Supplemental Information for

Folding non-homology proteins by coupling deep-learning contacts with I-TASSER assembly simulations

Wei Zheng, Chengxin Zhang, Yang Li, Robin Pearce, Eric Bell, Yang Zhang

Supplementary Figures

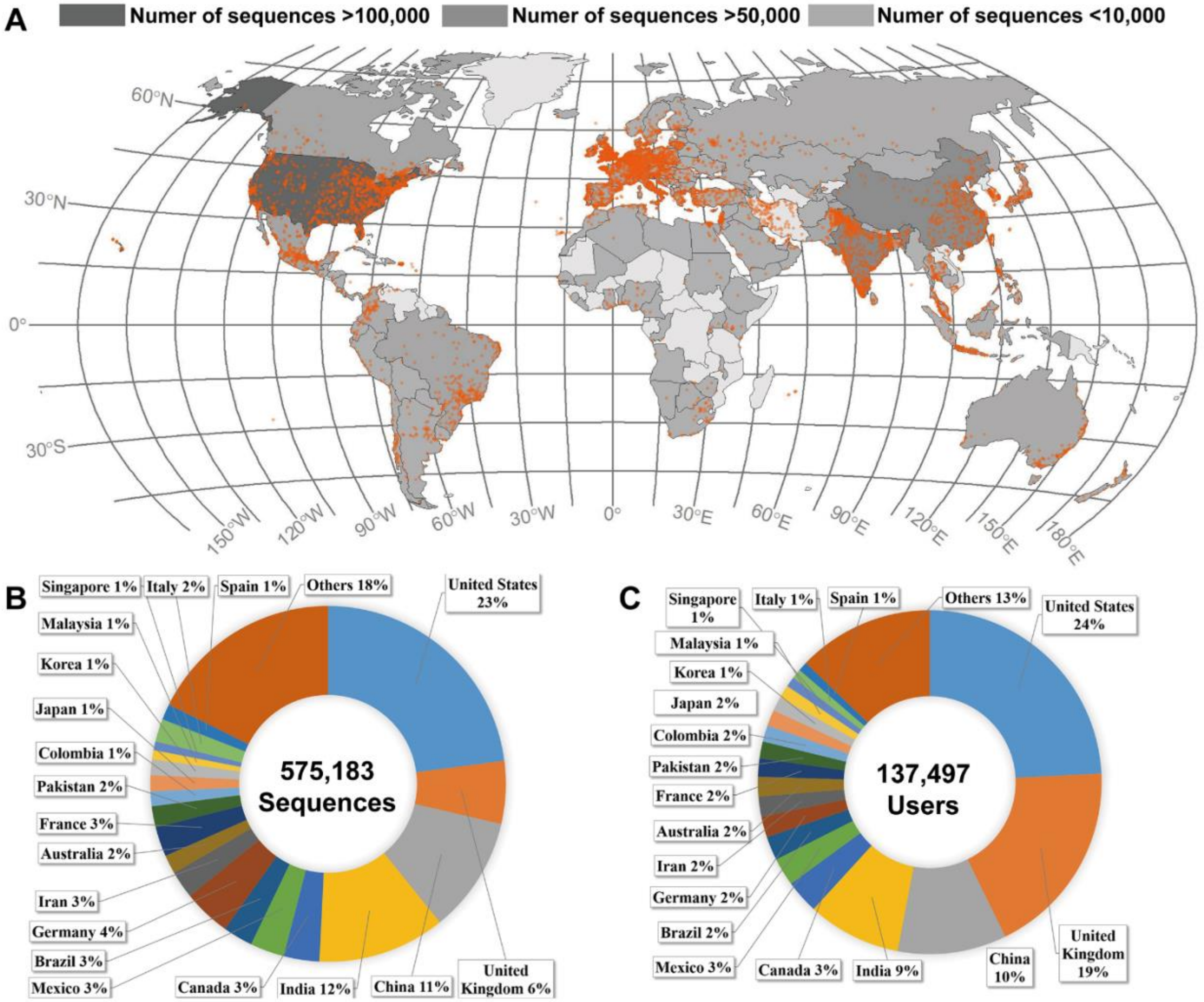

Figure S1. Geographical distribution of I-TASSER server users, Related to the STAR Methods section "Methods Summary". Overall, the I-TASSER server has completed predictions for 575,183 proteins submitted by 137,497 users from 149 countries or regions until Octobor, 2020. (A) Geographical distribution of I-TASSER server usage. In the world map, different countries are colored from dark to light gray in descending order of the number of sequences submitted to the I-TASSER server. Different cities are marked by orange points, whose size is proportional to the number of registered I-TASSER users in these cities. (B) The pie chart for the percentage of the number of sequences submitted to I-TASSER by different countries among all submitted sequences. (C) The pie chart for the percentage of the number of registered I-TASSER users in different countries among all registered users.

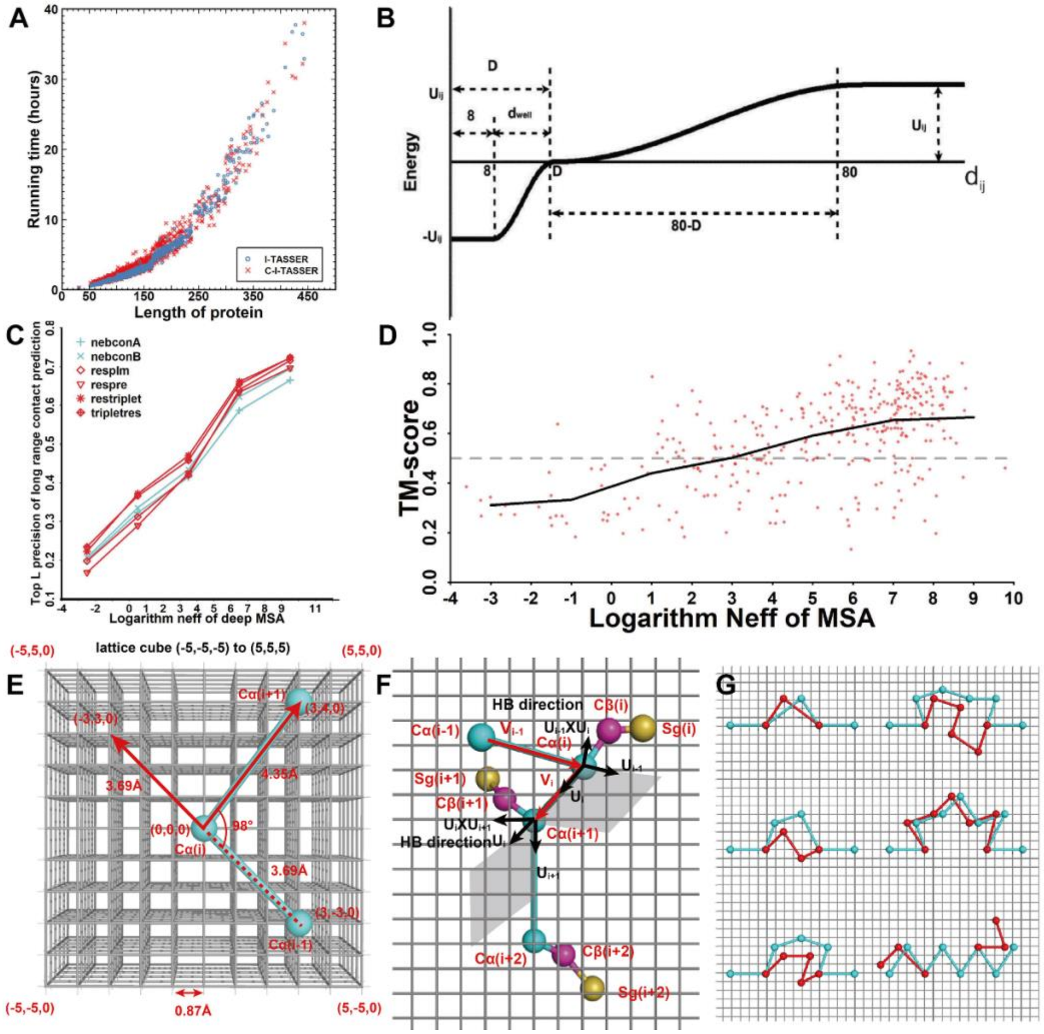

Figure S2. The Running time, MSA analysis and simulation illustration for C-I-TASSER, Related to the STAR Methods section "Replica-exchange Monte Carlo in C-I-TASSER". (A) The time complexity comparison between C-I-TASSER and I-TASSER. (B) Illustration of the sequence-based contact restraint. (C) The top *L* precision of long-range contact prediction for the 6 methods used in C-I-TASSER for MSAs with different logarithm *Neff* values at a base of 2. The 4 contact predictors colored in red utilize deep learning; and the 2 colored in cyan are meta-approaches. (D) TM-scores of the C-I-TASSER models for MSAs with different logarithm Neff values using a base of 2. The black line represents the average TM-scores under each logarithm Neff bin with a bin width of 2. Illustration of modeling and simulation setting in C-I-TASSER. (E) Reduced representation of an amino acid by a three-dimensional underlying cubic lattice system with a lattice grid of 0.87 Å. Only the alpha carbon (C_α) atom of each residue is treated explicitly. Considering the C_α of the *i*-th residue, $C_\alpha(i)$, the lattice cube is from (-5,-5,-5) to (5,5,5). The $C_\alpha(i)$ is located at (0,0,0). The C_α of the previous (*i*-1)-th residue, $C_\alpha(i\text{-}1)$ is located at (3,-3,0) and the C_α-C_α bond length between $C_\alpha(i\text{-}1)$ and $C_\alpha(i)$ is 3.69 Å. The C_α of the next (*i*+1)-th residue, $C_\alpha(i+1)$, is located at (3,4,0) and the C_α-C_α bond length between $C_\alpha(i+1)$ and $C_\alpha(i)$ is 4.35 Å. Additionally, the C_α-C_α bond angle is 98°. (F) Determination of the positions for the C_β and center of side-group heavy atoms. The positions of three consecutive C_α atoms are used to define a local coordinate system for the determination of the beta carbon (C_β) (except glycine), and the center of side-group heavy atoms (SG) (except glycine and alanine). $\overrightarrow{V_{\iota-1}}$ is the vector from $C_\alpha(i\text{-}1)$ to $C_\alpha(i)$, and $\overrightarrow{U_{\iota-1}}$ is the unit vector for $\overrightarrow{V_{\iota-1}}$. The cross product of $\overrightarrow{U_{\iota-1}}$ and $\overrightarrow{U_\iota}$, $\overrightarrow{U_{\iota-1}} \times \overrightarrow{U_\iota}$, is the direction of the hydrogen bond (HB). (G) Conformational movements in the C-I-TASSER Monte Carlo simulations. The cyan and red lines are the C_α traces before and after the movements, respectively. There are 6 types of conformational movements in the C-I-TASSER simulations: (1) 2-bond vector walk; (2) 3-bond vector walk; (3) 4- bond vector walk; (4) 5-bond vector walk; (5) 6-bond vector walk; (6) N- or C-terminal random walk.

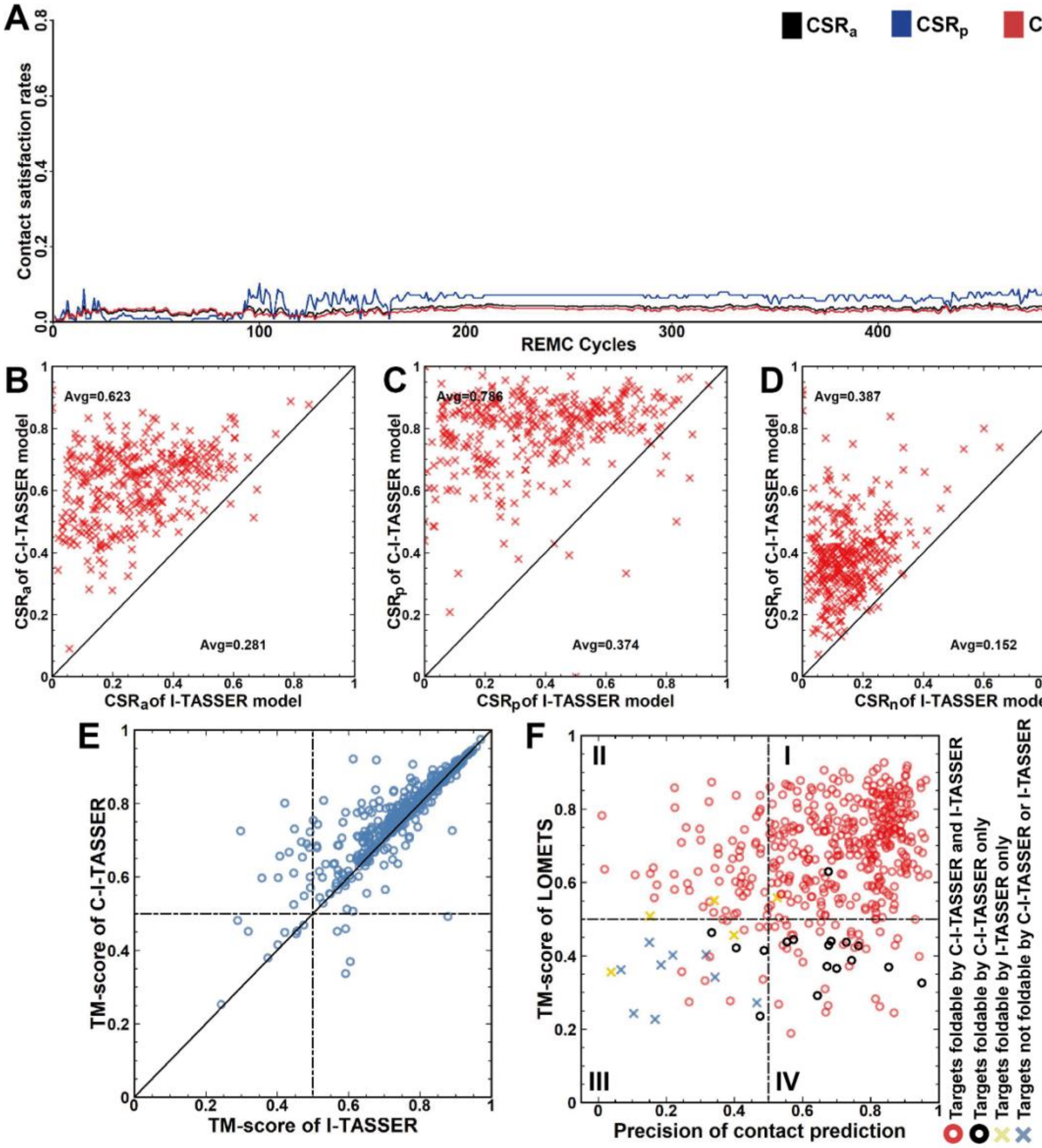

Figure S3. The simulation analysis and the performance on Easy targets for C-I-TASSER, Related to Figure 2 and Figure 3. (A) The trajectories of contact satisfaction rate (CSR) of the I-TASSER folding simulations on 4v00. The same scale as Figure 3B is used here. The comparison of contact satisfaction rate (CSR) in the final models by I-TASSER and C-I-TASSER. (B) CSR_a; (C) CSR_p; (D) CSR_n. C-I-TASSER modeling results on the 455 Easy targets in the benchmark dataset. (E) Comparison between TM-scores of the first models built by C-I-TASSER and I-TASSER for different target types on the 455 Easy target proteins. The blue circles represent Easy targets. (F) Impact of threading alignments and contact-map predictions on fold results for 455 Easy targets. Four regions are depicted based on whether or not the threading templates were good (TM-score ≥0.5) or the predicted contacts were accurate (Precision ≥0.5). The red circles denote the targets that can be folded by both C-I-TASSER and I-TASSER with a TM-score ≥0.5; the black points are the targets that can only be folded by C-I-TASSER and not I-TASSER; the yellow crosses are the targets that can only be folded by I-TASSER and not C-I-TASSER; the blue crosses indicate the targets that cannot be folded by either C-I-TASSER or I-TASSER.

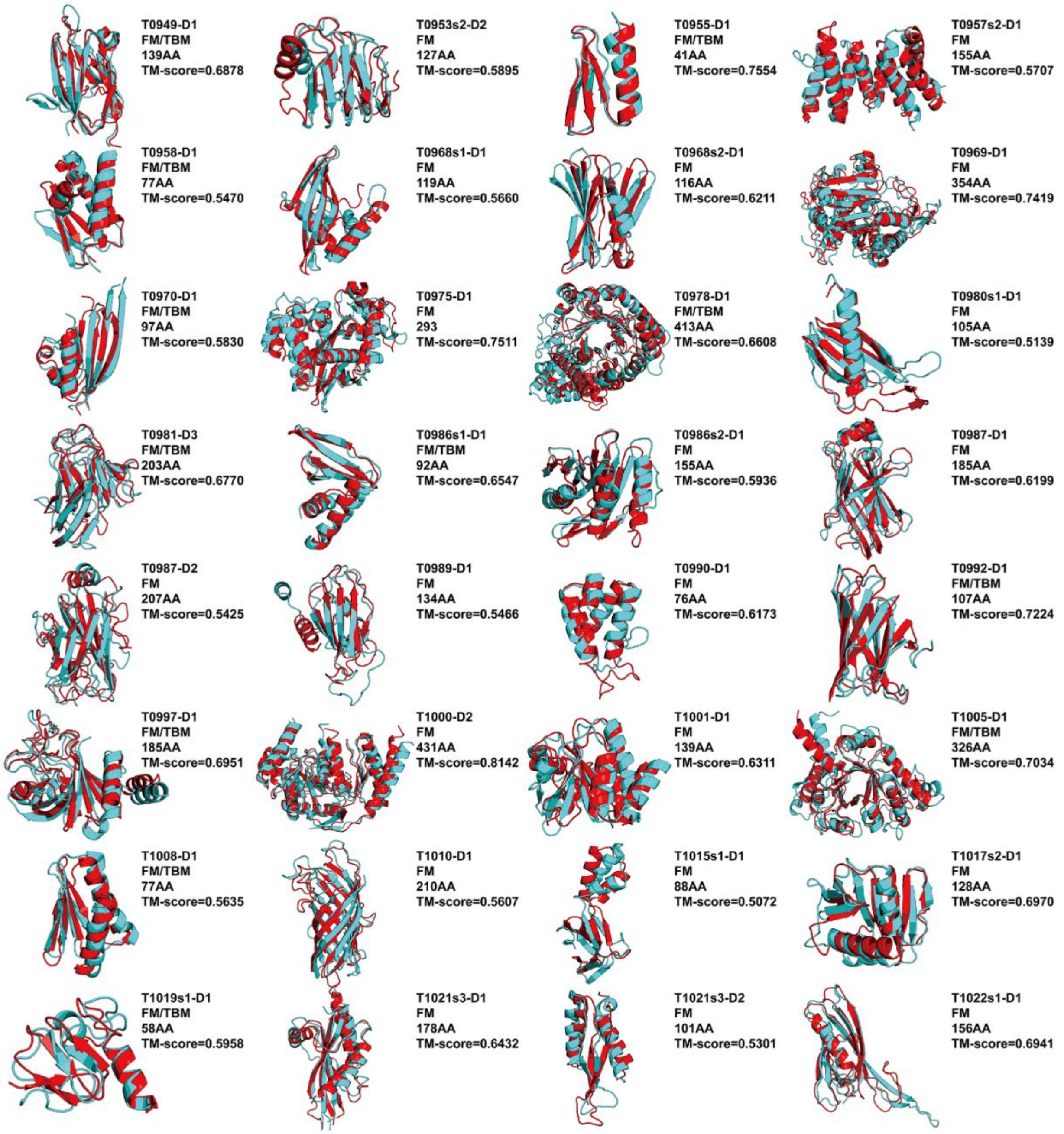

Figure S4. The 32 representative targets in CASP13 for which C-I-TASSER generated high-quality models, Related to the STAR Methods section "Methods Summary". The C-I-TASSER models are colored in red, while the experimental structures are in cyan.

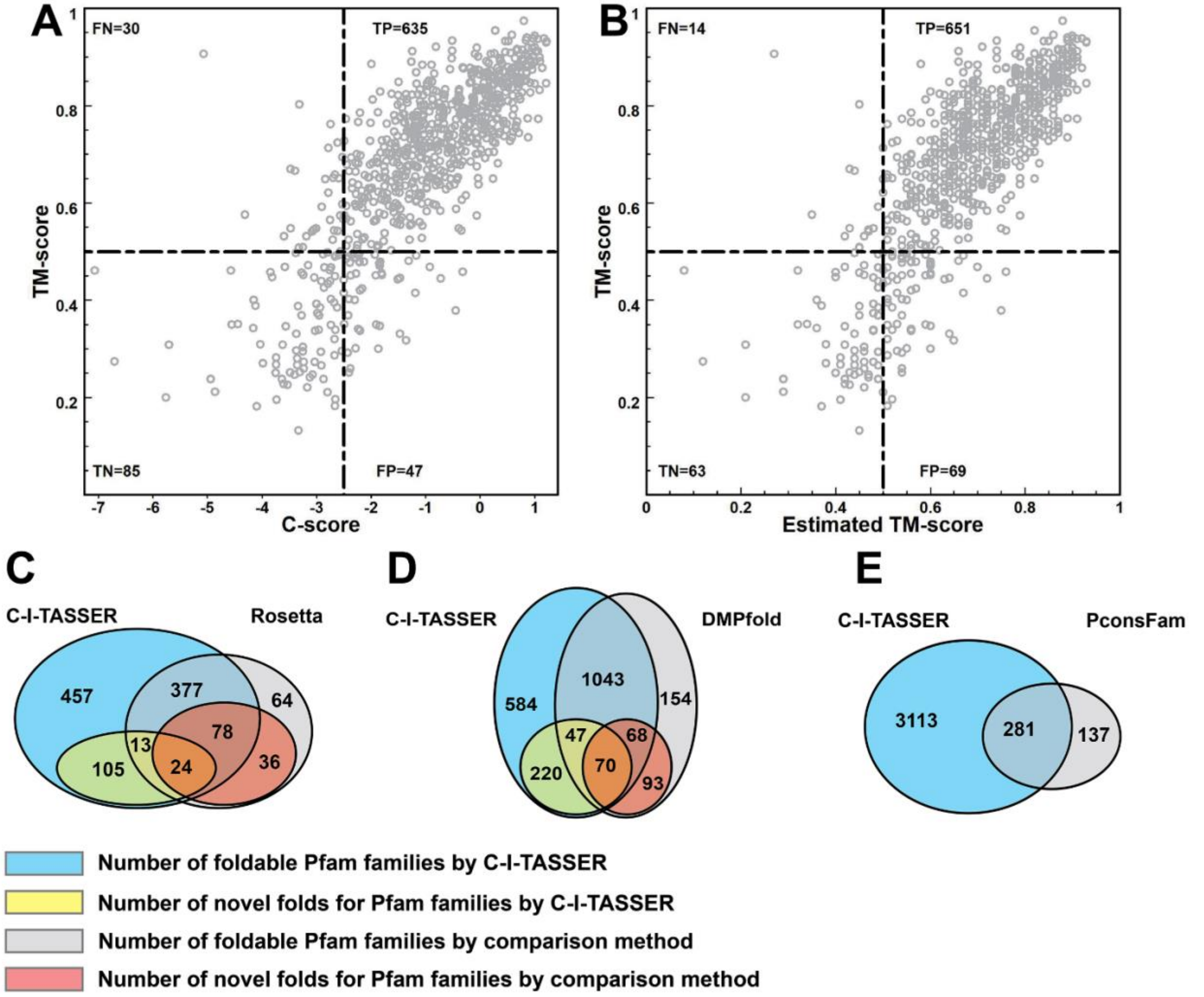

Figure S5. The model quality estimation and comparison with the state of the art for C-I-TASSER, Related to the STAR Methods section "Model quality estimation of C-I-TASSER" and Figure 4. The relationship between the TM-score of the first model generated by C-I-TASSER and two measures, (A) C-score, and (B) Estimated TM-score, for estimating the model quality. Venn diagrams for the number of successful models or novel folds for Pfam families modeled by C-I-TASSER and the three other selected methods: (C) Rosetta, (D) DMPfold and (E) PconsFam. Since our Pfam dataset includes the greatest number of Pfam families, we restricted the successful models and novel folds detected by C-I-TASSER to the Pfam datasets used by either Rosetta, DMPfold, or PconsFam in this comparison.

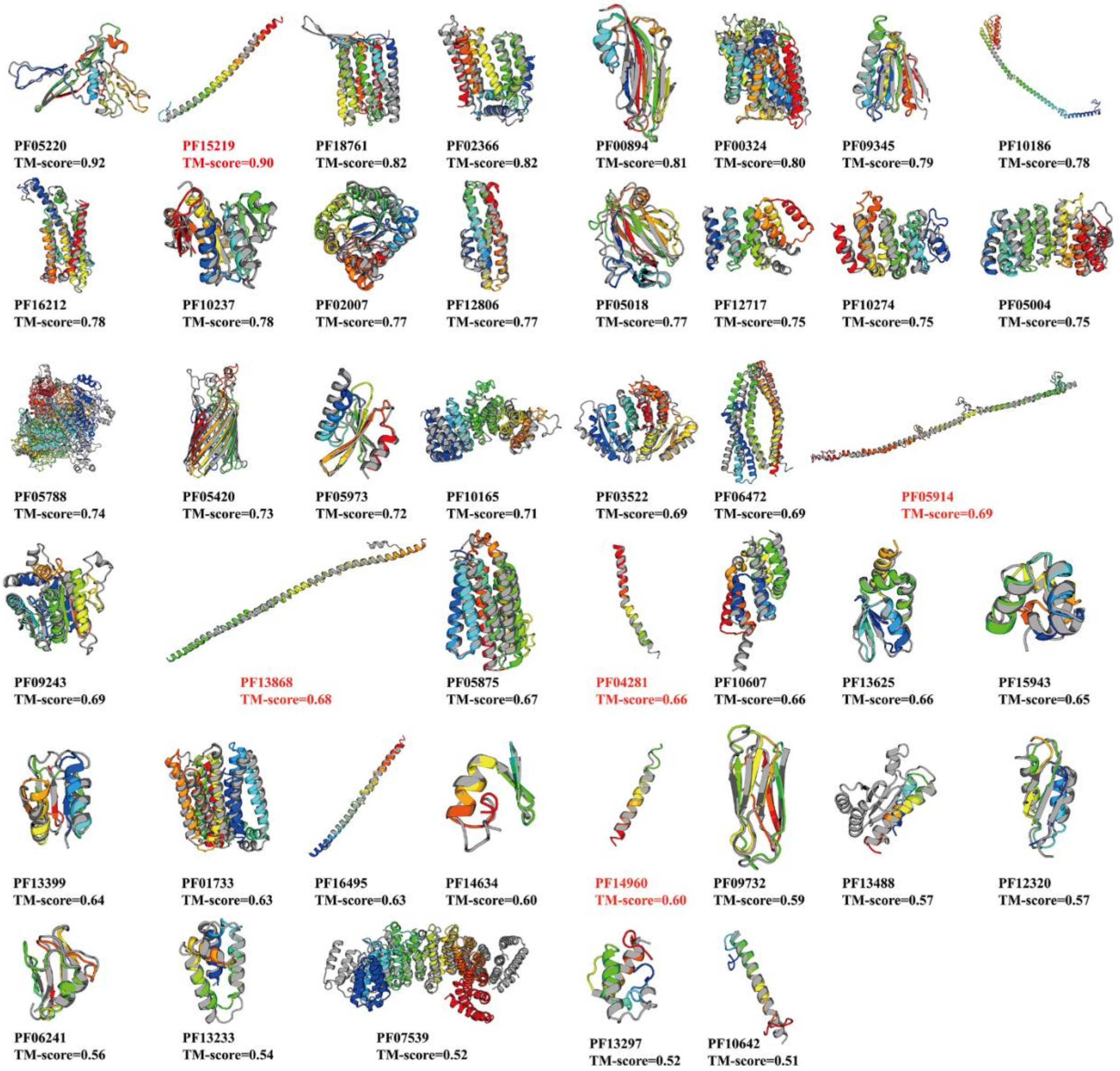

Figure S6. Case study on Pfam families, Related to Figure 4. 5 naïve folds that were regarded as Hard, i.e., only a single helix (red), and the other 38 families that were regarded as Easy (black) by LOMETS. In each case, the model is shown in rainbow color and the solved experimental structure of the member from the same Pfam family, if available, is shown in gray.

Supplementary Tables

Table S1. Comparison between the results of C-I-TASSER, I-TASSER, CNS, trRosetta models and LOMETS templates for different target types on the benchmark dataset, Related to Figure 2. *P*-values were calculated between the TM-scores for the C-I-TASSER models and others using paired one-sided Student's t-tests. #{TM-score ≥0.5} is the number of targets with a TM-score ≥0.5.

Target	Method	TM-score	*P*-value	#{TM-score ≥0.5}
Hard (342)	C-I-TASSER	0.573	*	224
	I-TASSER	0.392	5.07E-50	88
	LOMETS	0.289	1.20E-55	21
	CNS	0.498	7.35E-28	173
	trRosetta	0.500	5.51E-7	155
Easy (455)	C-I-TASSER	0.765	*	441
	I-TASSER	0.741	2.49E-28	429
	LOMETS	0.657	1.85E-68	382
	CNS	0.408	1.62E-76	113
	trRosetta	0.534	1.99E-53	221
All (797)	C-I-TASSER	0.683	*	665
	I-TASSER	0.591	1.32E-80	517
	LOMETS	0.499	1.55E-121	403
	CNS	0.446	8.15E-115	286
	trRosetta	0.519	1.17E-53	376

Table S2. Comparison between the results of C-I-TASSER, I-TASSER and LOMETS templates for targets on the membrane protein dataset, Related to the STAR Methods section "Collection of membrane protein dataset". Note that all targets are LOMETS Hard targets. *P*-values were calculated between the TM-scores for the C-I-TASSER models and others using paired one-sided Student's t-tests. #{TM-score ≥0.5} is the number of targets with a TM-score ≥0.5.

Target	Method	TM-score	*P*-value	#{TM-score ≥0.5}
	LOMETS	0.311	9.86E-15	10
All (80)	I-TASSER	0.429	3.74E-13	24
	C-I-TASSER	0.668	*	68

Table S3. Summary of the modeling results of the top-20 server groups in the CASP13 experiment, Related to the STAR Methods section "Methods Summary". Here, C-I-TASSER is registered as 'Zhang-Server'. QUARK from the Yang Zhang Lab was not listed because it utilized the C-I-TASSER models for some of the TBM domains. Data were taken from the official CASP13 webpage at https://predictioncenter.org/casp13/.

#	Groups	$N_{domains}$	TM-score	Z-score(TM)	GDT	Z-score(GDT)
1	Zhang-Server	112	0.685	1.143	0.625	1.180
2	RaptorX-DeepModeller	112	0.670	1.026	0.613	1.065
3	RaptorX-TBM	112	0.644	0.813	0.587	0.835
4	BAKER-ROSETTASERVER	111	0.606	0.692	0.553	0.750
5	RaptorX-Contact	112	0.603	0.700	0.533	0.675
6	MULTICOM-CONSTRUCT	112	0.597	0.534	0.547	0.565
7	MULTICOM_CLUSTER	112	0.590	0.516	0.539	0.550
8	MULTICOM-NOVEL	112	0.588	0.492	0.538	0.528
9	Yang-Server	109	0.593	0.489	0.535	0.493
10	Zhou-SPOT-3D	112	0.578	0.481	0.523	0.486
11	FALCON	112	0.565	0.387	0.516	0.387
12	IntFOLD5	112	0.566	0.378	0.514	0.385
13	Zhang-CEthreader	112	0.567	0.393	0.507	0.373
14	MESHI-server	57	0.683	0.342	0.615	0.361
15	Seok-server	112	0.575	0.330	0.526	0.355
16	CMA-align	107	0.564	0.346	0.505	0.321
17	AWSEM-Suite	111	0.527	0.210	0.459	0.147
18	slbio_server	99	0.524	0.071	0.480	0.117
19	Seok-assembly	81	0.476	-0.019	0.433	0.008
20	FALCON-TBM	112	0.478	-0.063	0.431	-0.058

Table S5. Summary of C-I-TASSER models for all 24 SARS-CoV-2 proteins, Related to Figure 6.

SARS-Cov-2	Length (AA)	Experimental (PDB ID)	Range	Neff of MSA	TM-score of Model	Estimated TM-score of Model	TM-score of LOMETS
Host translation inhibitor. (nsp1)	180	7K3N_A	1-180	2.1	0.85	0.87	0.81
Non-structural protein 2. (nsp2)	638			2.8		0.40	
Papain-like proteinase. (PL-PRO, nsp3)	1945	7KAG_A	1-111	1.8	0.74	0.90	0.73
		6W6Y_A	207-379	203.3	0.95		0.91
		6W9C_A	748-1060	5.9	0.97		0.96
			1260-1945 (d1:1260-1410;) (d2:1411-1576;) (d3:1577-1945;)				
Non-structural protein 4. (nsp4)	500			2.5		0.53	
Proteinase 3CL-PRO. (nsp5)	306	6LU7_A	1-306	2.4	0.98	0.96	0.90
Non-structural protein 6. (nsp6)	290			6.8		0.37	
Non-structural protein 7. (nsp7)	83	7BTF_C	1-83	2.5	0.67	0.63	0.38
Non-structural protein 8. (nsp8)	198	7CYQ_D	1-198	1.9	0.57	0.88	0.54
		7CYQ_D	(d1:1-83;)	2.4	0.82		0.78
		7BTF_D	(d2:84-132;)	3.0	0.95		0.94
Non-structural protein 9. (nsp9)	113	6W9Q_A	1-113	2.7	0.95	0.93	0.88
Non-structural protein 10. (nsp10)	139	6W75_B	1-139	2.1	0.92	0.90	0.88
RNA-directed RNA polymerase (RdRp). (nsp12)	932	6M71_A	1-932	2.0	0.96	0.80	0.91
Helicase (Hel).	601	5RL9_A	1-601	166.7	0.94	0.99	0.91
Guanine-N7 methyltransferase (ExoN).	527			1.1		0.99	
Uridylate-specific endoribonuclease (NendoU). (nsp15)	346	6VWW_A	1-346	3.5	0.99	0.99	0.94
2'-O-methyltransferase (2'-O-MT). (nsp16)	298	6W75_A	1-298	6.1	0.97	0.99	0.93
Surface glycoprotein (S).	1273	6VXX_A (closed state)	27-1146	2.3	0.97	0.98	0.86
ORF3a.	275	6XDC_A	1-275	0.4	0.30	0.28	0.20
E.	75	7K3G_A	8-39	4.5	0.46	0.60	0.40
M.	222			2.9		0.37	
ORF6.	61			0.4		0.54	
ORF7a.	121	6W37_A	16-82	0.2	0.97	0.72	0.90
ORF8. (ns8)	121	7JTL_A	1-121	0.4	0.27	0.45	0.19
N.	419	6M3M_A	50-174	4.2	0.95	0.67	0.75
		6YUN_A	249-364	4.9	0.88		0.77
ORF10.	38			0.2		0.49	