SI Appendix: Protein homology model refinement by large-scale energy optimization

Authors: Hahnbeom Park, Sergey Ovchinnikov, David E Kim, Frank DiMaio, and David Baker

Table of contents

SI Figures

SI Tables

SI Methods

References

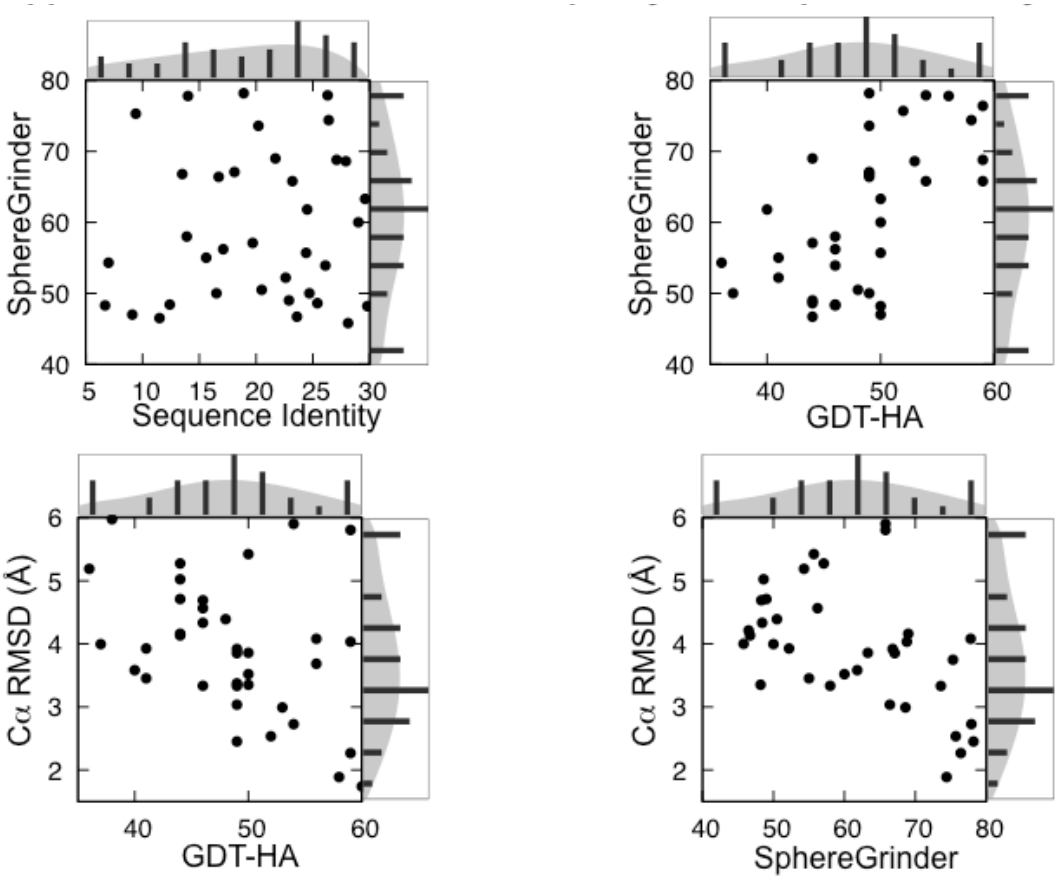

Fig. S1. **Correlations between the model quality metrics for the input models.** In the top-left panel, the correlation between SphereGrinder of the input structure and the sequence identity to the best template is shown. In the remaining panels, pairwise correlations between three metrics, SphereGrinder, GDT-HA, and Cα-RMSD, are shown. 44 proteins from benchmark set1 are used for the analysis.

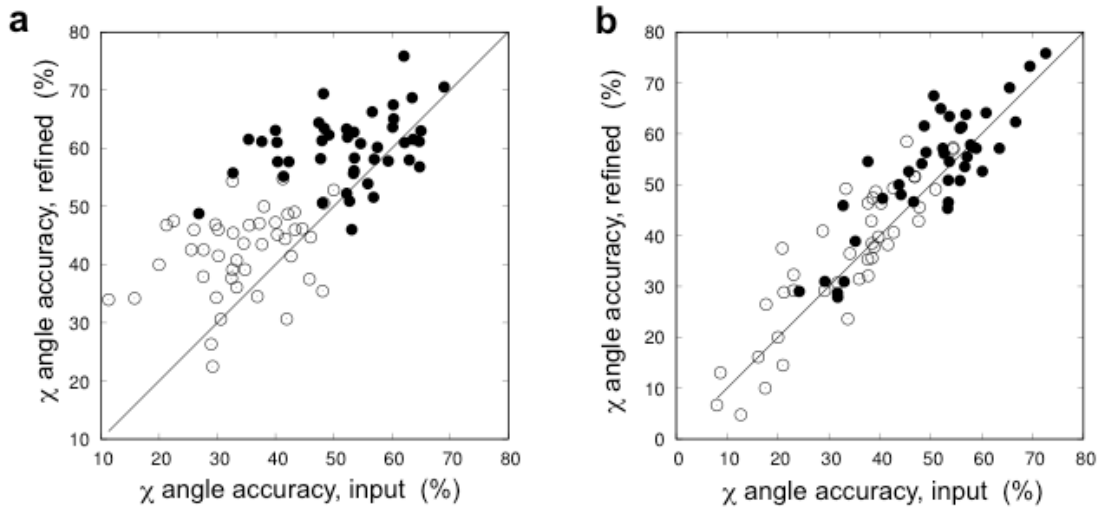

Fig. S2. Side-chain accuracy improvement by refinement, shown for (a) benchmark set1 and (b) set2. In each panel, χ_1 accuracy (measured as % of residues with χ_1 deviation from the native less than 30 degrees) is shown as solid circles, and χ_{1+2} accuracy (measured as % of residues with both χ_1 and χ_2 deviation from the native less than 30 degrees) as empty circles, respectively; results are compared between the input model (x-axis) and the final refined model (y-axis). Solvent exposed residues with relative accessible surface area > 50% are not counted.

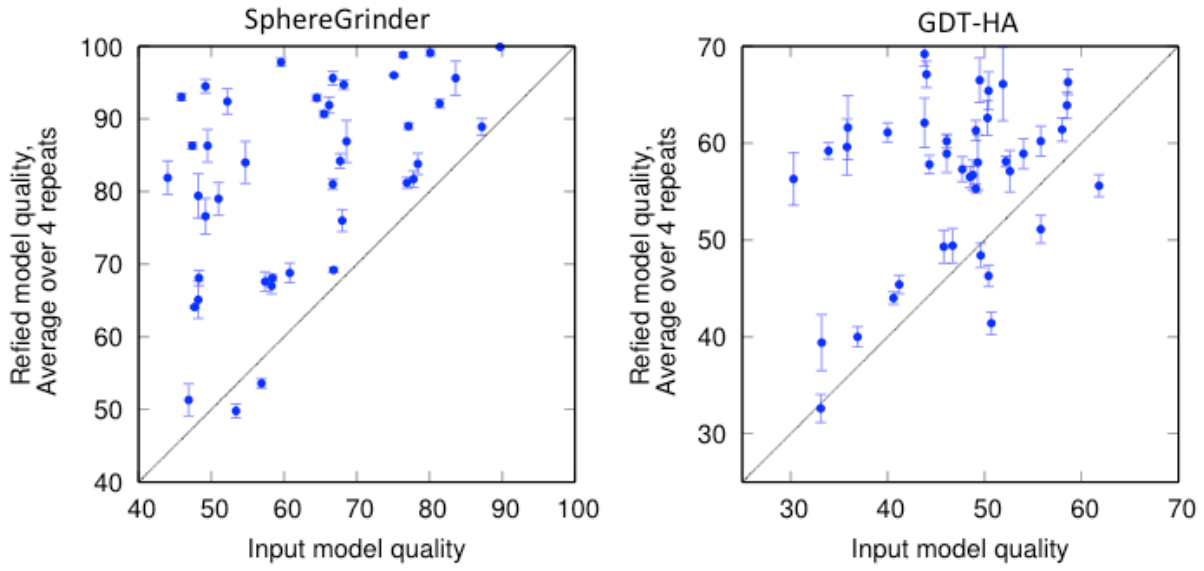

Fig. S3. **Convergence of the model quality within multiple repeats of refinement runs**. Variations in the model quality for 4 repeated runs are shown as error bars in the y-axis (points represent the average over 4 runs), compared to the input model quality in the x-axis, reported in two metrics, SphereGrinder (left panel) and GDT-HA (right panel).

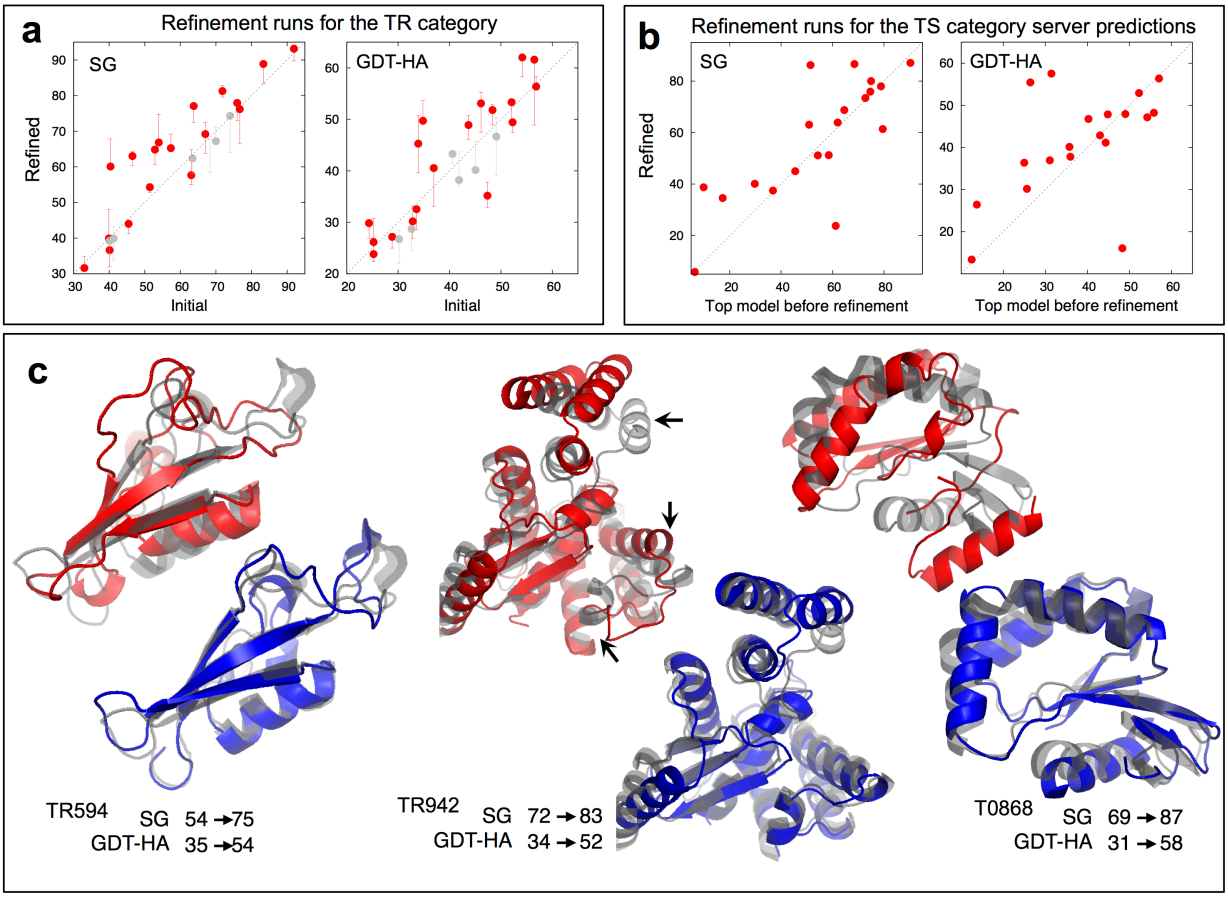

Fig. S4. **Blind prediction results in CASP12. (a,b)** Scatter plots comparing the model quality of the input model to the refined model: **(a)** 24 targets tested for the refinement category (TR) and **(b)** 20 targets tested for the regular tertiary structure prediction category (TS). For the TR category, the model quality range for five cluster representatives is shown as error bars, and the best-ranked model (model1) in dots. 4 targets in gray color are those forming heavy homo- or hetero-oligomers, or a membrane protein. For the TS category, a model ranked as the best before the refinement is selected for the comparison in the x-axis. **(c)** Examples with remarkable blind predictions in CASP12: two targets from the TR category, TR594 (top), TR942 (center), and one from the TS category, T0868 (bottom). Native, input (or best ranked before refinement for T0868), and refined structures are shown in gray, red, and blue colors, respectively.

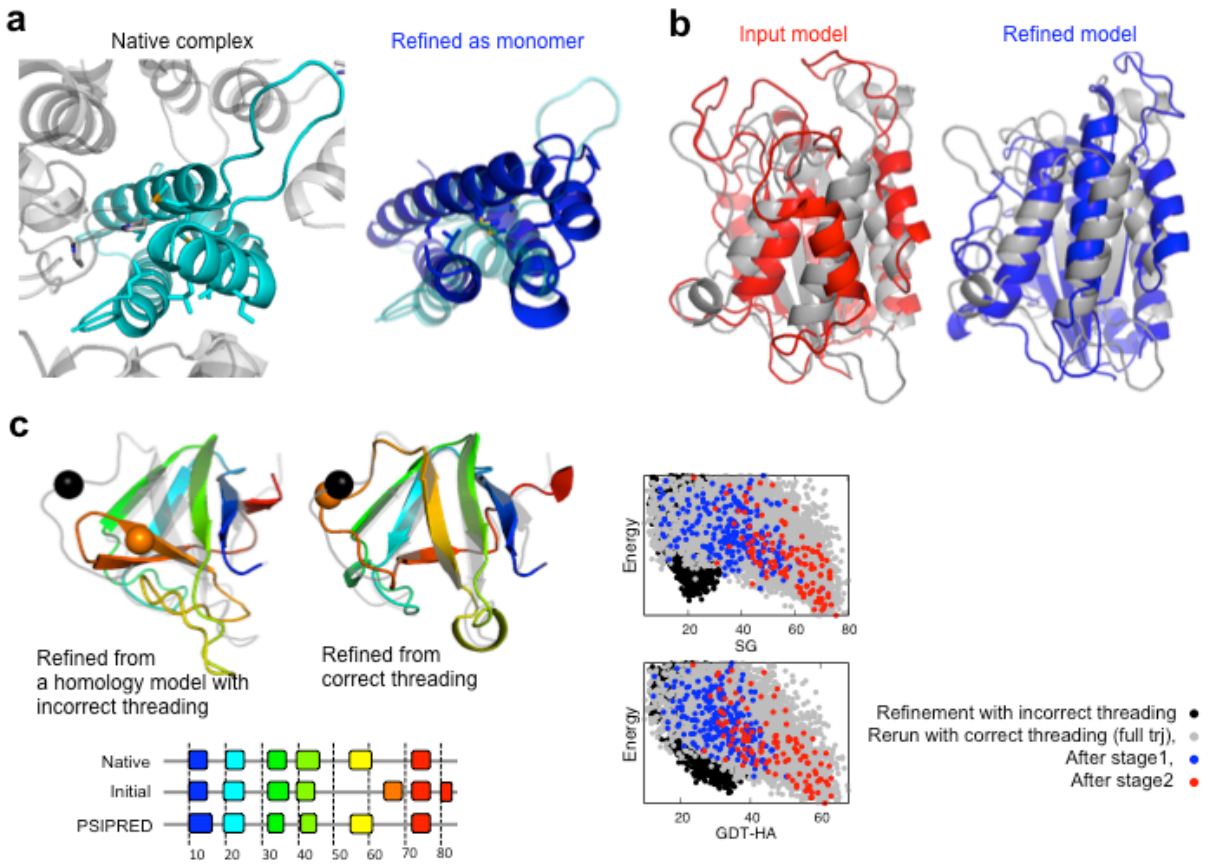

Fig. S5. Examples of refinement failures. (a) TR862, a failure due to native complex interactions not considered when refined as a monomer. Refinement of the input model as a monomer altered loop conformations and helical orientations to bury a set of exposed hydrophobic residues; in the native complex these residues are desolvated through inter-chain contacts. **(b)** TR901, a failure when a protein is larger than 200 residues and has a complex topology. **(c)** TR896, a failure when the input model contains significant register shifts from the native structure. In the top panel, the native structures are shown in gray and refined structures in rainbow color; on the left side, refined starting from the CASP12 input model (having incorrect threading on the template), and on the right side, refined starting from a re-threaded model according to PSIPRED (1) prediction. Residues forming (or predicted as) β-strands are shown at the bottom following the color of regions in the structures. On the right panel, energy landscapes for the refinement runs are shown in two metrics, SphereGrinder (SG) (top) and GDT-HA (bottom), when ran on the CASP12 input model (full trajectory as black dots) and on the re-threaded model (full trajectory as gray, first iteration as blue, and final iteration as red dots).

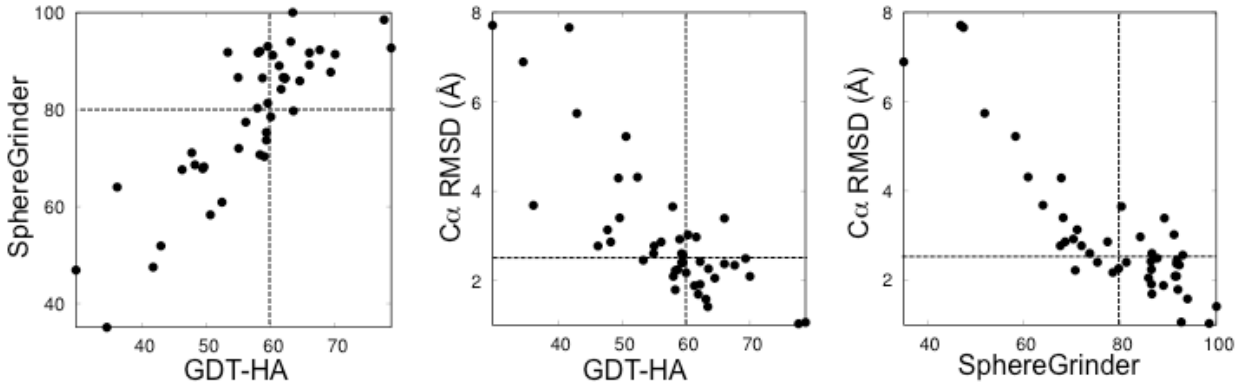

Fig. S6. Correlation between the model quality metrics for the final models, shown for 44 proteins in benchmark set1. Thresholds in model accuracy metrics used for the definition of "correct fold" (see main text) are shown as dashed lines in each panel.

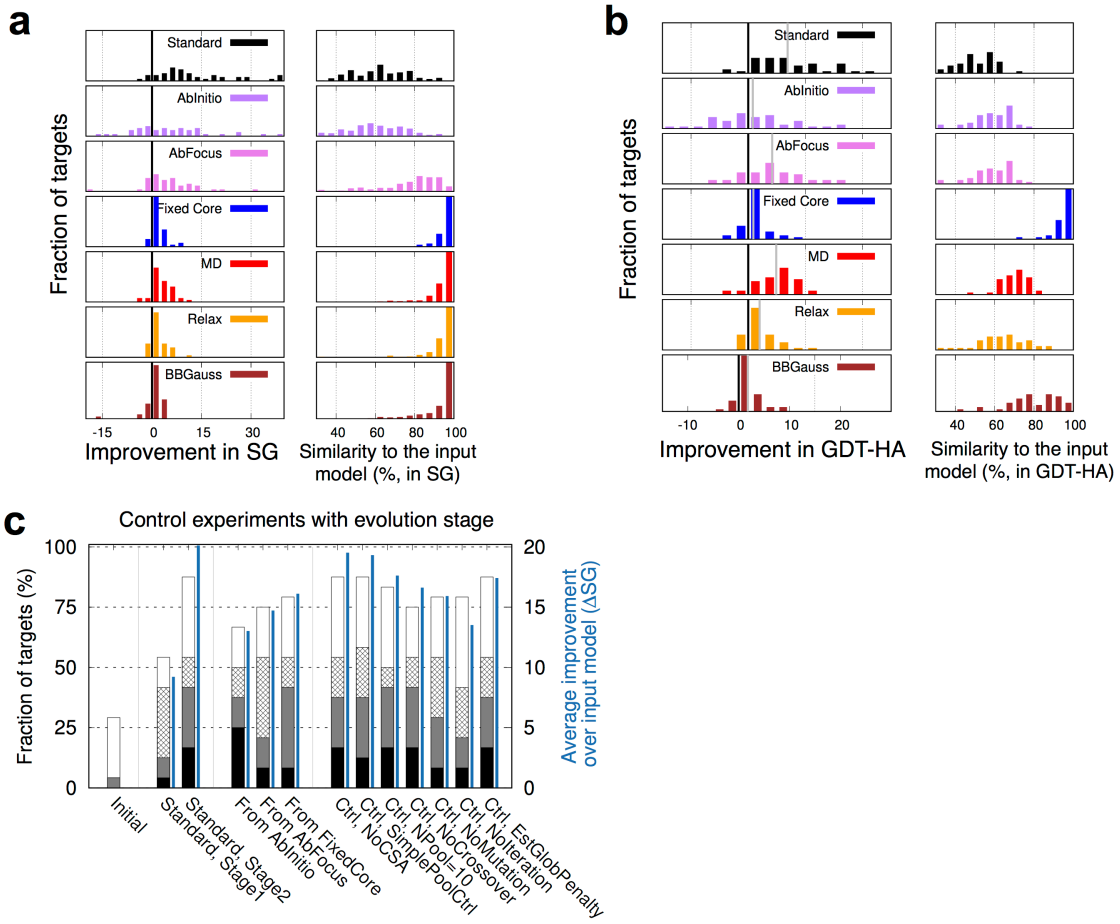

Fig. S7. **Supplementary to Fig. 3. a,b)** Supplementary to **Fig. 3a**, distribution of per-target improvements over the input models and structural similarity to the input models for the methods in Fig. 3a are shown in two metrics, a) SG and b) GDT-HA. Similarity to the input model is measured using the input model as reference (closer to 100 is more similar to the input model). *MD, Relax, BBGauss, and FixedCore* produce only small changes in the input models (right panels), *AbInitio* produces equivalent amount of structural variation as the standard protocol but more often degrades input model quality. **c)** Control experiment results with evolution stage; figure captions are as explained for the diversification stage in **Fig. 3a**. Strategies starting with “From” (in x-axis) refers to the experiments in which the evolution stage was run with a pool of structures generated with a different diversification logic (e.g. “From AbInitio” is an evolution stage result but starting from the pool of structures generated by “AbInitio” logic). The remaining control experiments starting with “Ctrl” took the output of the standard diversification logic (“Standard, Stage1”), and details are explained in the evolution stage section of **SI Methods**.

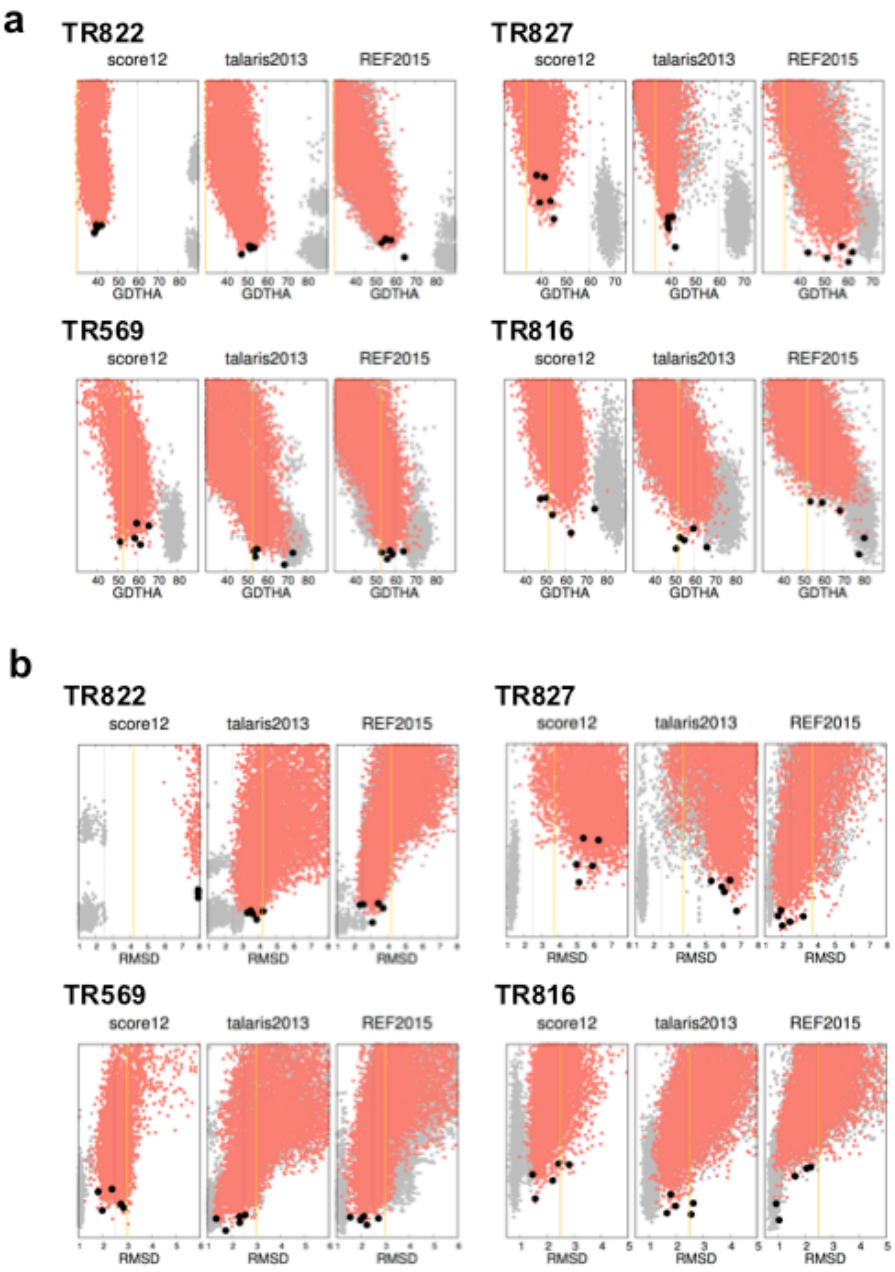

Fig. S8. Energy landscapes in other metrics for the selected targets in Fig. 4c: **a)** GDT-HA and **b)** RMSD.

Table S1. List of 44 targets in benchmark set 1.

Target			Used for parameter decision	Native structure			Starting model			Best template[1)]	
Source	Name	PDBID		Topology	nres	Expt. method[2)]	GDT-HA	SG	Source of the input model	PDB ID	Seq. Identity
CAMEO		3wvaA		αβ	170	1.4	49.1	67.7	Robetta	4uuxA	13.5
		3wz4A		αβ	144	2.2	46.7	66.8	Robetta	4jf8A	16.7
		4idiA		β	144	1.9	46.0	53.4	Robetta	1wgkA	26.1
		4ipuA		β	137	1.3	45.8	58.4	Robetta	2qv8A	13.9
		4ld6A		αβ	117	1.7	54.3	68.0	Robetta	3llrA	31.6
		4nofA		β	125	1.6	58.6	66.2	Robetta	2ocwA	23.2
		4oleA		β	122	2.5	40.6	58.3	Robetta	2l0dA	15.6
		4qprA		αβ	143	1.5	55.8	77.1	Robetta	2vjwA	14.0
		4u77A		αβ	134	2.0	55.8	87.2	Robetta	1buoA	14.9
		4uapA		β	152	2.0	45.8	56.9	Robetta	1t2xA	17.1
		4wxmA		αβ	143	2.3	48.8	77.8	Robetta	1l5yA	18.9
		4zhbA		α	114	1.3	49.0	76.9	Robetta	4rlvA	20.2
CASP TS category	T0540	3mx7A		β	90	1.8	46.1	48.3	FALCON-SWIFT	2kd2A	6.7
	T0552	2l3bA		β	111	NMR	43.8	45.9	FAMS03	1kb9A	25.4
	T0564	2l0cA		β	89	NMR	44.3	49.2	Distill	1jmcA	23.6
	T0572	2kxyA		β	93	NMR	36.9	47.7	QUARK	2qsvA	24.7
	T0579	2ky9A		β	64	NMR	46.1	49.2	MidwayFolding	2qqrB	12.4
	T0604D1	3nlcA		αβ	82	2.1	44.2	59.6	I-Tasser	1qo8D	19.7
	T0612	3o0lA[3)]		β	88	1.8	58.0	81.4	QUARK	2uxtB	26.4
	T0630	2kytA		αβ	123	NMR	47.7	51.0	gws	3kw0B	20.5
	T0643	3nzlA		α	75	1.2	44.0	65.5	STAT-PROTEUS	3dnvB	21.7
	T0669	2ltlA		αβ	109	NMR	48.5	47.4	BhageerathH	2ffmA	16.5
	T0724D1	4fmrA		β	115	2.2	49.5	54.7	Robetta	3t2lA	24.4
	T0743	4hyzA		αβ	114	2.2	49.3	66.7	PconsM	2cw9A	18.1

CASP TR category-	TR283	4cvhA		αβ	168	2.4	41	57.4	nns	1vpaA	22.6
	TR557	2kyyA	y	αβ	125	NMR	50	60.8	Robetta	3lmm	29.6
	TR568	3n6yA	y	β	97	1.5	35	34	Robetta	3cu7B	24.7
	TR569	2kywA	y	β	79	NMR	53	68.6	Robetta	1ftpB	27.9
	TR574	3nrfA	y	β	102	2.2	40	66.7	I-Tasser	3ivrB	24.5
	TR624	3nrlA	y	β	69	1.7	36	52.2	N/A	2zzeA	7.0
	TR663	4exrA	y	β	152	1.8	49	83.6	N/A	2gu3A	13.0
	TR696	4rt5A	y	β	100	1.5	50	64.5	Bilab-ENABLE	1lqkB	29.0
	TR705	4ftdA	y	β	96	1.9	44	49.5	AOBA-server	3s30B	22.9
	TR769	2mq8A		αβ	97	NMR	60	89.7	I-Tasser	2kl8A	25.8
	TR780	4qdyA		β	95	2.7	59	68.2	I-Tasser	3hs0I	26.3
	TR280	4qdyA		β	96	2.7	54	78.4	Robetta	3u4yB	27.1
	TR803	4ogmA		αβ	134	2.2	38	38.4	myprotein-me	3gn9A	20.9
	TR816	5a1qA		α	68	1.6	52	80.1	I-Tasser	3cazA	3.9
	TR822	5fu5A		β	117	1.5	30	48.2	TASSER	2w47A	11.5
	TR827	N/A		α	193	N/A	34	75.1	nns	3pkrA	9.4
	TR828	4z29A		β	84	2.0	50	48.2	I-Tasser	3qx3B	29.8
	TR829	4rgiA		αβ	67	1.7	50	44	QUARK	3k8rA	9.1
	TR854	4rn3A		α	70	2.1	59	76.4	Robetta	2ah5A	31.4
	TR857	2mqcA		β	96	NMR	33	46.9	eThread	2osxA	28.1

1) Based on sequence similarity to the target sequence. Template protein with highest structural similarity can differ.
2) Resolution shown for crystal structures in Å
3) Dimeric interface residues 20-41 are trimmed.

Table S2. Comparison of the refinement results to the best cherry-picked models by other methods in previous rounds of CASP.

Target	ΔSphereGrinder			ΔGDT-HA		
	Best by other groups[1)]	Current work[2)]	Difference	Best by other groups[1)]	Current work[2)]	Difference
TR280	18.2	24.5	6.3	12.8	17.0	4.2
TR283	6.7	10.6	3.8	4.2	6.1	1.9
TR557	4.2	7.5	3.3	1.8	0.4	-1.4
TR568	9.8	39.7	29.9	3.7	26.3	22.6
TR569	19.2	26.3	7.0	7.4	9.8	2.4
TR574	2.9	26.5	23.5	5.3	22.0	16.7
TR624	22.5	42.0	19.6	12.6	24.5	11.9
TR663	3.3	4.9	1.6	8.7	6.3	-2.4
TR696	17.0	27.0	10.0	8.0	17.0	9.0
TR705	15.6	37.0	21.4	10.7	24.8	14.1
TR769	10.3	10.3	0.0	12.7	6.8	-5.9
TR780	10.0	4.7	-5.3	6.8	2.6	-4.2
TR803	6.7	13.1	6.3	0.6	6.4	5.8
TR816	16.2	19.9	3.7	9.0	26.3	17.3
TR822	7.9	38.6	30.7	8.2	34.0	25.8
TR827	8.8	20.7	11.9	14.3	26.6	12.3
TR828	16.5	21.3	4.9	4.2	5.1	0.9
TR829	32.8	32.1	-0.7	8.2	28.7	20.5
TR854	7.1	22.9	15.7	7.4	12.8	5.4
TR857	8.3	2.1	-6.2	7.1	-3.1	-10.2
Average	12.2	21.6	9.4	7.7	15.0	7.3

1) Best submissions among all five models generated by all groups other than “Baker”
2) Best of five cluster representative

Table S3. Crystallographic phasing experiment using input and refined models.

Target	LLG, input model	LLG, refined models[1)]
TR568	**67**	**93**
TR574	24	51
TR624	42	41
TR663	6	53
TR696	20	37
TR780+TR280	38	**78**
TR803	27	26
TR816	35	**172**
TR822	25	**61**
TR829	16	34

1) Log likelihood gain (LLG) among the values from molecular replacement on five models. Models with LLG > 60 are highlighted by bold letters.

Table S4. List of 40 targets in benchmark set 2.

PDBID	Native structure					Starting model			Best template[1)]	
	Topology	nres	Modeled range	Evaluation range	Expt. method[2)]	GDT-HA	SG	RMSD	PDB ID	Seq. Identity
2mdpA	αβ	85	1-85	6-29,41-85	NMR	34.5	64.7	3.26	4x25A	8.1
2mx7A	α	100	277-388	289-388	NMR	37.8	49.5	7.23	1qjtA	23.5
2mzoA	α	93	1-93	6-85	NMR	49.7	74.7	2.58	1eo0A	27.2
2n12A	α	82	58-139	63-80,93-139	NMR	34.6	36.2	11.31	1qeyA	17.1
2n3dA	β	100	37-136	37-136	NMR	40.0	43.5	8.01	3a1yA	16.9
2n3lA	αβ	74	6-79	6-79	NMR	43.9	77.7	4.10	1x4aA	37.6
2n59A	β	100	1-100	1-100	NMR	48.0	79.5	2.40	2xskA	18.9
2n93A	β	130	1-130	1-130	NMR	34.0	63.1	3.76	4qgvA	29.5
2nanA	β	140	22-161	22-161	NMR	47.3	71.1	2.79	2ox8A	22.1
2nbsA	αβ	116	1-116	11-116	NMR	49.5	92.0	2.39	2vimA	29.8
2ncoA	α	84	124-207	124-164,170-207	NMR	44.1	59.0	5.18	1h4bA	20.0
2rv9A	β	130	1-130	1-130	NMR	44.6	74.2	2.76	3le0A	24.5
2rvaA	β	131	1-131	1-131	NMR	46.4	57.3	3.93	3le0A	20.9
4uwqB	β	126	27-152	32-79,90-146	3.28	58.1	76.2	2.83	2oxgB	31.6
4ybaA	α	99	1-99	5-81	1.70	55.2	50.0	5.70	2b5aA	26.0
4z3uA	β	181	1-181	4-174	2.71	43.3	56.3	3.85	5d5nA	18.8
4zuaA	αβ	178	1-178	11-165	2.50	39.4	63.9	4.91	1xjaA	11.8
4zv5A	α	91	2-92	2-92	1.57	31.0	37.4	13.1	2f77X	26.7
5aozA	β	140	405-544	405-539	1.14	44.6	73.0	3.48	1aohB	19.9
5azxA	β	103	30-132	30-129	1.58	48.7	73.0	3.56	2p9rA	10.7
5b1rA	β	127	228-354	229-306,311-348	1.20	50.0	31.5	7.75	4zesA	17.6
5c4pA	β	126	1-126	3-94,103-121	1.97	47.5	73.4	3.53	1xqaA	23.8
5cesA	αβ	96	202-297	202-292	2.10	44.8	38.5	10.29	1x7vA	10.1
5dyqA	β	152	1-152	10-140	1.66	48.8	66.8	4.30	5btuA	17.4

5e46A	β	169	16-184	16-184	1.85	49.8	65.1	6.99	3rt0C	17.6
5e6fA	αβ	152	1-152	2-113,128-145	2.60	48.5	74.2	2.72	1hjrA	15.0
5eliA	β	120	17-136	20-131	3.10	56.2	58.5	4.19	2q87A	27.6
5f3qA	α	193	1-193	2-35,44-193	2.10	32.3	38.0	18.20	3n3wA	16.0
5fidA	β	137	16-152	18-152	1.81	45.6	58.2	3.38	2i0wA	14.4
5forA	αβ	139	1-139	6-139	2.50	32.8	55.2	4.05	3hv2A	19.1
5fr7A	αβ	138	8-145	8-144	1.95	32.5	59.2	4.34	2xgaA	11.2
5fvjA	αβ	166	1-166	5-160	1.70	37.8	61.5	4.03	1wwzA	14.5
5g51A	αβ	139	260-398	260-398	1.45	43.5	60.8	3.26	4nwr0	10.0
5ghaE	β	75	-8-65	2-63	2.50	52.0	78.2	2.99	1ryjA	27.4
5gt1A	β	165	348-512	360-512	1.85	45.1	37.6	5.09	2b0pA	18.2
5i2qA	α	120	65-184	80-184	1.94	55.0	66.2	2.56	1c7vA	27.3
5jojA	α	97	1-97	8-97	NMR	47.8	73.3	2.89	2amiA	28.8
5lgmA	α	69	16-84	16-84	NMR	35.1	44.2	4.66	2hgcA	16.5
5m1mA	α	155	1-155	2-155	1.50	38.8	78.9	3.17	2w9yA	9.2
5xgaA	αβ	108	8-115	8-115	1.95	40.7	54.8	5.14	3lr4A	18.3

2) Based on sequence similarity to the target sequence. Template protein with highest structural similarity can differ.
3) Resolution shown for crystal structures in Å

Table S5. Decomposition of the energy terms contributing to the discrimination of native-like structures. Targets for which the energy landscapes significantly differ by the all-atom energy functions used are shown here. Energy gap (ΔE) between a relaxed native structure and the best scoring non-native structure (SphereGrinder < 0.8) is reported for total energy (ΔE,total), contribution to the energy by van der Waals interactions (ΔE_{vdw}), and contribution to the energy by Coulombic plus solvation interactions ($\Delta(E_{Coulomb} + E_{solv})$). Negative energy gap means better discrimination of the native-like conformation against false conformations.

	ΔE,total			ΔE_{vdw}			$\Delta(E_{Coulomb} + E_{solv})$		
	Talaris2013	Ref2015	Δ	Talaris2013	Ref2015	Δ	Talaris2013	Ref2015	Δ
4qprA	-1	-18	**-17**	+4	-8	**-12**	**-2**	**-11**	**-9**
4wxmA	-3	-24	**-21**	+2	-8	**-10**	-4	-3	+1
T0572	+3	-17	**-20**	+14	-5	**-19**	-8	-14	**-6**
T0579	-4	-16	**-12**	+5	-5	**-10**	-6	-9	-3
T0669	-5	-28	**-23**	+1	-15	**-16**	-2	-3	-1
TR574	-8	-19	**-11**	-2	-27	**-25**	-3	+12	+15
TR816	-3	-20	**-17**	0	-12	**-12**	-6	-5	+1
TR822	-10	-27	**-17**	-22	-19	+3	+9	+3	**-6**
TR827	-5	-27	**-22**	+13	-17	**-30**	-8	-2	+6

SI Methods

Model quality metrics

Three model quality metrics used in CASP refinement challenge assessment are used throughout the article. *RMSD* measures the root-mean-squared distance of Cα positions between two structures after superposition. *Global distance test - high accuracy* (GDT-HA) counts the percentage of residue Cα coordinates which are correctly positioned in a global frame when the entire model is superimposed onto the native structure. A residue is assigned as correct if the distance between corresponding Cα atoms in the model and the native stucture in the global frame is below thresholds of 0.5, 1.0, 2.0, and 4.0 Å; the fraction of correct residues is computed for each threshold and the four values are averaged. *SphereGrinde*r (SG) counts the percentage of residues with the correct local context and therefore is insensitive to global superposition. Correctness of the local context is measured by computing -- for the residue Cα of interest -- the RMSD of all atoms in a 6.0 Å sphere. Two thresholds in local RMSD, 2.0 and 4.0 Å, are used; the values reported are the average of the two. The correlations between the three metrics are reported in **Fig. S1** for input models, and in **Fig. S6** for final models. Structures are assessed as *correct folds* if two of three criteria are satisfied: RMSD equal or less than 2.5 Å, GDT-HA equal or higher than 60.0, or SG equal or higher than 80.0.

Benchmark set

We select sets of targets from CASP [2] and CAMEO [3] based on the criteria that i) the protein size is between 60 and 200 amino acids and ii) the best homology model is of low-resolution accuracy. Homology model quality is considered as low-resolution if i) GDT-HA is between 30 to 60 for proteins smaller than 100 residues, and ii) between 30 to 55 for larger proteins; models with even lower accuracy are have incorrect topologies and are not considered in this study. For the first benchmark set (set1) we applied the criteria on previous CASP rounds (CASP9 to CASP11) and also CAMEO rounds from July 2013 to September 2015. A total of 32 targets fall into this criteria through the entire refinement category (TR) targets from CASP9 through CASP11. Of these, we excluded 11 targets having considerable inter-domain (TR606, TR671, TR774, TR228) or inter-subunit interactions (TR517, TR576, TR622, TR698, TR722, TR759, TR772), as well as one target likely stabilized by the crystal lattice (TR837). In addition to the 20 targets selected from CASP TR targets, we added 12 targets from CAMEO and other 12 non-refinement CASP targets, subject to the same criteria on protein size, starting model quality, and biological assembly. Details of targets are listed in **Table S1**. 8 of these targets were used for making decisions on several options (marked in **Table S1**), including the functional form of the restraints and its relative strength in the coarse-grained modeling step, and the number of structures to sample in the diversification and evolutionary stages.

For the second benchmark set (set2), we applied the same criteria on CAMEO targets since October 2015 to August 2017.

Determining fraction of unreliable residues

The fraction of residues assigned as unreliable regions is determined as a function of both protein size *Nres* and target difficulty s_0:

$$f(Nres) = \min(1.0, 0.3 + \max(0.2, 70.0/Nres))$$
$$g(s_0) = \min(0.5, 1.0 – s_0/100)$$
$$minfrac = f \times (g-0.2),\ maxfrac = f \times g \qquad \text{-- Eq 1}$$

with (estimated) target difficulty s_0 [4] in GDT-HA scale; multiplication factor *f*, a function of target size ranging from 0.5 (Nres > 350) to 1.0 (Nres <100); multiplication factor *g*, a function of target difficulty ranging from 0.4 (s_0 > 60, close to native) and 0.5 (s_0 < 50, distant to native). The threshold starts from 2.0 times the lowest 40-percentile residue-level fluctuation, and adjusted until the fraction of residues with fluctuation higher than the threshold falls within the range of (minfrac,maxfrac).

Evolution stage

Here we describe the details of components in the evolution stage mentioned in the main text.

Parent selection rule: At the beginning of each iteration, a subset of current members in the pool (10 in this study) are selected as seed parents with a priority based on: i) the number of times the model was used as a seed and ii) energy values (if the former values are equal). A member used less often, or used an equal number of times but with lower energy is selected.

Pool update logic: At the end of each iteration, newly generated structures are first filtered based on their mutual *structural distance*, measured by *D = 1 - S-score* (with a reference deviation [5] of 2.5 Å); S-score [5] is a structural "similarity" measure, hence *1 - S-score* returns "distance" between two structures ranging from 0 (identical) to 1 (totally different). A structure is filtered out if there is any other structure in the newly generated pool having a better energy value and structurally similar ($D < D_{filter}$, D_{filter} = 0.2 in this study). This filtered set of newly generated models is then compared to the reference pool members for the current iteration. Replacement happens between similar structures ($D < D_{cut}$) if the new model is more favorable in energy than a reference pool member, and also between dissimilar structures ($D > D_{cut}$) if the new model is dissimilar to any reference pool members but still has a more favorable energy than the worst reference pool member [6]. In the pool update stage, the distance threshold D_{cut} is decreased in subsequent iterations [6], decreasing pool diversity as the algorithm proceeds. In this study, D_{cut} is defined as:

$$D_{cut} = \max(0.6 \lambda g'(s_0), D_{filter})$$

$$g'(s_0) = \min(1.0, 1.0 - (s_0 - 40.0)/40.0) \quad \text{-- Eq. 2}$$

with λ linearly decreasing from 1.0 to 0.5 through the 30th iteration (and kept at 0.5 thereafter). $g'(s_0)$ is a factor considering input model quality, ranging from 0.0 (very accurate) to 1.0 (very inaccurate), with s_0 representing (estimated) input model quality in GDT-HA scale. For instance, D_{cut} decreases from 0.6 to 0.3 for very difficult targets (s_0 < 40.0), and from 0.3 to 0.2 for relatively easier cases with s_0 = 60.0. We took s_0 as the actual input model's GDT-HA subtracted by 5.0 if the native is known; however, replacing it to an estimated value from the structural convergence in homologs [4] did not degrade the overall result much (*EstGlobPenalty* in **Figure S7c**).

Global structure deformation factor: When a structure drifts away too much from the input structure, objective function E'(x) becomes penalized from the original energy value E(x) by a multiplication factor $p(s_{toinit})$ ranging between 0 and 1:

$$E'(x) = p(s_{toinit}) E(x),$$

$$p(s_{toinit}) = 1.0 - 0.004 (\min(0, s_{toinit}-s_0))^2 \text{ -- Eq 3}$$

Here, the structural similarity between a model and the input structure, s_{toinit}, is measured in a

GDT-HA scale, and the penalty starts to apply when $s_{toinit} < s_0$. For instance, if s_0 is set to 40.0 (that is, GDT-HA of the input model is estimated as 45), the penalty starts at $s_{toinit} < 40.0$, which starts to significantly differ from identity (< 0.95) at $s_{toinit} < 34.0$. This factor ensures that the global search will focus on conformational space sharing the same topology with the input structure, not on completely different topologies.

Control experiments for the evolution stage: In the evolution stage, seven control experiments are tested, the names of which are listed at the bottom panel of **Figure S7c**. *NoCSA* eliminates distance annealing logic by fixing λ factor in Eq 2 to a constant of 0.5. *SimplePoolControl* replaces the priority rule for seed parent selection to an energy-weighted stochastic selection (also called the roulette-wheel algorithm) [7]:

$$\text{Priority for i} = \frac{\exp\left(\left(E_i - E_{\min}\right)/0.2\left(E_{\max} - E_{\min}\right)\right)}{Z},$$

$$Z = \sum_i \exp\left(\left(E_i - E_{\min}\right)/0.2\left(E_{\max} - E_{\min}\right)\right) \quad \text{-- Eq 4,}$$

Npool=10 uses only 10 structures for the pool size throughout (standard is 50). *NoCrossover* and *NoMutation* eliminate the crossover and mutation operations, while keeping the total number of models generated at each iteration the same as standard at 120. *NoIteration* generates 6,000 models from the initial structural pool rather than over 50 iterations. *EstGlobalPenalty* replaces the s_0 in Eq 2 and Eq 3 with an estimate of the starting model quality based on the structural variations from templates [4] adjusted into GDT-HA scale (which was given as the input model's GDT-HA minus 5).

Update iteration: During the iteration process, a regular iteration is replaced by a special iteration, *update iteration*, at which reliable regions and restraints are updated according to structural variations in the current population, followed by generation of 300 models following application of mutation operations to the top 10 models in the current population. The purpose of this special iteration is: a) to adapt the restraint set to the structure at the current iteration, and b) to introduce additional diversity through mutation operations. This special iteration occurs if more than 90% of members have served as seed parents, which generally happens at around 15-25 iterations.

Restraints used in the protocol

In the coarse-grained sampling in HybridizeMover, distance restraints are applied to residue Cβ pairs for those within distance d' in the input structure, where d' is a threshold of amino-acid-pair-specific Cβ distances (Cα for GLY) for the interacting residue pairs observed in general natural proteins [8]. A reasonable range of structural change is allowed to the input structure (even for a reliable part) rather than strictly preserving it: a flat-bottomed bounded function [9] is applied to the residue Cβ distances that begins to penalize if it gets larger than a reference distance of d'+2.0 Å (instead of the distance from input structure). Weight on each residue-pair restraint is determined depending on the purpose of the modeling: for *restrained_sampling* in diversification stage, a weight of 1.0 kcal/molÅ^2 is equally applied to all residue pairs. For *permissive_sampling* in the diversification stage, weights w_{ij} (in kcal/molÅ^2) are determined by the corresponding residue-level fluctuation δ_i:

$$w_{ij} = \min\left(1, p_{ij}/p_0\right),$$

$$p_{ij} = \frac{p(\delta_i)p(\delta_j)}{\sum_{i,j} p(\delta_i)p(\delta_j)}, \; p(\delta_i) = \exp\left(-k\left(\delta_i/\delta_0\right)\right) \quad \text{-- Eq 5}$$

Here, δ_0 is 30th percentile of the largest fluctuation, p_0 is 30th percentile of the largest p_{ij}, and k = 0.5. w_{ij} generally drops to 0 for a pair of residues including any residue from unreliable regions.

At the beginning of first and *update iteration* of the evolution stage (see above in the *Evolution stage* section), residue pair weights are re-learned from the current structural pool by measuring the distance deviations in the current structures from the input structure:

$$w_{ij} = \frac{1}{n}\sum_k \frac{1}{1+\left(\left(d_{ij,k} - d_{ij,0}\right)/\sigma\right)^2} \quad \text{-- Eq 6}$$

Here, $d_{ij,k}$ is the Cβ distance of residue i,j in the k-th structure in the population, $d_{ij,0}$ is the corresponding distance in the input structure, σ = 1.0, and n is the number of structures. Residue pairs showing greater deviations from the input structure have lower weight thus are allowed to move further in sampling with smaller penalties. Note that all the parameters in Eq 5 and 6 are arbitrarily assigned and could be further optimized.

Representative model selection by structural averaging

The basic concept of structural averaging simulation trajectory for the selection of a single representative model is explained elsewhere [10]; here we describe methodological differences from the previous studies. In contrast to conformational ensemble generated by restraining to a reference structure (as in previous studies), conformations produced by an iterative discontinuous search in this study may contain large structural variations. Therefore, we take a subset of the trajectory for structural averaging around the best scoring model (reference model) having sufficient structural similarity. Structural similarity threshold to the reference model is dynamically determined to get the optimal balance between the number of structures averaged and overall structural variation in the subset trajectory; initially set to S-score of 0.7, and is decreased by 0.1 until more than 20 similar conformations are found in the entire trajectory. Once the subset trajectory is collected, local fluctuations in residue Cα positions are measured, and structurally averaged only for the backbone atoms of the residues with Cα fluctuation less than 2.0 Å; for the remaining residues (with fluctuation greater than 2.0 Å) backbone coordinates are brought from the reference model. Side-chain optimization and minimization are followed with strong harmonic restraints (10 kcal/mol $Å^2$) on the backbone coordinates.

Computational cost

The majority of the computational cost is used for running mutation or crossover operations within HybridizeMover. Each mutation or crossover operator (including both coarse-grained and all-atom modeling) takes about 4 minutes for a protein with length of 100 residues (using single Intel E5-2650 core at 2.0GHz). Running the entire diversification and evolution stages for a single target requires approximately 10,000 such operations, or about 700 CPU hours, which translates into ~12 hours running in parallel using 64 cores, for a 100-residue protein.

CASP12 targets and protocol

Blind predictions with the protocol are made for the targets from two categories in CASP12. In the refinement category (TR) the best server model is selected as the starting model by the organizers and asked for further refinement. To decide whether to apply the protocol reported in this study, we used the rule for the input structure quality as described in the Benchmark section above, but extended the protein size limit to 400 residues, and also removed the condition on biological units; this was to evaluate the method for the cases more challenging than the benchmark targets. For the remaining targets, a high-resolution protocol is applied running mutation operations focusing on the unreliable regions (results not reported in this paper). 24 of 42 TR targets in CASP12 were tested with the protocol (the remaining with the high-resolution protocol) and submitted by the group name "BAKER".

According to the assessment, at least 12 of these 24 proteins refined by the protocol form hetero- or homo-oligomers; of these, 3 oligomers with large oligomeric contacts (TR862 and TR884, TR875 native not deposited yet but likely to have strong dimer interface) and 1 transmembrane protein (TR945**)**, were clearly not suitable to refine by the protocol (colored gray in **Figure S4a)**, i.e. the native monomeric structure may not have the lowest energy. Modeling conditions for the remaining 20 cases were also quite challenging compared to the benchmark targets; 3 had native conformation too distant from the starting model to be refined (starting GDTHA <= 30 and starting SG <= 40, TR869, TR870, TR898), and 5 were large proteins (TR694 263 residues, TR901 223 residues, TR905 242 residues, TR928 381 residues, TR942 387 residues).

In the tertiary structure prediction category (TS), we tested a fully automated, simplified version of the protocol at the final refinement stage of the homology modeling pipeline on the Robetta server, and submitted the resulting model as one of the models for group name "BAKER-ROSETTASERVER". We applied the protocol if the expected homology model quality [4] was not greater than 0.6 and the protein length was not greater than 200 residues. Among 57 domains that fall into the template-based modeling category (TBM), this protocol was applied to 20 domains, but excluding T0896-D1 from the analysis for which the models sampled had a completely different fold from the native (due to incorrect template selection). There are two main changes in the protocol for the TS category introduced for efficiency. Instead of running the diversification stage on a single input structure, models generated by different templates or from *de novo* models in preceding stages in the server pipeline were clustered and served as the starting population of the evolution stage; these structures generally had more diversity than those generated by the diversification stage in the standard protocol. The evolution stage was also simplified to finish the whole process within submission deadline, by running 30 iterations and generating 60 structures per iteration which uses only 30% of computation time required for the standard protocol.

Molecular replacement

Phaser [11] 2.7.1 version in the Phenix software suite [12] version dev-1616 was used for testing crystallographic molecular replacement (MR). A single input model before refinement and 5 refined models were tested for MR suitability. Initial model accuracy required for Phaser is estimated by setting RMSD=1.2 Å uniformly, and B-factors were uniformly set to 30. MR_RNP mode was applied to the model structure superimposed into the native crystal coordinate for the cases with single asymmetric unit (ASU). If more than one ASU exist, MR_AUTO mode is run with the correct number of asymmetric units assigned. 11 targets were selected from a total of 20 TR targets (listed in **Table 3**), excluding 4 cases forming a multi-domain protein for which a full-length high quality homology model is available at any other domain, and 5 cases for which crystallographic diffraction data are not available, such as structures determined by NMR. Of

these 11 targets tested, TR780 and TR280 are separate domains combined to form a full chain, hence 5 models from each domain are selected and combined (a total of 25 combinations) for MR on the whole protein and reported as a single target, instead of testing and reporting individually.

Instructions to run the refinement pipeline

Running the pipeline requires a compiled version of the Rosetta suite release version 3.9 or later. The overall iterative process is carried out by a master python script, which manages the system calls of a series of Rosetta executables, such as the Rosetta hybridization mover through an xml script and other Rosetta public apps for tasks such as model selection, clustering, and structural averaging.

The whole package containing the master python script and various files required for the pipeline is available in the Rosetta repository. Detailed instructions can be found in the Rosetta documentation webpage:

https://www.rosettacommons.org/docs/latest/IterativeHybridize

References

1. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 95-202.

2. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2016) Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* 84 Suppl 1:4–14.

3. Haas J, et al. (2013) The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database* 2013:bat031.

4. Song Y, et al. (2013) High-resolution comparative modeling with RosettaCM. *Structure* 21(10):1735–1742.

5. Wallner B, Elofsson A (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci* 15(4):900–913.

6. Lee J, Scheraga HA, Rackovsky S (1997) New optimization method for conformational energy calculations on polypeptides: conformational space annealing. *J Comput Chem* 18(9):1222–1232.

7. Genetic algorithms in search, optimization, and machine learning (1989) *Choice Reviews Online* 27(02):27–0936–27–0936.

8. Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proc Natl Acad USA* 110: 15674-15679.

9. Kim DE, Dimaio F, Yu-Ruei Wang R, Song Y, Baker D (2014) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* 82 Suppl 2:208–218.

10. Park H, DiMaio F, Baker D (2015) The origin of consistent protein structure refinement from structural averaging. *Structure* 23(6):1123–1128.

11. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40(Pt 4):658–674.

12. Adams PD, et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):213–221.