Automated Real-Space Refinement of Protein Structures Using a Realistic Backbone Move Set

Esmael J. Haddadian,[†] Haipeng Gong,[†] Abhishek K. Jha,[†‡§] Xiaojing Yang,[†] Joe DeBartolo,[†] James R. Hinshaw,[‡§] Phoebe A. Rice,[†] Tobin R. Sosnick,[†¶] and Karl F. Freed[‡§‖]

[†]Department of Biochemistry and Molecular Biology, [‡]Department of Chemistry, [§]James Franck Institute, [¶]Institute for Biophysical Dynamics, and [‖]Computation Institute, University of Chicago, Chicago, Illinois

Supplemental Methods

Statistical potentials

Four energy functions are employed, including our TSP along with native contact energy (NCE), a metric for the similarity to the input reference structure, a hydrogen bond potential (HB) (1), the repulsive portion of the C_β-level statistical potential (2) (called rDOPE-C_β) that is designed to prevent steric clashes, and a neighbor-independent torsional statistical potential TSP1. Each substage employs a slightly different combination of energy functions (E) as follows

$$E_I = 5 \times E_{TSP1} + 20 \times E_{TSP} + 2 \times E_{NCE}(\upsilon = 1) + 50 \times E_{HB}, \quad (1)$$

$$E_{II} = 10 \times E_{TSP1} + 10 \times E_{TSP} + 1 \times E_{NCE}(\upsilon = 0) + 100 \times E_{HB} + E_{rDOPE-C_\beta}, \quad (2)$$

$$E_{III} = 10 \times E_{TSP1} + 20 \times E_{TSP} + 1 \times E_{NCE}(\upsilon = 0) + 200 \times E_{HB} + E_{rDOPE-C_\beta}, \quad (3)$$

The NCE energy is defined as

$$E_{NCE}(\upsilon) = \sqrt{\sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\bar{b}^2}{b_i b_j} \cdot \frac{(d_{ij} - d_{ij}^0)^2}{(d_{ij}^0)^\upsilon}}, \quad (4)$$

where d_{ij} and d_{ij}^0 are the distances between the α carbons of residues i and j for the current conformation and for the reference conformation (e.g., the initial model), respectively. N is the total number of residues, and the adjustable parameter υ controls the relative weights of contributions from local and non-local separations. The b_i and b_j are the crystallographic temperature B factors for the C_α carbons of residue i and j, while $\bar{b}$ is the average B factor for all C_α carbons.

To optimize the backbone hydrogen bonds, a modified form of the geometry-dependent hydrogen bond potential of Kortemme et al. (1) is used. The DOPE-C_β statistical potential is a distance dependent potential which has been derived from high-resolution PDB structures. Here, only the repulsive terms in DOPE-C_β are retained to mimic a soft-sphere potential and prevent steric clashes. The attractive terms from DOPE-C_β are unnecessary as the NCE term maintains the backbone close to the starting structure.

Structure refinement against the electron density. Our real space refinement is performed with respect to the weighted 2mFo-DFc maps in one asymmetric unit. The program Phenix (3) is used to calculate the 2mFo-DFc maps in CNS file format, choosing the starting structure and reflections as the inputs. These maps are then converted to the Situs file format using the program Situs 2.5 (4), and the maps are then converted to a 3-D potential in NAMD in the DX file format using a grid spacing of 1 Å. If present, the electron density of any ligands is removed (including the DNA in the DNA-binding protein). The electron densities are retained for all explicit bound water molecules in the crystal structures. Secondary structures are maintained by constraining the backbone hydrogen bonds and dihedrals angles to the initial values of the first stage backbone-refined model. Peptide bonds are restrained to their pre-existing *cis* or *trans* configurations. All chiral centers are also restrained to their

original handedness. The file preparations are performed using the program VMD 1.8.8a1 (5).

The protein structures (without ligands) plus their explicit bound water molecules are energy minimized for 20,000 steps within a uniform solvent having a dielectric constant of 80, along with an extra energy term dependent on the overlap between the electron density and the model (6). Explicit water molecules are not moved during minimization. We modified the Charmm force field parameters to maintain the backbone bond lengths and angles closer to their ideal values by using two- to five-fold stronger force constants and slightly shifted mean positions. The relative weighting for the MDFF energy function is adjusted to optimize two competing metrics, ideal bond lengths and angles versus agreement with the density. Better agreement with the electron density improves the crystallographic R factors. During the initial stages of model building, maintaining an ideal bond lengths and angles may be more important than the detailed fit to the electron density because subsequent adjustments improves the fit to the density.

Disulfide brides are ignored in Stage 1. In Stage 2, they are handled according to standard molecular dynamics protocol, which puts in an explicit disulfide bond term (and can recover a bridge broken in Stage 1, if it existed in the input structure).

Calculation of crystallographic metrics.

Both manual intervention and TOP are real space refinements to the electron densities, usually using the 2mFo-DFc map; except that the former is a (tedious) residue-by-residue method, while the latter is a fully automated global procedure incorporating knowledge-based restraints. Moreover, manual refinement can be applied after the TOP procedure for further structural improvements. This extra step is not performed here because our goal is to assess the automated process although one example is provided to illustrate the possible improvement to a TOP-refined structure.

The R-work, R-free, and map correlation indices for the starting and ending structures are calculated using the program Phenix 1.7 (3). For each diffraction data set, 5% of the total observed reflections are randomly chosen and set aside for calculating the R-free for cross-validation. Phenix is applied to determine the bulk solvent correction and the temperature (B) factor refinement including individual atomic displacement parameters (ADP) and TLS, starting from randomized B values to remove bias (allowing Phenix to recalculate the B values). The TLS groups are assigned either by chain ID or by Phenix default (for 2E74.pdb). No positional refinements are involved. The same Phenix protocol is used to calculate the R-work, R-free, and map correlation indices for TOP's two stages, using the same initially assigned reflections.

Comparison to other methods

We compare TOP to the Phenix program implementing the "discard_psi_phi=false" option which keeps the dihedral angles restrained according to the CCP4 monomer library definitions. As a representative application, we chose a 1001 aa α/β protein at 2.85 Å

resolution currently undergoing refinement. The TOP & Phenix algorithms, respectively, increase the number of ϕ,ψ angles in the preferred region (according to COOT) from 76% → 87% & 81%; the average TSP score improves from 3.59 → 1.13 & 3.53, while the number of hydrogen bonds increases from 354 → 398 & 351 (Table S4). Only the TOP algorithm generates a map with distinct β and PPII basins, as is observed in high resolution structures (Fig. S6).

The CNS program is capable of constraining a given ϕ,ψ pair to stay near a user-specified value during its refinement procedure. Because *a priori* knowledge of the proper angles across the entire sequence is unfeasible (except in cases of molecular replacement where the model in question has previously been refined at high resolution), CNS's capabilities are not comparable to those of the TOP procedure.

The Coot program offers a real space refinement feature with torsional restraint option based on secondary structure. However, we find that it leads to locally distorted structures in regions with poor electron density and it is intended for segments shorter than 20 aa. To test Coots' regularize/real space refinement capability, we apply it to two regions of a 3.4 Å resolution protein at an early stage of refinement and obtain the following results. A 44 aa curved helix is refined/regularized in three separate segments due to the length limit. With "Use torsion restraints" and "alpha-helix restraints" options, the helix becomes straight and fails to fit into the electron density. For real space refinement against low-resolution map with poor side chain density such as in our starting model, Coot forces atoms of long side chains into the main chain densities, or even into densities of neighbor secondary structures. When Coot is applied to a second region, a 22 aa anti-parallel sheet-loop-sheet, the resulting densities in the 9 aa connecting loop are poor. Side chains atoms are completely misplaced after real space refinement and overlap with the main chain atoms, as observed with the curved helix. Consequently, the main chain hydrogen bonds are destroyed.

In the Program O, the Lego_loop tool allows the user to pick endpoints of a loop and scroll through possible options "rented" from a library of fragments from high-resolution structures. This procedure differs from our method in that it is manual, local, and relies on the existence of suitable fragments. Some regions of the protein may be amenable to this tool but not the entire protein.

Although the 'Backrub move" (7) appears similar to our double crank move, it is in fact entirely different, e.g., it does not involve ϕ,ψ moves taken from a library of values in non-redundant high-resolution PDB structures. Rather, it focuses on rigid body motions about two Cα-Cα atoms in two residues separated by 1-11 amino acids with a specific side chain compensation (e.g., see Figure 1 in (7)).

References

1. Kortemme, T., A. V. Morozov, and D. Baker. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J. Mol. Biol. 326:1239-1259.

2. Fitzgerald, J. E., A. K. Jha, A. Colubri, T. R. Sosnick, and K. F. Freed. 2007. Reduced Cbeta statistical potentials can outperform all-atom potentials in decoy identification. Protein Sci. 16:2123-2139.
3. Adams, P. D., R. W. Grosse-Kunstleve, L. W. Hung, T. R. Ioerger, A. J. McCoy, N. W. Moriarty, R. J. Read, J. C. Sacchettini, N. K. Sauter, and T. C. Terwilliger. 2002. PHENIX: building new software for automated crystallographic structure determination. A. Crys. D Biol. Crys. 58:1948-1954.
4. Wriggers, W. 2010. Using Situs for the integration of multi-resolution structures. Biophys. Rev. 2:21-27.
5. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. J. Mol. Graph. 14:33-38, 27-38.
6. Trabuco, L. G., E. Villa, K. Mitra, J. Frank, and K. Schulten. 2008. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure 16:673-683.
7. Smith, C. A., and T. Kortemme. 2008. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 380:742-756.
8. Emsley, P., B. Lohkamp, W. G. Scott, and K. Cowtan. 2010. Features and development of Coot. Acta Crys. D Biol. Crys. 66:486-501.

Table S1. Different moves and the resulting effects on the overall ubiquitin fold

Move type	*Example move*	Location: helix (1ubq.pdb)	RMSD to original structure
Pivot move on 1 aa (single ϕ,ψ change)	$(\phi_i,\psi_i)\rightarrow(\phi_i+70,\psi_i+17)$	K27:(**-134,-52**) →(**-64,-35**)	4.9 Å
Single-crank move on 2 as's: i-1,i	$(\psi_{i-1},\phi_i)\rightarrow(\psi_{i-1}-70,\phi_i+70)$	V26: (-98,**30**) →(-98,**-40**) K27:(**-134**,-52) →(**-64**,-52)	0.8 Å
Single-crank move on 2 as's: i,i+1	$(\psi_i,\phi_{i+1})\rightarrow(\psi_i+17,\phi_{i+1}-17)$	K27:(-134,**-52**) →(-134,**-35**) A28:(**-58**,-46) → (**-75**,-46)	0.2 Å
Double-crank move on 3 as's: i-1,i,i+1	$(\psi_{i-1},\phi_i)\rightarrow(\psi_{i-1}-70,\phi_i+70)$ $(\psi_i,\phi_{i+1})\rightarrow(\psi_i+17,\psi_{i+1}-17)$	V26: (--,**30**) →(--,**-40**) K27:(**-134,-52**) →(**-64,-35**) A28:(**-58**,--) → (**-75**,--)	0.7 Å

Table S2. TOP Structure refinement

Protein		**APC22750**					
		Initial	TOP[1]		Initial	TOP	
			Stage 1	Stage 2		Stage 1	Stage 2
Resolution		2.09-25.0 Å			2.09-25.0 Å		
Number of residues		465			480		
Starting model		During refinement			Deposited (1VR4)		
C_α-RMSD (Å)		N/A	0.71	0.42	N/A	0.46	0.14
<TSP>		3.31	-0.4	0.08	0.45	-1.2	-0.9
no. of hydrogen bonds[2]		162	190	214	239	249	263
R-work **R-free**		0.3091 0.3537	0.3880 0.4233	0.2979 (0.2983)[3] 0.3163 (0.3403)	0.2061 0.2647	0.3150 0.3507	0.2087 (0.2074)[3] 0.2372 (0.2589)
Map Correlation		0.76	0.66	0.77	0.85	0.76	0.85
RMSD from ideal	**Bond length (Å)** **Angle (°)**	0.045 2.567	0.047 2.904	0.040 3.130	0.014 1.692	0.017 1.935	0.016 1.810
Ramachandran Map statistics (%)							
TSP-Favored: [-6,0) **TSP-Allowed**: [0,5) **TSP-Generously-allowed:**[5,10) **TSP-Scarce: ≥ 10**		31 16 10 34	76 7 2 16	67 13 4 16	59 21 8 11	80 10 4 6	73 16 5 7
Preferred[4] **Allowed** **Outliers**		79 6 15	88 3 9	88 4 8	92 5 3	92 5 3	93 4 3
MolProbity Evaluation							
Clashscore, all atoms		98	182	25	18	87	6
Poor Rotamers		8%	8%	10%	14%	14%	5%
Ramachandran Outliers		11%	7%	7%	1%	1%	1%
Ramachandran Favored		81%	88%	89%	94%	93%	95%
Cβ Deviations > 0.25 Å		0	0	15	3	0	4
MolProbity Score		3.86	3.98	3.23	3.05	3.71	2.21
Residues with bad bonds		0.00%	0.00%	0.43%	0.21%	0.21%	0.00%
Residues with bad angles		0.86%	0.86%	1.08%	0.21%	0.21%	0.00%

[1] Stage 1 and Stage 2 refer to backbone refinement using MCSA/double-crank algorithm and all-atom energy minimization using the electron density, respectively.
[2]Backbone hydrogen bonds are defined when the amide nitrogen and carbonyl oxygen are within 3.5 Å and the angle between the N-H and O=C bond vectors exceeds 145°.
[3] The values in parentheses are the R and R-free values calculated where the more stringent maps generated after excluding the free reflections are used in the real space refinement stage of TOP.
[4] As defined by the program COOT (8).

Table S2. TOP Structure refinement (cont.)

Protein		CYTOCHROME b_6f COMPLEX			A-COBRATOXIN-ACHBP COMPLEX		
		Initial	TOP[1]		Initial	TOP[1]	
			Stage 1	Stage 2		Stage 1	Stage 2
Resolution		3.00- 39.30Å			4.20-25.0 Å		
Number of residues		959			1356		
Starting model		Deposited (2E74)			Deposited (1YI5)		
C_α-RMSD (Å)		N/A	0.78	0.40	N/A	0.8	0.62
<TSP>		2.92	-0.9	-0.7	4.35	0.09	0.68
no. of hydrogen bonds[2]		410	445	468	332	481	513
R-work **R-free**		0.2248 0.2704	0.3515 0.3760	0.2423 (0.2395)[3] 0.2726 (0.2802)[3]	0.2529 0.3128	0.3404 0.3863	0.2531 (0.2506)[3] 0.2857 (0.3107)[3]
Map Correlation		0.8	0.7	0.75	0.68	0.59	0.68
RMSD from ideal	**Bond length (Å)** **Angle (°)**	0.029 2.659	0.039 3.162	0.035 2.831	0.012 1.579	0.036 2.125	0.026 2.093
Ramachandran Map statistics (%)							
TSP-Favored: [-6,0) **TSP-Allowed**: [0,5) **TSP-Generously-allowed:**[5,10) **TSP-Scarce: ≥ 10**		41 18 15 27	76 10 5 10	72 13 5 11	31 17 18 34	67 15 8 10	59 21 9 11
Preferred [4] **Allowed** **Outliers**		83 11 6	91 5 4	91 6 4	85 10 5	91 6 2	91 6 3
MolProbity Evaluation							
Clashscore, all atoms		55	140	14	17	121	11
Poor Rotamers		9	9	7	17%	17%	7
Ramachandran Outliers		3	2	2	3%	2%	2
Ramachandran Favored		84	93	93	87%	93%	94
Cβ Deviations > 0.25 Å		9	0	2	4	0	8
MolProbity Score		3.60	3.78	2.73	3.28	3.93	2.62
Residues with bad bonds		0.00%	0.00%	0.00%	0.00%	0.00%	0.44
Residues with bad angles		0.94%	0.94%	0.52%	0.37%	0.37%	1.03

Table S2. TOP Structure refinement (cont.)

Protein		A-DNA-BINDING PROTEIN			A-PROTEIN-LIPID COMPLEX		
		Initial	TOP[1]		Initial	TOP[1]	
			Stage 1	Stage 2		Stage 1	Stage 2
Resolution		3.35-20.0			2.60-43.5 Å		
Number of residues		364			376		
Starting model		Early stage			Final stage (3OV6)		
C_α-RMSD (Å)		N/A	0.85	0.97	N/A	0.74	0.32
<TSP>		3.33	-2.0	-1.5	2.92	-0.1	0.13
# of hydrogen bonds[2]		147	187	192	168	181	193
R-work **R-free**		0.2890 0.3722	0.3874 0.4418	0.3136 (0.3079)[3] 0.3562 (0.3830)	0.2285 0.2851	0.3611 0.4076	0.2323 (0.2341)[3] 0.2551 (0.2760)
Map Correlation		0.75	0.68	0.74	0.88	0.78	0.86
RMSD from ideal	**Bond length (Å)** **Angle (°)[3]**	0.007 1.218	0.014 1.689	0.016 1.837	0.009 1.401	0.026 2.147	0.017 1.944
Ramachandran Map statistics (%)							
TSP-Favored: [-6,0) **TSP-Allowed**: [0,5) **TSP-Generously-allowed:**[5,10) **TSP-Scarce: ≥ 10**		38 18 16 28	86 8 2 5	75 16 4 5	42 19 18 22	69 15 8 8	64 19 7 10
Preferred[4] **Allowed** **Outliers**		81 11 8	94 3 3	94 3 3	93 5 1	94 4 2	95 3 2
MolProbity Evaluation							
Clashscore, all atoms		29	119	14	65	248	5
Poor Rotamers		6%	5%	13 %	8%	38%	4%
Ramachandran Outliers		1%	1%	2%	4%	2%	1%
Ramachandran Favored		95%	96%	96%	82%	95%	96%
Cβ Deviations > 0.25 Å		2	0	5	1	0	2
MolProbity Score		2.89	3.31	2.82	3.68	4.38	1.99
Residues with bad bonds		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Residues with bad angles		0.80%	0.80%	0.00%	0.27%	0.27%	0.80%

Table S3. TOP Structure refinement[1]

Protein			PaBphP-PCD			
Resolution Number of residues Starting model			2.60-49.0 Å 3827 Final stage			
			Initial	TOP[1]		Manual Refinement After Top
				Stage 1	Stage 2	
C_α-RMSD (Å)			N/A	0.73	0.33	0.3
<TSP>			2.91	-0.6	-0.2	1.28
# of hydrogen bonds[2]			1352	1814	1909	1754
R-work R-free			0.2244 0.2820	0.3551 0.3854	0.2338 (0.2375)[3] 0.2567 (0.2803)[3]	0.2198 0.2613
Map Correlation			0.80	0.69	0.80	0.82
RMSD from ideal		Bond length (Å) Angle (°)	0.004 0.946	0.017 1.894	0.016 1.893	0.005 0.973
Ramachandran Map statistics (%)	TSP-Favored: [-6,0) TSP-Allowed: [0,5) TSP-Generously-allowed:[5,10) TSP-Scarce: ≥ 10		41 19 15 25	74 11 4 11	70 15 4 11	54 18 12 16
	Preferred[4] Allowed Outliers		90 8 2	94 4 2	94 4 2	95 5 0.3
MolProbity Evaluation	Clashscore, all atoms Poor Rotamers Ramachandran Outliers Ramachandran Favored Cβ Deviations > 0.25 Å MolProbity Score Residues with bad bonds Residues with bad angles		37 11% 1% 92% 2 3.32 0.00% 0.50%	133 11% 2% 94% 0 3.77 0.00% 0.52%	7.3 9% 1% 95% 28 2.47 0.03% 1.36%	22 8 0.2 96 3 2.83 0.00% 0.52

Table S4. Comparison between TOP and Phenix applied to a bacteriophytochrome

Structure	Rwork/Rfree	Ramachandran Statistics[1]	<TSP>	H-bonds
Starting model	0.266/0.305	76/11/13	3.59	354
Post TOP	0.251/0.289	87/6/7	1.13	398
Post Phenix using discard Phi_Psi =False[2]	0.244/0.300	81/10/9	3.53	351
Post Phenix using discard Phi_Psi =true[2]	0.245/0.300	79/11/1	3.79	356
Phenix Autobuild	0.279/0.364	71/15/14	4.64	334

[1]As defined by the program COOT (Preferred /Allowed/Outliers).
[2]Using Wxc_scale=0.5 and Wxu_scale=2.0

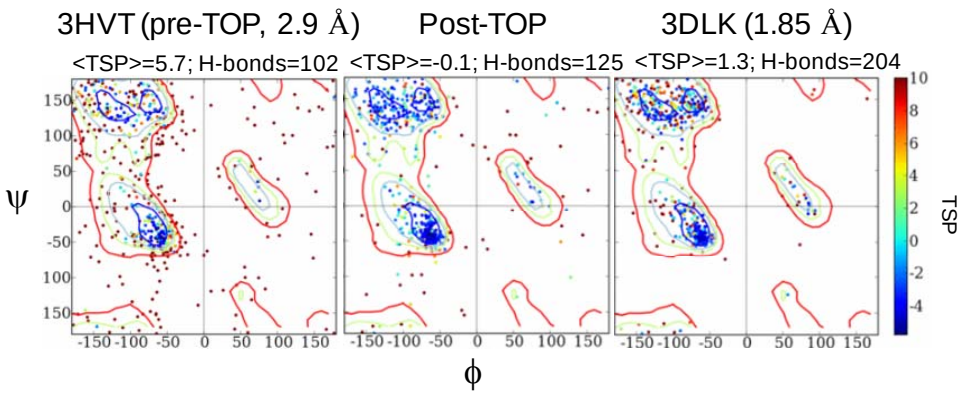


Figure S1. TOP selects native-like angles. Starting from a low resolution crystal structure of HIV reverse transcriptase (3HVT), the first, backbone-only refinement, stage of TOP selects angles that on average are closer to those observed in a medium resolution crystal structure, as illustrated with RamaMaps.

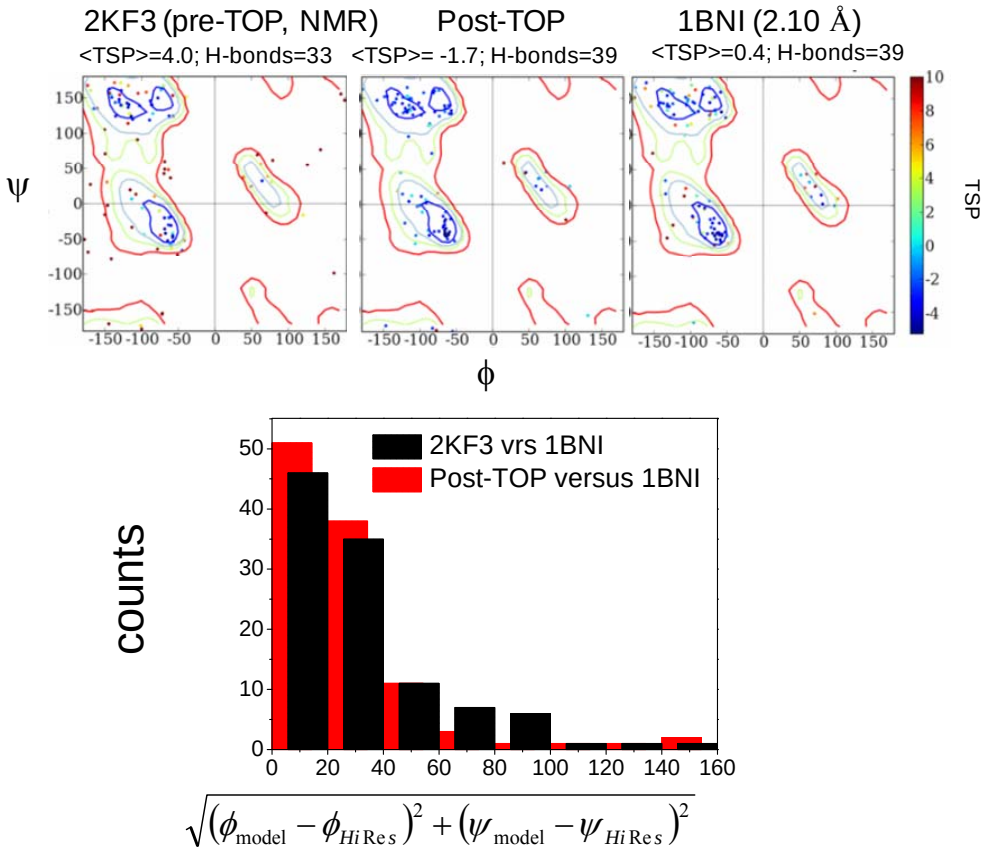


Figure S2. TOP selects native angles. Starting from an NMR structure of barnase (2KF3), the first, backbone-only refinement, stage of TOP selects angles that on average are closer to those observed in a medium resolution crystal structure, as illustrated with Ramachandran maps and histograms.

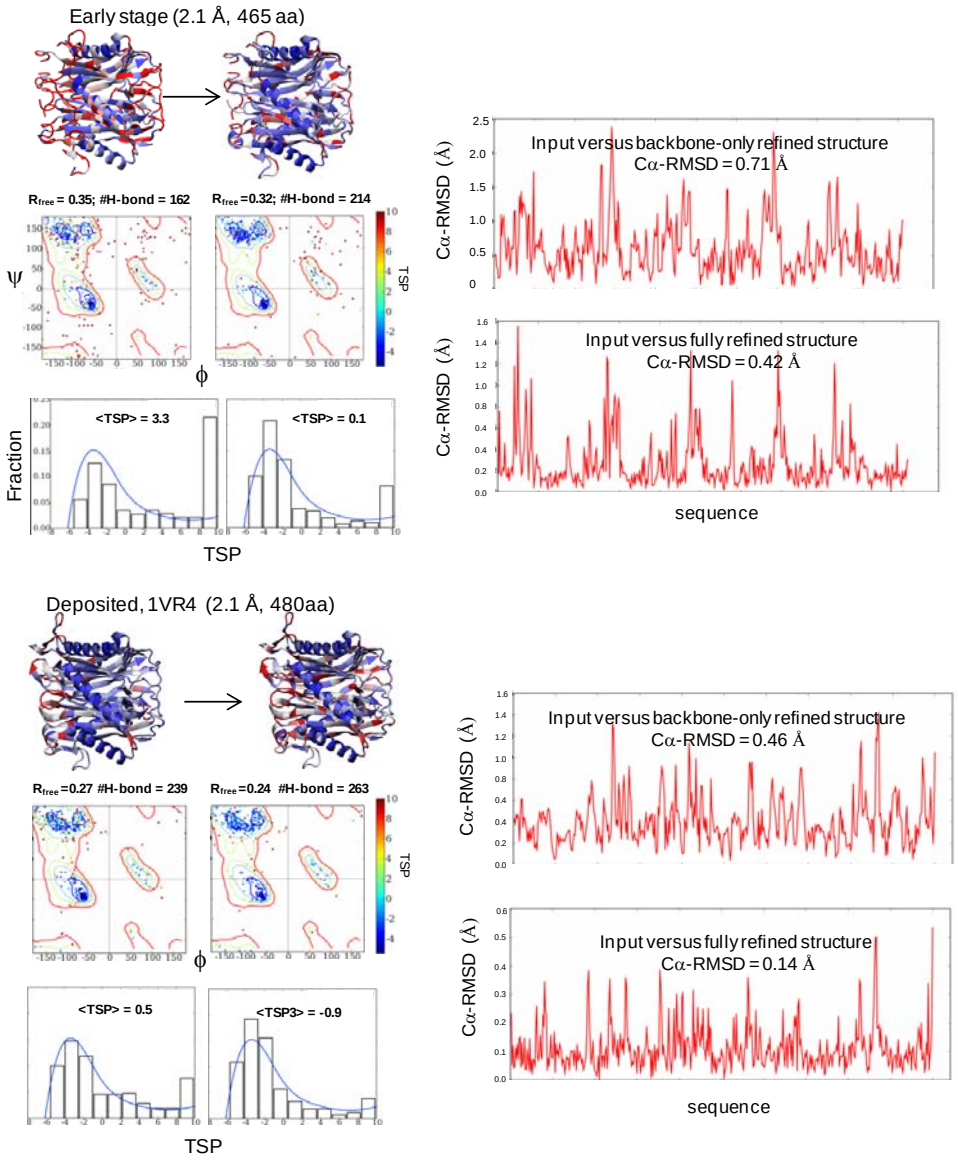


Figure S3. TOP is applied to APC22750 at an early stage (465 aa) and late stage (480 aa) of refinement. Right side: The backbone moves during the torsional refinement stage but returns closer to the initial structure during the real space refinement using the electron density. The C_α-RMSD between the initial and final refined structures is 0.42 and 0.14 Å when starting from the early stage and the deposited structures, respectively. Variability exists across the protein, and there are regions with poor TSP scores that move by up to ~2 Å during the refinement of the early stage structure. But most displacements are under 0.5 Å after refinement against the electron density. Displacements starting from the deposited structure are significantly less.

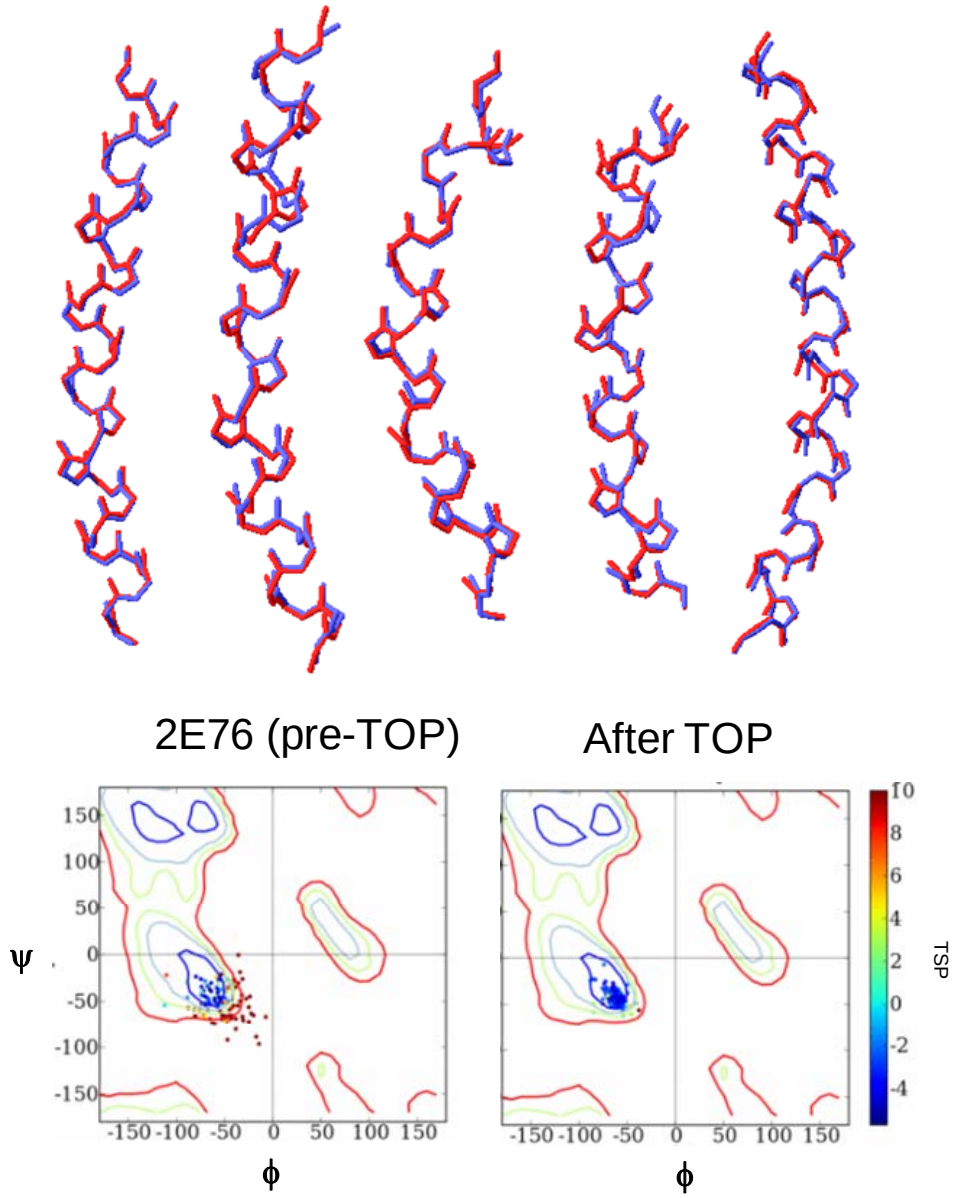
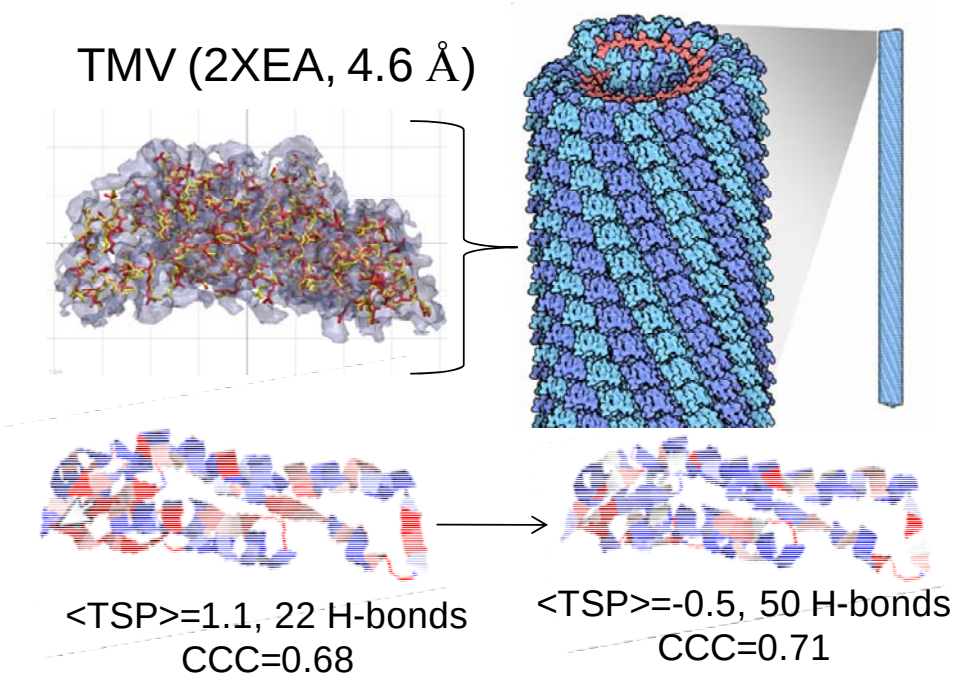


Figure S4. Testing of applicability to membrane proteins. The five kinked helices present in 2E76 (blue) are superimposed onto the helices produced by TOP (red) for residues in Chain A, 79-105; Chain B, 94-116; Chain D, 13-42; Chain F, 3-29; Chain H, 3-25.



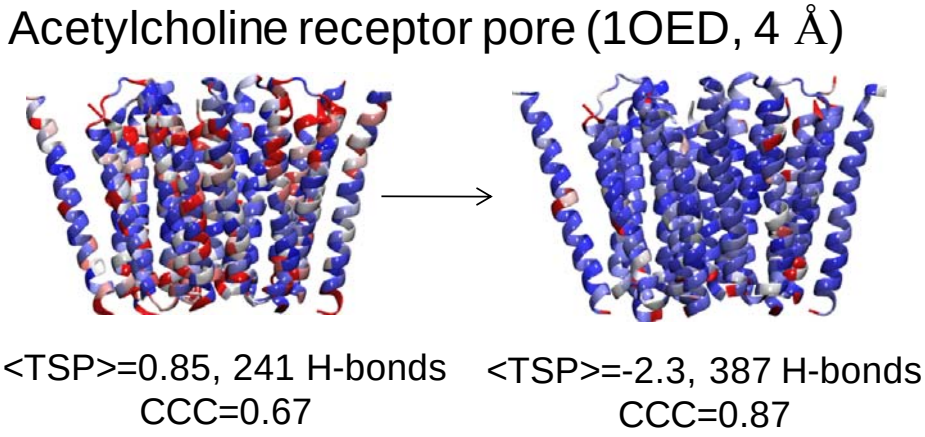


Figure S5. Improved cryoEM structures for TMV and Acetylcholine receptor pore using TOP. Coloring of structures reflects TOP score (see Figure S1, red =10, blue = -6).

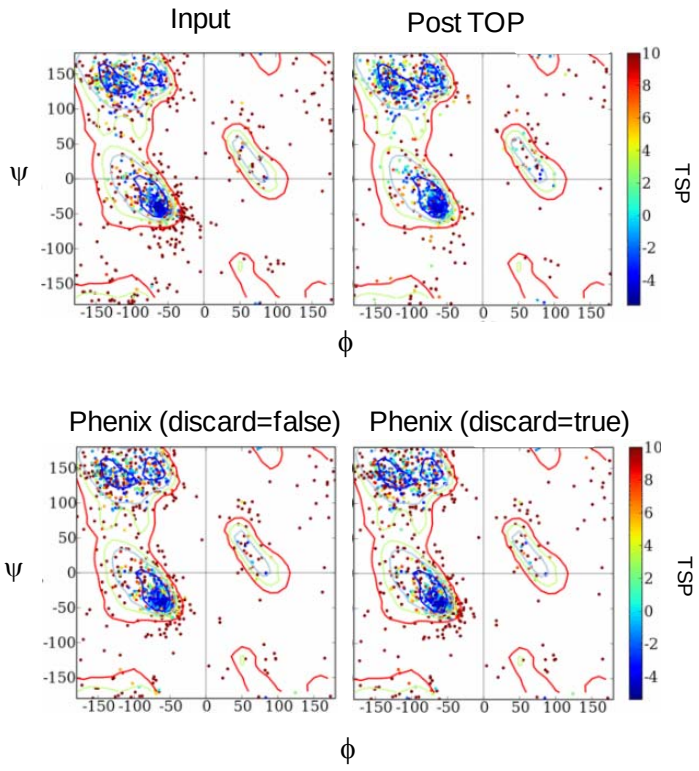


Figure S6. The Ramachandran map of a 1001 aa α/β protein is noticeably improved after application of TOP as compared to that generated by Phenix using the discard_phi_psi option=false or true options (see Table S4).