Biophysical Journal, Volume *112*

Supplemental Information

Flexible Fitting of Atomic Models into Cryo-EM Density Maps Guided by Helix Correspondences

Hang Dou, Derek W. Burrows, Matthew L. Baker, and Tao Ju

Supporting Material

A. Laplacian-based surface deformation

A common problem in computer graphics is how to deform a surface so that a subset of its vertices (called *handles*) go to their pre-defined locations (called *targets*) while the rest of the surface maintains its shape as much as possible. This is useful in interactive character animation where the user can control the deformation of the characters by dragging a few handles. Specifically, consider a triangulated surface mesh of n vertices $\{v_1, \dots, v_n\}$. Let $\{h_1, \dots, h_m\}$ be the indices of $m(< n)$ handle vertices whose target locations are $\{t_1, \dots, t_m\}$. The goal is find deformed locations of each vertex, $\{v'_1, \dots, v'_n\}$ (see Figure S1a).

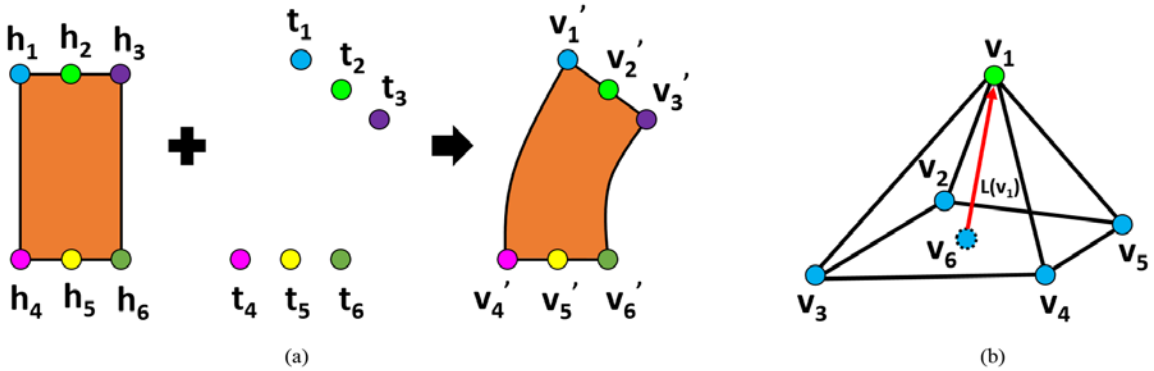


Figure S1. Illustration of Laplacian-based deformation. (a) Given handle points (h_i) and corresponding target points (t_i), the original vertices are transformed into deformed locations (v_i'). (b) Vertex v_1's Laplacian vector $L(v_1)$ is the vector (red) from vertex v_6 (mean of v_1's neighboring vertices) to vertex v_1.

Laplacian-based deformation solves the problem by minimizing the following energy,

$$E = w_{fit}E_{fit} + w_{shape}E_{shape}\ , \tag{1}$$

where E_{fit} and E_{shape} respectively measures the deviation of handles from the targets and the distortion of the shape, and w_{fit} and w_{shape} are scalar weights. Specifically, the fitting term measures the squared Euclidean distances between the handles and targets,

$$E_{fit} = \sum_{i=1}^{m} \left\| v'_{h_i} - t_i \right\|_2^2 \ . \tag{2}$$

The shape term E_{shape} measures the change in the local geometry after deformation. The local geometry at each vertex v_i is defined by the linear Laplacian operator L (see Figure S1b), which is the vector from the centroid of v_i's neighboring vertices to v_i:

$$L(v_i) = v_i - \frac{1}{|N_i|}\sum_{j \in N_i} v_j \tag{3}$$

Here, N_i denotes the indices of those vertices that are connected to v_i by some triangle edge. The shape term is expressed as the squared difference between the original and deformed Laplacian vectors,

$$E_{shape} = \sum_{i=1}^{n} \|L(v'_i) - T_i L(v_i)\|_2^2 \quad . \quad (4)$$

Since the Laplacian is not invariant under scaling and rotation, the transformation T_i estimates the scaling and rotation of local neighborhood of v_i after deformation. There are many ways to compute T_i, one of which (that we adopt) is to express it as the minimizing transformation,

$$T_i = \text{argmin}_T \left(\|v'_i - T v_i\|_2^2 + \sum_{j \in N_i} \left\|v'_j - T v_j\right\|_2^2 \right) \quad , \quad (5)$$

which in turn can be approximated as a linear expression of the unknowns, v'_i (see (1) for details). The resulting shape term (E_{shape}) approximately measures the amount of non-linear distortion to the original surface due to the deformation.

The combined energy (E) is a quadratic form of the unknowns ($\{v'_1, \ldots, v'_n\}$), and hence has a global minimum that can be found by solving a system of linear equations (see (1) for details). Such a system can be solved efficiently using tools such as Matlab and Eigen (2).

B. Fitting weight in helix-guided fitting stage

Figure S2 shows the fitting results with varying w_{fit}, while w_{shape} is fixed to 1.0. Observe that if w_{fit} is too small (a), the shape term dominates the energy function and there is not enough flexible to achieve the desired deformation. Good results are obtained in this example for $w_{fit} > 0.5$, and the fitting does not change significantly with larger values of w_{fit} (Figure S2 (c) and (d) are almost the same). In our experiments, we observed that setting both the fitting weight (w_{fit}) and shape weight (w_{shape}) both to 1.0 achieve good results in all our test proteins.

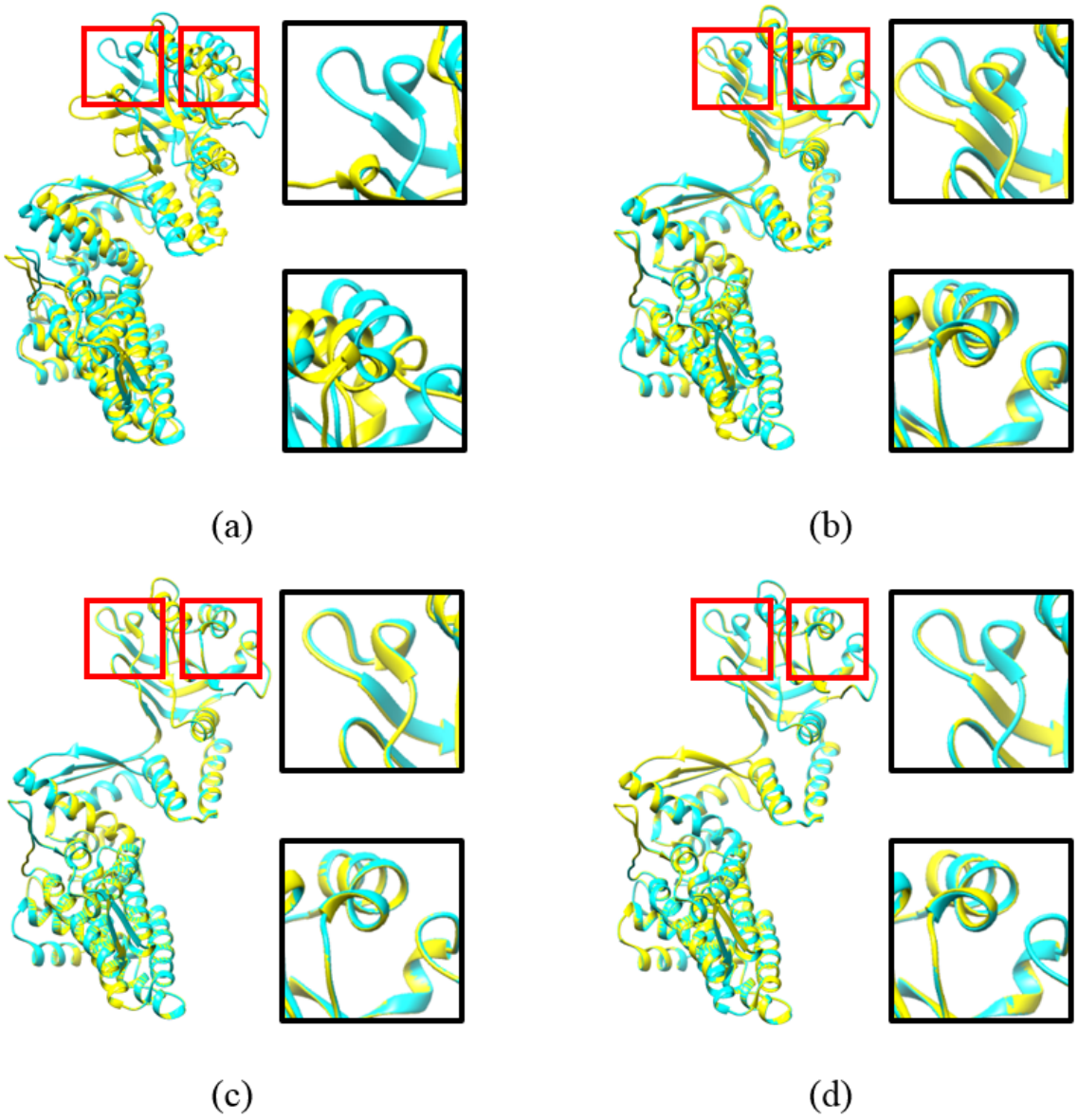


Figure S2. Helix-guided fitting of GroEL (protein PDB ID 2C7C chain M to EMDB Map ID 1180) with $w_{shape} = 1.0$ and w_{fit} set to 0.05 (a), 0.3 (b), 0.7 (c), and 10.0 (d). The ground-truth model is shown in cyan and the fitted model is shown in yellow. (c) and (d) are expected to be similar, which shows that the fitting result does not change significantly with w_{fit} larger than 1.0.

C. Additional tables and figures for the results

Data set	Our method time in seconds				
	Helix-guided	Skeleton-guided		All atoms	Total
		Iterations	Time		
Adenylate kinase	0.127	7	1.267	0.01	1.404
Triacylglycerol acylhydrolase	0.261	6	2.321	0.012	2.594
Maltodextrin binding protein	0.314	5	2.361	0.013	2.688
Aspartate aminotransferase	0.371	5	2.715	0.02	3.106
GroEL	0.427	3	2.098	0.025	2.55
Lactoferrin	1.013	5	8.921	0.033	9.967

Table S1. Running time of our algorithm on the data set with simulated density maps, showing timing break-down for each step of our method as well as the total time. From data set 1 to data set 6, the number of amino acid residues keeps increasing, as shown in column (d) of Table 1.

Data set	RMSD (Å)		
	Rigid fitting[a]	Helix-guided fitting[b]	Helix-and-skeleton[c]-guided fitting
Adenylate kinase	11.513	5.713	3.208
Triacylglycerol acylhydrolase	4.062	1.546	1.954
Aspartate aminotransferase	7.556	2.458	1.399

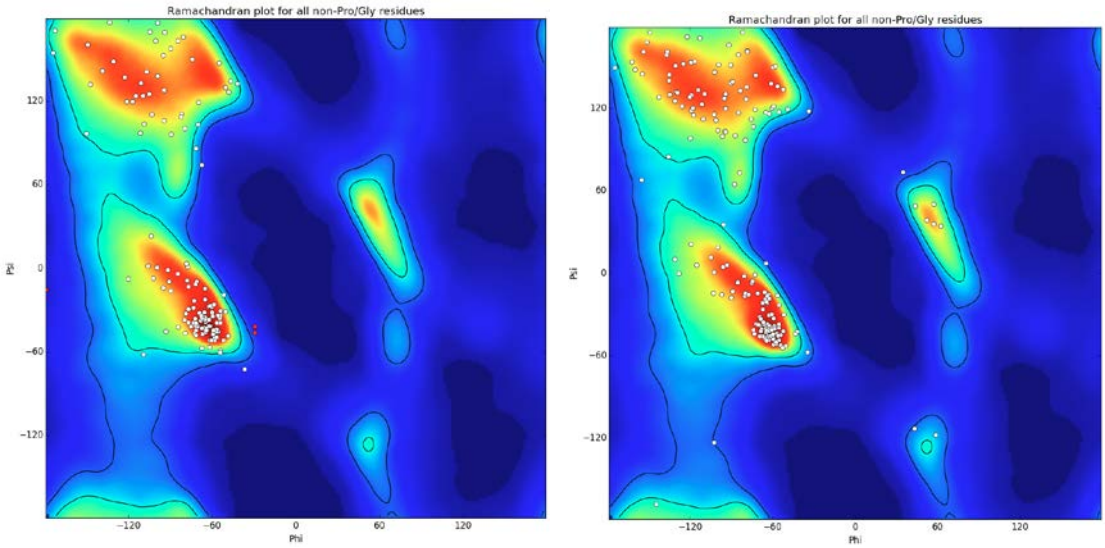
Table S2. All-atom fitting accuracy on the data set with simulated density maps which are generated at 9Å. We selected only those data sets with one-to-one atom correspondence. The metrics are: root-mean-square-deviation (RMSD) between the target model and fitted source model after rigid-body fitting (a), helix-guided fitting (b) and helix-and-skeleton-guided fitting (c). The residue ID we use to compute the C-α atoms RMSDs are listed in column (c) of Table 1.

Data set	Rigid fitting[a] RMSD (Å)	Helix-skeleton guided fitting[b] RMSD (Å)
Ribosome maturation protein sbds	20.085	5.592
Chaperonin	5.399	2.232
60 kda chaperonin	14.677	2.662
DNA polymerase iii subunit alpha	12.268	3.062

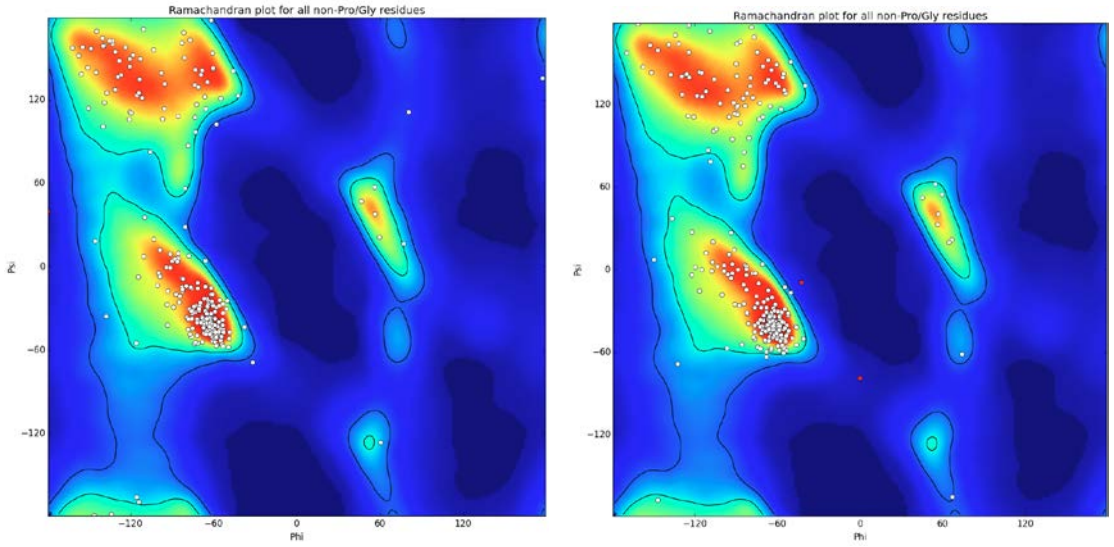
Table S3. All-atom fitting accuracy on the data set with experimental density maps. We selected only those data sets with one-to-one atom correspondence. The metrics are root-mean-square-deviation (RMSD) between the target model and fitted source model after rigid-body fitting (a) and our flexible fitting (b). The residue ID we use to compute the C-α atoms RMSDs are column (d) of Table 4.

RMSD (Å) of helix-and-skeleton-guided fitting of Adenylate Kinase				
Map resolution (Å)	All residues[a]	Identified helix residues[b]	Strands residues[c]	Loop residues[d]
9	2.865	2.651	2.572	3.546
7	2.958	2.921	2.48	3.42
5	2.867	2.507	2.647	3.751
3	3.292	2.386	3.714	4.551

Table S4. Accuracy of fitting source model (Adenylate kinase, PDB ID: 4AKE, chain A) to simulated density maps generated at different resolution from the target model (Adenylate kinase, PDB ID: 1AKE, chain A). The metrics include: RMSD of all the residues (a), identified helix residues (b), strand residues (c) and loop residues (d) between the target model and the fitted source model using helix-and-skeleton-guided fitting.

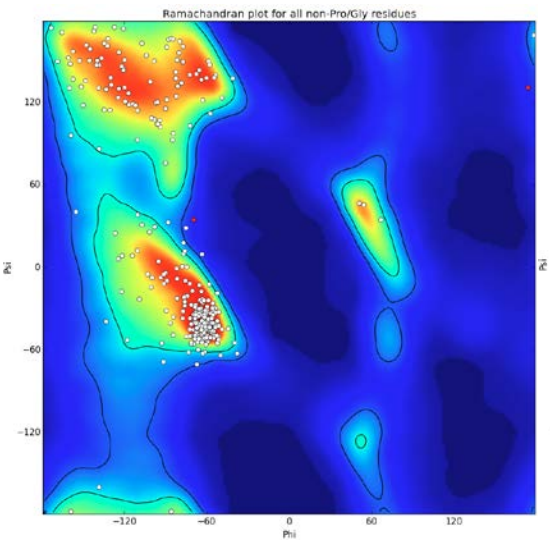


(a) 2.8% outliers (b) 0% outliers

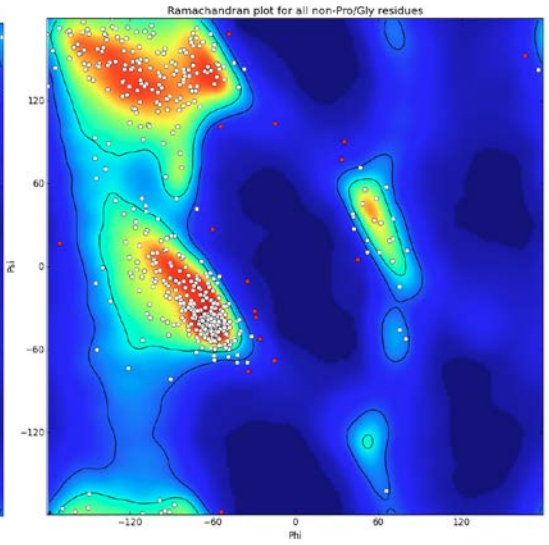


(c) 0.5% outliers

(d) 0.5% outliers

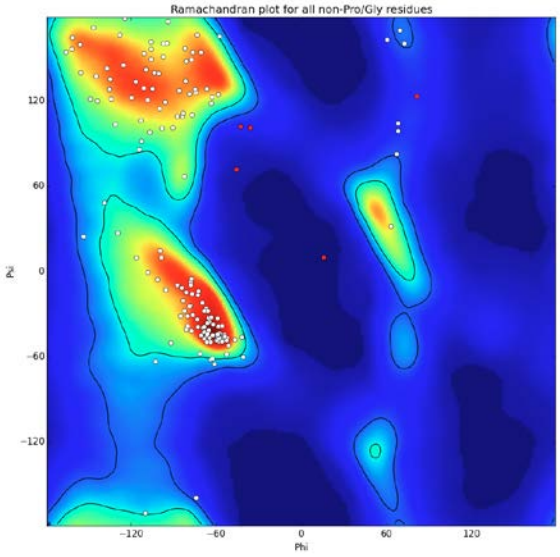


(e) 1.3% outliers

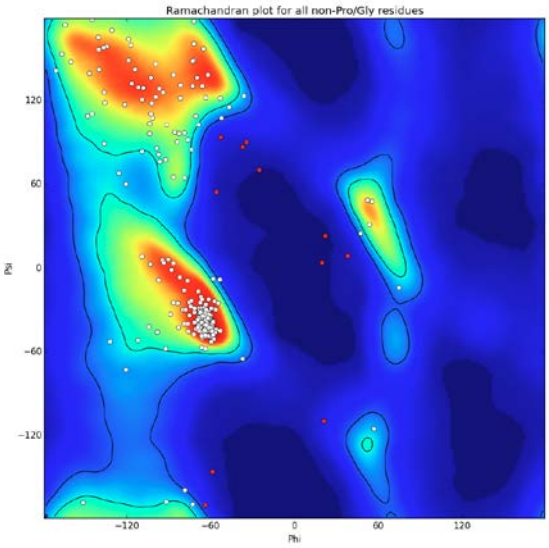


(f) 3.2% outliers

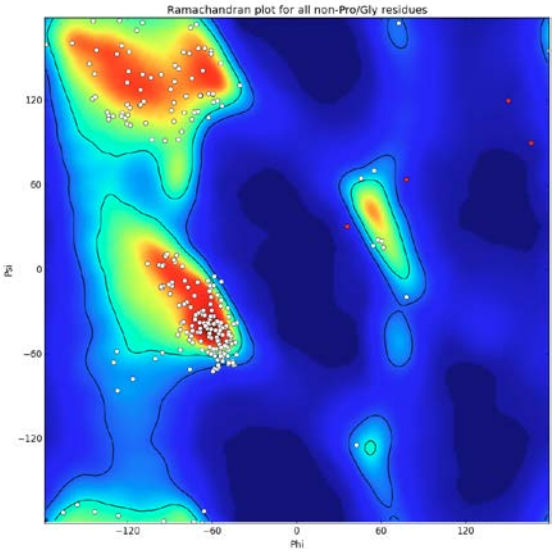
Figure S3. Ramachandran plots for all non-Pro/Gly residues (Psi for y axis and Phi for x axis). The represented data are: (a) Adenylate kinase (source PDB ID: 4AKE chain A, target PDB ID: 1AKE chain A); (b) Triacylglycerol acylhydrolase (source PDB ID: 3TGL chain A, target PDB ID: 4TGL chain A); (c) Maltodextrin binding protein (source PDB ID: 1OMP chain A, target PDB ID: 1ANF chain A); (d) Aspartate aminotransferase (source PDB ID: 9AAT chain A, target PDB ID: 1ANF chain A); (e) GroEL (source PDB ID: 1OEL chain A, target PDB ID: 2C7C chain A); (f) Lactoferrin (source PDB ID: 1LFG chain A, target PDB ID: 1LFH chain A).



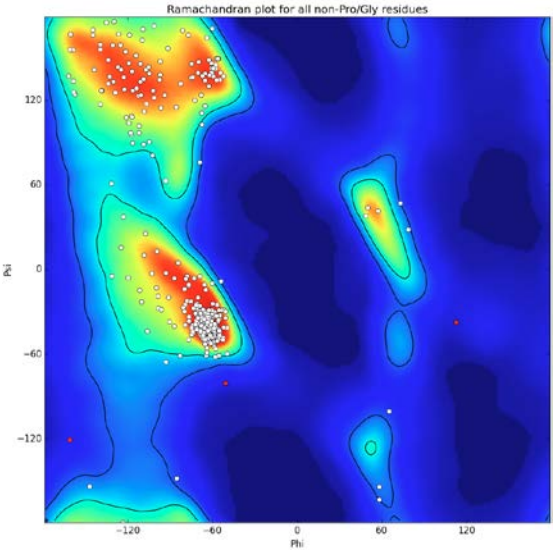
(a) 2.4% outliers



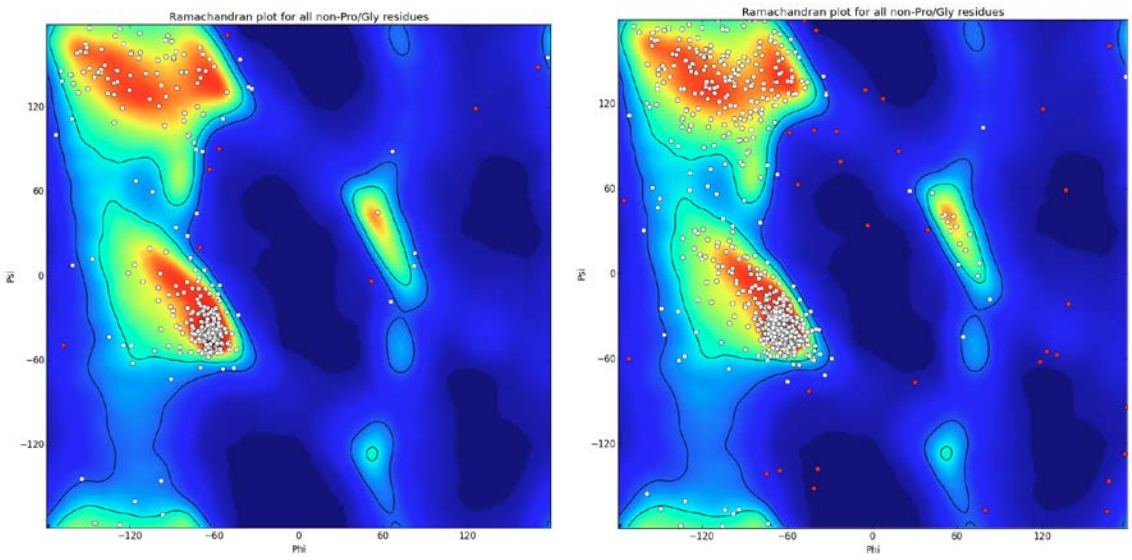
(b) 0.9% outliers



(c) 2.5% outliers



(d) 0.6% outliers



(e) 1.5% outliers (f) 3.0% outliers

Figure S4. Ramachandran plots for all non-Pro/Gly residues (Psi for y axis and Phi for x axis). The represented data are: Ribosome maturation protein SBDS (source PDB ID: 5AN9 chain J, target EMDB ID: 3146); (b) Magnesium transport protein CorA (source PDB ID: 3JCF chain E, target EMDB ID: 6552); (c) 26s protease regulatory subunit 6b homolog (source PDB ID: 3JCO chain K, target EMDB ID: 6575); (d) Chaperonin (source PDB ID: 3IZH chain C, target EMDB ID: 5645); (e) 60 KDA chaperonin (source PDB ID: 2C7C chain M, target EMDB ID: 1180); (f) DNA polymerase iii subunit alpha (source PDB ID: 5FKV chain A, target EMDB ID: 3201).

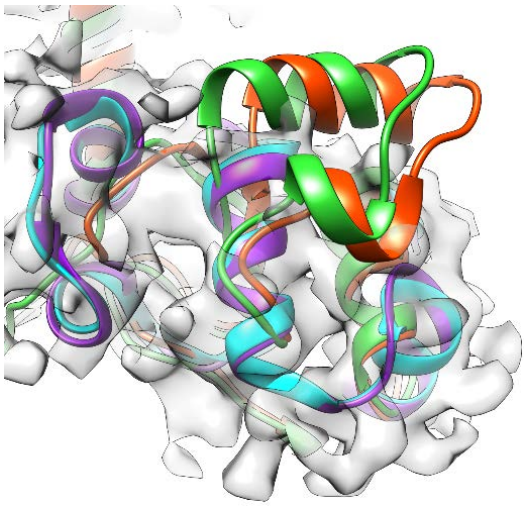
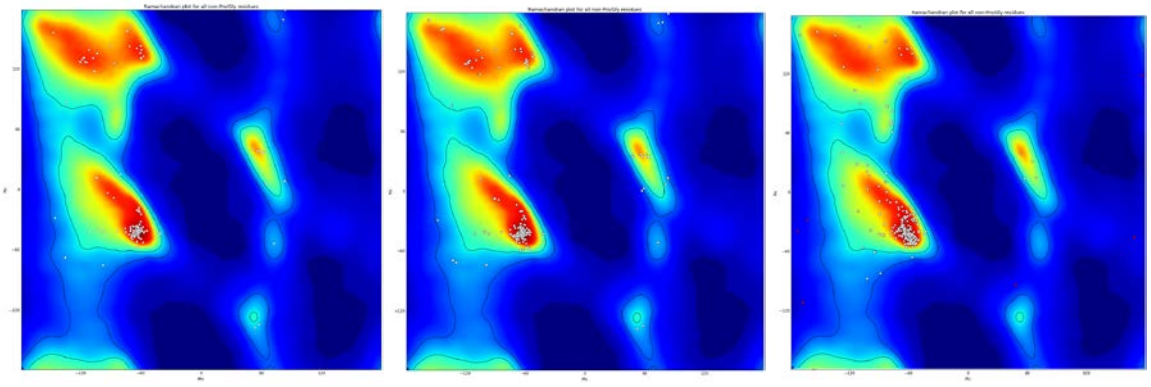
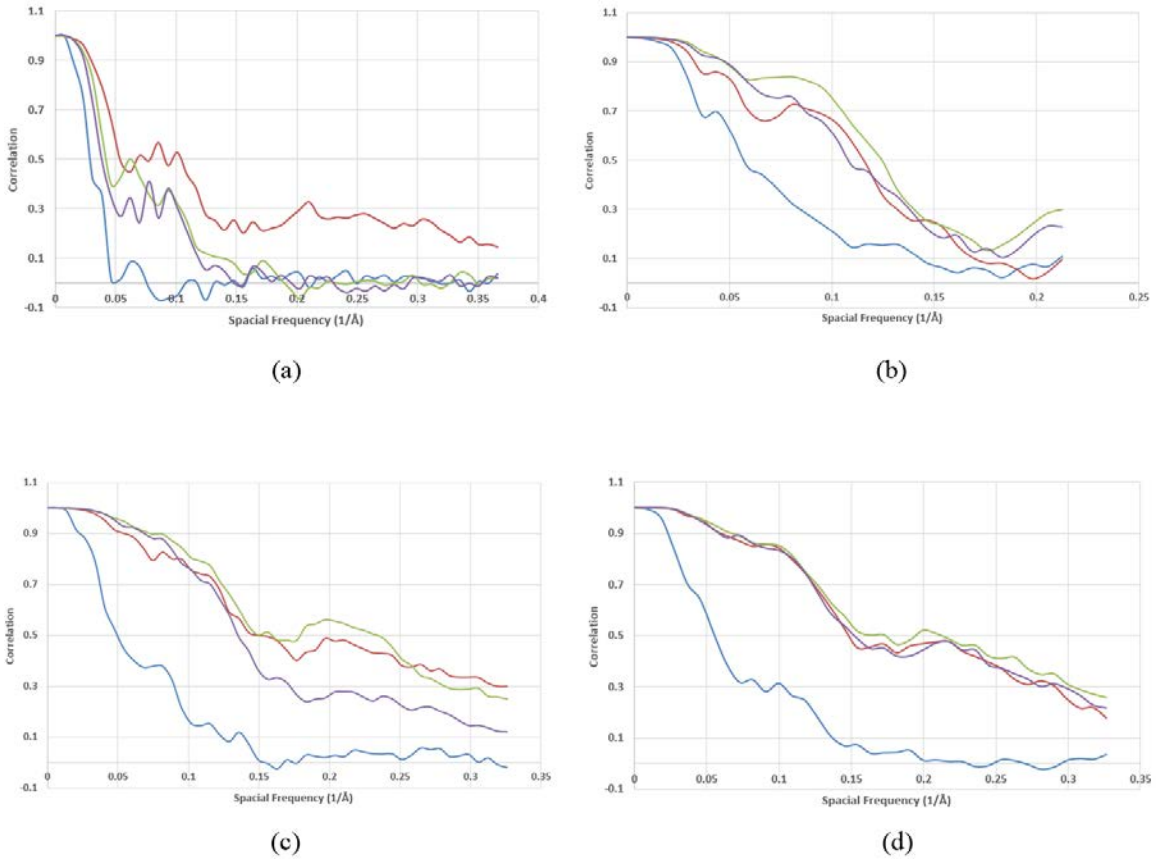


Figure S5. Fitting result of Ribosome maturation protein SBDS. A zoom-in view of the target model (cyan, PDB ID: 5ANB chain J) and the fitted model (PDB ID: 5AN9 chain J) of Flex-EM (red), MDFF (green), and our method (purple).



(a) 1.0% outliers (b) 1.0% outliers (c) 2.2% outliers

Figure S6. Ramachandran plots for all non-Pro/Gly residues (Psi for y axis and Phi for x axis) of TRPV1. The represented data are: (a) Soure model (PDB ID: 3J5Q chain D); (b) Fitted model of our method; (c) Target model (PDB ID: 3J9J).



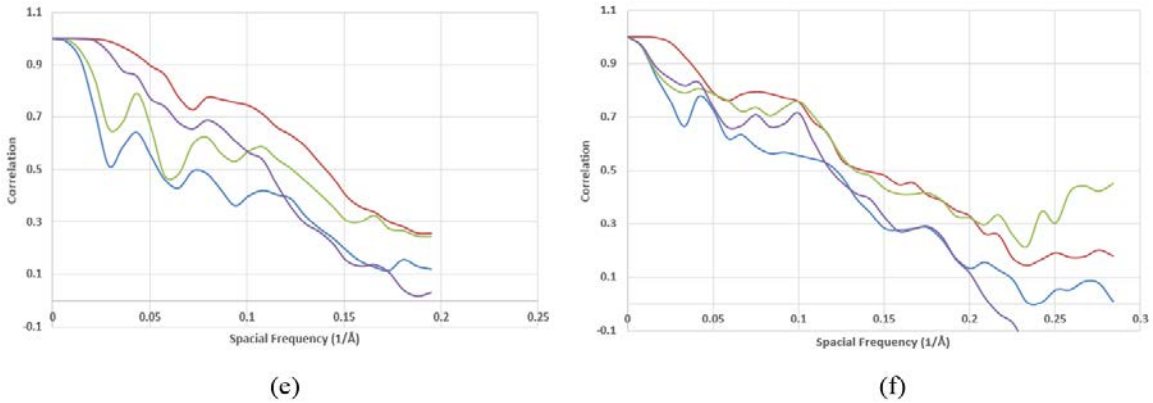


Figure S7. Fourier shell correlation plots (correlation for y axis and spacial frequence for x axis). The blue, red, green and purple curves denote rigid fiting, helix-and-skeleton guided fitting, MDFF and FlexEM respectively. The represented data are: (a) Ribosome maturation protein SBDS (source PDB ID: 5AN9 chain J, target EMDB ID: 3146); (b) Magnesium transport protein CorA (source PDB ID: 3JCF chain E, target EMDB ID: 6552); (c) 26s protease regulatory subunit 6b homolog (source PDB ID: 3JCO chain K, target EMDB ID: 6575); (d) Chaperonin (source PDB ID: 3IZH chain C, target EMDB ID: 5645); (e) 60 KDA chaperonin (source PDB ID: 2C7C chain M, target EMDB ID: 1180); (f) DNA polymerase iii subunit alpha (source PDB ID: 5FKV chain A, target EMDB ID: 3201).

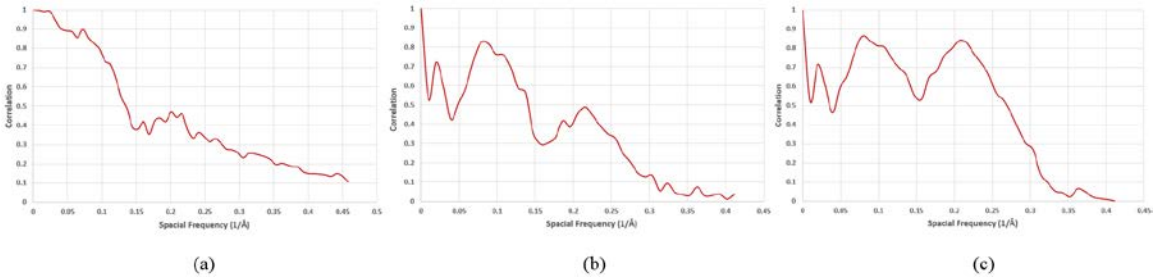


Figure S8. Fourier shell correlation plots (correlation for y axis and spacial frequence for x axis) of TRPV1. (a) The plot shows the FSC between the map simulated (to the resolution of the experimental cryo-EM map, EMDB ID 5778) from the model fitted by helix-and-skeleton guided fitting and the map simulated from the atomic model (PDB ID 3J9J). (b) The plot shows the FSC between the map simulated (to the resolution of the experimental cryo-EM map, EMDB ID 5778) from the model fitted by helix-and-skeleton guided fitting and the experimental cryo-EM map (EMDB ID 5778). (c) The plot shows the FSC between the map simulated (to the resolution of the experimental cryo-EM map, EMDB ID 5778) from the helix-and-skeleton guided fitting model refined by Phenix and the experimental cryo-EM map (EMDB ID 5778).

D. Parameter settings for MDFF and Flex-EM

Flexible fitting using MDFF was carried out performed with the MDF GUI in VMD 1.9.2 as described in Computational Biophysics Workshop:

http://www.ks.uiuc.edu/Training/Tutorials/science/mdff/tutorial_mdff-html/

More specifically, PSF files were first generated using the AutoPSF function in the VMD Modeling Extensions. Corresponding map and fit PDB, PSF files were loaded into the MDFF GUI; chirality and secondary structure restraints were enabled. Simulation parameters were set as follows:

Temperature=300K, Final Temperature=300K, Minimization steps=200, Time steps=50000 and system environment=vacuum. NAMD files were generated and executed using NAMD 2.11. Unless otherwise noted, all simulations were performed using a single core. GPU accelerated runs of NAMD were performed using the multicore-CUDA version of NAMD 2.11.

Flexible fitting using FlexEM followed the instructions in the project page of Protein Structure Fitting and Refinement Guided by Cryo-EM Density:

http://topf-group.ismb.lon.ac.uk/flex-em/

As the models to flexibly fit were already rigidly fit to their corresponding density using Chimera, the optimization process was set to MD. 4 iterations were performed. In map/model pairs with large initial iterations 20 runs using CG optimization were performed. Secondary structure elements were defined in the rigid bodies file and cap_shift was set to 0.15. Additionally, box size, apix and resolution were set based on the corresponding map parameters.

References

1. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.-P. 2004. Laplacian surface editing. Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing. pp. 175—184.
2. Guennebaud, G., Jacob, B.. 2010. Eigen v3. http://eigen.tuxfamily.org.