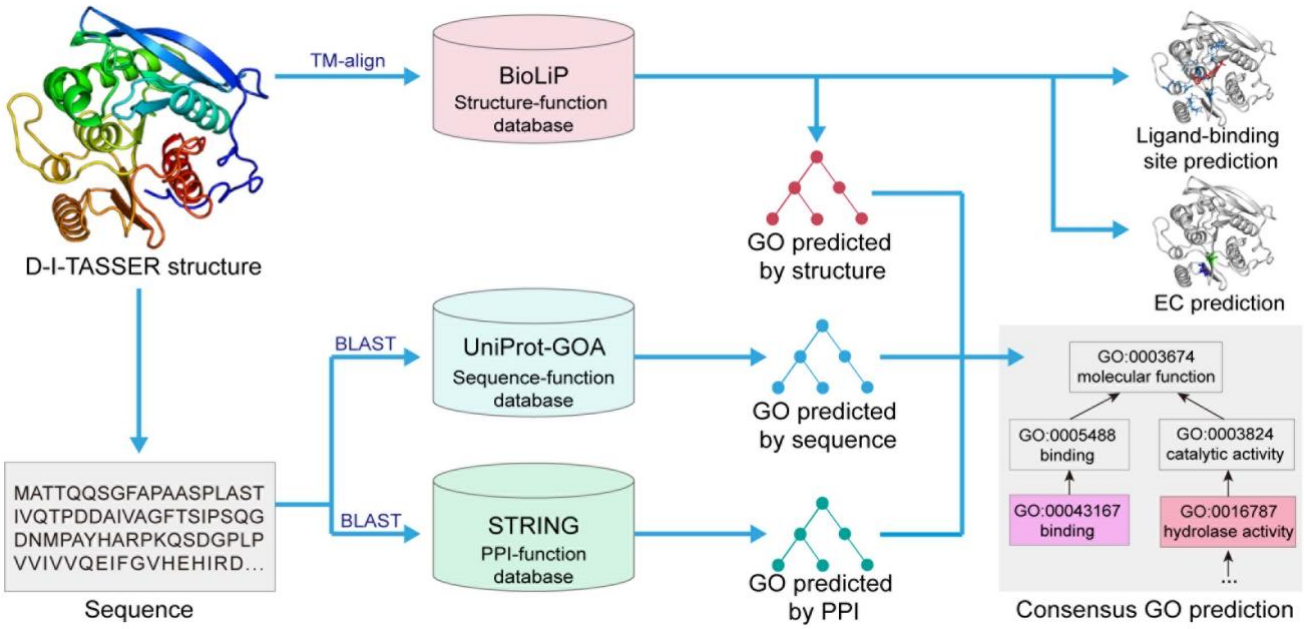
Comparison between models generated by using contacts predicted by the individual predictor and combined contacts from different predictors.

(a) Comparison between TM-scores of models generated by using ResPre contacts and that by using combined contacts of different predictors. (**b**) An example (T1092-D2) to show models generated by using contacts predicted by different methods, where the thin blue line and color cartoons represent the native structure and predicted models, respectively.



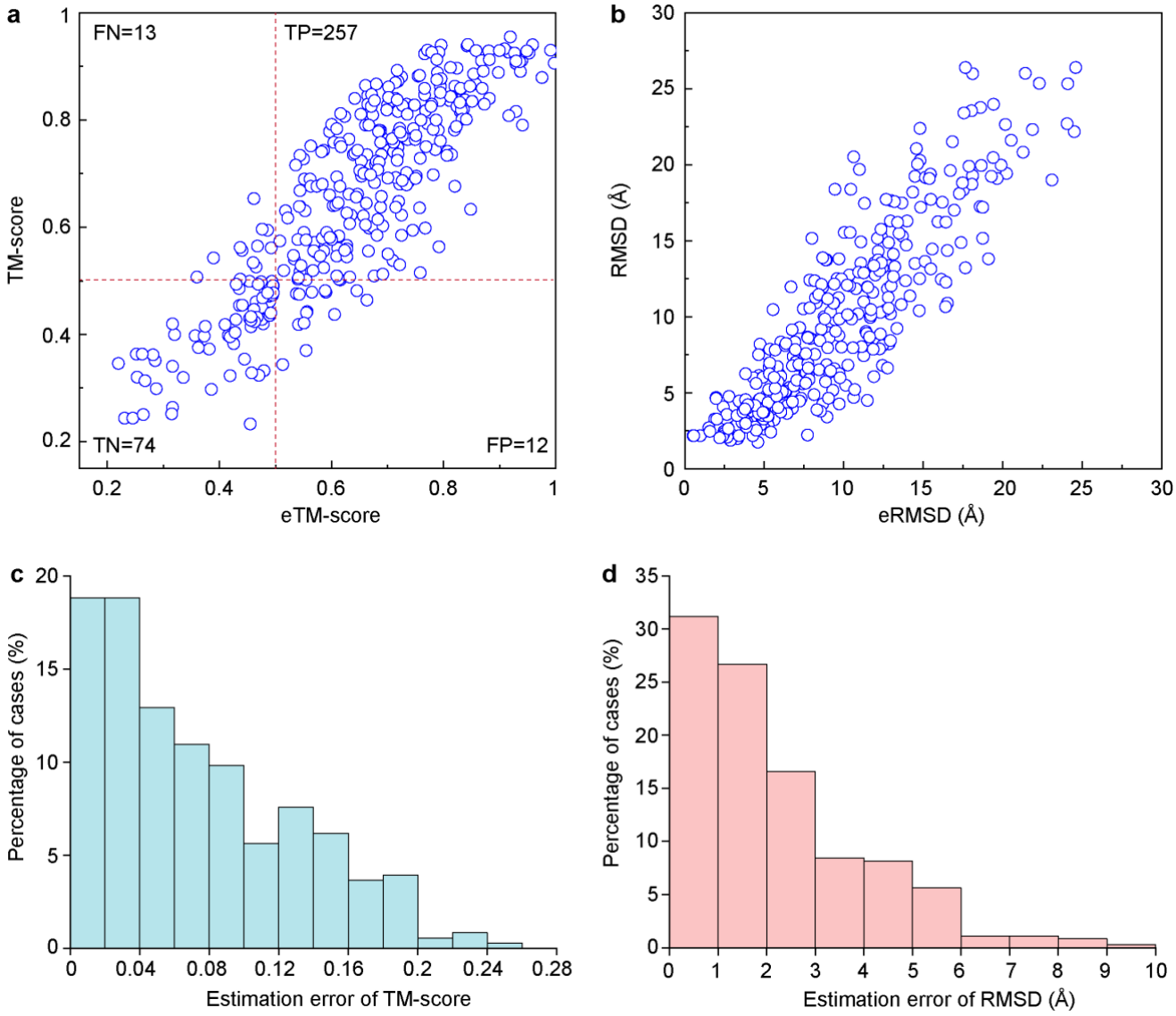
Supplementary Figure 7

Pipeline of DEMO for domain assembly. Starting from individual domain structures, templates are first identified by structurally threading the domains through a non-redundant multi-domain structural library using TM-align. Meanwhile, the inter-domain distance map and interface map are predicted by DeepPotential. Replica-exchange Monte Carlo simulations are then used to assemble the domain structures under the guidance of inter-domain distance profiles and orientations deduced from the templates, inter-domain distance and interface restraints predicted by DeepPotential, and the inherent knowledge-based force field. Finally, the model with the lowest energy is selected for linker reconstruction and side-chain refinement.



Supplementary Figure 8

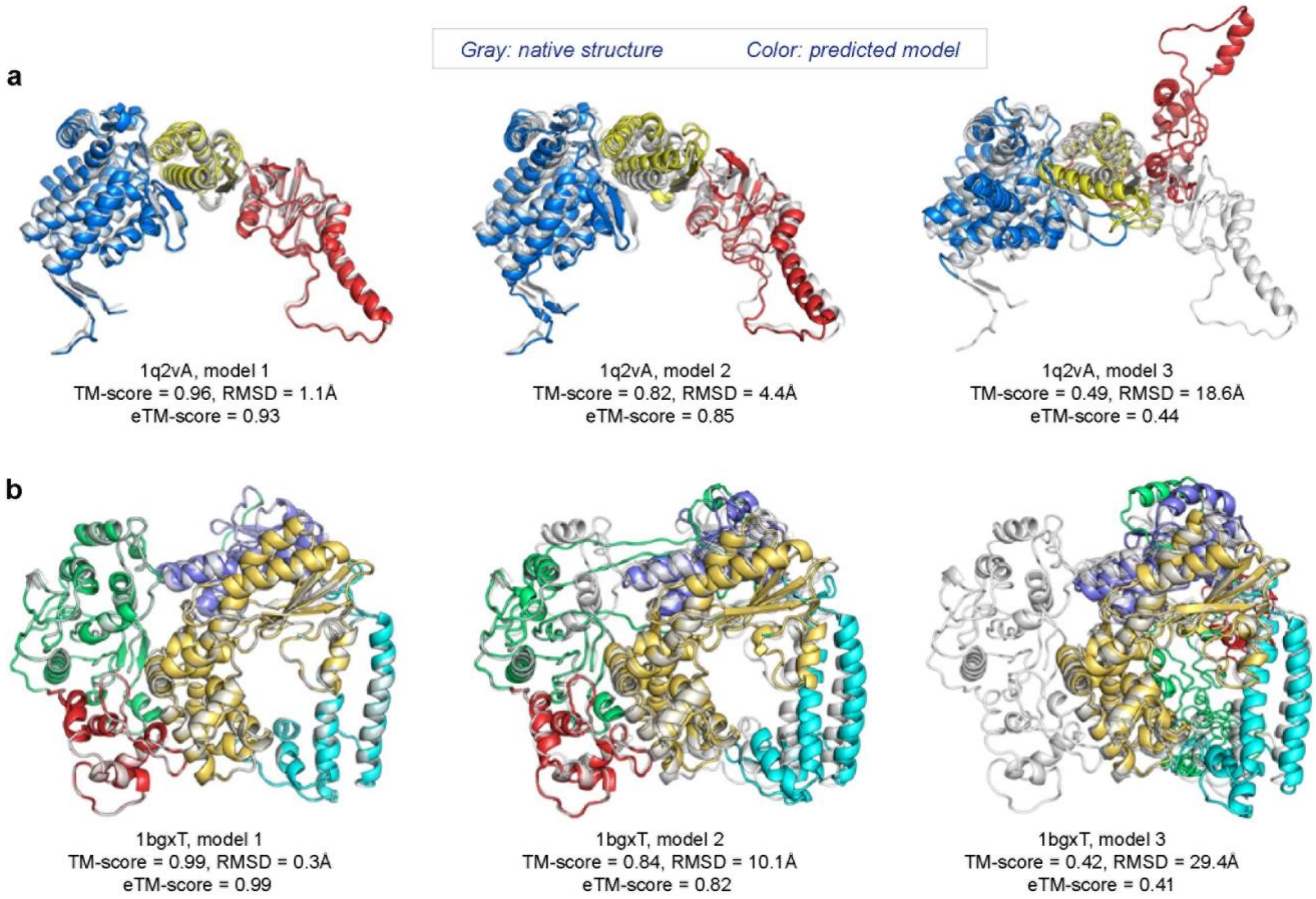
Pipeline of COFACTOR for protein function prediction based on the D-I-TASSER predicted structure, sequence, and protein-protein interaction (PPI). The Gene Ontology (GO) results are determined by a consensus of the structure-, sequence- and PPI-based predictions, while the ligand-binding site and Enzyme Commission (EC) are predicted by structure-based template transfer.



Supplementary Figure 9

Relationship between the eTM-score/eRMSD and the actual TM-score/RMSD to the native, and the distribution of the estimation error for TM-score and RMSD.

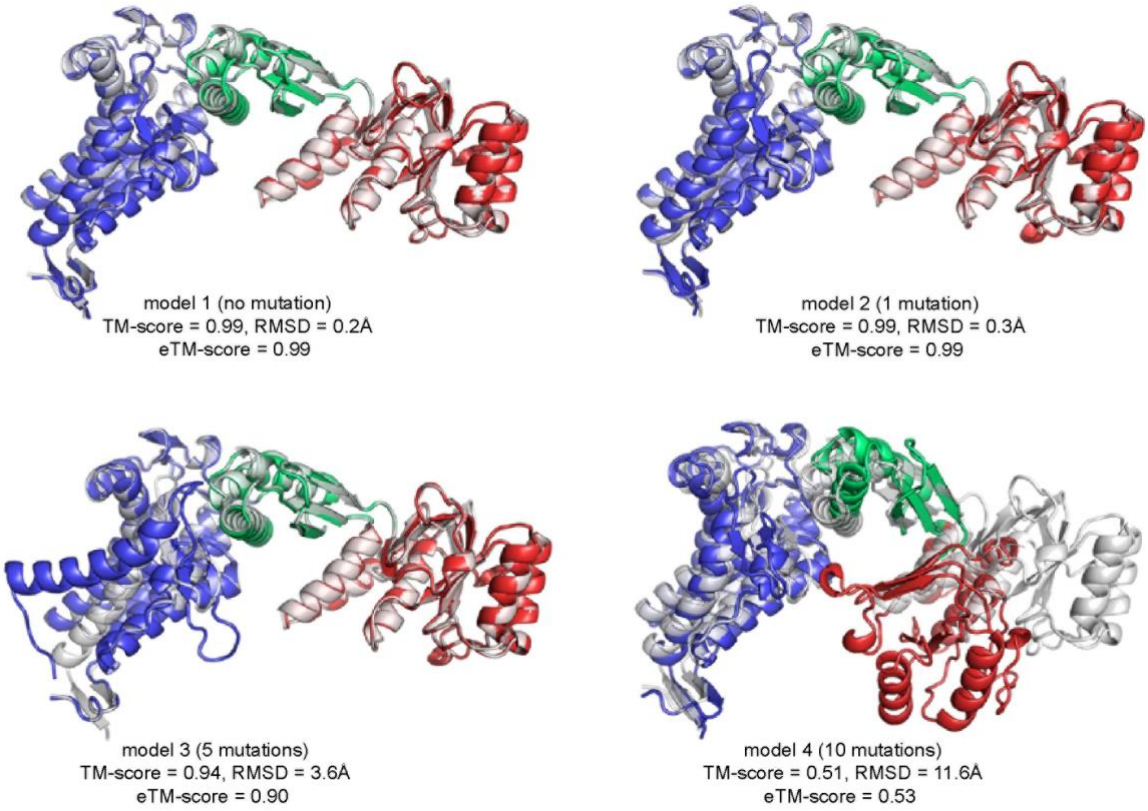
(**a**) The relationship between the actual TM-score and the eTM-score of the first model generated by I-TASSER-MTD, where TP, FP, TN, and FN represent the number of true positive, false positive, true negative, and false negative cases with correct global folds (TM-scores > 0.5). (**b**) The relationship between the eRMSD and the actual RMSD of the first model generated by I-TASSER-MTD. (**c**) Distribution of the estimation error for TM-score. (**d**) Distribution of the estimation error for RMSD.



Supplementary Figure 10

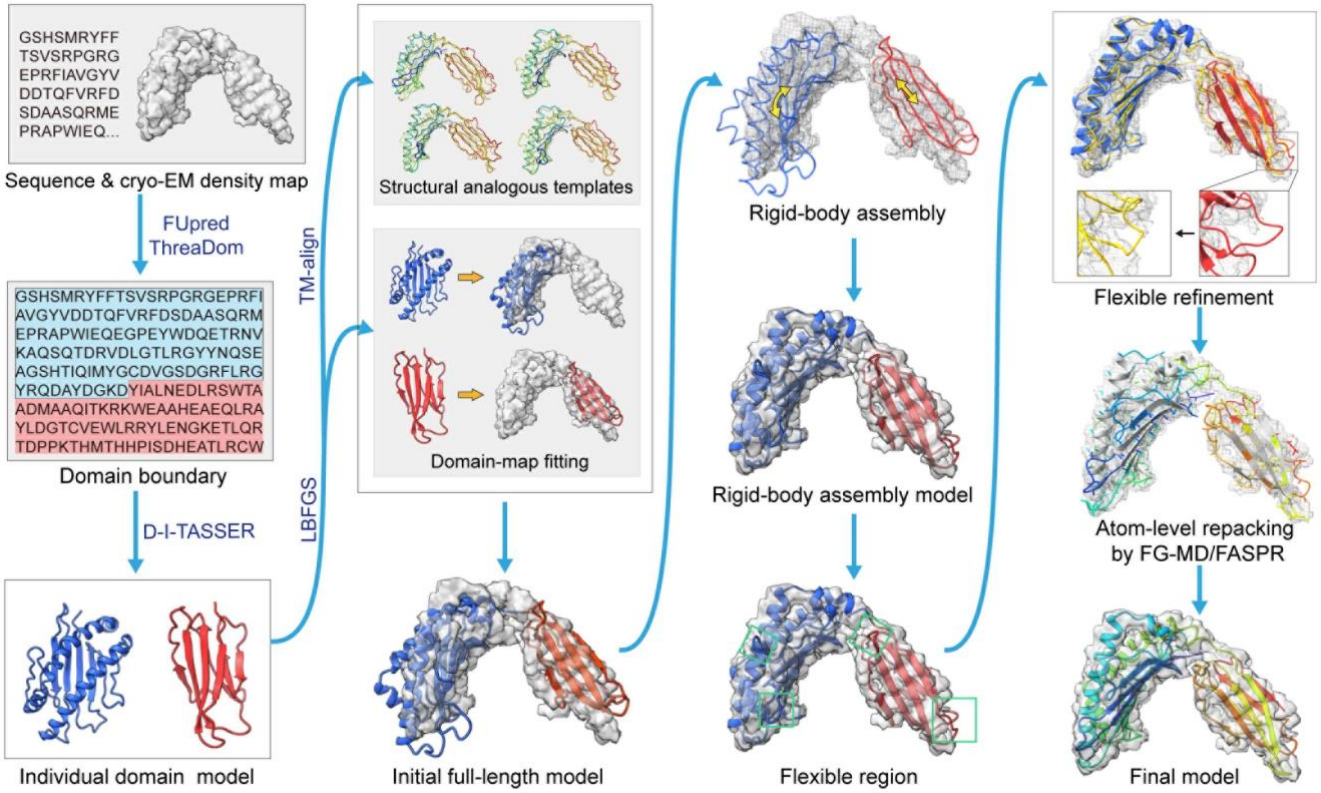
Representative examples show different predicted models with different eTM-scores, where gray and color models indicate native structure and predicted models, respectively, and different domains in predicted models are marked by different colors.

(**a**) A protein (PDBID: 1q2vA) consists of 2 discontinuous domains and 1 continuous domain. All domains were assembled with correct orientations in the first model, and thus obtaining both high TM-score = 0.96 and eTM-score = 0.93. The red domain has some shift in the second model compared to the native, resulting in reduced TM-score = 0.82 and eTM-score = 0.85 for the full-length model. In the third model, the yellow and red domains were assembled with completely wrong orientations, which results in lower TM-score = 0.49 and eTM-score = 0.44 for the full-length model. (**b**) A protein (PDBID: 1bgxT) consists of 2 discontinuous domains and 2 continuous domains. In the first model, all domains were assembled in completely correct orientations, resulting in a full-length model with very high TM-score = 0.99 and eTM-score = 0.99. In the second model, the cyan and purple domains were assembled with some shifts compared to the native, which results in a reduced TM-score = 0.84 and eTM-score = 0.82 for the full-length model. The red and green domains assembled in completely wrong orientations in the third model, leading to a low TM-score = 0.42 and a low eTM-score = 0.41 for the full-length model.



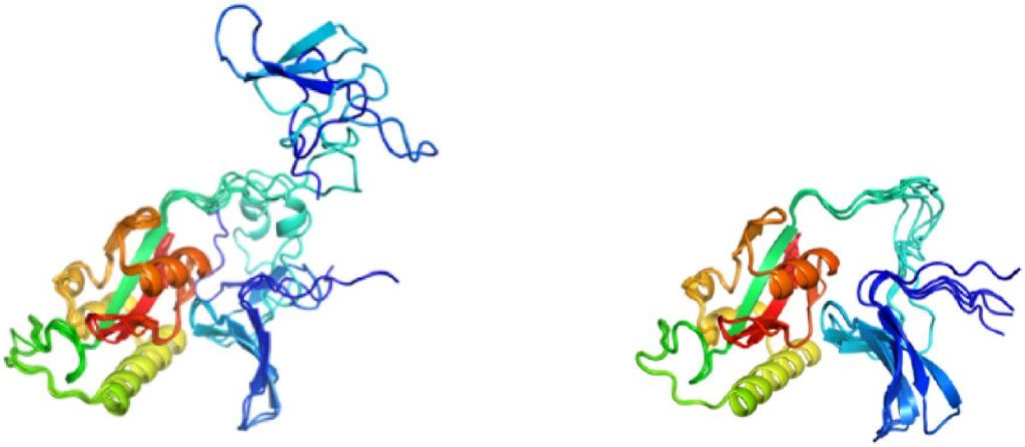
Supplementary Figure 11

A representative example (PDBID: 1we3F) showing different models generated by introducing different numbers of random mutations. Model 1 is generated by the original sequence without introducing any mutations, which obtains a high eTM-score = 0.99 and real TM-score = 0.99. When 1 random mutation is introduced, the model (model 2) is not impacted and also obtains a very high eTM-score = 0.99 and TM-score = 0.99. When 5 random mutations are introduced, the accuracy of one of the domain models (blue) is reduced with some local incorrect folding regions, resulting in a model (model 3) with reduced eTM-score = 0.90 and TM-score = 0.94. When 10 random mutations are introduced, the accuracy of the red domain is reduced and the inter-domain contacts/distances predicted by DeepPotential is impacted, which results in a model (model 4) with a low eTM-score= 0.53 and a low TM-score = 0.51.



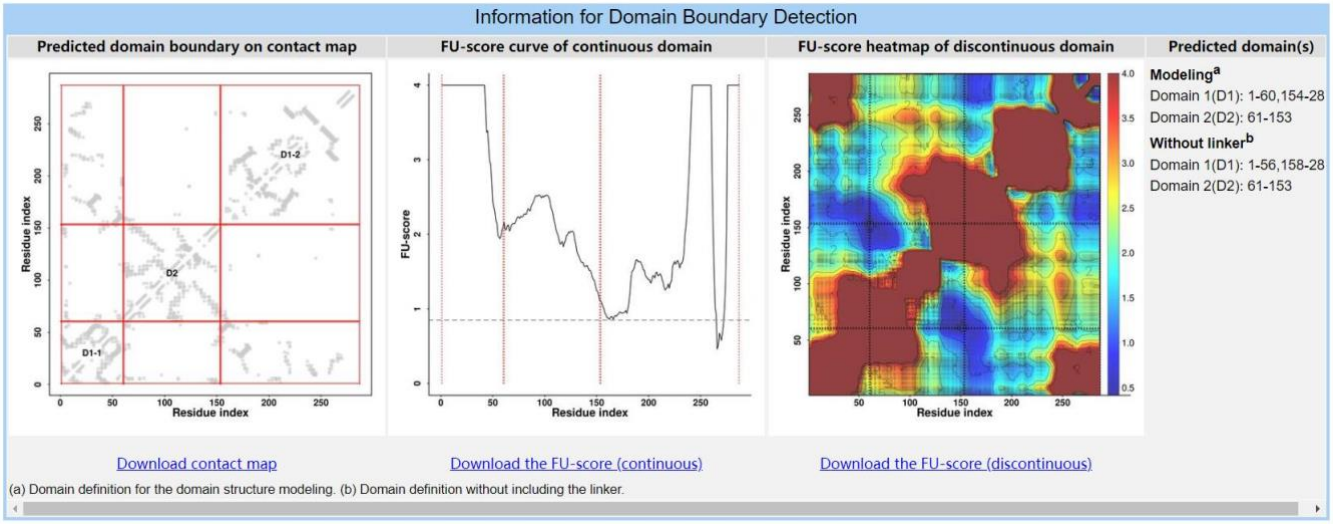
Supplementary Figure 12

Pipeline for multi-domain protein structure modeling using cryo-EM density maps in I-TASSER-MTD. Starting from the query sequence, the domain boundary is first predicted by FUpred and ThreaDom, and the model of each individual domain is generated by D-I-TASSER. Each domain model is then independently fit into the density map using L-BFGS guided by the density correlation between the domain model and the density map. Meanwhile, the analogous full-length structure templates are identified by TM-align. The initial full-length model is created according to models generated by domain-map fitting and the full-length templates. The domain rigid-body assembly is subsequently performed to optimize the position and orientation of each domain in the density map, and the flexible regions for remodeling are determined according to the density correlation score of each region. Next, the atom-, segment-, and domain-level refinement are performed using REMC simulations to simultaneously improve the full-length model and the individual domain models. Finally, the full-length model with the lowest energy is selected for side-chain repacking by FASPR and FG-MD.



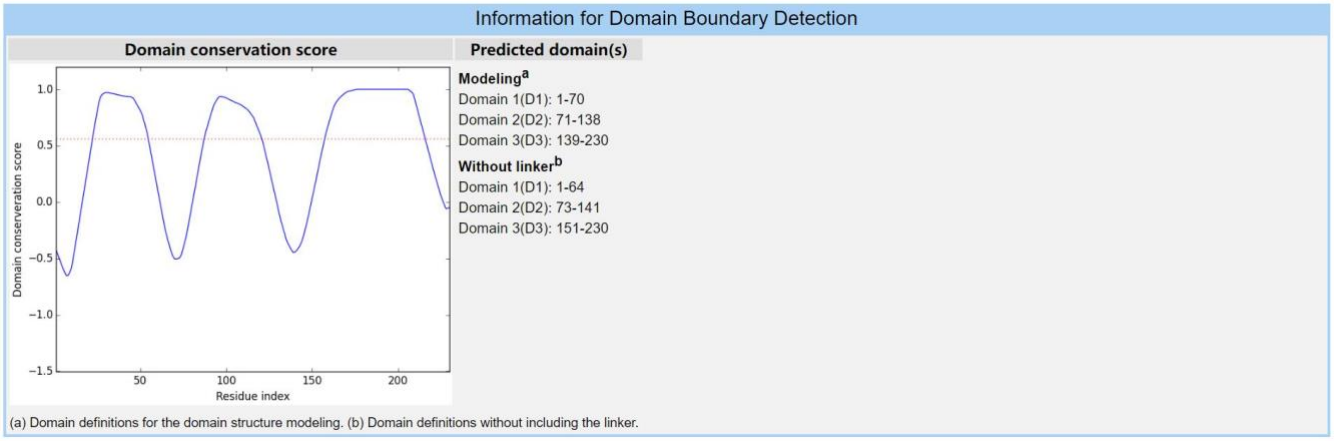
Supplementary Figure 13

The 5 models generated by I-TASSER-MTD and AlphaFold2 for the human protein Pin1. Models are represented in cartoons with blue to red running from N- to C-terminal. The 5 models generated by I-TASSER-MTD are highly diverged and contain both 'extend' and 'compact' states which mimic the alternative conformational distributions as observed by Born et al[6], while the 5 models constructed by AlphaFold2 all converge to a single 'compact' state.



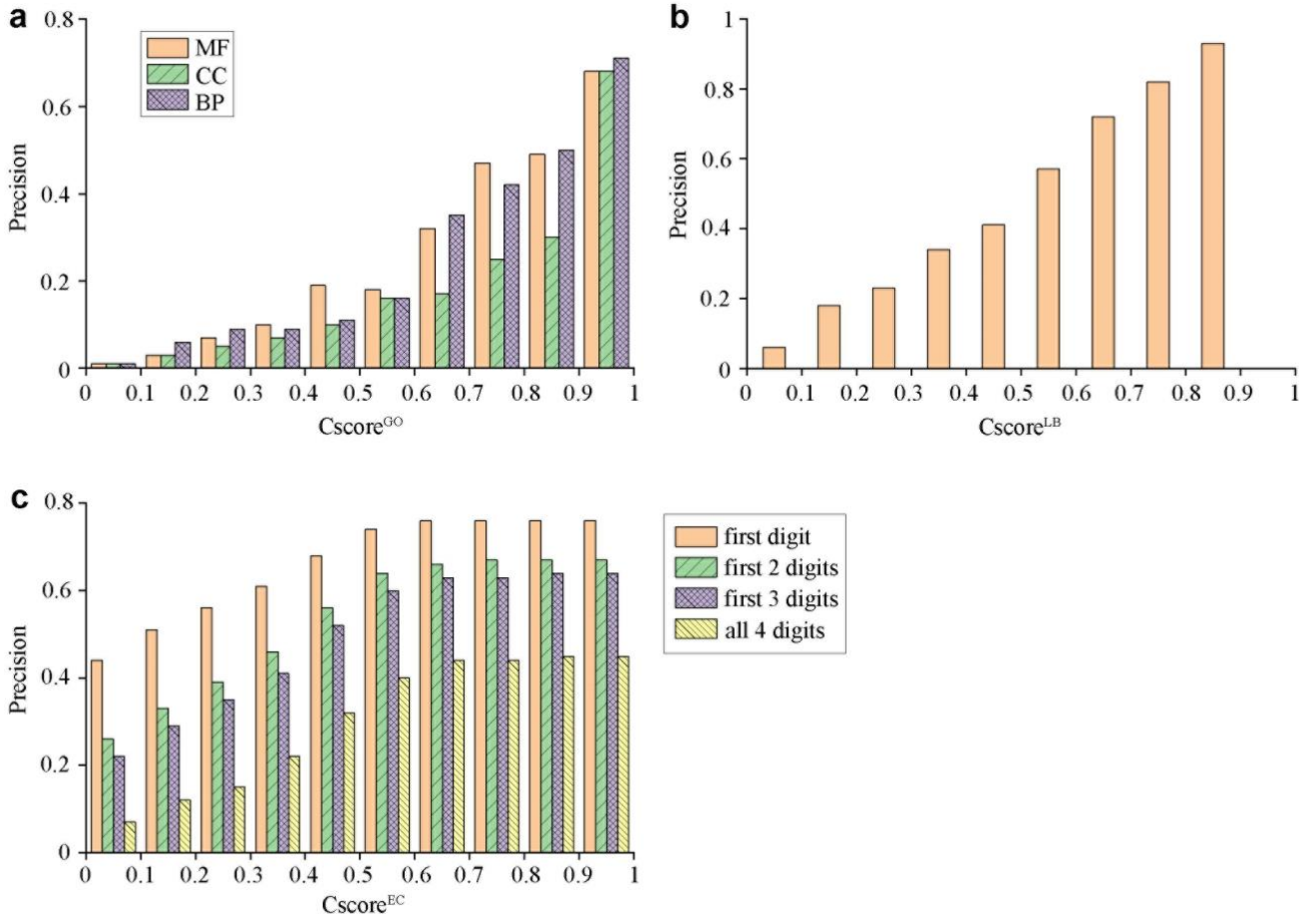
Supplementary Figure 14

Example of the domain boundary prediction (PDBID: 7m6b), which contains a discontinuous domain (1-60,154-287). Residues 1-60 belong to the first segment of the discontinuous domain (marked as D1-1 in the contact map). The second segment of the discontinuous domain contains the residues from 154 to 287 (marked as D1-2 in the contact map).



Supplementary Figure 15

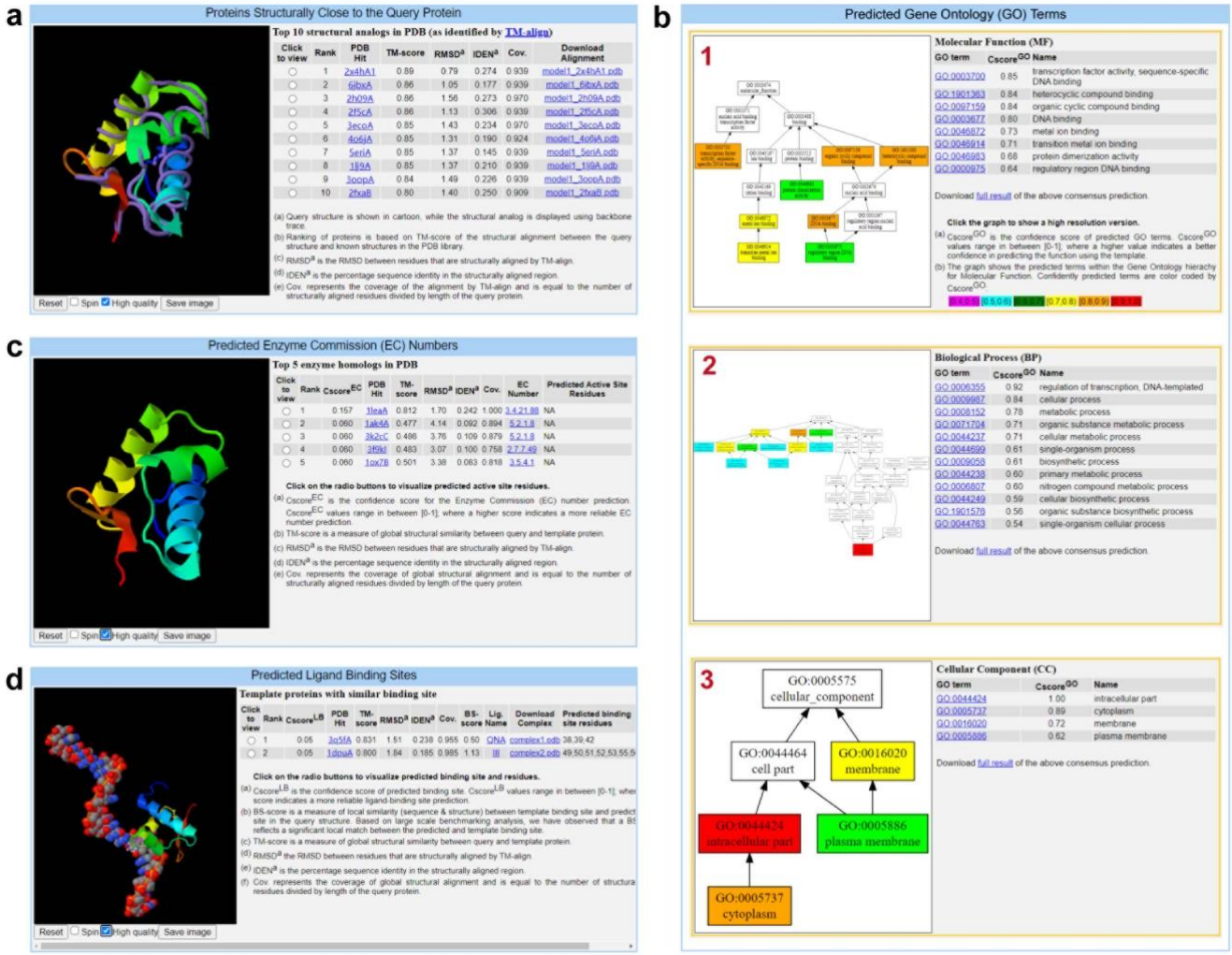
Example of the domain boundaries predicted by ThreaDom. The first column is the curve of the domain conservation score (DCS), where the blue line and red line indicate the DCS and the cutoff of the DCS. The second column shows the predicted domain boundaries.



Supplementary Figure 16

The precision of COFACTOR predictions versus the confidence score for each category of function annotation.

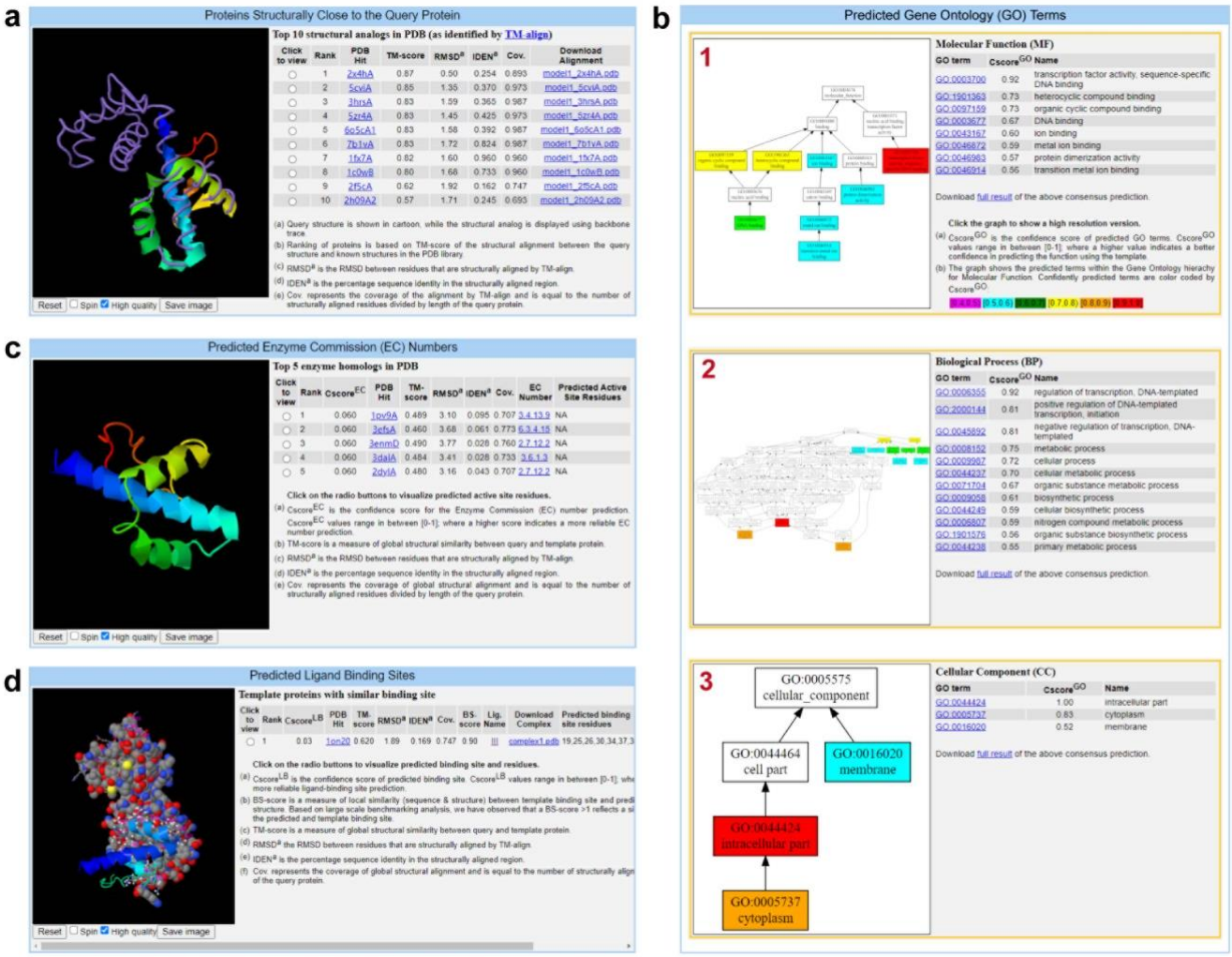
(**a**) The relationship between the precision and the confidence score of GO terms. (**b**) The relationship between the precision and the confidence score of ligand-binding (LB) sites. (**c**) The relationship between the precision and the confidence score of EC numbers, where different columns represent different numbers of digits of the EC number.



Supplementary Figure 17

Predicted function of the first domain (1-66) of the example protein (PDBID: 1fx7A)

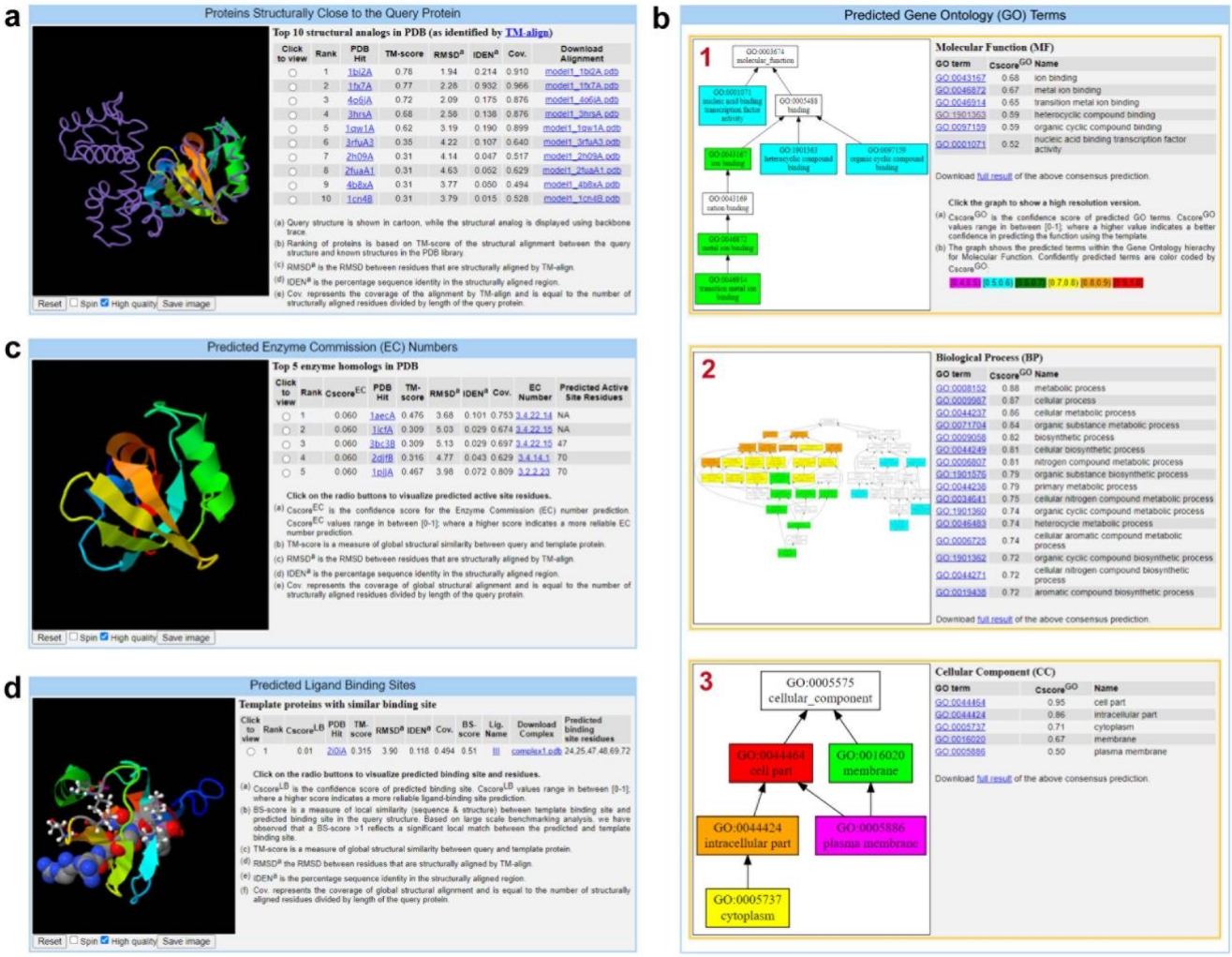
(**a**) Top ten analogous structures that are structurally close to the domain. (**b**) Results of the predicted GO terms including molecular function (1), biological process (2), and cellular component (3). (**c**) Results of the predicted EC numbers from the top-five homologous enzyme templates. (**d**) Results of the predicted ligand-binding site from the top 5 homologous templates.



Supplementary Figure 18

Predicted function of the second domain (67-141) of the example protein (PDBID: 1fx7A)

(**a**) Top ten analogous structures that are structurally close to the domain. (**b**) Results of the predicted GO terms including molecular function (1), biological process (2), and cellular component (3). (**c**) Results of the predicted EC numbers from the top-five homologous enzyme templates. (**d**) Results of the predicted ligand-binding site from the top 5 homologous templates.



Supplementary Figure 19

Predicted function of the third domain (142-230) of the example protein (PDBID: 1fx7A)

(**a**) Top ten analogous structures that are structurally close to the domain. (**b**) Results of the predicted GO terms including molecular function (1), biological process (2), and cellular component (3). (**c**) Results of the predicted EC numbers from the top-five homologous enzyme templates. (**d**) Results of the predicted ligand-binding site from the top 5 homologous templates.

Supplementary Tables

Supplementary Table 1. Description of the main individual programs employed in the I-TASSER-MTD pipeline.

Name	Description
LOMETS2[7]	A meta-threading program for protein template identification from the PDB by combining 11 state-of-the-art threading algorithms including deep-learning contact-based method (CEthreader[8]), sequence profile-based methods (SP3[9], PPAS[10], FFAS3D[11], MUSTER[12], Neff-MUSTER, SparksX[13], and PROSPECT2[14]), and profile HMM-based methods (HHpred[15], HHsearch[16], and PRC[17]).
ThreaDom[18,19]	A method for domain boundary prediction based on the domain information conserved in the threading templates identified by LOMETS2.
FUpred[20]	A protein domain boundary predictor by maximizing the number of intra-domain contacts and minimizing the number of inter-domain contacts that are predicted by a deep residual convolutional neural network model.
DeepMSA[21]	A method to generate high quality multiple sequence alignment (MSA) by iteratively searching the query through multiple whole-genome and meta-genome sequence databases.
DeepPotential[22]	A deep residual network-based predictor for predicting the protein intra-domain and inter-domain residue-residue spatial restraints including contacts, distances, torsion angles, and hydrogen-bonding networks.
DEMO[23]	An algorithm for assembling component domain models into full-length protein structures by coupling deep-learning inter-domain restraints with analogous templates identified by domain-level structural alignments using TM-align[24] from the PDB.
I-TASSER/D-I-TASSER[25]/I-TASSER-MTD	I-TASSER[26,27] is a method for the protein structure modeling through iterative threading fragment assembly simulations, where the threading templates are identified from the PDB by LOMETS[10]. D-I-TASSER is a significantly improved version of I-TASSER by incorporating the deep-learning spatial restraints to guide the simulations. I-TASSER-MTD is a fully automated pipeline built on D-I-TASSER and DEMO for multi-domain protein structure prediction and structure-based function annotation from sequence alone.
COFACTOR[28,29]	A program to infer three categories of protein functions including gene ontology, enzyme commission and ligand-binding sites from analogous and homologous function templates, sequence profile alignments, and protein–protein interaction networks.

Supplementary Table 2. Performance of D-I-TASSER on the CASP14 individual domain targets when only using contacts predicted by different methods.

	ResPre	DeepPLM	ResTriplet	TripletRes	NeBcon	Combined
TM-score	0.552	0.553	0.559	0.567	0.572	**0.586**
RMSD (Å)	9.1	9.0	8.9	8.8	8.7	**8.0**
#TM-score>0.5	54	55	55	57	57	**59**

Supplementary Notes

Supplementary Note 1. Combining multiple contact restraints for D-I-TASSER

Five different in-house contact predictors, ResPre[1], DeepPLM[2], ResTriplet[3], TripletRes[4], and NeBcon[5], are employed and combined together to guide the D-I-TASSER modeling simulation. Due to the different scoring schemes used by different contact predictors, we chose different confidence score cutoffs for different predictors that correspond to a contact accuracy of at least 0.5 for different ranges, including short-, medium-, and long-range contacts with sequence separation $|i-j| \leq 11$, $12 \leq |i-j| \leq 23$, and $|i-j| \geq 24$, respectively. For each individual contact predictor p, all of the residue-residue pairs are firstly ranked in descending order according to the confidence scores predicted by the predictor. A residue-residue pair (i,j) is selected into the contact pool if $con^p(i,j) > con^p_{cut}(r)$, where $con^p(i,j)$ is the confidence score of the residue-residue (i,j) predicted by predictor p, and $con^p_{cut}(r)$ is the confidence score cutoff for the predictor p for the range type $r \in$ {short-, medium-, and long-range}; or if $Lc^p < L^p_{cut}$, where Lc^p is the currently selected number of contacts by predictor p, and L^p_{cut} is the cutoff for the minimum number of selected contacts by predictor p. All the confidence cutoffs and parameter sets were determined over a set of 243 non-redundant training proteins. $L^p_{cut} = L$ for all predictors; $con^p_{cut}(short)$ = 0.647, 0.809, 0.607, 0.604, 0.483, and 0.512; $con^p_{cut}(medium)$ = 0.622, 0.789, 0.581, 0.598, 0.626, and 0.652; $con^p_{cut}(long)$ = 0.678, 0.806, 0.654, 0.652, 0.849, and 0.906 for TripletRes, ResTriplet, ResPRE, DeepPLM, NeBconB, and NeBconA, respectively.

The confidence score of the selected contacts from different predictors are renormalized. For the contact of each residue pair (i,j), for example, the new normalized confidence score is calculated according to the confidence score of different predictors:

$$C_{i,j} = \frac{1}{N}\sum_{p=1}^{N} w_p(i,j) \tag{S1}$$

where

$$w_p(i,j) = \begin{cases} 2.5 \times F \times \left[1 + con^p(i,j) - con^p_{cut}(r)\right], & \text{if predictor } p \text{ selects out } (i,j) \\ 0, & \text{otherwise} \end{cases} \tag{S2}$$

N is the number of contact predictors; F = 0.62, 1.25, 6.25, and 5.0 for Trivial, Easy, Hard, and Very hard target type determined by LOMETS2[7], respectively, when the number of effective sequences in the MSAs (Neff)[30] > 50; while F = 0.62, 1.5, 3.0, and 3.75 when Neff $\leq$50.

To verify the performance of the combined contacts, we test D-I-TASSER on all the 91 CASP14 individual domain targets by using contacts predicted by different predictors. Here, we only use the contact to guide the modeling simulation of D-I-TASSER to remove the impact of other restraints, and templates with a sequence identity >30% to the query are excluded. Supplementary Table 2 indicates that the average TM-score and RMSD of the model generated by D-I-TASSER with combined contacts are 0.586 and 8.0Å, which is better than that when using contacts predicted by any individual predictors as it obtains higher TM-score for the majority cases (Supplementary Fig. 6a). For example, D-I-TASSER with combined contacts achieves higher TM-score on 79.1% cases compared to D-I-TASSER using ResPre contacts. In addition, D-I-TASSER with combined contacts correctly predicted the global folds with TM-score >0.5 on 59 out of 91 cases, which is also higher than that by using contacts from the individual predictor. Supplementary Fig. 6b shows an example (T1092-D2) of models generated by using different contacts. The model constructed by employing combined contacts from different methods obtains a TM-score/RMSD of 0.93/1.5Å, which is better than the best model (TM-score =0.89, RMSD = 2.0Å) built by using contacts predicted by the individual predictor (ResTriplet).

Supplementary Note 2. Force Filed for domain assembly simulations

The force field for domain assembly is a sum of the 7 terms:

$$E = \sum_{m=1}^{N_{\mathrm{dom}}} \sum_{n=1}^{N_{\mathrm{dom}}} \left(w_1 E_{dt}(m,n) + w_2 E_{cl}(m,n) + w_3 E_{ct}(m,n) + w_4 E_{dp}(m,n) + w_5 E_{bd}(m,n) + w_6 E_{it}(m,n) \right) + w_7 E_{tr} \quad \text{(S3)}$$

where *m* and *n* are domain index, and N_{dom} is the total number of domains.

The first term is the *inter-domain* C_β *distance map* as predicted by DomainDist:

$$E_{\mathrm{dt}}(m,n) = -\sum_{i=1}^{L_m} \sum_{j=1}^{L_n} \log\left(P\left(i,j,k\left(d_{ij}\right)\right) + \varepsilon \right) \quad \text{(S4)}$$

where L_m and L_n represent the sequence length of the *m*-th and *n*-th domain, respectively. d_{ij} is the distance between the *i*-th C_β (C_α for Glycine) atom in the *m*-th domain and *j*-th C_β atom in the *n*-th domain, $P\left(i,j,k\left(d_{ij}\right)\right)$ is the predicted probability of the distance d_{ij} located in the *k*-th distance bin, and $\varepsilon = 1E-4$ is the pseudo count to offset low-probability bins. In the calculation, we only consider atom pairs with probability peak located in [2Å, 20Å], and these atom pairs with predicted probabilities >0.5 in the last bin [>20 Å], which represents a low prediction confidence in [2Å, 20Å], are excluded.

The second term is designed to eliminate *steric clashes* between domains, i.e.,

$$E_{cl}(m,n) = \sum_{i=1}^{L_m} \sum_{j=1}^{L_n} \begin{cases} \dfrac{1}{d_{ij}}, & if\ d_{ij} < d_{cut} \\ 0, & otherwise \end{cases} \quad \text{(S5)}$$

where $d_{\mathrm{cut}} = 3.75$ Å is set as the clash distance cutoff.

The third term is the *generic domain-domain contact energy* computed by:

$$E_{ct}(m,n) = \sum_{i=1}^{L_m} \sum_{j=1}^{L_n} \begin{cases} -u_{ij}, & \text{if } d_{ij} < 8\text{Å} \\ -\dfrac{1}{2}u_{ij}\left[1 - \sin\left(\dfrac{d_{ij}-9}{2}\pi\right)\right], & \text{if } 8\text{Å} \le d_{ij} \le 10\text{Å} \\ \dfrac{1}{2}u_{ij}\left[1 - \sin\left(\dfrac{d_{ij}-45}{70}\pi\right)\right], & \text{if } 10\text{Å} < d_{ij} \le 80\text{Å} \\ u_{ij}, & \text{otherwise} \end{cases} \quad \text{(S6)}$$

where the scale parameter u_{ij} depends on the hydrophobic and hydrophilic features of the residue pairs. $u_{ij} = 0.1$, if both of the residues are hydrophobic (ALA, CYS, VAL, ILE, PRO, MET, LEU, PHE, TYR, TRP); $u_{ij} = 0.01$, if the two residues are hydrophilic (SER, THR, ASP, ASN, LYS, GLU, GLN, ARG, HIS); or $u_{ij} = 0.05$, otherwise. This energy item is used to control the inter-domain distance, which will push the two domains together if they are two far away each other.

The fourth term is the *domain-domain distance profile* deduced from the templates identified by TM-align, which is calculated by:

$$E_{dp}(m,n) = -\sum_{i=1}^{L_m} \sum_{j=1}^{L_n} \frac{1}{T_{ij}} \sum_{t=1}^{T_{ij}} \frac{1}{\left|d_{ij} - D_{ij}^{t}\right|} \quad \text{(S7)}$$

For a residue pair (*i* and *j*, with *i* from N-terminal domain and *j* from C-terminal domain), T_{ij} is the number of templates that satisfy the following two conditions: (1) the template has both residue *i* and *j* aligned by TM-align; (2) $0.6|i-j| <$

$|a_i - a_j| < 1.5|i-j|$, where a_i and a_j are the indexes of the aligned residues of i and j on the template. D_{ij}^t is the C_α distance between the residue a_i and a_j in the t-th template.

The fifth term is the *domain boundary energy* is defined as

$$E_{bc}(m,n) = (b_{mn} - b_0)^2 \tag{S8}$$

where b_{mn} is the Cα distance between two consecutive domains, and $b_0 = 3.8$ Å is the standard length of Cα-Cα bond.

The sixth term is the *domain interface energy* is defined as

$$E_{it}(m,n) = \sum_{i=1}^{L_m}\sum_{j=1}^{L_n} \begin{cases} -U_{ij}, & \text{if } d_{ij} < 18\text{Å} \\ -\frac{1}{2}U_{ij}\left[1 - \sin\left(\frac{d_{ij}-19}{2}\pi\right)\right], & \text{if } 18\text{Å} \le d_{ij} \le 20\text{Å} \\ \frac{1}{2}U_{ij}\left[1 - \sin\left(\frac{d_{ij}-50}{60}\pi\right)\right], & \text{if } 20\text{Å} < d_{ij} \le 80\text{Å} \\ U_{ij}, & \text{otherwise} \end{cases} \tag{S9}$$

where U_{ij} is the confidence score of the i-th residue and j-th residue with the Cα distance <18 Å.

The last term in is the *local domain distance restraint*:

$$E_{tr} = \frac{1}{L}\sum_{i=1}^{L} d\,(S_i, S'_i) \tag{S10}$$

where $d(S_i, S'_i)$ represents the distance between the i-th C_α atom (S_i) and its corresponding atom S'_i in the initial structure generated in the template superposition process, and L is the length of the protein. This term is to prevent the assembly deviating too much from the orientation obtained from the template.

The weighting parameters in Eq. (S3) are determined by maximizing the correlation between total energy and RMSD to the native on the structure decoys over a training set of 425 non-redundant proteins. This resulted in $w_1 = 5$, $w_2 = 0.2$, $w_3 = 0.1$, $w_4 = 0.02$, $w_5 = 0.03$, $w_6 = 3.0$, and $w_7 = 0.15$.

Supplementary Note 3. Estimated TM-score of D-I-TASSER predicted models

The accuracy of the D-I-TASSER structure models is calculated by the estimated TM-score (eTM-score):

$$\text{eTM-score} = w_1 \ln\left(\frac{M(m)}{M_{\text{tot}}} \times \frac{1}{\langle \text{RMSD} \rangle_m}\right) + w_2 \ln\left(\frac{1}{K}\sum_{i=1}^{K}\frac{Z(t)}{Z_0(t)}\right) + w_3 w_{\text{neff}} \ln\left(\frac{O_m^{\text{pre}}}{N_{\text{pred}}}\right)$$

$$+ w_4 w_{\text{neff}} \ln\left(\frac{1}{n}\sum_{(i,j)}^{n}\left|d_{i,j}^{\text{pre}} - d_{i,j}^{m}\right|\right) + w_5 \tag{S11}$$

$$w_{\text{neff}} = \min(\max(0.66, 0.1(\log_2(Neff) - 3)), 1) \tag{S12}$$

$$Neff = \frac{1}{\sqrt{L}}\sum\nolimits_{n=1}^{N}\frac{1}{1 + \sum_{m=1, m\neq n}^{N} I\left[S_{m,n} \geq 0.8\right]} \tag{S13}$$

where M_{tot} is the total number of structure decoys used in the SPICKER clustering; M is the number of decoys in the top cluster; $\langle \text{RMSD} \rangle_m$ is the average RMSD of the decoys to the m-th cluster centroid. These terms are used to evaluate the degree of convergence of the structure assembly simulations. $Z(t)$ is the score of the top template identified by the t-th threading method in LOMETS2; $Z_0(t)$ is the Z-score cutoff of the t-th threading method to distinguish between good and bad templates; K is the total number of threading methods employed in LOMTES. These Z-score related measures describe the significance of the LOMETS2 threading templates and alignments. N_{pred} is the number of predicted contacts applied to guide the REMC simulations of D-I-TASSER, and O_m^{pre} is the number of overlapped contacts between the final predicted model and the predicted contacts. These two terms is used to measure the contact satisfaction rate. n is the number of distances of the residue pairs used to guide the D-I-TASSER simulation; $d_{i,j}^{\text{pre}}$ and $d_{i,j}^{m}$ are the distances of the residue pair (i,j) in the predicted distance map and the reported model, respectively. These terms are applied to assess how closely the distances in the reported model match the predicted ones by DeepPotential. w_{neff} is a weight calculated according to the number of effective sequences in the MSAs (*Neff*), where *Neff* is calculated by Eq. (S13). L is the length of a query protein, N is the number of sequences in the MSA, $S_{m,n}$ is the sequence identity between the m-th and n-th sequences, and I[ ] represents the Iverson bracket, which means $I\left[S_{m,n} \geq 0.8\right] = 1$ if $S_{m,n} \geq 0.8$, and 0 otherwise. $w_1 = 0.047$, $w_2 = 0.063$, $w_3 = 0.077$, $w_4 = -0.185$, and $w_5 = 0.740$ are free parameters.

Supplementary Note 4. Estimated RMSD of I-TASSER-MTD predicted models

The estimated RMSD (eRMSD) of the I-TASSER-MTD predicted models can be calculated by

$$\mathrm{eRMSD}(k) = w_1 \ln\left(\frac{M(k)}{M_{\mathrm{tot}}} \times \frac{1}{\langle \mathrm{RMSD} \rangle_k}\right) + w_2 \ln\left(\frac{1}{10}\sum_{i=1}^{10}\frac{T\text{-score}(i)}{T\text{-score}_0}\right) + w_3 w_{\mathrm{neff}} \ln\left(\frac{1}{T}\sum_{t=1}^{T}\left|d_t^{\mathrm{pre}} - d_t^{\mathrm{model}}(k)\right|\right)$$

$$+ w_4 w_{\mathrm{neff}} \ln\left(\frac{O(I^{\mathrm{pre}}, I^{\mathrm{model}})_k}{N(I^{\mathrm{pre}})}\right) + w_5 \frac{1}{N_{\mathrm{dom}}}\sum_{D=1}^{N_{\mathrm{dom}}} \mathrm{eTM\text{-}score}_{\mathrm{dom}}(D) + w_6$$

$$+ w_7 \ln(L) \qquad \text{(S14)}$$

The first term evaluates the degree of convergence of the domain assembly simulations, where M_{tot} is the total number of full-length decoys generated in the domain assembly simulations, $M(k)$ is the number of structure decoys with RMSD <1.5Å to the *k*th full-length model, and $\langle \mathrm{RMSD} \rangle_k$ denotes the average RMSD between these decoys and the *k*th reported model. The second term assesses the quality of the full-length template, where $T\text{-score}(i)$ is the template score of the *i*th full-length template, which is calculated as the harmonic mean of the TM-scores between the domain models and the full-length template that is used for DEMO-based domain assembly, and $T\text{-score}_0$=0.85 is the cutoff used to distinguish good from bad templates. The third term assesses how closely the distances in the reported model match the predicted distances by DeepPotential, where *T* is the number of predicted inter-domain distances used to guide the domain assembly, and d_t^{pre} and $d_t^{\mathrm{model}}(k)$ are the distances of the *t*th residue pair in the predicted distance map and the *k*th reported model, respectively. The fourth term accounts for the domain-domain interface satisfaction rate of the predicted interface map in the reported model, where $N(I^{\mathrm{pre}})$ is the number of predicted domain-domain interfaces and $O(I^{\mathrm{pre}}, I^{\mathrm{model}})_k$ is the number of overlapped interfaces between the predicted interface map and the *k*th reported model. Since restraints in the third and fourth terms are predicted using MSAs, w_{neff} is a weight associated with the quality of the MSA and calculated based on the number of effective sequences (neff, see Supplementary Eq. S12). Finally, the fifth term accounts for the quality of individual domain models from D-I-TASSER, where N_{dom} is the total number of domains and $\mathrm{eTM\text{-}score}_{\mathrm{dom}}(D)$ is the estimated TM-score of the *D*th domain model from D-I-TASSER (Supplementary Note 3). *L* is the sequence length. $w_1 = -1.40$, $w_2 = -2.74$, $w_3 = 4.78$, $w_4 = -1.19$, $w_5 = -16.43$, $w_6 = 0.0$, and $w_7 = 2.66$ are the weighting factors, which are optimized using an improved differential evolution algorithm[31] to minimize the error between the eRMSD and the actual RMSD of the decoys to the native structure on the DEMO training set of 425 non-redundant multi-domain proteins.

Supplementary Note 5. A 3-gradient contact potential

This energy term was developed to use the restraints from the predicted contacts or user provided contact/distance restraints. It is defined as the 3-gradient contact potential, which can be calculated by

$$E_{\mathrm{con}}(d_{ij}) = \begin{cases} -U_{ij}, & d_{ij} < d_{\mathrm{cut}} \\ -\frac{1}{2}U_{ij}\left[1-\sin\left(\frac{d_{ij}-\left(\frac{d_{\mathrm{cut}}+D}{2}\right)}{D-d_{\mathrm{cut}}}\pi\right)\right], & d_{\mathrm{cut}} \le d_{ij} < D \\ \frac{1}{2}U'_{ij}\left[1+\sin\left(\frac{d_{ij}-\left(\frac{D+80}{2}\right)}{(80-D)}\pi\right)\right], & D \le d_{ij} < 80\text{Å} \\ U'_{ij}, & d_{ij} \ge 80\text{Å} \end{cases} \tag{S15}$$

where d_{ij} is the C_α distance between the *i*-th and *j*-th residues of the model, d_{cut} is the distance cutoff of the contact or defined by the user, and D=d_{cut}+2.0 is a constant. U_{ij} and U'_{ij} are the weights of residue pair, which are defined as

$$U_{ij} = \ln\left(\frac{C_{ij}}{0.22}\right), U'_{ij} = \ln\left(\frac{C_{ij}}{0.7}\right) \tag{S16}$$

where C_{ij} is the confidence score of the residue pair (ij).

Supplementary Note 6. Confidence score of Gene Ontology, Enzyme Commission, and ligand-binding sites prediction

The Gene Ontology (GO) term, Enzyme Commission (EC) numbers, and ligand-binding sites of the full-length protein and the individual domains are predicted by our latest version of COFACTOR[28,29]. Here, we briefly describe the definition of the confidence score for each prediction.

GO term: The GO prediction consists of three pipelines for structure-, sequence- and PPI-based predictions (Supplementary Fig. 7). In the structure-based GO prediction, the query structure is compared to proteins in the BioLiP library[32], which contains a non-redundant set of entries annotated with known GO terms, through the local and global structural alignments based on TM-align[24]. The structure-homology based confidence score for a particular GO term λ is calculated by

$$C\text{score}^{\text{GO}}_{\text{structure}}(\lambda) = 1 - \prod_{i=1}^{N(\lambda)}\left(1 - FC\text{score}_i(\lambda)\right) \tag{S17}$$

where $N(\lambda)$ is the number of templates associated with the GO term λ, $FC\text{score}_i(\lambda)$ is the confidence score of the *i*th template associated with the GO term λ, which is defined as

$$\text{FCscore} = \frac{2}{1 + \exp\left(-(0.25 \times L_{\text{sim}} \times SS_{\text{bs}} + TM + 2.5 \times ID)\right)} - 1 \tag{S18}$$

where *ID* is the sequence identity between the query protein and template in the aligned region determined by TM-align, and SS_{bs} is the sequence identity at the binding site. *TM* is the TM-score between the structure of the query protein and the template. L_{sim} is the local structure similarity between the query protein and template, which can be calculated by

$$L_{\text{sim}} = \frac{1}{N_t}\sum_{i=1}^{N_{\text{ali}}}\left(\frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2} + M_i\right) \tag{S19}$$

where N_t is the number of residues in the active/binding sites of the template, N_{ali} is the number of aligned residue pairs, d_i is the C_α atom distance between the *i*th aligned residue pair, $d_0 = 3\text{Å}$ is the distance cutoff, and M_i is the BLOSUM62 substitution matrix score[33] between the *i*th residue pair.

In the sequence-based GO prediction, the query sequence is searched against the UniProt-GOA database through both sequence and sequence-profile alignments by BLAST[34] and PSI-BLAST[35], respectively. The sequence-based confidence score for a particular GO term λ is defined as

$$C\text{score}^{\text{GO}}_{\text{sequence}}(\lambda) = w \times \text{GOfreq}_{\text{blast}}(\lambda) + (1 - w) \times \text{GOfreq}_{\text{psiblast}}(\lambda) \tag{S20}$$

where w is the weight and equals to the maximum sequence identity between the query and all the templates. $\text{GOfreq}_{\text{blast}}(\lambda)$ is the confidence score for the GO term λ resulting from the BLAST search, which can be calculated by

$$\text{GOfreq}_{\text{blast}}(\lambda) = \frac{\sum_{k=1}^{N(\lambda)} s_k(\lambda)}{\sum_{k=1}^{N} s_k} \tag{S21}$$

where N is the number of identified templates, s_k is the sequence identity between the query and the *k*th template, $N(\lambda)$ and $s_k(\lambda)$ are those associated with the GO term λ. $\text{GOfreq}_{\text{psiblast}}(\lambda)$ is the confidence score for the GO term λ calculated based on the template identified by PSI-BLAST, and it is defined in the same way as in BLAST.

In the protein-protein interaction (PPI) based GO prediction, the query is first mapped to the STRING PPI database[36] by BLAST; only the BLAST hit with the most significant E-values is subsequently considered. GO terms of the interaction partners, as annotated in the STRING database, are then collected and assigned to the query protein. The confidence score for GO term λ mapped by PPI is defined as

$$C\text{score}_{\text{PPI}}^{\text{GO}}(\lambda) = S_q \times \frac{\sum_{k=1}^{N(\lambda)} str_k(\lambda)}{\sum_{k=1}^{N} str_k} \tag{S22}$$

where N is the number of interacting partners, str_k is the confidence score of interaction between the query and the kth interaction partner as assigned by the STRING database, S_q is the sequence identity in the first step of BLAST alignment between the query sequence and the mapped STRING entry, and $N(\lambda)$ and $str_k(\lambda)$ are those associated to a specific GO term λ.

The final confidence sore of the GO prediction is obtained by combining the structure-, sequence- and PPI-based confidence sore:

$$C\text{score}^{\text{GO}}(\lambda) = 1 - \prod_{m}(1 - C\text{score}_{m}^{\text{GO}}(\lambda)^{w_m}) \tag{S23}$$

where $m \in \{\text{structure}, \text{sequece}, \text{PPI}\}$, w_m is the weight for each of the three methods, with $w_{\text{structure}} = 1 - w$, $w_{\text{sequence}} = w_{\text{PPI}} = 1$, where w equals to the maximum sequence identity among identified function templates.

EC number: The EC number prediction is similar to the structure-based GO prediction. Enzymatic homologs are identified by aligning the target structure to the enzyme structures in the BioLiP library according to TM-align, where the active site residues mapped from the Catalytic Site Atlas database[37]. The confidence score for each predicted EC number is estimated based on the global and local similarity between the target and the top template hits:

$$C\text{score}^{\text{EC}} = \frac{2}{1 + \exp\left(-(0.25 \times L_{\text{sim}} \times SS_{\text{bs}} + TM + 2.5 \times ID)\right)} - 1 \tag{S24}$$

where TM is the TM-score between the structure of the query and the template, ID is the corresponding sequence identity, SS_{bs} is the sequence identity at the active sites, and L_{sim} is the local structure similarity as defined in Eq. (S19).

Ligand-binding site: The ligand-binding prediction includes three steps[28,29]. First, functional homologies are identified by matching the query structure through the BioLiP library, which contains a non-redundant set of structure templates harboring in the known ligand-binding sites for interaction between receptor proteins and small molecule compounds, short peptides and nucleic acids. The initial binding sites are then mapped to the query from the individual templates based on the structural alignments. Next, the ligands from each individual template are superposed to the predicted binding sites on the query structure using superposition matrices from a local alignment of the query and template binding sites, and the ligand poses are refined by a short Metropolis Monte Carlo simulation through rigid-body rotation and translation. Finally, the consensus binding sites are obtained by clustering of all ligands that are superposed to the query structure, based on distances of the centers of mass of the ligands using a cutoff of 8Å. Different ligands within the same binding pocket are further grouped by the average linkage clustering with chemical similarity, using the Tanimoto coefficient[38] with a cutoff of 0.7. The model with the highest ligand-binding confidence score $C\text{score}^{\text{LB}}$ defined as follow among all the clusters is selected:

$$C\text{score}^{\text{LB}} = \frac{2}{1 + \exp\left(-\frac{M}{M_{\text{tot}}}(0.25 \times L_{\text{sim}} + TM + 0.25 \times ID + \frac{2}{1+D})\right)} - 1 \tag{S25}$$

where M is the number of ligands in the ligand cluster, M_{tot} is the total number of ligands collected from all homologous templates, L_{sim} is the local similarity at the binding site defined in Eq. (S19), TM is TM-score between query and template, ID is the sequence identity between query and template in the structurally aligned region and D is the average distance between ligands within the cluster.

Supplementary References

1. 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* **35**, 4647-4655 (2019).
2. Zheng, W. *et al.* Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, Bioinformatics* **87**, 1149-1164 (2019).
3. Li, Y., Zhang, C., Bell, E.W., Yu, D.J. & Zhang, Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins: Structure, Function, Bioinformatics* **87**, 1082-1091 (2019).
4. 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).
5. He, B., Mortuza, S., 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).
6. Born, A. *et al.* Reconstruction of Coupled Intra-and Interdomain Protein Motion from Nuclear and Electron Magnetic Resonance. *Journal of the American Chemical Society* **143**, 16055-16067 (2021).
7. Zheng, W. *et al.* LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic acids research* **47**, W429-W436 (2019).
8. Zheng, W. *et al.* Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLoS computational biology* **15**, e1007411 (2019).
9. Zhou, H. & Zhou, Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Structure, Function, Bioinformatics* **58**, 321-328 (2005).
10. Wu, S. & Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* **35**, 3375-3382 (2007).
11. Xu, D., Jaroszewski, L., Li, Z. & Godzik, A. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* **30**, 660-667 (2014).
12. Wu, S. & Zhang, Y. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, Bioinformatics* **72**, 547-556 (2008).
13. Yang, Y., Faraggi, E., Zhao, H. & Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **27**, 2076-2082 (2011).
14. Xu, Y. & Xu, D. Protein threading using PROSPECT: design and evaluation. *Proteins: Structure, Function, Bioinformatics* **40**, 343-354 (2000).
15. Meier, A. & Söding, J. Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. *PLoS computational biology* **11**, e1004343 (2015).
16. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951-960 (2005).
17. Madera, M. Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics* **24**, 2630-2631 (2008).
18. Xue, Z., Xu, D., Wang, Y. & Zhang, Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **29**, i247-i256 (2013).
19. Wang, Y. *et al.* ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. *Nucleic acids research* **45**, W400-W407 (2017).
20. Zheng, W. *et al.* FUpred: Detecting protein domains through deep-learning based contact map prediction. *Bioinformatics* **36**, 3749-3757 (2020).
21. Zhang, C., Zheng, W., Mortuza, S., 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 (2020).

22. Li, Y. *et al.* Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins: Structure, Function, Bioinformatics* **89**, 1911-1921 (2021).
23. 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* **116**, 15930-15938 (2019).
24. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **33**, 2302-2309 (2005).
25. Zheng, W. *et al.* Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins: Structure, Function, Bioinformatics* **89**, 1734-1751 (2021).
26. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **9**, 40 (2008).
27. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nature methods* **12**, 7-8 (2015).
28. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research* **40**, W471-W477 (2012).
29. Zhang, C., Freddolino, P.L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic acids research* **45**, W291-W299 (2017).
30. Zheng, W. *et al.* Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Reports Methods* **1**, 100014 (2021).
31. Zhou, X., Peng, C., Liu, J., Zhang, Y. & Zhang, G. Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. *IEEE Transactions on Evolutionary Computation* **24**, 536-550 (2020).
32. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research* **41**, D1096-D1103 (2012).
33. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915-10919 (1992).
34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).
35. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
36. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research* **47**, D607-D613 (2019).
37. Furnham, N. *et al.* The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic acids research* **42**, D485-D489 (2014).
38. Rogers, D.J. & Tanimoto, T.T.J.S. A computer program for classifying plants. **132**, 1115-1118 (1960).