



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

Xiaogen Zhou^{1,2}, Wei Zheng¹, Yang Li¹, Robin Pearce¹, Chengxin Zhang¹, Eric W. Bell¹, Guijun Zhang² and Yang Zhang^{1,3}✉

Most proteins in cells are composed of multiple folding units (or domains) to perform complex functions in a cooperative manner. Relative to the rapid progress in single-domain structure prediction, there are few effective tools available for multi-domain protein structure assembly, mainly due to the complexity of modeling multi-domain proteins, which involves higher degrees of freedom in domain-orientation space and various levels of continuous and discontinuous domain assembly and linker refinement. To meet the challenge and the high demand of the community, we developed I-TASSER-MTD to model the structures and functions of multi-domain proteins through a progressive protocol that combines sequence-based domain parsing, single-domain structure folding, inter-domain structure assembly and structure-based function annotation in a fully automated pipeline. Advanced deep-learning models have been incorporated into each of the steps to enhance both the domain modeling and inter-domain assembly accuracy. The protocol allows for the incorporation of experimental cross-linking data and cryo-electron microscopy density maps to guide the multi-domain structure assembly simulations. I-TASSER-MTD is built on I-TASSER but substantially extends its ability and accuracy in modeling large multi-domain protein structures and provides meaningful functional insights for the targets at both the domain- and full-chain levels from the amino acid sequence alone.

Introduction

Much progress has been made in protein structure prediction as a result of decades of effort^{1–4}. The progress has been particularly notable in recent years owing to the introduction of coevolution-based contact prediction^{5–7} and deep neural-network learning techniques^{8–10}. In particular, the end-to-end sequence-to-structure training approaches, such as AlphaFold2 (ref. 11), built on the attention and equivariant transformer networks, have achieved unprecedented modeling accuracy in the protein structure prediction as witnessed in the recent CASP14 experiment¹². However, most of the advanced methods have mainly focused on the modeling of individual domain structures, which are the minimum folding units of proteins that fold and function independently. In fact, more than two-thirds of prokaryotic proteins and four-fifths of eukaryotic proteins contain two or more domains¹³, where many proteins perform higher-level cellular functions through cooperative domain interactions^{14,15}. Therefore, determining the full-length structures of multi-domain proteins is a crucial step towards elucidating their full functions and designing new drugs to regulate these functions.

A common approach for multi-domain protein structure modeling is to split the query sequence into domains and generate models for each individual domain separately^{16,17}. The individual domain models are subsequently assembled into full-length models, usually under the guidance of other homologous multi-domain proteins from the Protein Data Bank (PDB)¹⁸. However, many multi-domain proteins have been solved only as single-domain proteins, and just 35.3% of proteins in the PDB contain multi-domain structures. The lack of homologous multi-domain structures makes the template-based domain-assembly approach infeasible for most multi-domain protein targets. On the other hand, template-free (or *ab initio*) domain structure assembly is challenging, owing to the fact that multi-domain proteins have a high degree of freedom in domain-orientation space and we do not have reliable force fields to accommodate the domain–domain interactions. In a recent study,

¹Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA. ²College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. ³Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA.

✉e-mail: zhng@umich.edu

Zhou et al. proposed a hybrid approach, called DEMO¹⁶, which first detects analogous structure templates through domain-complex structure alignment¹⁹. Next, a knowledge-based force field is combined with deep-learning contact/distance restraints to refine the multi-domain analogous templates through replica-exchange Monte Carlo (REMC) simulations. As structure alignment can often detect more distant templates beyond sequence-based alignments, DEMO remarkably enhances the coverage and ability of the homology-based approaches. Large-scale benchmark tests demonstrated advantages of DEMO over other state-of-the-art approaches (either homology or *ab initio*) on proteins containing both continuous and discontinuous domain structures¹⁶.

Although several methods have been proposed for domain boundary prediction^{20–22} and domain model assembly^{16,17,23}, there have been very few protocols dedicated to automated multi-domain protein structure and function prediction from sequence alone. One reason is that the complexity of modeling multi-domain proteins, which involves various levels of continuous and discontinuous domain assembly and linker refinement, makes the pipeline development difficult to automate. In addition, although domains are usually basic units with distinct biological functions, the function of multi-domain proteins is often predicted using the same strategies used for single-domain proteins; this is especially true for the modeling of protein-level functions such as Gene Ontology (GO) terms, while the commonly used GO predictors, such as MetaGO²⁴, NetGO²⁵ and INGA²⁶, do not attempt to split the sequence into domains and annotate functions at individual domain level. Some studies²⁷ have attempted to assign GO terms to specific regions of the protein, but they are either not specifically optimized for multi-domain structure models²⁸ or do not use protein structure at all²⁷. These methods are therefore less useful for the function annotation of multi-domain proteins. Furthermore, several important questions should be considered to improve the modeling accuracy when developing such unified multi-domain protein structure and function prediction pipelines. First, coevolutionary analysis and deep learning have been successfully applied to single-domain protein structure prediction and have notably improved its accuracy²⁹. How can we extend similar techniques to guide domain parsing and domain model assembly? Second, is there any intrinsic correlation between the similarity of single-domain proteins and the global similarity of multi-domain proteins? If yes, how can such correlation be exploited through analogous multi-domain protein template detection informed by the structural alignment of individual domains? Third, how can we incorporate the sparse restraints, such as cryo-electron microscopy (EM) density maps and cross-linking data, into the simulation to guide the full-length structure modeling? Finally, given that the distinct functions of individual domains usually contribute towards the overall protein function¹⁵, how can we draw a more comprehensive function annotation by combining domain-level and full-length function prediction?

To address the above questions and meet the high demand of the community, we developed I-TASSER-MTD (with MTD standing for ‘multi-domain’), a fully automated pipeline for multi-domain protein structure assembly and structure-based function annotation from sequence alone (the I-TASSER-MTD platform is freely available at <https://zhanggroup.org/I-TASSER-MTD/>)^{16,30,31}. The core of I-TASSER-MTD is built on I-TASSER⁴ and DEMO¹⁶, where the former has consistently been ranked as the top method for automatic protein structure prediction in the last eight iterations of the community-wide CASP experiments^{31–39}, while the online server for I-TASSER has completed structure and function predictions for >680,000 proteins submitted by >160,000 registered users from 159 countries⁴⁰. With the integration of cutting-edge inter-domain assembly methods, as well as advanced deep neural-network techniques⁴¹, I-TASSER-MTD dramatically extends the ability and capacity of I-TASSER for modeling large multi-domain protein structures. Meanwhile, unlike pure deep-learning methods, which largely act as a block box, the accessibility and interpretability of the domain structure assembly process allows for the I-TASSER-MTD server to effectively incorporate the restraints from cryo-EM and cross-linking experiments as input by users. As a validation of the pipeline, I-TASSER-MTD participated in the most recent CASP14 (under the group name ‘Zhang-Server’) and achieved the best performance on modeling the multi-domain proteins among all server groups in the experiment³⁰. To the best of our knowledge, this represents the first protocol devoted specifically to the modeling of multi-domain proteins, which we believe will benefit the biological and biomedical communities.

The I-TASSER-MTD pipeline

I-TASSER-MTD is built on multiple tools that we recently developed for sequence-based domain boundary prediction, deep-learning spatial restraint prediction, single-domain structure folding, multi-domain structure assembly and structure-based function annotation. The pipeline is depicted

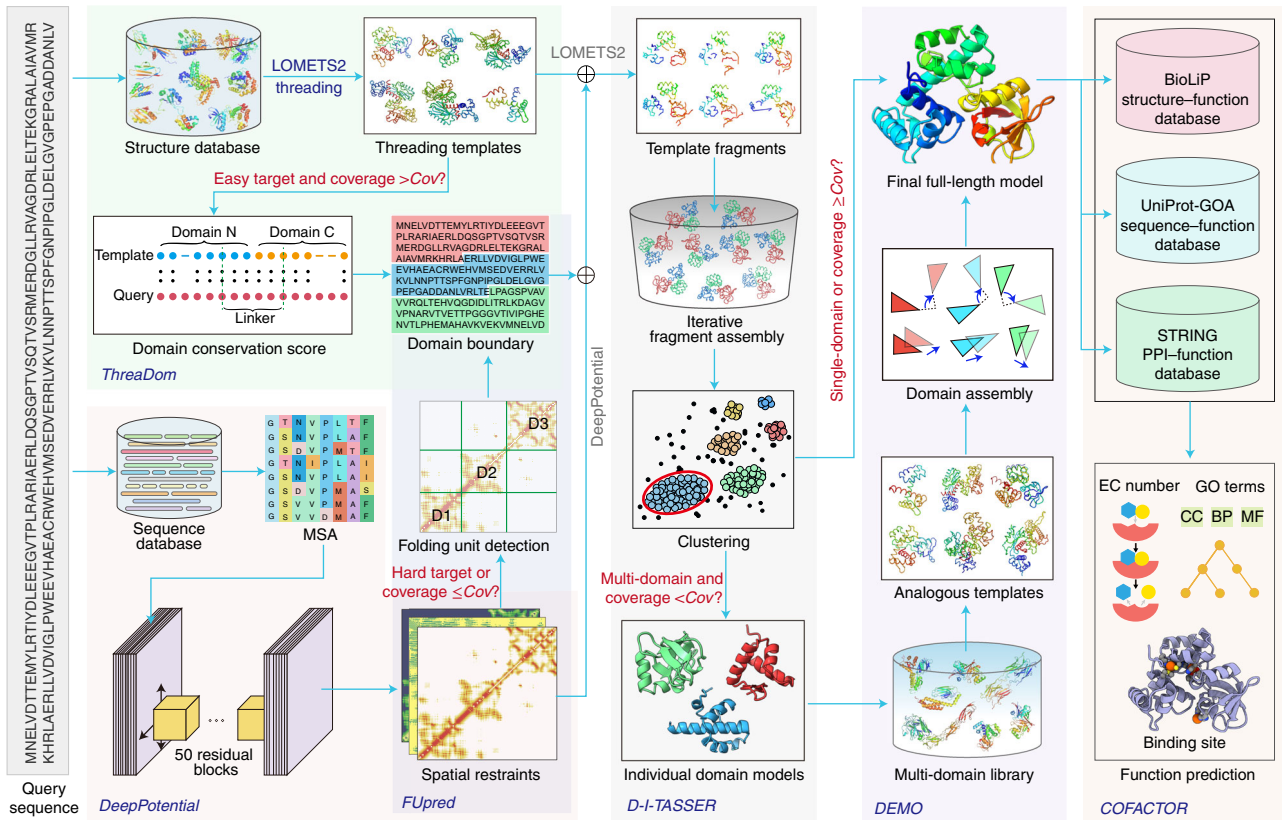


Fig. 1 | Overview of the I-TASSER-MTD protocol for multi-domain protein structure and function prediction. Cov is the cutoff of the alignment coverage for assessing if the query needs to be modeled as a single-unit or multi-unit target, where a unit can contain a single domain or multiple domains if the latter is fully covered by the LOMETS2 threading alignments. CC, BP and MF represent, respectively, the cellular component, biological process and molecular function in GO. It should be noted that, following their implementation on full-length sequence, LOMETS2 and DeepPotential will be run again in D-I-TASSER to generate templates/restraints for each individual domain if the query is deemed as a multi-unit target.

in Fig. 1. Starting from the query sequence, multiple threading templates are first collected by LOMETS2 (ref. 42) a meta-server approach that combines up to 11 sequence/hidden Markov model profile-based methods and deep-learning threading programs—from the PDB, and domain boundaries are predicted by FUpred²² and ThreaDom^{20,43}. Meanwhile, residue–residue spatial restraints are created by DeepPotential⁴⁴ through residual convolutional network training. If the query sequence is deemed a multi-domain protein by FUpred or ThreaDom, and none of the top ten template alignments can cover all domains (i.e., one or more domains with an alignment coverage below the cutoff $Cov = 95\%$), the ‘multi-domain assembly’ mode will be initiated, where models for each domain will be independently constructed by D-I-TASSER³⁰, a crucially improved version of I-TASSER powered by the DeepPotential spatial restraints. Subsequently, the domain models will be assembled into the full-length model by DEMO¹⁶ based on the structurally analogous templates. Otherwise, if the query is deemed a single-domain protein or one or more top template alignments can cover all domains, the ‘multi-domain assembly’ mode will be turned off and the full-length structure will be directly modeled by D-I-TASSER. Finally, the protein function annotations, including the Enzyme Commission (EC) numbers, GO terms and ligand-binding sites are predicted by COFACTOR⁴⁵ for all individual domains and the full-chain protein, based on the modeled structures, sequences and protein–protein interactions (PPIs). The main individual programs employed in the I-TASSER-MTD pipeline are listed in Supplementary Table 1. Next, we briefly describe the main components of the I-TASSER-MTD pipeline.

Sequence-based domain boundary prediction

Accurate domain boundary assignment is crucial for multi-domain protein structure and function prediction. In the first stage of I-TASSER-MTD, the query sequence is split into domains by combining a threading template-based method, ThreaDom^{20,43}, with a deep-learning contact-based

program, FUpred²² (Supplementary Fig. 1). For doing this, the query sequence is first threaded through the PDB by LOMETS2 (ref. ⁴²), which utilizes and combines 11 state-of-the-art threading programs, to create multiple template alignments, where the query is assigned as an Easy (or Hard) target if outstanding templates are (or are not) identified. If the protein is defined as an Easy target by LOMETS2 and the alignment coverage is >95%, ThreaDom will be applied to predict the domain boundary. In ThreaDom, the domain conservation score (DCS), which linearly combines the template domain linker score and the gap penalty score, is calculated according to the LOMETS2 template alignments. The domain linker score is evaluated on the basis of the domain boundary definition of the templates in CATH⁴⁶ or defined by DomainParser⁴⁷, while the gap penalty score is measured by the number of gaps in the multiple template alignments. Finally, the domain boundaries are determined by the DCS using a target-specific scoring cutoff.

If the protein is defined as a Hard target by LOMETS2 or the alignment coverage is ≤95%, the domain boundary will be predicted by FUpred 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. To create the deep-learning model, the multiple sequence alignment (MSA) is first created by iteratively searching the query against four metagenome sequence databases (Metaclust⁴⁸, BFD⁴⁹, Mgnify⁵⁰ and IMG/M⁵¹) and two whole-genome databases (Uniclust30⁵² and Uniref90⁵³) using DeepMSA⁵⁴. Then the secondary structure is predicted by PSSpred⁵⁵, and the contact map is generated by the deep-learning-based predictor ResPRE¹⁰, which has been included in DeepPotential, using features directly extracted from the MSA. To deduce the domain boundaries from the contact map, a folding unit score (FU-score)²² is calculated for continuous domain and discontinuous domain detection (for continuous and discontinuous domain definition, see Supplementary Fig. 2). For a continuous two-domain protein, the FU-score of each residue is defined in relation to the number of intra-domain contacts and the number of inter-domain contacts for the two domains by considering the residue as a domain boundary. To calculate the FU-score of a residue pair (i, j) in a discontinuous two-domain protein, the C-terminal contact map ($j + 1, L$) is shifted to the N-terminal contact map ($1, i$) to convert the discontinuous domain into a continuous one to form a new contact map, where $i < j$, and L is the sequence length of the query protein (Supplementary Fig. 3). The FU-score for the residue pair (i, j) is calculated according to the number of intra-domain contacts and inter-domain contacts in the new map. The domain boundary is determined according to a trained cutoff of the FU-score, and the predicted secondary structure is also used to avoid the boundary being located within a helix or a strand. As the FU-score is defined for the two-domain case, the above procedure is recursively performed for the determined domains until no additional domains can be detected.

Deep neural network-based spatial restraints prediction

DeepPotential⁴⁴, a newly developed deep residual neural-network-based predictor, is used to create multiple spatial restraints, including distance maps for both C_α and C_β atoms, C_α -based hydrogen-bonding networks⁴ and C_α - C_β torsion angles. Different from the contact map prediction, DeepPotential, whose architecture is depicted in Supplementary Fig. 4, predicts the probability that inter-residue distances fall within 36 equal-width bins from [2, 20] Å, as well as two additional bins with distances <2 Å and >20 Å. The DeepPotential program was trained on a nonredundant dataset of 26,151 solved protein structures obtained from the PDB. Starting from the query sequence, three 1D features (i.e., field parameters of the Potts model, a one-hot representation of the sequence, and a hidden Markov model) and two 2D coevolutionary features (i.e., the pseudolikelihood maximized Potts model⁵⁶ and mutual information) are extracted from the MSA created by DeepMSA. The 1D features are tiled in two dimensions and concatenated with the 2D features before being fed into the neural network, which includes ten 1D residual blocks⁴¹ and ten 2D residual blocks. The neural network model was trained by the Adam optimization algorithm to marginally minimize cross-entropy loss. In addition to the 20 Å C_α/C_β distance cutoff, distance maps with multiple thresholds (i.e., 10, 13, 16 and 19 Å), inter-residue torsion angles⁵⁷ and hydrogen-bonding networks are also considered in a multi-task learning strategy⁵⁸. The contact maps and domain–domain interface maps are extracted from the predicted distances by the summation of the cumulative probability of distances <8 Å and distances <18 Å, respectively.

Individual domain structure modeling with D-I-TASSER

The structural model for each domain is built using D-I-TASSER, which is an extended version of I-TASSER^{4,40,59–61} that integrates the DeepPotential spatial restraints with iterative threading

assembly simulations (Supplementary Fig. 5). For each domain, LOMETS2 is used again to identify domain-level structural templates from a nonredundant PDB structural library. Meanwhile, distance maps, hydrogen-bonding networks⁴ and inter-residue torsion angles are predicted by DeepPotential. The contact maps are also predicted by four deep-learning-based methods (ResPre¹⁰, DeepPLM³¹, ResTriplet⁶² and TripletRes⁶³) and a naïve Bayes-based contact predictor (NeBcon⁶⁴) (see the description and benchmark results reported in Supplementary Note 1). Then, the domain-level structure models are constructed using REMC simulations under the guidance of the deep-learning-based spatial restraints and the I-TASSER potential⁴, the latter of which contains generic statistical potentials and threading template-based restraints. Five REMC simulations are performed in parallel for each protein, where the structural decoys from 8 (or 3 for hard targets) low-temperature replicas are clustered by SPICKER⁶⁵. Finally, the decoy of the center of the largest cluster is selected as the final model, and the side-chains of the final model are then repacked by FASPR⁶⁶, which is further refined by FG-MD⁶⁷. It should be noted that some programs (e.g., LOMETS2 and DeepPotential) that have been implemented at full-chain level are performed again in this step to create D-I-TASSER models for each domain. As the programs and the D-I-TASSER force field have been trained mainly on the domain level and many template structures solved in the PDB contain only single domains, the domain-level modeling often generates more reliable results than that starting from the full-chain sequences; this is also one of the major motivations for the development of the I-TASSER-MTD pipeline.

Domain assembly through analogous global structural alignments

After all individual domain models are created, they are assembled into full-length models using DEMO¹⁶ under the guidance of the analogous templates detected by domain-complex structural alignments¹⁹ (Supplementary Fig. 7). For this, a nonredundant multi-domain protein library, which is updated weekly, has been constructed by collecting all multi-domain protein structures from the PDB. The individual domain models are first aligned to each template of the library by TM-align¹⁹, and the harmonic mean of the TM-score⁶⁸ of all domains is defined as the score of the template (*T*-score). Then, five initial full-length models are respectively constructed on the basis of the top five templates with the highest score by searching the best position of each domain on the template through a sliding-window-based procedure. For each initial model, REMC simulations with all domains randomly rotated and translated as rigid bodies are performed to optimize the orientation of each domain under the guidance of an energy function (Supplementary Note 2) that includes C_{α} clashes between domains, domain boundary connectivity, inter-domain distances and domain–domain interfaces predicted by DeepPotential, a generic inter-domain contact potential, local domain distance restraints from the initial domain-template superpositions and inter-domain distance restraints deduced from the top templates. Owing to errors in the tail of domain models, linkers between domains may be disconnected after the assembly. To reconnect the domain structures, the linker residues of the full-length model with the lowest energy are relaxed, where the C_{α} atoms are regenerated by self-avoiding random walks, followed by adding the N, C, O atoms, and side-chain centers using FASPR⁶⁶. Next, Metropolis Monte Carlo simulations are performed to refine the linker model guided by a potential containing a C_{α} clash term between the linker and domain structures, a statistical torsion-angle potential from Ramachandran plots^{3,69}, an orientation-dependent side-chain contact potential, and a statistical N– C_{α} –C bond angle potential. Finally, the model with the lowest linker energy is selected for side-chain reconstruction and refinement by FASPR⁶⁶ and FG-MD⁶⁷.

Structure-, sequence- and PPI-based function annotation

The protein function, including GO terms, EC numbers and ligand-binding sites of the full-length protein structure and the individual domains are annotated by our latest version of COFACTOR⁴⁵. This new version of COFACTOR includes not only the structure-based function prediction pipeline inherited from the old COFACTOR program⁷⁰, but also new methods for sequence- and PPI-based function prediction (Supplementary Fig. 8). In the structure-based pipeline, the predicted model is searched by TM-align through the BioLiP library⁷¹ for structure templates with known GO terms, EC numbers and/or ligand-binding residues. Template residues in known active sites and/or ligand-binding sites are realigned by local structure similarity. The functions are transferred from structure templates to target structure by both global and local structural similarity. In the sequence-based pipeline, the target sequence is used to search through the UniProt-GOA database for BLAST and PSI-BLAST hits with GO annotations. GO terms are transferred to the target through a weighted

K -nearest-neighbor approach, where K is the number of (PSI-)BLAST hits with E -values ≤ 0.01 and the weight for each template is equal to the global sequence identity of the target-template alignment. The PPI-based pipeline of the latest version of COFACTOR is ported from MetaGO²⁴. In this pipeline, the target sequence is mapped by BLAST to the STRING database to identify PPI partners. The sequences of PPI partners are then mapped to UniProt-GOA by BLAST. GO terms are also predicted by a weighted K -nearest-neighbor method, where the weight of each UniProt-GOA sequence is calculated as the product of two values: the sequence identity between the UniProt-GOA sequence and the STRING PPI partner, and the STRING score between the target sequence and the PPI partner. Finally, the structure-, sequence- and PPI-based predictions are combined using weighted averaging to derive the final consensus function prediction.

Estimation of model quality

Estimating the accuracy of a predicted model is essential to decide how users will use the model in their research. In I-TASSER-MTD, the accuracy of the k th predicted model is estimated by the estimated TM-score, eTM-score(k), which is calculated on the basis of the convergence of the domain assembly simulations, the confidence of the full-length templates identified for domain assembly, the satisfaction rate of the predicted inter-domain distances and the estimated accuracy of the individual domain model by D-I-TASSER, using the following equation:

$$\begin{aligned} \text{eTM-score}(k) = & w_1 \ln \left(\frac{M(k)}{M_{\text{tot}}} \times \frac{1}{\langle \text{RMSD} \rangle_k} \right) + w_2 \ln \left(\frac{1}{10} \sum_{i=1}^{10} \frac{T\text{-score}(i)}{T\text{-score}_0} \right) + w_3 w_{\text{neff}} \ln \left(\frac{1}{T} \sum_{t=1}^T |d_t^{\text{pre}} - d_t^{\text{model}}(k)| \right) \\ & + w_4 w_{\text{neff}} \ln \left(\frac{O(I^{\text{pre}}, I^{\text{model}})_k}{N(I^{\text{pre}})} \right) + w_5 \frac{1}{N_{\text{dom}}} \sum_{D=1}^{N_{\text{dom}}} \text{eTM-score}_{\text{dom}}(D) + w_6 \end{aligned} \quad (1)$$

The first term in Equation (1) 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 root-mean-square deviation (RMSD) $< 1.5 \text{ \AA}$ to the k th full-length model and $\langle \text{RMSD} \rangle_k$ 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 $d_t^{\text{model}}(k)$ 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 $N(I^{\text{pre}})$ is the number of predicted domain–domain interfaces and $O(I^{\text{pre}}, I^{\text{model}})_k$ is the number of overlapped interfaces between the predicted interface map and the k th reported model. As 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 on the basis of the number of effective sequences (neff; 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 $\text{eTM-score}_{\text{dom}}(D)$ is the estimated TM-score of the D th domain model from D-I-TASSER (Supplementary Note 3). $w_1 = 0.065$, $w_2 = 0.063$, $w_3 = -0.08$, $w_4 = 0.01$, $w_5 = 0.96$ and $w_6 = 0.1$ are the weighting factors, which are optimized using an improved differential evolution algorithm^{72–74} to minimize the error between the eTM-score and the actual TM-score of the decoys to the native structure on the DEMO training set of 425 nonredundant multi-domain proteins. In addition, the estimated RMSD (eRMSD) of I-TASSER-MTD models is also calculated on the basis of the same terms in Equation 1 but with an additional term to count for the protein length (L) (Eq. (S14) in Supplementary Note 4), where the weighting factors for eRMSD are $w_1 = -1.40$, $w_2 = -2.74$, $w_3 = 4.78$, $w_4 = -1.19$, $w_5 = -16.43$, $w_6 = 0.0$ and $w_7 = 2.66$. Meanwhile, we define a new score to quantitatively assess the relative populations of the assembled conformations for the k th reported model by

$$\text{P-score}(k) = \frac{p(k)}{\sum_k p(k)} \quad (2)$$

where $p(k) = \frac{M(k)}{M_{\text{tot}}}$ is the normalized number of structural decoys of k th model in the I-TASSER-MTD assembly simulations.

The accuracy of eTM-score and eRMSD were examined on the DEMO benchmark set, which includes 356 multi-domain proteins with different domain types that are nonhomologous to the DEMO training dataset. As shown in Supplementary Fig. 9, the eTM-score has a high Pearson correlation coefficient (PCC = 0.85) with the actual TM-score, where the average error of the eTM-score is 0.07. Compared with the eTM-score, the eRMSD and RMSD have a slightly lower correlation (PCC = 0.82), where the average error between eRMSD and RMSD (2.2 Å) is relatively high (see the distribution in Supplementary Fig. 9d). It should be noted that RMSD is not the best measurement for the accuracy of predicted models when the modeling accuracy is low; as all residue pairs have the same weight in the RMSD calculation, this renders the RMSD value sensitive to local variations, such as tails or loops, rather than the global fold. In this regard, it is recommended to use TM-score as a more reliable measurement for model accuracy assessment; because the residue pairs with smaller distance errors are weighted more heavily than the residues with larger errors in the TM-score calculation, the TM-score value is generally more sensitive to the accuracy of the global fold of the predicted models⁶⁸.

As shown in Supplementary Fig. 9a, if we use an eTM-score cutoff of 0.5 to select models with correct global topologies, both the false-negative and false-positive rate are <0.15, indicating that the fold-level prediction by the eTM-score is correct in >85% of the cases. As illustrative examples, six models from two targets are included in Supplementary Fig. 10, showing that the eTM-score is highly correlated with the actual TM-score of the models in different ranges of model quality. In Supplementary Fig. 11, we also show the eTM-score and thus the model quality will be reduced when more and more random mutations are introduced to the target sequence of PDBID 1we3F.

In addition to the global quality assessment, to show the local accuracy of each individual domain model, I-TASSER-MTD estimates the residue-level distance error of the predicted model relative to the native structure using ResQ⁷⁵, a method for estimating residue-level quality in protein structure prediction on the basis of local variations of modeling simulations and the uncertainty of homologous template alignments.

Experimental design incorporating user-specific data

User-specified domain boundaries

While the I-TASSER-MTD server is capable of fully automated domain partition, it optionally allows users to provide their own domain definition. If a complete and effective domain definition is provided, the server will use the provided domain information to split the query sequence into multiple parts for independent structure modeling, rather than running FUpred or ThreaDom for the domain boundary prediction. If only a partial domain definition is provided, for example, with a mixed input sequence of defined and unknown domains, the server will keep the user-defined domains and predict the domains of unknown regions by FUpred or ThreaDom. After inputting the domain definition, the server will automatically check the effectiveness of the given domain definition. Specifically, a domain or a segment should be represented by a starting residue index and ending residue index with a hyphen between them. Domains should be separated by semicolons, while segments of a discontinuous domain should be separated by commas. The length of a domain should be between 30 and 1,000 amino acids, which is the range of domain lengths for the majority of CATH domains⁴⁶. Missing or overlapping residue indices are not allowed in the domain definition if users want to completely use their own domain definition rather than one determined by the server. All domains should be written on one line and end with semicolons (for details, see Box 1).

User-specified full-length structure templates

The I-TASSER-MTD server assembles multiple domain structures using analogous full-length templates followed by knowledge-based domain structure refinement simulations. By default, the server detects the analogous full-length templates from the multi-domain protein library through structural alignments by TM-align, where the library includes multi-domain proteins at a pair-wise sequence identity <70% or sequence identity ≥70% but with TM-scores <0.5. Alternatively, the server allows users to specify solved protein structures or separately predicted models as templates. The user-specified templates must be in PDB format. Users can provide up to 20 templates and compress them into one file to upload to the server. If more than five templates are provided, each template will be evaluated by the harmonic mean of the TM-score between all domain models and the template, and

Box 1 | Providing the domain definition

The three cases listed below should be considered in the provided domain definition.

- 1 Proteins with a complete continuous domain definition can be pasted in the text box like this example (PDBID: 2qbuA, Supplementary Fig. 2a):

1 – 131;132 – 228;

where 1–131 indicates that residue 1 to residue 131 belong to the first domain and residues 132–228 to the range for the second domain. Note that domains must be separated by semicolons, and a semicolon should be included at the end of the definition.

- 2 Proteins containing discontinuous domains in a complete domain definition should be prepared similarly to the following example (PDBID: 1atgA, Supplementary Fig. 2b):

1 – 81,191 – 232;82 – 190;

where '1–81,191–232' represents the definition of the first domain, which is a discontinuous domain; 1–81 indicates that residues from positions 1 through 81 make up the first domain segment; and 191–232 indicates that residues from 191 to 232 belong to its second segment. 82–190 is the range of the second continuous domain, which is inserted in the first discontinuous domain. Note that different parts of a discontinuous domain must be separated by commas.

- 3 The partial domain definition should be prepared like the following example (PDBID: 1h88C, Supplementary Fig. 2c):

1 – 51;

where '1–51' indicates that one of the domains should include residues from 1 to 51. The total sequence length of the protein is 152, and the complete domain definition should be '1–51;52–105;106–152;' as defined in the CATH database. The server will keep the domain provided by users when predicting the domain boundary. When providing the defined domain, please ensure that the length of every undefined domain region is be larger than the minimum length (30) of a domain.

the top five templates with the highest scores will be selected for the initial full-length model generation. If only one template is specified, the top four analogous templates detected from the library and the uploaded template will be used to create the initial full-length model. For both cases, the domain-template alignments are determined by a sliding-window procedure based on TM-score as used by DEMO¹⁶. Furthermore, all templates provided by the user will be utilized to extract the inter-domain distance profiles, which will be used to mainly guide the domain assembly simulations along with restraints deduced from the templates identified from the library. If the provided template comes from computationally modeled structures (such as those from AlphaFold2 predictions) and includes the confidence score (pLDDT)¹¹ in the 'Temperature factor' column, only the distances of the residue pairs with pLDDT >70 for both residues are extracted from the template to guide the assembly simulations.

Incorporating experimental cross-linking data or contact/distance restraints

Distance restraints and spatial contact information for a protein subunit or domain derived from chemical cross-linking by mass spectrometry can be very useful for protein structure modeling. Given this, the I-TASSER-MTD server provides an option to incorporate user-specified cross-linking data as an additional restraint to guide the domain model assembly. Users can upload the cross-linking data to the server, and the server will automatically check and confirm that the data are in the correct format: the first column and the second column in the provided file should be the residue indices with C_{α} distances less than a cutoff, which should be specified in the third column; headers or footers are not allowed in the data file. Once effective cross-linking data are given, they are converted to a three-gradient contact potential¹⁶ and added to the DEMO assembly energy function, where the weight of the cross-linking potential was optimized over a training set. The other parts of the DEMO assembly process will remain unchanged. It should be noted that, since the cross-linking data are implemented as residue–residue contacts, the provided data are not limited to cross-linking. Instead, contact or distance restraints obtained from any methods can be written in the specified format and input to the server to guide the domain model assembly (for details, see Box 2). To allow I-TASSER-MTD to better implement the restraints, users are recommended to provide a confidence score of each residue pair in the fourth column when providing the predicted contact or distance restraints. All confidence scores should be in the range of [0, 1], and they will be considered as 1 if not provided. The confidence score of each residue pair will be used to determine the weight of the residue pair in the contact potential (Supplementary Note 4). As residue pairs with low confidence scores usually

Box 2 | Incorporating residue-residue contact restraints

Residue-residue contacts from experimental cross-linking data or predicted by any other programs can be prepared and pasted in the text box on the main page of the server similar to the following example (PDBID: 1fx7A):

15	165	28	0.85
63	220	21	0.96
94	168	17	0.79

where the first and the second columns are the residue indices, and the third column is the maximum C_{α} distance for the residue pair. The fourth column is the confidence score that the distance of the residue pair is less than the given distance listed in the third column. The confidence score is optional, and it will be set to 1 if not provided. Values in the same row should be separated by tabs or spaces. For example, the first row '15 165 28 0.85' indicates that the C_{α} distance between residues 15 and 165 has a confidence of 0.85 being less than 28 Å. Note that the residue index is for the full-chain sequence rather than each individual domain, and the residue index starts from 1 rather than 0.

have relatively low accuracy in the contact/distance prediction, users are recommended to provide restraints with confidence score >0.5 to reduce noise.

Integrating experimental cryo-EM data

Cryo-EM has become established as powerful method for macromolecular structure determination in recent years⁷⁶. The I-TASSER-MTD server is equipped to integrate cryo-EM density maps to assist domain assembly and refinement (Supplementary Fig. 12). Specifically, when a cryo-EM density map is available, each domain model is first matched into the density map by performing limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization with different starting positions to identify the best location and orientation with the highest density correlation to the density map. The top poses of each domain are combined to build multiple initial full-length models. The full-length models generated on the basis of the templates identified by TM-align are also fit into the density map using the same L-BFGS algorithm. According to the density correlation score⁷⁷ between the density probed from the full-length model and the experimental density, the top full-length models are selected to optimize the orientations of all domains by DEMO rigid-body assembly under the guidance of the density correlation score and the inherent DEMO assembly force field. To further improve the individual domain models and the full-length model, the top models are selected from the rigid-body assembly results according to the density correlation score, and atom-, segment- and domain-level refinement are performed using REMC simulations guided by a knowledge-based force field, inter-domain distances predicted by DeepPotential, and the density correlation score. Finally, the full-length model with the lowest energy is selected for side-chain repacking by FASPR and FG-MD.

Comparison with other methods

Although a number of webservers have been developed to predict the structures and functions of proteins, very few are devoted to fully automated multi-domain protein structure and function prediction from sequence alone. The I-TASSER-MTD server distinguishes itself from other protein structure prediction servers in several major aspects pertaining to multi-domain proteins (Table 1). First, unlike other protein structure prediction servers where all query proteins are considered as single-domain proteins or the domain boundaries are roughly determined using external online programs such as PFAM⁷⁸ or the NCBI Conserved Domain Database search service⁷⁹, I-TASSER-MTD automatically detects the domains according to a combination of threading- and deep-learning based domain boundary prediction algorithms depending on the target type (homologous or non-homologous). Of particular significance, I-TASSER-MTD can effectively predict complex discontinuous domains that consist of multiple separate segments at the sequence level. Second, for multi-domain proteins, almost all servers can generate individual domain models by submitting the individual domain sequences independently, but few can automatically assemble the full-length model. The I-TASSER-MTD server not only can construct the models of individual domains, but can also assemble all domain models into the full-length model. Third, an important unique feature of I-TASSER-MTD is that it provides options for several types of experimental restraints to guide the domain model assembly, including homologous templates, cross-linking data, cryo-EM density maps and other sources of residue-residue contact or distance information. Fourth, unlike other servers

Table 1 | Comparison of I-TASSER-MTD with other protein structure prediction servers

Name	URL	DBP	IDM	MDA	DF, FF	EDM
I-TASSER-MTD	https://zhanggroup.org/I-TASSER-MTD/	Yes	Yes	Yes	DF + FF	Yes
RoseTTAFold ⁸²	http://rosetta.bakerlab.org/	No	Yes	No	No	No
RaptorX ^{103,104}	http://raptorx.uchicago.edu/	Yes	Yes	No	DF/FF	No
trRosetta ^{57,105}	https://yanglab.nankai.edu.cn/trRosetta/	No	Yes	No	No	No
PSIPRED ¹⁰⁶	http://bioinf.cs.ucl.ac.uk/psipred/	Yes	Yes	No	DF/FF	No
HHpred ¹⁰⁷	https://toolkit.tuebingen.mpg.de/tools/hhpred	No	Yes	No	No	No
Phyre ¹⁰⁸	http://www.sbg.bio.ic.ac.uk/phyre2/	Yes	Yes	No	DF/FF	No

DBP, domain boundary prediction; DF, domain level function prediction; EDM, experimental data-assisted modeling; FF, full-length level function prediction; IDM, individual domain modeling; MDA, multi-domain protein structure assembly.

that mostly predict the function of the individual domains, I-TASSER-MTD performs the function annotation of the query protein at both the domain level and full-length level.

As a blind test, I-TASSER-MTD (as ‘Zhang-Server’) participated in the most recent community-wide CASP experiment for fully automated protein structure prediction. In Fig. 2a, we present a summary of the five best performing servers in CASP14, in which we sorted the servers according to the average global distance test (GDT) score of the full-length models for all multi-domain proteins with one or more template-free modeling (FM) or template-free modeling/template-based modeling (FM/TBM) domain. The average GDT score of I-TASSER-MTD for the multi-domain proteins was the highest among all participating servers. The accuracy of the individual domain models for multi-domain proteins was also higher than that of other servers. For example, I-TASSER-MTD achieved an average GDT score of 61.4 for all individual domain models of the multi-domain proteins, which was 19.4% higher than that of the second-best server, ROSETTA (51.4). This is mainly due to the incorporation of the highly accurate deep-learning-based restraints from DeepPotential in the I-TASSER-MTD simulations. The average GDT score of I-TASSER-MTD was also ranked the highest for the structure modeling of single-domain proteins in CASP14.

Owing to the employment of FUPred and ThreaDom, I-TASSER-MTD can accurately distinguish multi-domain proteins from single-domain proteins and predict the domain boundaries with reasonable accuracy. Here, since we cannot obtain the domain definitions used by other servers in CASP, I-TASSER-MTD is compared with two state-of-the-art methods ConDO²¹ and DoBo⁸⁰ on all CASP14 targets. As shown in Fig. 2b, the accuracy of I-TASSER-MTD domain boundary prediction is significantly higher than the that of the two control methods in terms of normalized domain overlap (NDO) score⁸¹ for the protein domain boundary prediction, as well as accuracy (ACC) and Matthew’s correlation coefficient (MCC) for protein classification. For example, the NDO score of I-TASSER-MTD for multi-domain protein was 0.86, which was 65.4% and 79.2% higher than that of ConDO (0.52) and DoBO (0.48), respectively.

We also compared I-TASSER-MTD with AlphaFold2 (ref. 11) and RoseTTAFold⁸² on all CASP14 targets. As the AlphaFold2 results reported in CASP14 are based on the human-expert group, while the results of I-TASSER-MTD (as ‘Zhang-Server’) are based on the automated server group, we regenerated all models by running the standalone AlphaFold2 package for a fair comparison. The average TM-score of AlphaFold2 is 0.84, which is considerably higher than that of I-TASSER-MTD (0.65). For RoseTTAFold, which has two options, I-TASSER-MTD’s TM-score is slightly higher than the RoseTTAFold end-to-end version (0.63) but slightly lower than the RoseTTAFold pyRosetta version (0.69).

To highlight the effectiveness of I-TASSER-MTD on some protein targets, we list in Fig. 2c–e three examples of multi-domain models built by I-TASSER-MTD that had a significantly higher TM-score than models built with the state-of-the-art programs. First, Fig. 2c shows the comparison between the native and predicted structures of *human complement component C6* (Uniprot ID: P13671). Although AlphaFold2 almost correctly predicted all domain models (TM-score 0.78, 0.93, 0.93, 0.88 and 0.87), the domain orientations were not correctly generated, resulting in a full-length model with a poorer TM-score/RMSD of 0.63/31.1 Å. The I-TASSER-MTD model obtained a TM-score/RMSD of 0.95/3.2 Å since it correctly generated both domain models and inter-domain orientations after the assembly. Figure 2d shows the second example from *Sarcoplasmic/endoplasmic reticulum calcium ATPase 2* (Uniprot ID: P16615), where I-TASSER-MTD generated a better-quality model (TM-score/

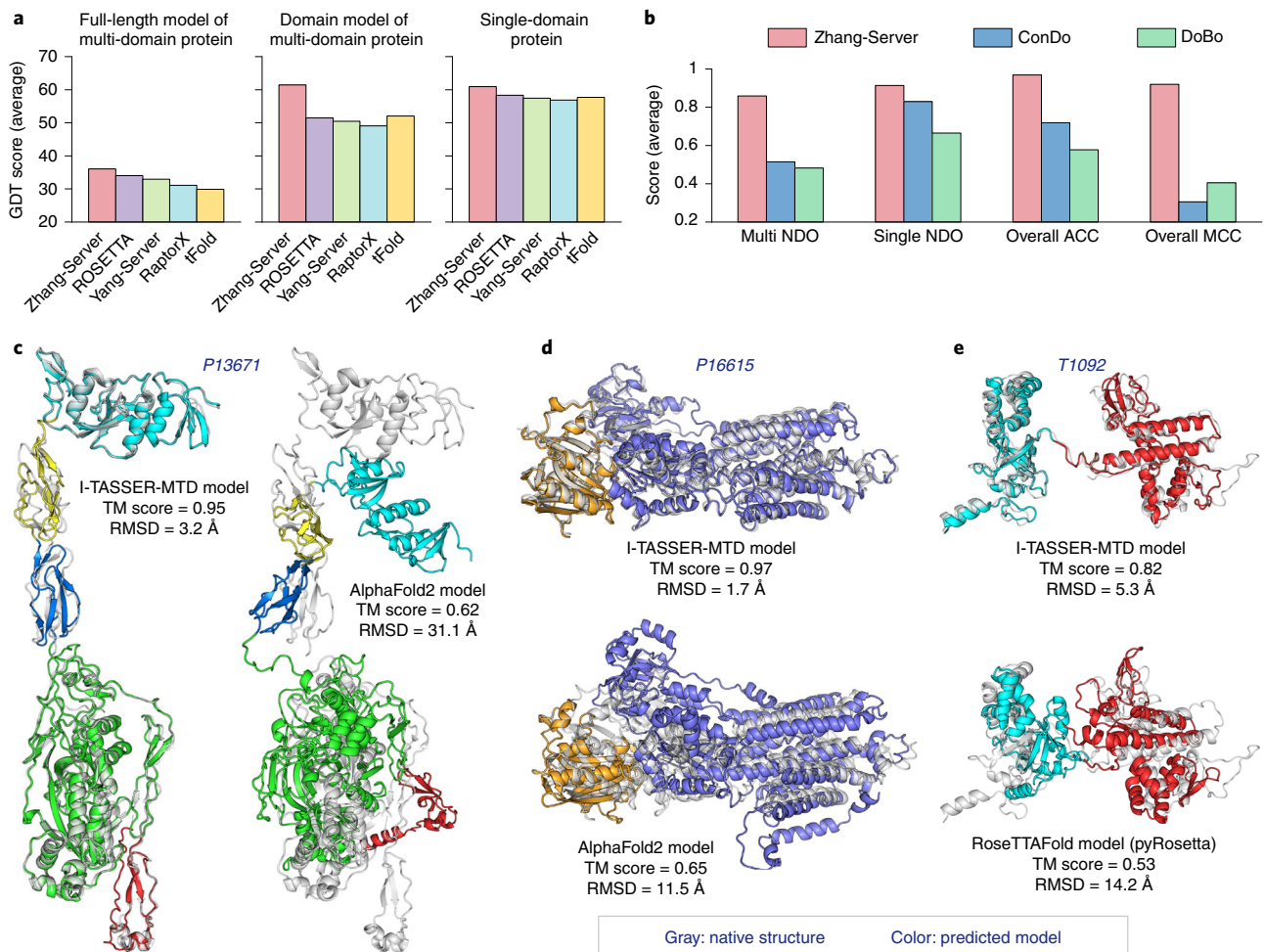


Fig. 2 | Comparison between I-TASSER-MTD and other methods. **a**, Comparison between I-TASSER-MTD (Zhang-Server) with the other top four servers of CASP14 on modeling the full-length multi-domain targets assessed by the GDT score. **b**, Comparison of I-TASSER-MTD with ConDo and DoBo for the protein domain boundary prediction on the CASP14 targets, where the y-axis is the NDO score of the multi-domain protein, the NDO score of the single-domain protein, and the ACC and MCC for the four subparts from left to right. **c–e**, Representative examples in which I-TASSER-MTD generated better quality full-length models than AlphaFold2 and RoseTTAFold, where the gray and color cartoon represent native structure and predicted model, respectively, and different colors indicate different domains: human complement component C6 (P13671) (**c**); human sarcoplasmic/endoplasmic reticulum calcium ATPase 2 (P16615) (**d**); DNA-directed RNA polymerase beta' subunit (T1092 of CASP14) (**e**).

RMSD 0.97/1.7 Å) than AlphaFold2 (0.65/11.5 Å), while the latter misfolded the larger-size domain resulting in an incorrect overall domain orientation. Finally, Fig. 2e presents an example of the CASP14 target (T1092) from *DNA-directed RNA polymerase beta' subunit*. Although the RoseTTAFold pyRosetta version generated a correct fold for the domain models (TM-score 0.77 and 0.87), the domain orientations were not correctly modeled, resulting a full-length TM-score/RMSD of 0.53/14.2 Å. Again, I-TASSER-MTD correctly constructed both the domain models and domain orientations, thus obtaining a full-length model with TM-score/RMSD of 0.82/5.3 Å. These data demonstrate that I-TASSER-MTD is complementary to these state-of-the-art programs, especially for long protein sequences with multiple domains, although the overall performance on many other targets of I-TASSER-MTD, the average TM-score of which is still lower than AlphaFold2, needs further improvement.

In addition, compared with the deep-learning-based end-to-end models (i.e., AlphaFold2 (ref. 11) and RoseTTAFold⁸²), which are largely a black box to both developer and users⁸³, I-TASSER-MTD has several advantages due to the fact that its simulation process is accessible and interpretable. First, I-TASSER-MTD reports the templates used to model each of the regions (domains) and full-length protein, which can help users better understand where the predictions come from and therefore provide functional insights for further studies on the protein. In fact, I-TASSER-MTD offers a separate section for protein function annotation built on the structural modeling results.

Second, I-TASSER-MTD simulations often generate different models for a query protein, which can be crucial for the protein folding and function study since there are some proteins that adopt alternative conformations for different states. For example, the human protein Pin1 contains separate regulatory and catalytic domains which sample ‘extended’ and ‘compact’ states with different structures⁸⁴. Experimental studies have shown that there is an equilibrium between ‘compact’ and ‘extended’ states in a rough approximation with populations of 50:50 (ref. ⁸⁵). Supplementary Fig. 13 shows that the top five models generated by I-TASSER-MTD are highly diverse and include both ‘extended’ and ‘compact’ states, while models constructed by AlphaFold2 converge to a single ‘compact’ state. Accordingly, I-TASSER-MTD provided a P-score to quantitatively assess the relative populations of assembled conformations (Equation 2)). Finally, owing to the accessibility of the modeling process, different sources of user-provided information, including cross-linking and cryo-EM data, could be readily incorporated into the I-TASSER-MTD pipeline to improve the multi-domain structure assembly.

Further applications of the I-TASSER-MTD protocol

In addition to its capacity as a powerful protein structure prediction server, the advancements for multi-domain structure prediction by I-TASSER-MTD provide several additional useful applications.

First, the domain definition and the corresponding domain models predicted by I-TASSER-MTD can be used for protein family detection. The classification of evolutionary relationships of proteins is crucial for protein structure and function studies. Several databases, such as the widely used SCOPe⁸⁶ and CATH⁸⁷ databases, have been developed to classify protein structures on the basis of individual domains, which are considered to be the structural, functional and evolutionary units of proteins⁸⁸. As structure is more conserved than sequence⁸⁹, almost all databases group protein domains in the light of both structural similarity and sequential information. However, the domain definition and classification for them are only performed on the basis of the solved proteins in the PDB. Therefore, for proteins without solved structures, the domain boundaries and models predicted by I-TASSER-MTD can be used to assist protein classification.

Second, I-TASSER-MTD supports cryo-EM data-assisted modeling, which can be used to help experimental biologists construct protein structures from cryo-EM density maps. Cryo-EM has become an indispensable method for determining structures of large proteins. For high-resolution cryo-EM density maps, atomic structures can be constructed by programs traditionally used for X-ray crystallography⁹⁰, but these programs perform relatively poorly for medium-to-low-resolution density maps⁹¹. A common method for the structure modeling of these challenging density maps is to fit a homologous structure into the density map, followed by atomic-level structural refinement. However, the success of the approach highly depends on the quality of the initial models, while many proteins, especially multi-domain proteins, have no homologous proteins with previously solved structures. For these cases, I-TASSER-MTD can be employed for modeling as it constructs full-length models by assembling independently predicted domain models, without requiring homologous full-length structures.

Finally, an important aspect of I-TASSER-MTD is the annotation of protein functions at both the domain- and full-chain level based on predicted structures using its integrated COFACTOR algorithm. COFACTOR was shown to be accurate and scalable for full proteome-scale structure-based annotation of microbes⁹² and higher organisms⁹³ alike. For this purpose, the I-TASSER/COFACTOR pipeline was incorporated into the neXtProt database for automated human protein function modeling⁹⁴. It was also applied to the JCVI-syn3.0 minimal genome where it found that a substantial number of previously unannotated proteins are putative vitamin transporters⁹². In this regard, COFACTOR predictions can be useful for function hypothesis generation when planning low-throughput experiments. For example, the COFACTOR GO prediction was recently used to guide the characterization of C9orf72, a guanine nucleotide exchange factor, which lead to a better understanding of its molecular role in amyotrophic lateral sclerosis⁹⁵. Similarly, the ligand-binding prediction of COFACTOR was used for identifying heme-binding sites in HemJ, the poorly studied Protoporphyrinogen IX oxidase in cyanobacteria⁹⁶.

Limitations

The domain boundary prediction method employed in the I-TASSER-MTD server for Hard targets with template alignment coverage $\leq 95\%$ is based on the deep-learning predicted contact maps that

require MSA collection. For extremely large proteins (>2,000 residues), contact map prediction and MSA collection will require a high amount of random access memory (RAM) that the current computing server cannot provide for some cases, which will result in failed domain partitioning by FUpred due to memory limit. However, this limitation can be overcome by using ThreaDom for all targets with sequence lengths >2,000 residues. Users can download the FUpred standalone package and run it locally to predict the domain boundaries for a query sequence if they want to use the FUpred-predicted domain definition for these cases. Furthermore, users also can provide the domain definition predicted by other external programs such as NCBI Conserved Domain Database or PFAM. This will also speed up the I-TASSER-MTD structure modeling process as the server will not take time to predict the domain boundaries. See the 'Experimental design' section for detailed instructions on how to provide the domain definition.

One highlight of the I-TASSER-MTD server is that it independently creates the model of each individual domain and assembles all domain models into the full-length model. However, the quality of the final full-length model is dependent on the accuracy of the individual domain models. Although the domain assembly process and scoring function can accommodate some degree of structural uncertainty, an incorrect domain model (e.g., TM-score <0.5) may affect the full-length template identification based on structural alignment and misguide the domain assembly. This will probably result in a poor final full-length model with a low eTM-score because each model is considered as a rigid body during the domain assembly. For cases with low eTM-scores (e.g., <0.5), users are advised to provide other sources of structural information, such as cross-linking experimental data, restraints from mutagenesis, high-confidence full-length templates or inter-domain contacts/distances determined by alternative programs, to guide the domain assembly.

Materials

Equipment

- A personal computer with Internet connection and a web browser with JavaScript enabled (the I-TASSER-MTD server is compatible with popular web browsers, including Google Chrome, Firefox, Microsoft Edge and Safari)
- The amino acid sequence of the protein of interest **▲ CRITICAL** The amino acid sequence should be in FASTA format, in which only characters from the single-letter code of the 20 standard amino acids are allowed. Spaces, line breaks and header lines starting with '>' will be ignored and will not affect the prediction.

Software

- A web browser such as Google Chrome, Firefox, Microsoft Edge or Safari
- (Optional) A molecular visualizing software, such as Jmol, RasMol or PYMOL, for viewing the 3D structure of the modeled protein and the predicted functional sites locally

Procedure

Query sequence submission ● Timing 5 min

- 1 Navigate to the I-TASSER-MTD website at <https://zhanggroup.org/I-TASSER-MTD/>.
- 2 Provide the amino acid sequence by copying and pasting the sequence into the provided form or directly uploading a plain text file containing the sequence.
▲ CRITICAL STEP The sequence should contain only one chain. If the provided sequence includes multiple chains, only the first chain will be used. At present, the I-TASSER-MTD server accepts protein sequences with a length between 30 and 3,000 amino acids.
- 3 Input an email address in the text box to receive the results when the job is completed.
▲ CRITICAL STEP It is crucial to provide a correct email address. Otherwise, the user will not be able to receive any notifications or the results.
- 4 (Optional) Enter a name for the protein. The protein will be named 'query protein' if no name is provided.
- 5 (Optional) Provide the domain definition. Enter the domain definition in the corresponding form. The maximum length of a sequence accepted by the server will be increased to 3,000 if the domain definition is provided. Read about specifying the domain definition in the 'Experimental design' section or click the question mark to read the explanation.

? TROUBLESHOOTING

- 6 (Optional) Upload the full-length templates to guide the assembly. Put all templates (each template in one independent PDB file) in one compressed file (*.tar.gz, *.zip, *.tar or *.tar.bz2), and upload it by clicking the corresponding button. Read about specifying the full-length templates in the 'Experimental design' section of this protocol or click the question mark to read the explanation.
! CAUTION The server uses a maximum of 20 templates. If users provide more than 20 templates, only the first 20 templates will be utilized during the domain assembly simulations.
? TROUBLESHOOTING
- 7 (Optional) Exclude some templates from the library for both individual domain modeling and domain model assembly. Users are advised to keep the default option to use all templates. In rare scenarios for benchmarking purposes, users can choose 'remove templates sharing >30% sequence identity with target' to remove all the templates that are highly homologous to the target sequence.
! CAUTION Users are advised to keep the default selection of using all templates during the modeling. In general, excluding homologous templates will make structure prediction harder. Therefore, this option is only for benchmarking purposes.
- 8 (Optional) Decide whether or not to predict the protein function at the domain and full-length level. If users need to analyze the function of the protein, choose 'YES' to predict the function. Otherwise, keep the default option.
! CAUTION Selecting the option to predict the protein function will greatly increase the job runtime because the protein function prediction cannot be performed until the model is generated.
- 9 (Optional) Provide experimental data to assist the domain assembly. Prepare the cross-linking data as described in the 'Experimental design' section, and select the corresponding box to paste the data in the form or directly upload the file containing the data. To provide cryo-EM data, select the corresponding box to show the 'upload' button, and upload the cryo-EM density map (in MRC or CCP4 format) by clicking the button. Users should also enter the resolution of the density map in the given text box.
▲ CRITICAL STEP As we showed using the DEMO benchmark set, cross-linking data and cryo-EM density map will significantly improve the quality of the final full-length model^{16,77}. In addition, cross-linking data can also be replaced by contact or distance restraints determined by any contact or distance prediction program.
? TROUBLESHOOTING
- 10 Click the 'Run I-TASSER-MTD' button to submit the job.
? TROUBLESHOOTING

Job monitoring ● Timing 6–12 h

- 11 Monitor the job status. Once the job is submitted, the browser will be directed to a new page displaying a confirmation of the length of the submitted sequence, a job identification number and an estimated time to complete the job. The page is automatically refreshed every 10 s, and the results will be shown in this page when the job is finished. Users may choose to bookmark this link to retrieve the results. Meanwhile, users will receive an email confirmation containing a link to the page when the job is successfully submitted.
■ PAUSE POINT Once the sequence is successfully submitted, it will be put in a queue until all other jobs before it are processed on the computer cluster. Users may choose to close the job status page. When the prediction is done, an email notification containing the link to the results page will be sent to the user. The results can be accessed through this link or the bookmarked page.
- 12 Click the link in the email notification or open the link bookmarked in Step 11 to visit the page containing the results. The page starts with a title and a link to download the tarball file including all results listed on the page (Fig. 3a). An example results page is available at <https://zhanggroup.org/I-TASSER-MTD/example/>.
! CAUTION The results will be stored on the server for 1 month. Users are recommended to download the results to their computers.

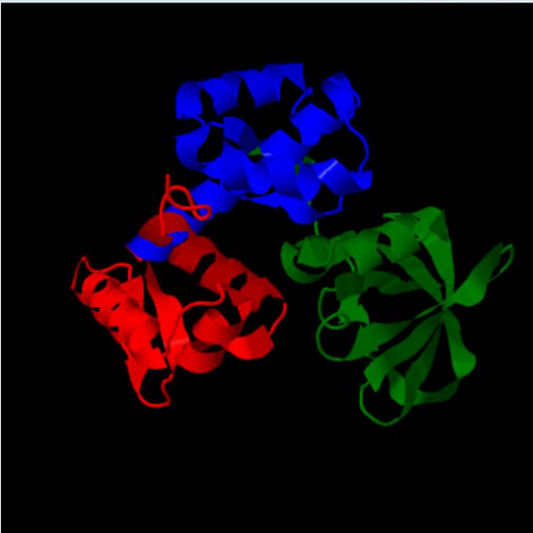
Query sequence and predicted domain definition ● Timing 3 min

- 13 View the first section of the results page to check the submitted amino acid sequence and the predicted domain definitions (Fig. 3b). The sequence of each predicted domain is represented by different colors in the full-length query sequence. The illustration of the color and the sequence range for each domain are listed below the query sequence. See Box 1 to follow the representation of the domain range.

a 1 I-TASSER-MTD Results for job ITM118
2 [\[Click on ITM118_results.tar.bz2 to download the tarball file including all results listed on this page\]](#)

b Query Sequence and Predicted Domain Definition
 >1fx7A (230 residues)
 MNELVDTTEMYLRTIYDLEEEGVTLPLRARIARLDQSGPTVSQTVSPMERDGLLRVAGDRHLELTKGRALAIIVMRKHLAERLLVDVIGLPWEEVHAEACRWEHVMSMEDVERRLVKLVNPTSPFGNPIPLDELGVG
 PEPGADDANLVRLTELPAGSPVAVVVRQLTEHVQGDIDLITRLKDGAVVFNARVTVETTPGGGVTVIPGHENVTLPHEMAHAVKVEKV
■ domain 1: 1-66 ■ domain 2: 67-141 ■ domain 3: 142-230
 3 domains in total. Different colors represent different domains. The domain definition is used for the domain structure modeling, see the domain boundary prediction part for details.

c Final Full-length Models Predicted by I-TASSER-MTD



Reset Spin High quality White background

Top 5 models constructed by I-TASSER-MTD


Click to view	eTM-score	eRMSD(Å)	P-score	PDB file
<input checked="" type="radio"/> Model 1	0.92+-0.08	2.8+-0.7	0.25	Download model
<input type="radio"/> Model 2	0.78+-0.11	4.2+-1.2	0.24	Download model
<input type="radio"/> Model 3	0.74+-0.15	4.8+-1.1	0.20	Download model
<input type="radio"/> Model 4	0.72+-0.14	5.0+-1.4	0.19	Download model
<input type="radio"/> Model 5	0.70+-0.16	5.3+-1.6	0.12	Download model

Probabilities of inter-domain interactions 1
 domain1-domain2: 1.00; domain1-domain3: 0.17; domain2-domain3: 0.85;

(a) Colored by domain: domain 1 in red; domain 2 in blue; domain 3 in green.
 (b) For each target, I-TASSER-MTD generates an ensemble of structural conformations by starting from a set of initial models generated by different templates. The server reports up to five final models sorted by the energy. The accuracy of each model is quantitatively evaluated by estimated TM-score (eTM-score) and estimated RMSD (eRMSD) that are calculated based on the significance of the structural analogous templates for domain models assembly, convergence parameters of the domain assembly simulations, satisfaction degrees of the inter-domain distances/interfaces, and the estimated accuracy of the individual domain model. eTM-score is typically in the range of [0,1], where an eTM-score of higher value signifies a model with a high confidence and vice-versa.
 (d) Since the top 5 models are ranked by the energy or cluster size, it is possible that the lower-rank models have a higher eTM-score in rare cases. Although the first model has a better quality in most cases, it is also possible that the lower-rank models have a better quality than the higher-rank models as seen in our benchmark tests.
[More about eTM-score 2](#)
 (e) P-score is used to estimate the population formed in the modeling simulations based on the structural similarity or SPICKER clustering. P-score ranges from 0 to 1, and a higher value means the structure occurs more often in the simulation trajectory.
 (f) The inter-domain interaction is defined as ≥1 residue pairs with distance <8Å apart from the linker region. The probability ranges from 0 to 1, and a large value indicates the two domains have a large probability of interaction.

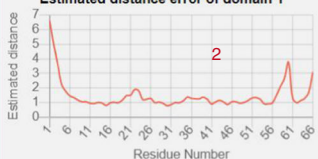
d Predicted Individual Domain Structures

Structure of domain 1



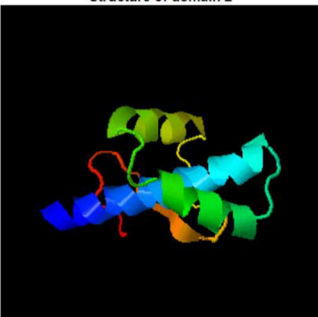
[Download dom1.pdb](#)
eTM-score=0.76 1

Estimated distance error of domain 1



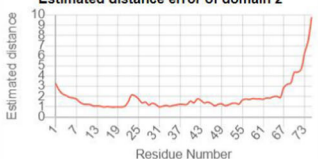
[Download distance error of domain 1](#)
[Click to view the predicted function](#)

Structure of domain 2




[Download dom2.pdb](#)
eTM-score=0.77

Estimated distance error of domain 2



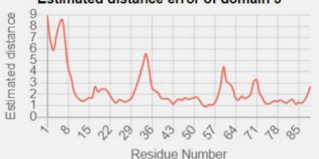
[Download distance error of domain 2](#)
[Click to view the predicted function](#)

Structure of domain 3



[Download dom3.pdb](#)
eTM-score=0.7

Estimated distance error of domain 3



[Download distance error of domain 3](#)
[Click to view the predicted function](#)

Fig. 3 | Example of the I-TASSER-MTD results page (Sections 2 and 3). **a**, Title of the results page and the link to download all results shown in the page. **b**, Query sequence in FASTA format submitted by the user, where the predicted domains are marked by different colors, and the range of each domain is listed below the sequence. **c**, Top final full-length models predicted by the server and their estimated accuracy (right). Model 1 is shown (left), where different domains are represented by different colors. **d**, Predicted individual domain models that are used to assemble the full-length model and the estimated distance error of each domain model.

Full-length structure prediction ● Timing 10 min

14 View the second section of the results page to analyze the full-length structure predictions (Fig. 3c). The default structure shown in the JSmol applet⁹⁷ is the first model, where different domain models are marked by different colors. Users can rotate and zoom the model by dragging the mouse on the image. The table to the right of the image summarizes the top five predicted models. In the first column, click the circle in front of the model name to display the corresponding full-length model in the JSmol applet.

▲ CRITICAL STEP If only one full-length model is reported in this section, it indicates that the model was directly generated by D-I-TASSER as the threading templates could cover all domains, and the top templates identified by LOMETS2 had consistent topologies. In these cases, the final model usually has a relative high eTM-score, indicating a high-quality final model.

15 View the second column of the table to analyze the eTM-score for each full-length model. As defined in Equation. 1, the eTM-score is calculated on the basis of the confidence of individual domain models and the confidence of the inter-domain assembly simulations. The eTM-score usually ranges from 0 to 1, where a higher score indicates a model of better quality. In general, models with eTM-score >0.5 have a correct global fold.

? TROUBLESHOOTING

16 View the eRMSD to the native structure shown in the third column of the table. As defined in Eq. (S14), eRMSD is estimated in a similar way as the eTM-score but with the sequence length incorporated.

! CAUTION Since the top five models are ranked by energy or by cluster size, in rare cases it is possible that the lower-rank models have a higher eTM-score or lower eRMSD. Accordingly, although the first model has a better quality on average, it is also possible that the lower-rank models have a better quality than the higher-rank models as seen in our benchmark tests. Users can also estimate the total local distance difference test (LDDT) of the model for reference by using deep-learning based quality assessment methods, such as DeepAccNet⁹⁸ and DeepUMQA⁹⁹.

17 View the P-score reported in the fourth column of the table. Here, P-score is used to assess the relative populations of complex conformations under the assumption that the relative populations of the complex conformations are approximately proportional to their entropy variations in the domain structural assembly simulations. The P-score value ranges between [0, 1], and a higher value indicates the structure occurs more often in the simulation trajectory.

18 Download the predicted models in PDB format by clicking on the 'Download model' link shown in the corresponding row of the column marked as 'PDB file'. Users can interactively view the predicted structure on their computer using the programs mentioned in the 'Materials' section.

19 View the predicted probability of the inter-domain interaction for every two domains, which is listed under the table (see label 1 in Fig. 3c). Here, an inter-domain interaction is defined as one or more residue pairs with distance <8 Å apart from the linker region. The domain interaction probability is estimated by the average probability score of the top 1% of the inter-domain residue pairs predicted by the deep-learning models. It ranges from 0 to 1, with a higher value indicating that the two domains have a larger probability of interaction.

20 Click the link labeled 'More about eTM-score' (see label 2 in Fig. 3c) to open a new page containing more information about the eTM-score and eRMSD.

Individual domain modeling ● Timing 5 min

21 Scroll down further to see the individual domain models that are independently created by D-I-TASSER (Fig. 3d) and used to assemble the full-length models shown in Fig. 3c. The 3D model of each domain is independently displayed using the JSmol applet.

22 Download the PDB file of the predicted individual domain structure by clicking on the 'download domX.pdb' link below the image of each individual domain model.

23 View the eTM-score of the domain modeling shown below the download link for each domain model (see label 1 in Fig. 3c), and the distance error to native (in angstrom) for each residue estimated by ResQ⁷⁵ which is displayed in the chart (see label 2 in Fig. 3c) under the structure. Users can view the estimated error of each residue by moving the mouse on the chart and download the file summarizing the predicted distance error of all residues by clicking on the 'Download distance error of domain X' link (see label 3 in Fig. 3c) below the chart.

24 View the predicted function of each individual domain. If the query protein is predicted to be a multi-domain protein and the user chose the option for function prediction, the predicted function

reported in four columns. The first column displays the contact map applied to guide the prediction of domain boundaries (see label 1 in Fig. 4c), according to which different domains (or segments of the discontinuous domain) are separated by red lines. Each domain or segment is marked by a name on the contact map. For example, 'D1' indicates the first domain (continuous). If a 'D1-1' is marked on the figure (Supplementary Fig. 14), it indicates that the second domain is a discontinuous domain, and 'D1-1' represents its first segment (Box 1). Users can click 'Download contact map' to download the contact data for further study.

! CAUTION If the domain boundary is predicted by ThreaDom, only two columns will be shown in this section (Supplementary Fig. 15), where the first column is the curve of the DCS, and the second column is the predicted domain definition.

- 28 View the second column of the domain boundary prediction results (see label 2 in Fig. 4c). The figure shows the FU-score curve of all residues for the continuous domain detection, on which vertical red lines correspond to the indices of the predicted domain boundary residues with FU-scores below the cutoff (the horizontal dotted line). Users can download the FU-score file by clicking 'Download the FU-score (continuous)'.

! CAUTION For some cases, the domain boundary does not correspond to the residue index with the lowest FU-score. This is because the predicted boundary is located within a strand or a helix, and so the boundary is shifted to the coil region as domain boundaries occur mainly at loop regions rather than on regular secondary structures.

- 29 View the FU-score heatmap for discontinuous domains shown in the third column of the domain boundary prediction results (see label 3 in Fig. 4c). The figure is generated according to the FU-score matrix for the discontinuous domain detection, where the colors ranging from blue to red indicate low to high scores, and the black dotted lines represent the predicted continuous domain boundaries. Users can download the FU-score matrix by clicking 'Download the FU-score (discontinuous)'.
- 30 View the predicted domains reported in the fourth column of the domain boundary prediction results (see label 4 in Fig. 4c), where each domain is written in one line. The 'Modeling' and 'Without linker' parts refer to the domain definitions used for the domain structure modeling and the domain boundary definition without including the linker, respectively.

Full-length templates for domain assembly ● Timing 5 min

- 31 View the next section of the results page to analyze the top ten full-length templates identified by the TM-align-based structural alignments on the basis of the domain models (Fig. 5a). These templates are employed to construct the initial full-length model, and the structurally aligned regions of these templates are used to deduce the inter-domain distance profiles to guide the domain assembly simulations.

- 32 View the sequence identity between the query and the template proteins (column SeqId), where a higher value indicates a strong evolutionary relationship between the query and template proteins.

- 33 View the score of the full-length template (column 'TplScore'), which is the harmonic mean of the TM-scores between the domain models and the template. The higher the score, the higher the structural similarity between the template and the query protein.

! CAUTION As described in the section entitled 'The I-TASSER-MTD pipeline', the full-length model will be directly created by D-I-TASSER if the LOMETS2 threading templates cover all domains. Therefore, in this case, the templates shown here will be those identified by LOMETS2 threading, and the TplScore will be the harmonic mean of the TM-scores between all domain models and the LOMETS2 template.

- 34 View the structural alignments to identify similar regions in the query and the template proteins. The query sequence is colored by domains, and the aligned residues in the template that are identical to the corresponding query residues are colored on the basis of their amino acid properties in the alignment.

- 35 Click on the PDB code and chain identifier for the templates in the 'Template' column. The browser will be directed to the Research Collaboratory for Structural Bioinformatics website showing information about the template protein.

Predicted distance/interface map for domain assembly ● Timing 5 min

- 36 Scroll down to see the predicted distance/interface map section for the domain model assembly (Fig. 5b). The first column and the second column show the deep-learning predicted distance maps

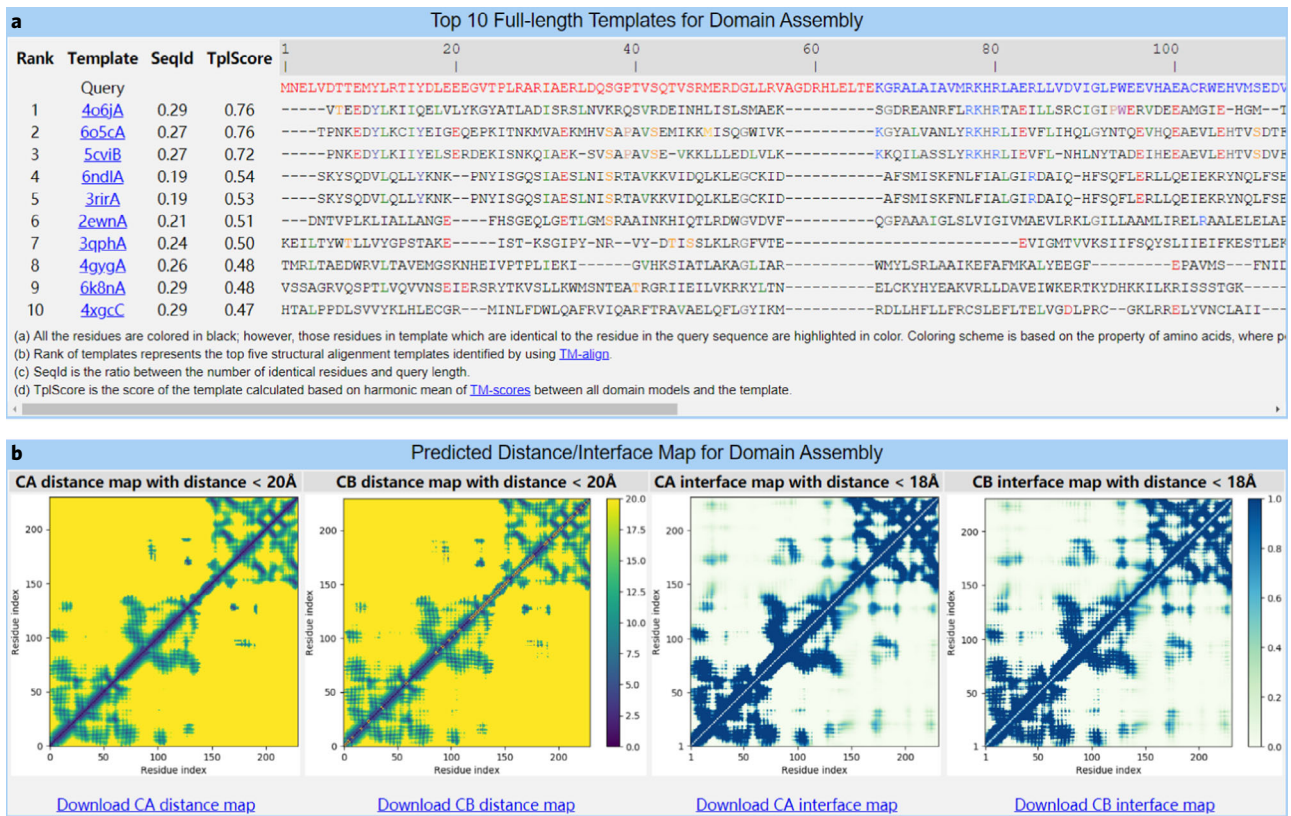


Fig. 5 | Example of the I-TASSER-MTD results page (Sections 7 and 8). **a**, Top ten full-length templates identified by the global structural alignment, which are used to guide the domain model assembly. **b**, Predicted residue-residue distance maps and domain-domain interface maps for domain assembly.

with C_{α} and C_{β} distances <20 Å, respectively. The subsequent two columns show the interface maps with C_{α} and C_{β} distances <18 Å predicted by deep learning, where all residue pairs are included in the map. In these maps, only the inter-domain distances and inter-domain interface probabilities are employed to guide the domain model assembly.

? TROUBLESHOOTING

- 37 Click the link below the figure to download the corresponding predicted distance map for further analysis. For example, click the link labeled 'Download CA distance map' to download the predicted distance map with C_{α} distances <20 Å.

Analogous proteins of the predicted full-length model ● Timing 5 min

- 38 View the structurally analogous proteins of the predicted top full-length model, which is identified from the PDB by TM-align structural alignment (Fig. 6a). The table to the right of the model summarizes the analogous proteins, where the proteins are ranked according to the TM-score (shown in the fourth column) between the predicted full-length model and the TM-align detected templates. The analogous proteins with a TM-score >0.5 can be used to determine the structural family of the query protein¹⁰⁰.
- 39 View the 'RMSD', 'IDEN' and 'Cov.' columns in the table to analyze the RMSD, sequence identity and coverage of the aligned regions determined by TM-align, which indicates the conservation of spatial motifs in the model and the structurally analogous proteins.
▲ CRITICAL STEP Users are recommended to read the explanation of each criterion, which is provided below the table.
- 40 Click the link in the last column of 'Download Alignment' to download the PDB file with the top model structurally aligned to the corresponding analogous protein.

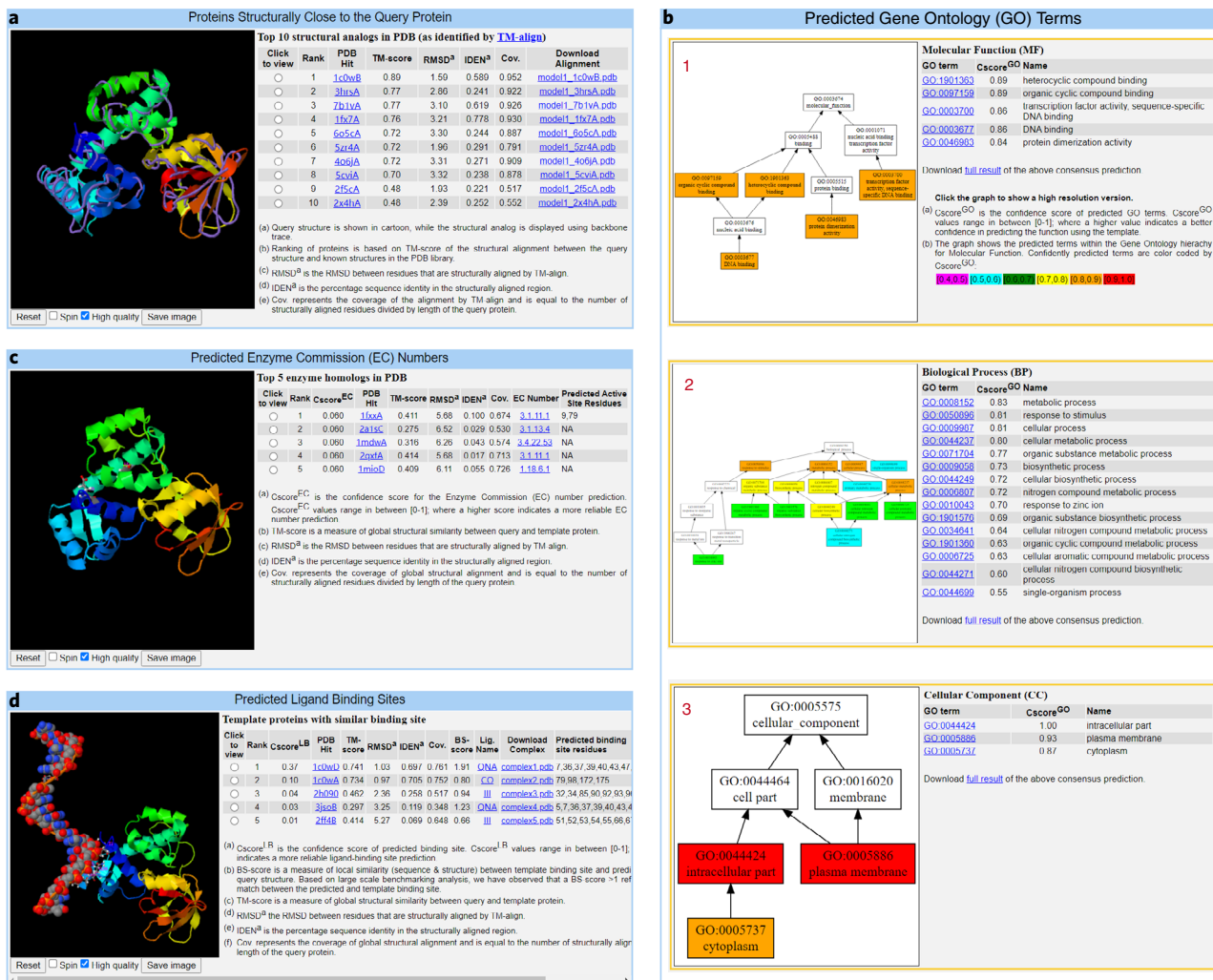


Fig. 6 | Example of the I-TASSER-MTD results page (Sections 9-12). **a**, Top ten analogous structures that are structurally close to the top model of the query protein. **b**, Results of the predicted GO terms including MF (1), BP (2) and CC (3). **c**, Results of the predicted EC numbers from the top five homologous enzyme templates. **d**, Results of the predicted ligand-binding site from the top five homologous templates.

GO term prediction ● Timing 5 min

41 View the GO terms predicted according to the I-TASSER-MTD full-length model, full-chain sequence and PPI networks (Fig. 6b). The predicted results include three subsections: molecular function (MF) (Figs. 6b-1), biological process (BP) (Figs. 6b-2) and cellular component (CC) (Figs. 6b-3).

42 For each GO aspect, view the directed acyclic graph by Graphviz¹⁰¹ shown in the left panel of the corresponding subsection, which is plotted by combining the predicted GO terms and their parent terms. On the graph, the predicted GO terms are colored by the confidence score Cscore^{GO} (Supplementary Note 6), where the color code for the range of Cscore^{GO} is illustrated in the bottom of the right panel. The Cscore^{GO} values range between 0 and 1, and a higher value usually indicates a better confidence in predicting the function using the template. The precision of each range of Cscore^{GO} is shown in Supplementary Fig. 16a. For example, BP has a 71% probability of being correctly predicted if Cscore^{GO} >0.9.

43 View the table of the summary of predicted results shown on the right of the graph. The three columns of the table show the predicted GO terms, the confidence score and the corresponding common name of the GO term. Click on each GO term in the first column to

visit the Amigo website (<http://amigo.geneontology.org/amigo>) to analyze the definition and lineage of the term.

- 44 Click on the link 'full result' below the table to download the results presented in the table.

EC number prediction ● Timing 5 min

- 45 Scroll down further to analyze the top five EC number predictions (Fig. 6c). The left panel shows the first I-TASSER-MTD model, as well as the predicted active sites in case that they exist, in an accompanying JSmol applet.
- 46 View the table in the right panel, which summarizes the results of the predictions. In the table, the EC predictions are ranked on the basis of the $Cscore^{EC}$ (confidence score of EC, Supplementary Note 6) reported in the third column. Here, $Cscore^{EC}$ values range between 0 and 1, where a higher score indicates a more reliable EC number prediction. The relationship between the precision and $Cscore^{EC}$ is reported in Supplementary Fig. 16c. For example, the $Cscore^{EC}$ in the range of [0.6, 0.7] indicates that the first digit, first two digits, first three digits and all four digits of the EC number have probabilities of 76%, 66%, 63% and 44% of being correctly predicted, respectively. Users can click the circle in the first column to see the corresponding predicted active sites in the left JSmol panel.
- 47 View the TM-score between the predicted model and the enzyme analogs (column 'PDB Hit'), RMSD of aligned regions, sequence identity, coverage of aligned regions, predicted EC number and predicted active sites in the table.
▲ CRITICAL STEP Although the $Cscore^{EC}$ is used to rank the predicted EC number, users are advised to consult both the $Cscore^{EC}$ and the TM-score. For example, if most of the identified functional analogs with similar folds (i.e., TM-score >0.5) have the same EC number digits and the $Cscore^{EC}$ is relatively high, the likelihood of the prediction being correct is very high.
- 48 Click on the EC numbers to visit the ExpASY enzyme database (<https://enzyme.expasy.org/>) to further learn about the enzyme families, such as reactions catalyzed by the enzyme, the cofactors required and the metabolic pathway in which they function.

Ligand-binding site prediction ● Timing 5 min

- 49 In case that there exist predicted ligand-binding sites, view the last section of the results page to analyze the best predicted ligand-binding sites in the predicted full-length model (Fig. 6d). The full-length model, together with the predicted binding site residues and the corresponding ligand, is displayed in the left panel using the JSmol applet. The binding site residues of the query protein are highlighted as 'ball and stick' with the corresponding residue numbers labeled in magenta, whereas ligand atoms are represented by colored spheres.
- 50 View the table reporting the top identified functional analogs and the derived binding site residues in the right panel. The predicted binding site residues are ranked according to the confidence score $Cscore^{LB}$ (confidence score of ligand-binding site, Supplementary Note 6) listed in the third column of the table. The $Cscore^{LB}$ value ranges from 0 to 1, with a higher value corresponding to a more reliable ligand-binding site. The precision versus $Cscore^{LB}$ is reported in Supplementary Fig. 16b to show how reliable the results are. For example, $Cscore^{LB} > 0.7$ indicates the ligand-binding site has >82% probability of being correctly predicted. Users can click on the PDB links to track the bound ligands in the functional analogs.
- 51 View the TM-score, RMSD, sequence identity and coverage of the aligned regions between the query model and the identified functional analogs, as well as the binding site score and predicted binding site residues in the table. The binding site score is a measure of local similarity calculated according to the local structural similarity of the ligand-binding sites and the sequence identity between the I-TASSER-MTD model and the structural analogs. A binding site score >1 indicates an outstanding local match between the predicted and template binding sites based on a large-scale benchmark study^{45,70}.
- 52 Download the ligand–protein complex. Users can click on the link listed in the column 'Download Complex' to download the PDB file containing the predicted model and the bound ligand. The file can be viewed interactively using any molecular visualization tool. Rendering the ligand as colored spheres and depicting predicted binding site residues as 'ball and stick' will aid in visualizing the binding site cleft and help with analyzing the ligand–protein interactions.

Troubleshooting

Troubleshooting advice can be found in Table 2.

Step	Error message or problem	Possible reason	Solution
5	Error! Some residues are missed in the domain definition	Some residue indices are not included in the domain definition	Check the domain definition to make sure that the domain definition includes all residue indices
	Error! Residue overlaps in the domain definition	Some residue indices are included in multiple domain ranges	Check the domain definition to make sure each residue index is included in only one domain
	Error! Wrong format of the domain definition	The domain definition cannot be recognized by the server	Read the instructions on how to correct the format of the domain definition or follow the formatting shown in the example input
	Error! Too many domains in the domain definition	Maximum number of domains accepted by the server is 20	Merge short domains in the domain definition or model the first 20 domains and the rest of the domains independently using the server, and assemble them using DEMO
	Error! Too large domain in the domain definition	Maximum length of the domain accepted by the server is 1,000	Further split large domains into multiple parts
	Error! Too short domain in the domain definition	Minimum length of the domain accepted by the server is 30	Merge short domains with other domains
6	Error! Incorrect template name	The template name cannot be recognized by the server	Rename the template according to the instructions
	Error! Wrong tarball format (XXX) of the templates	The tarball format is not supported by the server	Repackage the templates as *.tar.bz2, *.tar.gz, *.tar or *.zip
9	Error! Wrong cross-linking data	The cross-linking data cannot be read by the server	Check the instructions or the example input to ensure the cross-linking data is in the correct format. Ensure that the residue index is within the sequence length range
	Error! Wrong density map file	The cryo-EM density map data cannot be read by the server	Make sure that the density map is in the 8-bit, 16-bit, or 32-bit MRC and CCP4 format, and fix the error according to the instructions
10	The protein sequence is too short or too long	The range of sequence length is not within [30, 2000]	Check the sequence to make sure that the length is within [30, 2000]. Users who want to model larger proteins with sequence lengths >2,000 residues should specify their own domain boundary
15	Low eTM-score full-length model	Low-quality individual domain models, full-length templates or distance map	As described in the 'Limitations' section, users are advised to seek other sources of structural information, such as experimental data, full-length templates and contact/distance restraints predicted by other tools
35	No predicted distance maps	The sequence length is large, which causes a larger distance map than can be predicted by the server owing to RAM limitations	Users can download the standalone package and predict the distance locally

Timing

Steps 1–10, query sequence submission: 5 min
 Steps 11–12, job monitoring: 6–12 h
 Step 13, query sequence and predicted domain definition: 3 min
 Steps 14–20, predicted full-length structures: 10 min
 Steps 21–24, individual domain modeling: 5 min
 Step 25, secondary structure prediction: 2 min
 Step 26, solvent accessibility prediction: 2 min
 Steps 27–30, domain boundary prediction: 5 min

Steps 31–35, full-length templates for domain assembly: 5 min
Steps 36–37, predicted distance/interface map for domain assembly: 5 min
Steps 38–40, analogous proteins of the predicted full-length model: 5 min
Steps 41–44, GO term prediction: 5 min
Steps 45–48, EC number prediction: 5 min
Steps 49–52, ligand-binding site prediction: 5 min

Prediction of the 3D structure and function for a medium-size multi-domain protein (~300–500 residues) requires 6–10 h using the I-TASSER-MTD server. Longer proteins with more domains will take more time as the server independently models the structure of each domain, and the domain assembly can be performed only when all individual domain models are generated. The actual processing time, however, also depends on the number of jobs in the queue. Currently, the I-TASSER-MTD server is working on a supercomputer cluster that consists of 59 20-core IBM NeXTScale nx360 M4 nodes running at 2.8 GHz, and users can receive the results within 1–3 d in most cases.

Anticipated results

Here we used the iron-dependent regulator from *Mycobacterium tuberculosis* (PDB ID: 1fx7A) as an example to predict the structure and function of the protein using the I-TASSER-MTD server. In Step 8, we chose the ‘Option IV’ as ‘YES’ to predict the function of all individual domains and the full-length protein, while other options were kept as default. Once the job is finished, the user will receive an email containing a link to the results page. Clicking on the link will open the results page containing the following results, which are shown in Figs. 3–6:

- The title of the results page: I-TASSER-MTD results for job *jobid* (see label 1 in Fig. 3a).
- The link to download the results files. Click on the link to download a compressed file containing all results shown on the results page (see label 2 in Fig. 3a).
- The user submitted sequence in FASTA format with predicted domains marked by different colors (Fig. 3b). Users can confirm the sequence length shown in the parentheses after the query name.
- The top five full-length models predicted by I-TASSER-MTD (Fig. 3c). The left panel shows the predicted model using the JSmol applet, while the right panel summarizes the information of the predicted models.
- The predicted individual domain models (Fig. 3d). Each domain model is shown in an independent JSmol applet, with a link to download the model, and the predicted functions are given below the model. The predicted functions for each individual domain are shown in Supplementary Figs. 17–19.
- The predicted secondary structure of the full-length sequence (Fig. 4a). The results contain three rows. The first row shows the query sequence, while the second and the third row report the predicted secondary structure and the confidence score, respectively.
- The predicted solvent accessibility of the full-length sequence, which contains two rows (Fig. 4b). The first row and the second row display the query sequence and the confidence score for the predicted solvent accessibility, respectively.
- The predicted domain boundary (Fig. 4c). This section contains the deep-learning-predicted contact map for the domain boundary prediction, the FU-score curve for the continuous domain detection, the FU-score heatmap for the discontinuous domain detection and the predicted domain definition.
- The top ten full-length templates identified by TM-align structural alignment (Fig. 5a). This section reports the sequence identity, template score and the alignment between the query and the template.
- The distance/interface map predicted by deep learning (Fig. 5b). The C_{α}/C_{β} distance maps with distances <20 Å, the C_{α}/C_{β} interface maps with distances <18 Å and the corresponding link to download the distance/interface maps are shown in this section.
- The top ten proteins structurally close to the query protein (Fig. 6a). The predicted full-length models with the aligned analogous models are depicted in the left panel using the JSmol applet, while the information for the structural analogs is shown in the right panel.
- The predicted GO terms (Fig. 6b). This section contains the predicted MF, BP and CC. For each section, the predicted GO terms are plotted using a directed acyclic graph displayed in the left panel, while a table with the predicted GO terms is shown on the right panel.
- The predicted EC numbers (Fig. 6c). The JSmol panel on the left shows the full-length model and the predicted active site residues. The table on the right summarizes the predicted active sites.
- The predicted ligand-binding sites (Fig. 6d). The full-length model, predicted binding sites and corresponding ligands are displayed on the left panel using the JSmol applet, and a summary of the predicted ligand-binding sites is reported in a table on the left panel.

It should be noted that only the domain definition will be shown in the predicted domain boundary section if the user provides the domain definition. In addition, the link to the predicted functions of each individual domain will not be generated if the query protein is predicted as a single-domain protein or the user did not choose the option for function prediction. Proteins structurally close to the query protein, predicted GO terms, EC numbers and ligand-binding sites for the full-length protein will not be reported in the results page if the user kept the default option to not include protein function prediction.

Data availability

The raw data and example files are available at <https://zhanggroup.org/I-TASSER-MTD/> or from the corresponding author upon reasonable request.

Code availability

The I-TASSER-MTD standalone package is freely available for academic use at <https://zhanggroup.org/I-TASSER-MTD/>.

References

1. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
2. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209–225 (1997).
3. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).
4. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
5. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl Acad. Sci. USA* **106**, 67–72 (2009).
6. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
7. Mortuza, S. et al. Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nat. Commun.* **12**, 5011 (2021).
8. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
9. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
10. 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).
11. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
12. Kryzhanovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins* **89**, 1607–1617 (2021).
13. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300**, 1701–1703 (2003).
14. Apic, G., Huber, W. & Teichmann, S. A. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics* **4**, 67–78 (2003).
15. Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. J. N. R. M. C. B. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319 (2007).
16. Zhou, X. G., Hu, J., Zhang, C. X., Zhang, G. J. & Zhang, Y. Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl Acad. Sci. USA* **116**, 15930–15938 (2019).
17. Xu, D., Jaroszewski, L., Li, Z. & Godzik, A. AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction. *Bioinformatics* **31**, 2098–2105 (2015).
18. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
19. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
20. Xue, Z., Xu, D., Wang, Y. & Zhang, Y. ThreaDom: extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **29**, i247–i256 (2013).
21. Hong, S. H., Joo, K. & Lee, J. ConDo: protein domain boundary prediction using coevolutionary information. *Bioinformatics* **35**, 2411–2417 (2019).
22. Zheng, W. et al. FUpred: detecting protein domains through deep-learning based contact map prediction. *Bioinformatics* **36**, 3749–3757 (2020).
23. Wollacott, A. M., Zanghellini, A., Murphy, P. & Baker, D. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci.* **16**, 165–175 (2007).

24. Zhang, C., Zheng, W., Freddolino, P. L. & Zhang, Y. MetaGO: predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping. *J. Mol. Biol.* **430**, 2256–2265 (2018).
25. Yao, S. et al. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res.* **49**, W469–W475 (2021).
26. Piovesan, D. & Tosatto, S. C. INGA 2.0: improving protein function prediction for the dark proteome. *Nucleic Acids Res.* **47**, W373–W378 (2019).
27. Koo, D. C. E. & Bonneau, R. Towards region-specific propagation of protein functions. *Bioinformatics* **35**, 1737–1744 (2019).
28. Gligorijević, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 1–14 (2021).
29. Pearce, R. & Zhang, Y. Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* **297**, 100870 (2021).
30. Zheng, W. et al. Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins* **89**, 1734–1751 (2021).
31. Zheng, W. et al. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164 (2019).
32. Battey, J. N. et al. Automated server predictions in CASP7. *Proteins* **69** (Suppl.), 68–82 (2007).
33. Croll, T. I., Sammito, M. D., Kryshchuk, A. & Read, R. J. Evaluation of template-based modeling in CASP13. *Proteins* **87**, 1113–1127 (2019).
34. Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* **69** (Suppl.), 108–117 (2007).
35. Zhang, Y. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins* **77** (Suppl.), 100–113 (2009).
36. Xu, D., Zhang, J., Roy, A. & Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* **79** (Suppl.), 147–160 (2011).
37. Zhang, Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* **82** (Suppl.), 175–187 (2014).
38. Zhang, W. et al. Integration of QUARK and I-TASSER for ab initio protein structure prediction in CASP11. *Proteins* **84** (Suppl.), 76–86 (2016).
39. Zhang, C., Mortuza, S. M., He, B., Wang, Y. & Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins* **86** (Suppl.), 136–151 (2018).
40. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
41. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).
42. Zheng, W. et al. LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res.* **47**, W429–W436 (2019).
43. Wang, Y. et al. ThreaDomEx: a unified platform for predicting continuous and discontinuous protein domains by multiple-threading and segment assembly. *Nucleic Acids Res.* **45**, W400–W407 (2017).
44. Li, Y. et al. Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins* **89**, 1911–1921 (2021).
45. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res.* **45**, W291–W299 (2017).
46. Sillitoe, I. et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
47. Xu, Y., Xu, D. & Gabow, H. N. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics* **16**, 1091–1104 (2000).
48. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).
49. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manifold. *Nat. Methods* **16**, 603–606 (2019).
50. Mitchell, A. L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**, D570–D578 (2020).
51. Chen, I.-M. A. et al. The IMG/M data management and analysis system v. 6.0: new tools and advanced capabilities. *Nucleic Acids Res.* **49**, D751–D763 (2021).
52. Mirdita, M. et al. UniClust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
53. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
54. 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).
55. 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. *Sci. Rep.* **3**, 1–9 (2013).

56. Ekeberg, M., Hartonen, T. & Aurell, E. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences. *J. Comput. Phys.* **276**, 341–356 (2014).
57. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496–1503 (2020).
58. Thrun, S. in *Advances in Neural Information Processing Systems* 640–646 (Morgan Kaufmann Publishers, 1996).
59. Zheng, W., Zhang, C., Bell, E. W. & Zhang, Y. I-TASSER gateway: a protein structure and function prediction server powered by XSEDE. *Future Gener. Comput. Syst.* **99**, 73–85 (2019).
60. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinforma.* **9**, 40 (2008).
61. Zheng, W. et al. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep. Methods* **1**, 100014 (2021).
62. 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* **87**, 1082–1091 (2019).
63. Li, Y. et al. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLOS Comput. Biol.* **17**, e1008865 (2021).
64. 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).
65. Zhang, Y. & Skolnick, J. SPICKER: a clustering approach to identify near-native protein folds. *J. Comput. Chem.* **25**, 865–871 (2004).
66. Huang, X., Pearce, R. & Zhang, Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **36**, 3758–3765 (2020).
67. Zhang, J., Liang, Y. & Zhang, Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* **19**, 1784–1795 (2011).
68. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
69. Ramachandran, G. T. & Sasisekharan, V. in *Advances in Protein Chemistry*, 23 283–437 (Elsevier, 1968).
70. Roy, A., Yang, J. & Zhang, Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res.* **40**, W471–W477 (2012).
71. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103 (2012).
72. Zhou, X. G., Peng, C. X., Liu, J., Zhang, Y. & Zhang, G. J. Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. *IEEE Trans. Evol. Comput.* **24**, 536–550 (2020).
73. Zhou, X. G. & Zhang, G. J. Abstract convex underestimation assisted multistage differential evolution. *IEEE Trans. Cybern.* **47**, 2730–2741 (2017).
74. Zhou, X. G. & Zhang, G. J. Differential evolution with underestimation-based multimutation strategy. *IEEE Trans. Cybern.* **49**, 1353–1364 (2018).
75. Yang, J., Wang, Y. & Zhang, Y. ResQ: an approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *J. Mol. Biol.* **428**, 693–701 (2016).
76. Glaeser, R. M. How good can cryo-EM become? *Nat. Methods* **13**, 28–32 (2016).
77. Zhou, X. G. et al. Progressive assembly of multi-domain protein structures from cryo-EM density maps. *Nat. Comput. Sci.* **2**, 265–275 (2022).
78. Mistry, J. et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
79. Lu, S. et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).
80. Eickholt, J., Deng, X. & Cheng, J. DoBo: protein domain boundary prediction by integrating evolutionary signals and machine learning. *BMC Bioinforma.* **12**, 1–8 (2011).
81. Tai, C. H., Lee, W. J., Vincent, J. J. & Lee, B. Evaluation of domain prediction in CASP6. *Proteins* **61**, 183–192 (2005).
82. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
83. Pearce, R. & Zhang, Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol.* **68**, 194–207 (2021).
84. Born, A., Henen, M. A. & Vögeli, B. Activity and affinity of Pin1 variants. *Molecules* **25**, 36 (2020).
85. Born, A. et al. Reconstruction of coupled intra- and interdomain protein motion from nuclear and electron magnetic resonance. *J. Am. Chem. Soc.* **143**, 16055–16067 (2021).
86. Chandonia, J.-M., Fox, N. K. & Brenner, S. E. SCOPe: manual curation and artifact removal in the structural classification of proteins—extended database. *J. Mol. Biol.* **429**, 348–355 (2017).
87. Lam, S. D. et al. Gene3D: expanding the utility of domain assignments. *Nucleic Acids Res.* **44**, D404–D409 (2015).
88. Yu, L. et al. Grammar of protein domain architectures. *Proc. Natl Acad. Sci. USA* **116**, 3636–3645 (2019).
89. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826 (1986).
90. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D. Biol. Crystallogr.* **66**, 486–501 (2010).
91. DiMaio, F. et al. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods* **12**, 361–365 (2015).

92. Zhang, C. et al. Functions of essential genes and a scale-free protein interaction network revealed by structure-based function and interaction prediction for a minimal genome. *J. Proteome Res.* **20**, 1178–1189 (2021).
93. Zhang, C., Wei, X., Omenn, G. S. & Zhang, Y. Structure and protein interaction-based gene ontology annotations reveal likely functions of uncharacterized proteins on human chromosome 17. *J. Proteome Res.* **17**, 4186–4196 (2018).
94. Zhang, C., Lane, L., Omenn, G. S. & Zhang, Y. Blinded testing of function annotation for uPE1 proteins by I-TASSER/COFACTOR pipeline using the 2018–2019 additions to neXtProt and the CAFA3 challenge. *J. Proteome Res.* **18**, 4154–4166 (2019).
95. Iyer, S., Subramanian, V. & Acharya, K. R. C9orf72, a protein associated with amyotrophic lateral sclerosis (ALS) is a guanine nucleotide exchange factor. *PeerJ* **6**, e5815 (2018).
96. Skotnicová, P. et al. The cyanobacterial protoporphyrinogen oxidase HemJ is a new b-type heme protein functionally coupled with coproporphyrinogen III oxidase. *J. Biol. Chem.* **293**, 12394–12404 (2018).
97. Hanson, R. M., Prilusky, J., Renjian, Z., Nakane, T. & Sussman, J. L. JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia. *Isr. J. Chem.* **53**, 207–216 (2013).
98. Hiranuma, N. et al. Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **12**, 1–11 (2021).
99. Guo, S.-S., Liu, J., Zhou, X. & Zhang, G. DeepUMQA: ultrafast shape recognition-based protein model quality assessment using deep learning. *Bioinformatics* **38**, 1895–1903 (2022).
100. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
101. Ellson, J., Gansner, E.R., Koutsofios, E., North, S.C. & Woodhull, G. in *Graph Drawing Software* 127–148 (Springer, 2004).
102. Towns, J. et al. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **16**, 62–74 (2014).
103. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl Acad. Sci. USA* **116**, 16856–16865 (2019).
104. Källberg, M. et al. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **7**, 1511–1522 (2012).
105. Du, Z. et al. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **16**, 5634–5651 (2021).
106. Lobley, A., Sadowski, M. I. & Jones, D. T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **25**, 1761–1767 (2009).
107. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
108. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

Acknowledgements

This work is supported in part by the National Institute of General Medical Sciences (GM136422 and S10OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.), the National Science Foundation (IIS1901191 and DBI2030790 to Y.Z.), the National Nature Science Foundation of China (62173304 and 61773346 to G.Z.), the ‘New Generation Artificial Intelligence’ major project of Science and Technology Innovation 2030 of the Ministry of Science and Technology of China (2021ZD0150100 to G.Z.) and the Key Project of Zhejiang Provincial Natural Science Foundation of China (LZ20F030002 to G.Z.). This work used the Extreme Science and Engineering Discovery Environment (XSEDE)¹⁰², which is supported by the National Science Foundation (ACI1548562).

Author contributions

Y.Z. conceived and designed the project. X.Z. developed the pipeline and performed the test. W.Z. developed the method for domain boundaries prediction. Y.L. developed the method for contacts and distances prediction. C.Z. developed the method for protein function prediction. Y.Z., W.Z., Y.L., C.Z. and R.P. developed the method for individual domain modeling. X.Z. developed the method for multi-domain protein structure assembly. X.Z. and E.B. tested the server. G.Z. helped supervise the research. X.Z. and Y.Z. wrote the manuscript, and all authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41596-022-00728-0>.

Correspondence and requests for materials should be addressed to Yang Zhang.

Peer review information *Nature Protocols* thanks Ruben Sánchez-García, Beat R. Vogeli and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Received: 4 February 2022; Accepted: 24 May 2022;
Published online: 5 August 2022

Related links

Key references using this protocol

Zhou, X. et al. *Proc. Natl Acad. Sci. USA* **116**, 15930–15938 (2019): <https://doi.org/10.1073/pnas.1905068116>

Zhang, C. et al. *Nucleic Acids Res.* **45**, W291–299 (2017): <https://doi.org/10.1093/nar/gkx366>

Zheng, W. et al. *Cell Rep. Methods* **1**, 100014 (2021): <https://doi.org/10.1016/j.crmeth.2021.100014>

Hermes, C. et al. *Nat. Commun.* **12**, 144 (2021): <https://doi.org/10.1038/s41467-020-20418-3>