Fitting PolIII heterotrimer subunit C82 in a cryo-EM map representing elongating PolIII (EMD-3178)

In this tutorial, we will use flexible fitting to model the structure of one subunit (C82) of PolIII. We'll use a predicted structure from AlphaFold as a starting model. The data for today's tutorial is available here:

ccpem_tutorial/data

There is an AlphaFold prediction for the C82 subunit of PolIII available on the AlphaFold Protein Structure database, available here: https://alphafold.ebi.ac.uk/entry/P32349 and pre-downloaded into the data folder: **C82_AlphaFoldDB.pdb**.

To fit the AlphaFold (AF) model in the EM map, the model has to be placed at the location of the subunit and further refined to accommodate any conformational changes pertaining to elongating PolII complex.

1. **Examining the AlphaFold predicted model (Chimera)**

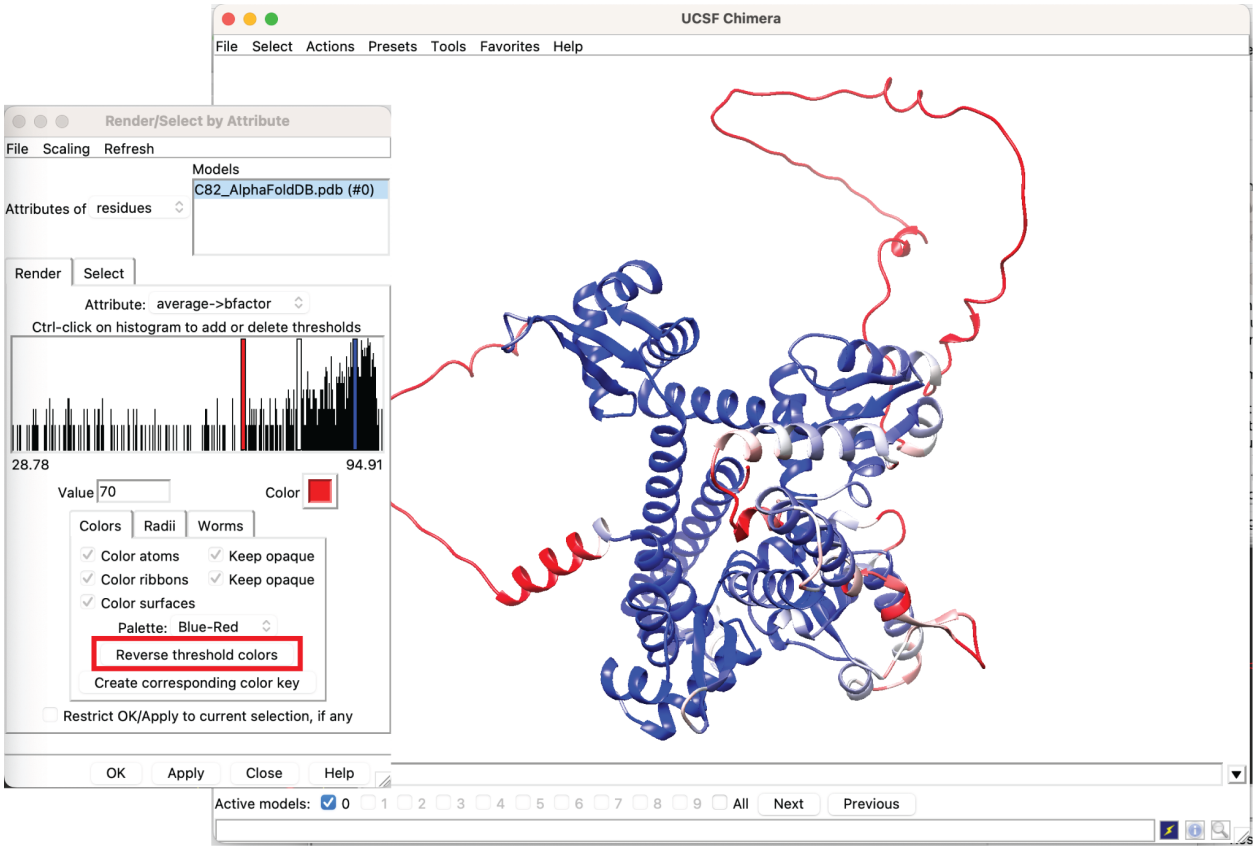
First, however, we need to examine the AF predicted model, to assess whether it will be a useful starting point for modelling.

We can get a good initial estimate of an AF predicted model's quality by examining the pLDDT score, a per-residue estimate of the AF network's confidence in local model quality. The pLDDT score is on a scale of 0 - 100, where a higher score is better. Regions with a pLDDT > 90 are expected to be very accurately modelled (e.g. accurate side chain positions) and those with a score between 70 and 90 should still be reasonably well modelled (i.e. a good backbone prediction). Regions with a pLDDT below 70 are not likely to be well modelled, and should be interpreted with caution.

We can view the pLDDT scores for the **C82_AlphaFoldDB.pdb** model in Chimera. Open the predicted model in Chimera by typing the following command:

$ chimera data/C82_AlphaFoldDB.pdb

The pLDDT scores are stored in the B-factor column of PDB files of AF predicted models. We can use Chimera to colour the structure based on pLDDT scores using the "Render by Attribute" tool. Open it by clicking "Tools->Structure Analysis->Render By attribute" and colour the residues by choosing "Attributes of: residues" and "Attribute: average bfactor". We recommend using "Reverse Threshold Colours" to comply with the standard colour scaling for displaying pLDDT (i.e blue for high scores, red for low). The colour range can be adjusted manually, a sensible range would be: minimum (red) 70, middle (white) 80, maximum (blue) 90.



This shows the model is mostly confidently predicted, but that there are large loops with a low pLDDT score. It has been shown that a low pLDDT score can correlate with unstructured, intrinsically disordered regions, which is probably the case here. To simplify the modelling today, we will remove these areas. We can do this in Chimera with the following command:

del #0:1-26,375-442

Save the model (File->Save PDB) as **C82_AlphaFoldDB-noloops.pdb** and close this Chimera session.

2. Rigid-body docking of the AlphaFold model (Chimera)

Now we have prepared our model, we need to dock it into the cryo EM map.

Analysis of local resolution of this part of the volume suggests that the local resolution in the region of the C82 subunit ranges from 4-6Å. The map is globally sharpened using an optimal B-factor. However, global sharpening may lead to over-sharpening in areas of lower local resolution (or inadequate sharpening in locally higher resolution regions). As this subunit density has relatively lower local resolution, one may benefit from blurring or low-pass filtering the map. The secondary structure features become more evident in a low-pass filtered map

(**emd_3178-lp5.mrc**) than in the globally sharpened map (**emd_3178.map**)[1]. We will use the low-pass filtered map for fitting the atomic model of C82.

To reduce the search space (and improve accuracy of the search), cut out a section of the map volume corresponding to the heterotrimer domain.

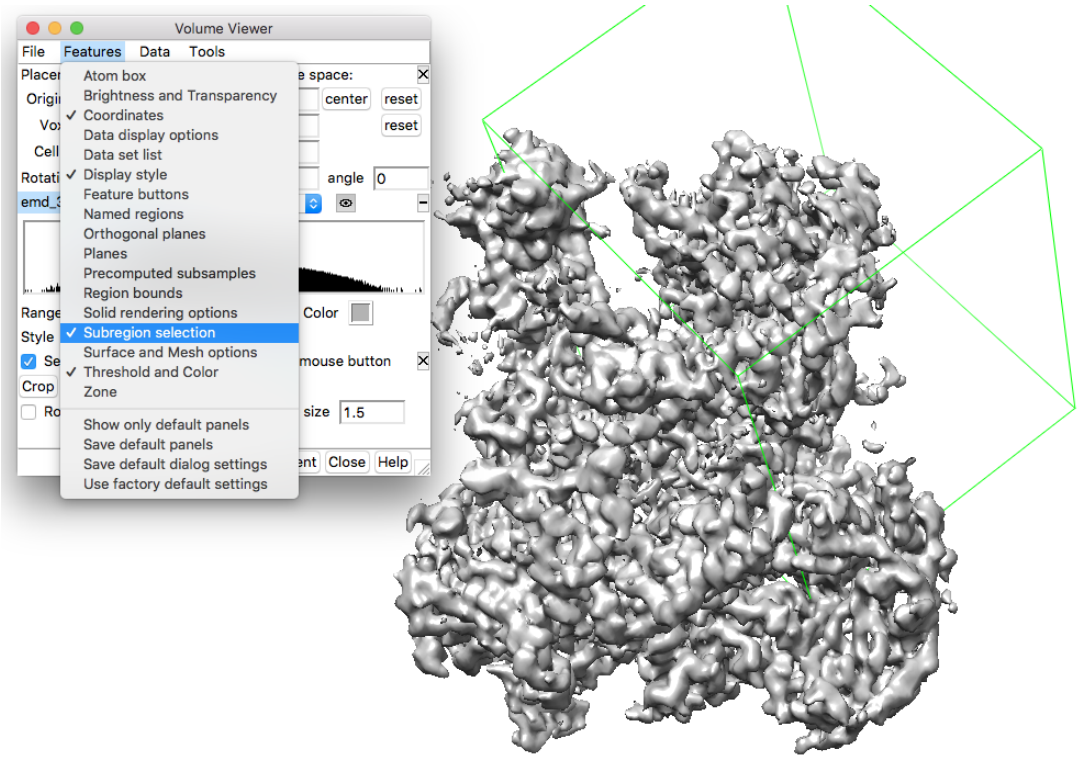
1.1. Crop a sub-volume corresponding to the heterotrimer domain

Open the map in Chimera using the command:

```
$ chimera data/emd_3178-lp5.mrc
```

And crop out the area corresponding to C82. This can be done with tools in the *Volume Viewer:*

Volume Viewer -> *Features* -> *Subregion selection*



Enable mouse middle click to adjust the crop box. Centre the crop box at the heterotrimer domain and adjust the sides (middle click on the sides and drag) to cover the required region. Click *Crop* on the *Subregion selection* window to get the cropped map. The cropped map can be saved (**emd_3178_lp5_cropped.mrc**) using *File/Save map as* option in Volume viewer.

[1] You can open the two maps (emd_3178.mrc and emd_3178-lp5.mrc) and compare them in Chimera. Do you find it easier to identify secondary structure in the density for the lowpass filtered map?

1.2. Fit the C82 model in the cropped map

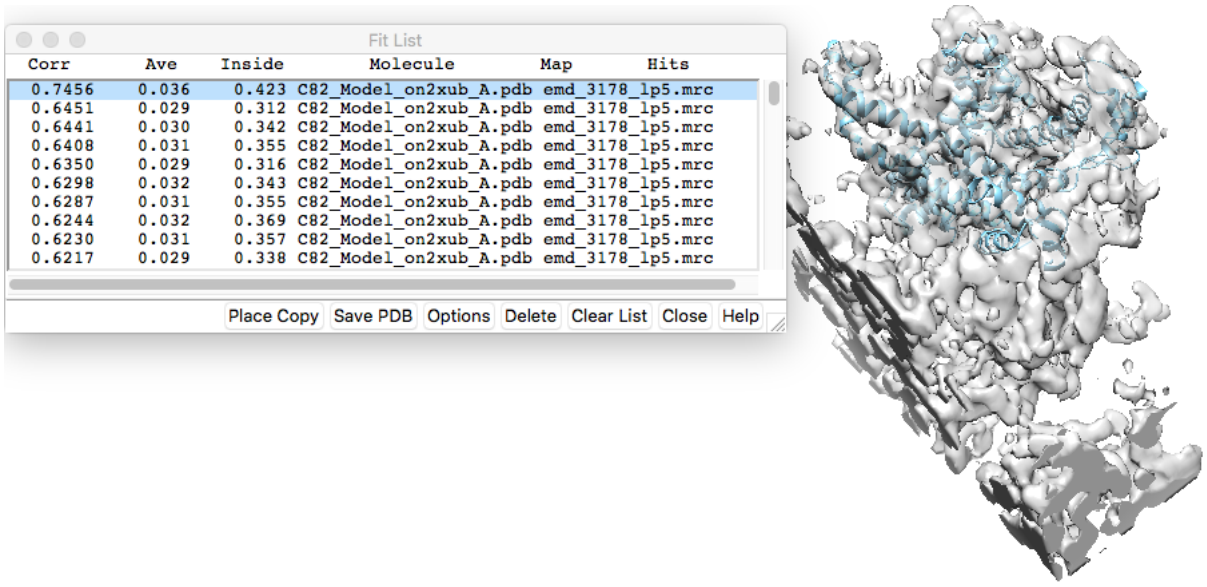
Open your newly cropped map (**emd_3178_lp5_cropped.mrc)** and the edited AF model **C82_AlphaFoldDB-noloops.pdb** using the command:

$ chimera data/emd_3178-lp5-cropped.mrc data/C82_AlphaFoldDB-noloops.pdb

To rigidly fit the model in the cropped volume, run the following command on Chimera command line (*Favourites/Command Line*)

fitmap #1 #0 search 200 resolution 5.0

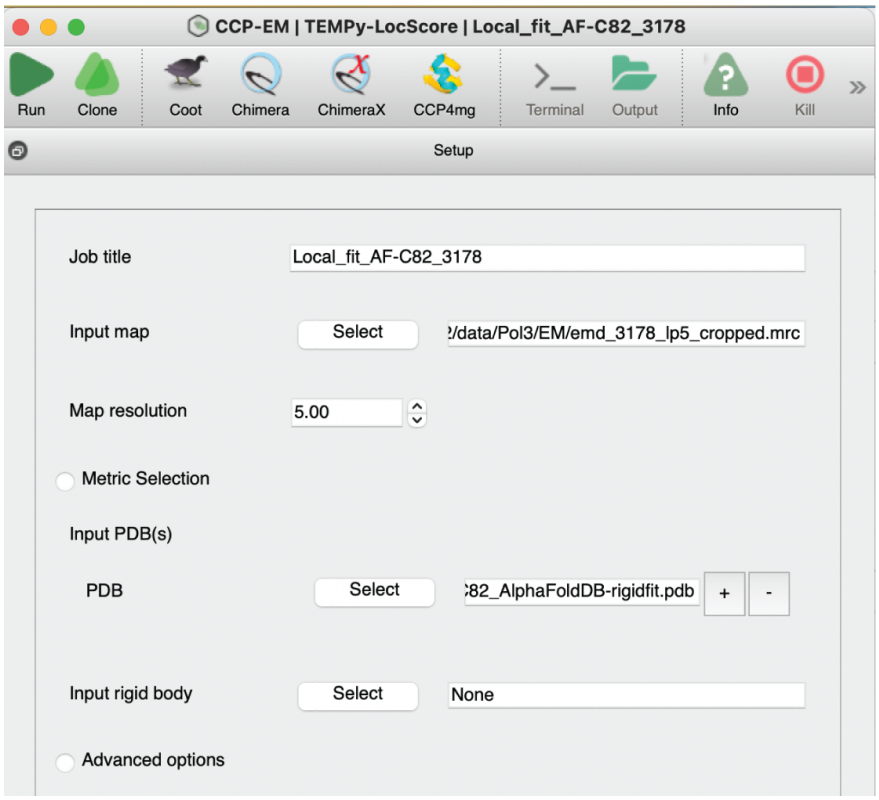
The model will be placed at 200 random starting points in the map and locally searched to maximize agreement with the density. Once the search is complete, a *Fit List* window appears with a list of model locations ranked by fit to density.



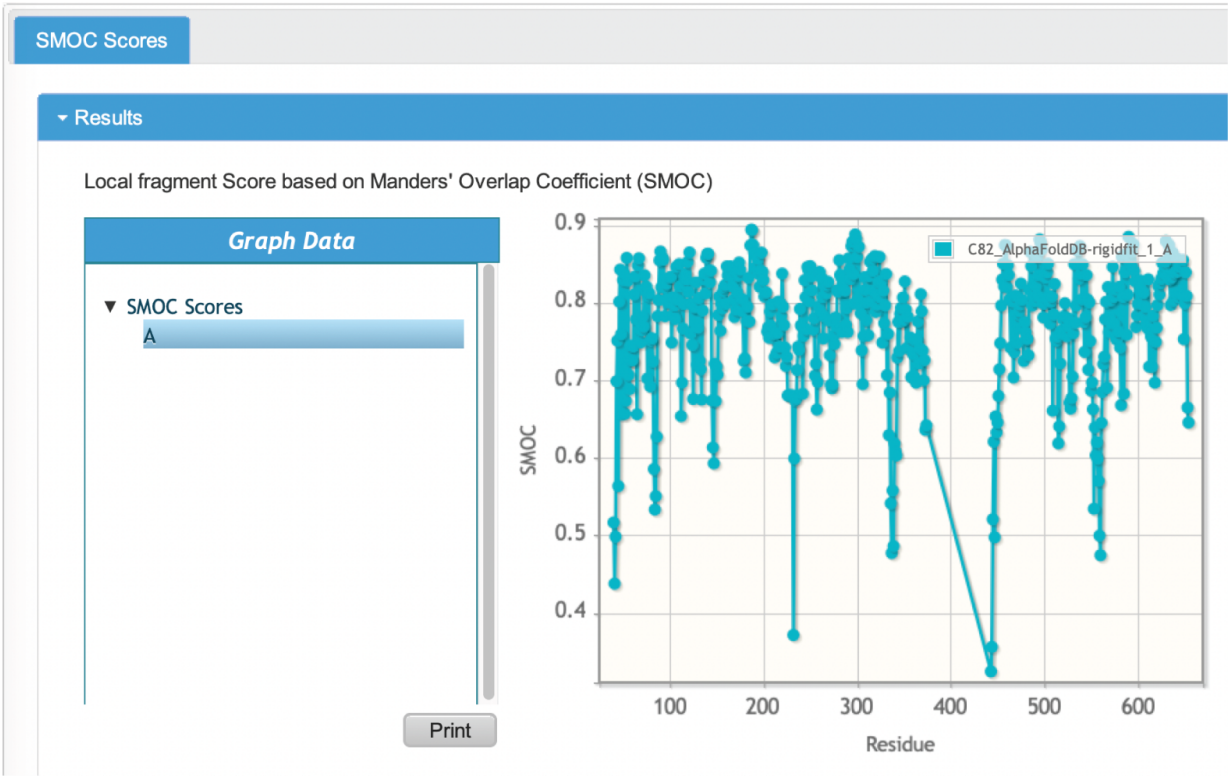
The fitted model can be saved (**C82_AlphaFoldDB-rigidfit.pdb)** relative to the map from *File/Save PDB* (with *Save relative to model* enabled and the volume selected as the model relative to which the coordinates are saved).

3. Examining model fit in density

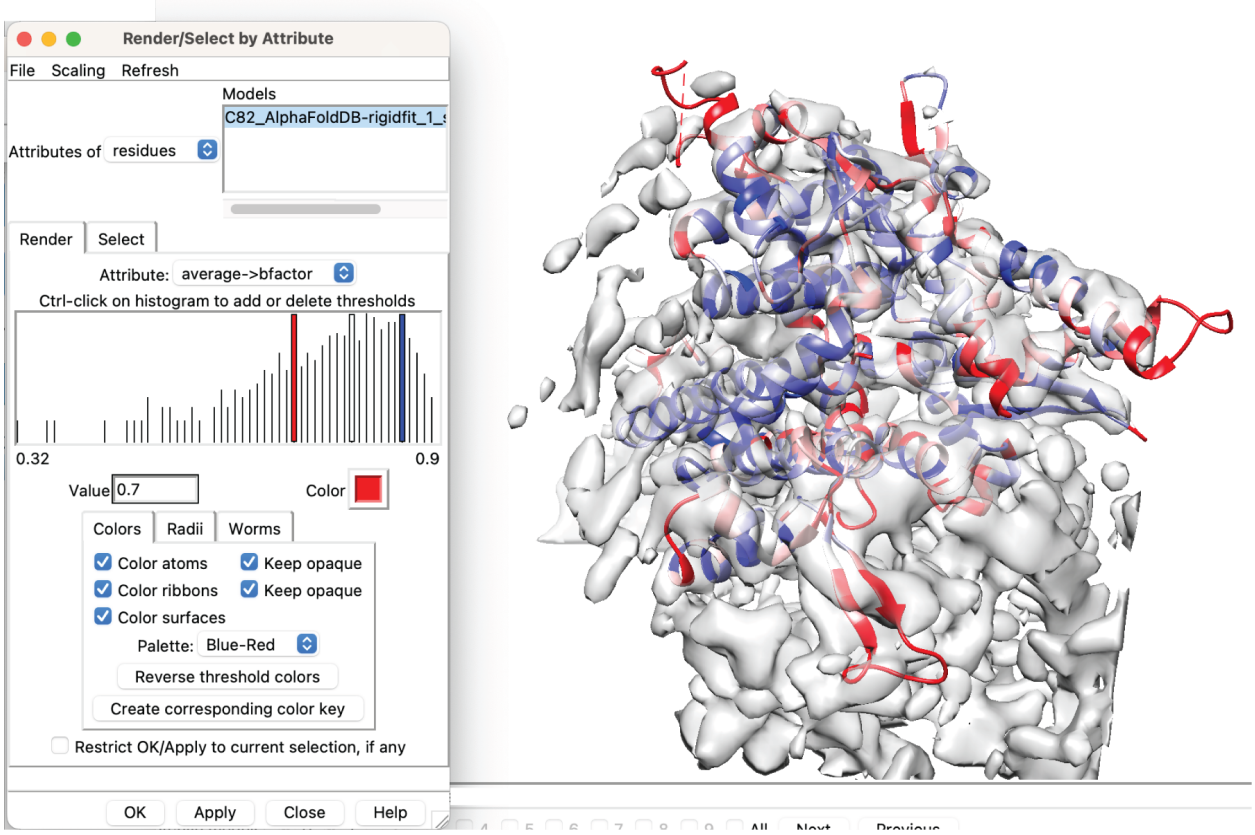
The local fit of the atomic model in density can be examined using the “TEMPy: Local score” tool in the CCP-EM interface.



The plot of local score per residue appears on the *Results* tab. The plot shows that the range of scores vary from 0.32 to 0.93. Hence there are local regions of the atomic model that are not in good agreement with the density.



This can also be visualised in Chimera by clicking the *View in Chimera* button. This opens the model and map in Chimera with the model backbone colored by SMOC scores. The range of scores used for coloring can be adjusted further by looking at the score distribution using Render/Select by Attribute window (Tools->Structure Analysis->Render by Attribute, choose residues/average bfactor). Here we use 0.7 (red), 0.8 (white) and 0.85 (blue) for the colour range.



4. Flexible fitting of the homology model

The rigidly fitted model has regions that are not fitted well in the density and hence require further flexible fitting and refinement.

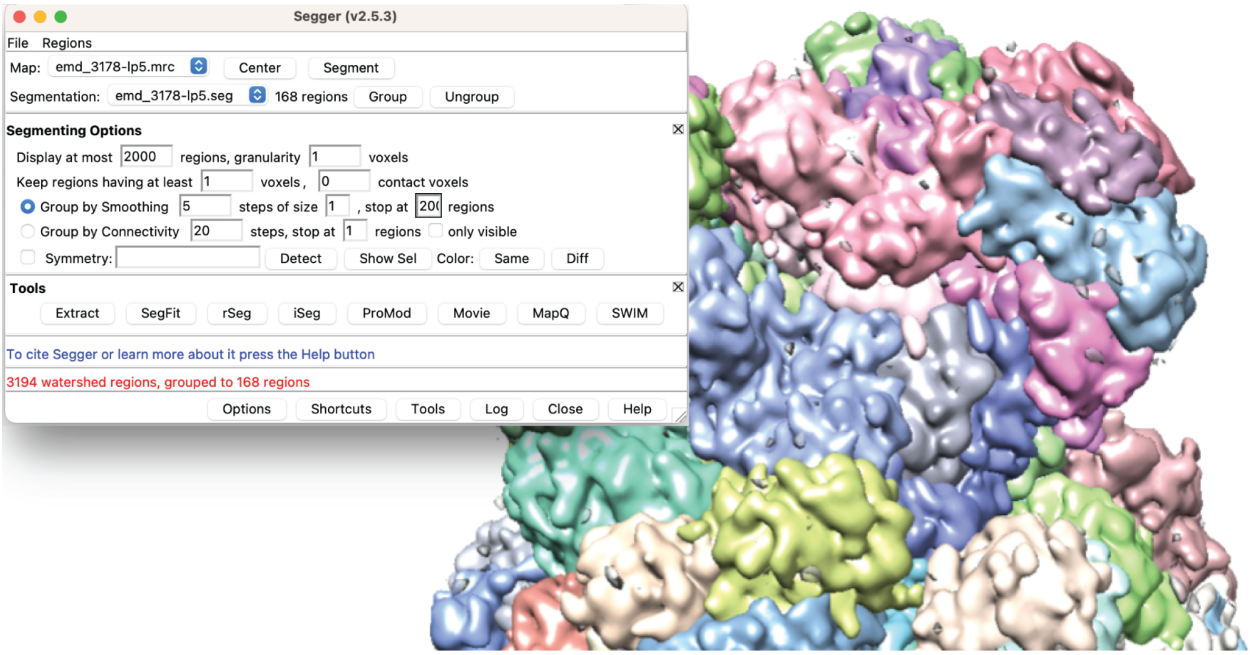
For flexible fitting of the atomic model in volume density, it is often necessary to extract a volume segment corresponding to the molecule of interest. Using a segment instead of the full map, makes the fitting process computationally efficient and minimises chances of fitting into the density of neighbouring subunits. We will extract a segment from the map including density *regions* traced by the rigidly fitted model.

3.1. Map segmentation with Segger

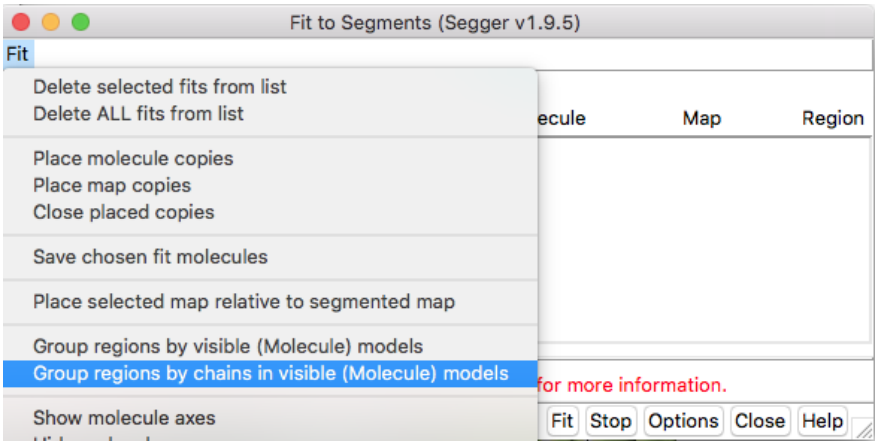
Open **emd_3178_lp5.mrc** and **C82_AlphaFoldDB-rigidfit.pdb** in a new chimera window. Set contour *Level* to 0.04 in the *Volume Viewer* to cover most of the non-background density. Set volume opacity to 0.5 (*Color* button).

The Segger tool in Chimera (*Tools/Volume Data/ Segger*) uses a watershed algorithm to segment volume starting from local maximas and iteratively filters the volume to groups neighboring maximas into larger segments. The grouping terminates when the number of segments falls below a cutoff (provided by user).

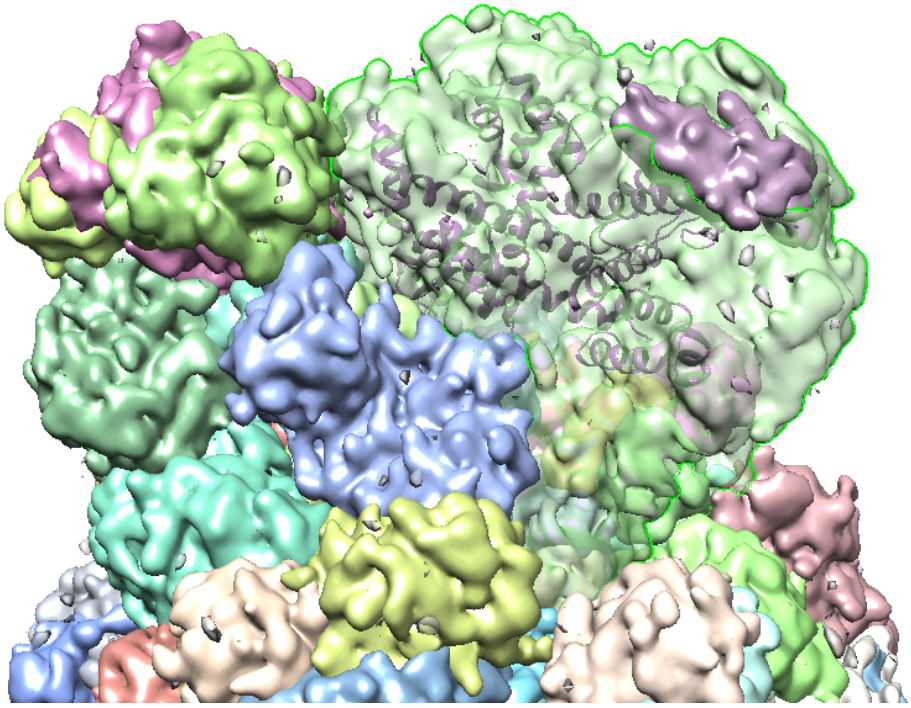
We need to customise the segger run, click the *Options* tab and use the parameters in the figure below to run Segger. Here we stop grouping the segments if the number of regions <= 200 (and we get 168 segment regions). One may use 100 instead of 200 to get a coarser segmentation (larger segments), if the conformational changes required to fit the model are larger.



On the *Segment Map* window, click on *Tools* in the segger window, and then *SegFit.* A *Fit to Segments* window appears. Use the *Group regions by chains in visible models* option to group the segments that cover the starting model.



This will group the segments based on the rigidly fitted model and provide a single segment that covers the model.



Ctrl click on the grouped segment to select it and in the *Segment Map* window, use *Regions/Make transparent* to make the segment transparent. *File/Save selected regions to .mrc file* to save the segment map (**emd_3178_lp5_regions_bymodel.mrc**).

3.2. Flexible fitting with Flex-EM

Next, we will use Flex-EM[1,2] to fit the homology model of C82 in the map of elongating PolIII (EMDB ID: 3178). Flex-EM uses MODELLER[3] for the molecular dynamics runs, with the density score added to the calculations.

Input files:

1. We will use the volume around C82 segmented using Segger[4] tool in Chimera[5] (**emd_3178_lp5_regions_bymodel.mrc**).

2. The starting AF model (**C82_AlphaFoldDB_rigidfit.pdb**) that we rigidly fitted in the segmented density using Chimera (*fitmap*).

We will use Flex-EM with rigid-body restraints in a hierarchical way starting with larger rigid bodies (sub-domains) in the initial run, followed by another Flex-EM run with relatively smaller

rigid bodies (secondary structures). The initial step simulates large body movements whereas the secondary structure fits are optimized in the second stage.

Rigid body restraints are listed in a text file and this file has to be added as input for Flex-EM. Each line in the rigid body restraint file has the set of segments which are part of this rigid body, where each segment is defined by the start and end residue of the segment.

For example:

10:A 20:A 50:A 70:A

100:B 130:B 100:A 120:A

adds residues 10 to 20 and 50 to 70 of chain A to one single rigid body, and 100 to 130 of chain B and 100 to 120 of chain A to another rigid body.

We can also use RIBFIND[6] to generate rigid body files, which can then be used as an input file to Flex-EM. RIBFIND clusters secondary structures that are closely in contact along with intermittent loops into rigid bodies. Cluster cut-offs are used to consider the percent of residues in a secondary structure that are expected to be in contact (with those in another secondary structure) to form rigid bodies. E.g. cluster cut-off of 100% considers each secondary structure as a separate rigid body while a cut-off of 50% groups together secondary structure where half of residues are in contact.

We will be using CCP-EM GUI interface to run RIBFIND and Flex-EM.

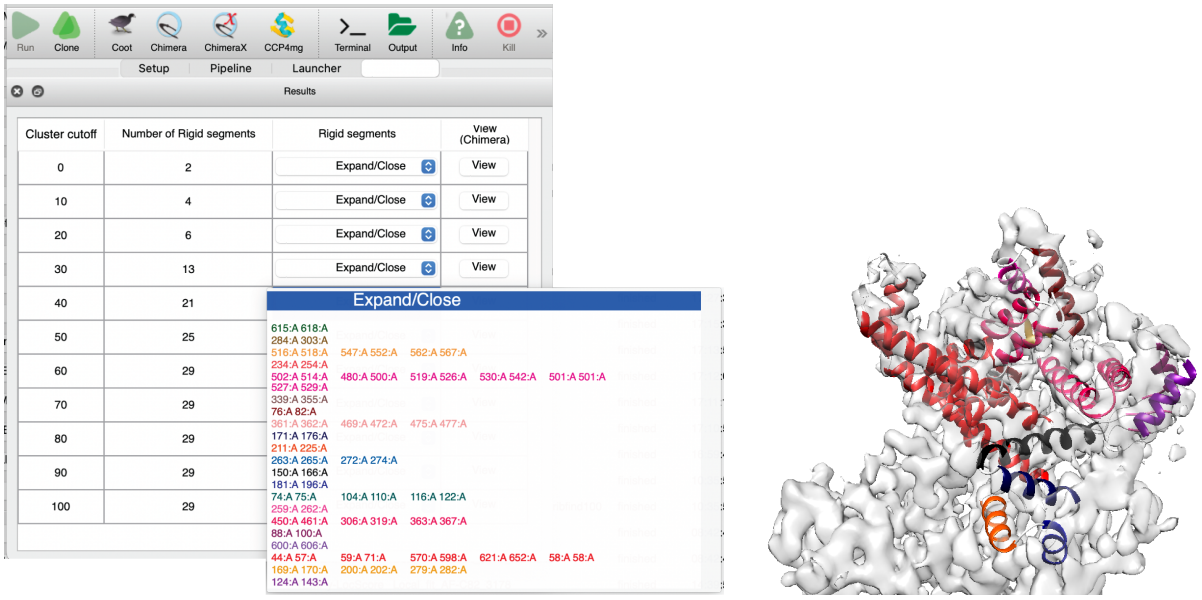
A. Running RIBFIND

1. Launch the CCP-EM GUI

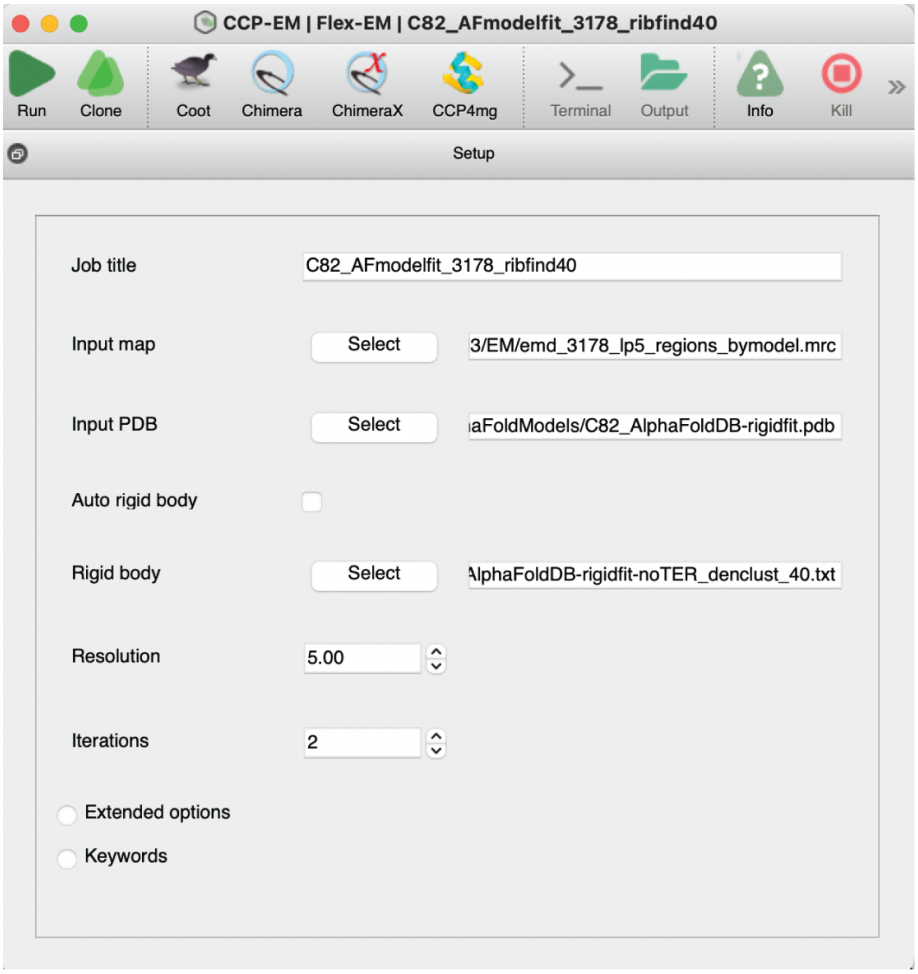
2. Use the model C82_AlphaFoldDB_noTER.pdb[2] as input for the RIBFIND task in CCPEM interface.

3. Click ‘Run’ to start RIBFIND and the results will appear in the ‘Results’ tab. A list of rigid bodies identified at different cluster cutoffs will be listed along with a ‘view’ button to see the rigid bodies colored in Chimera and the parts of the protein which do not form any rigid bodies are colored in white. A cluster cutoff of 40% groups together secondary structures into compact subdomains (lower cutoffs start to group these sub-domains together).

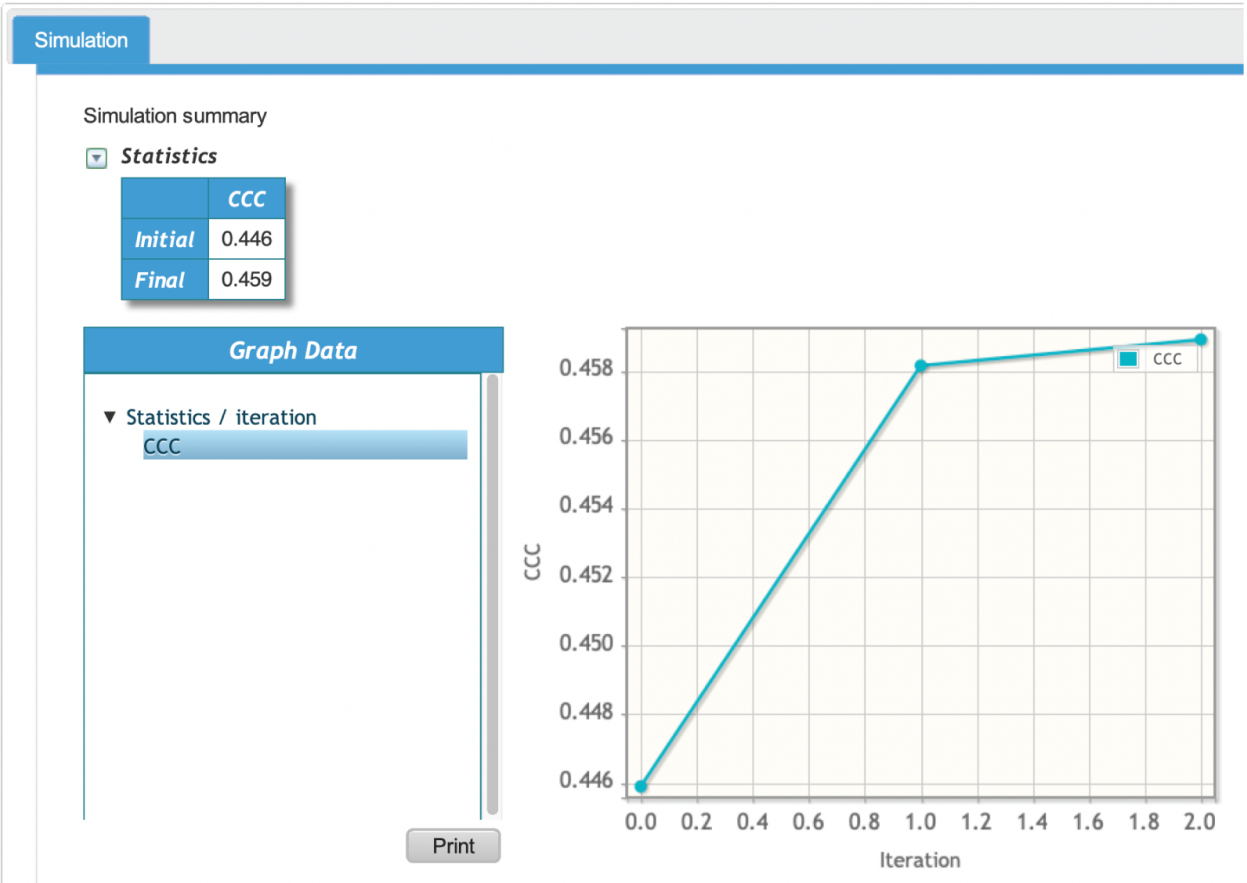
[2] We use this model to avoid a bug in RIBFIND, where residues after a TER entry in the pdb are not included in clustering.



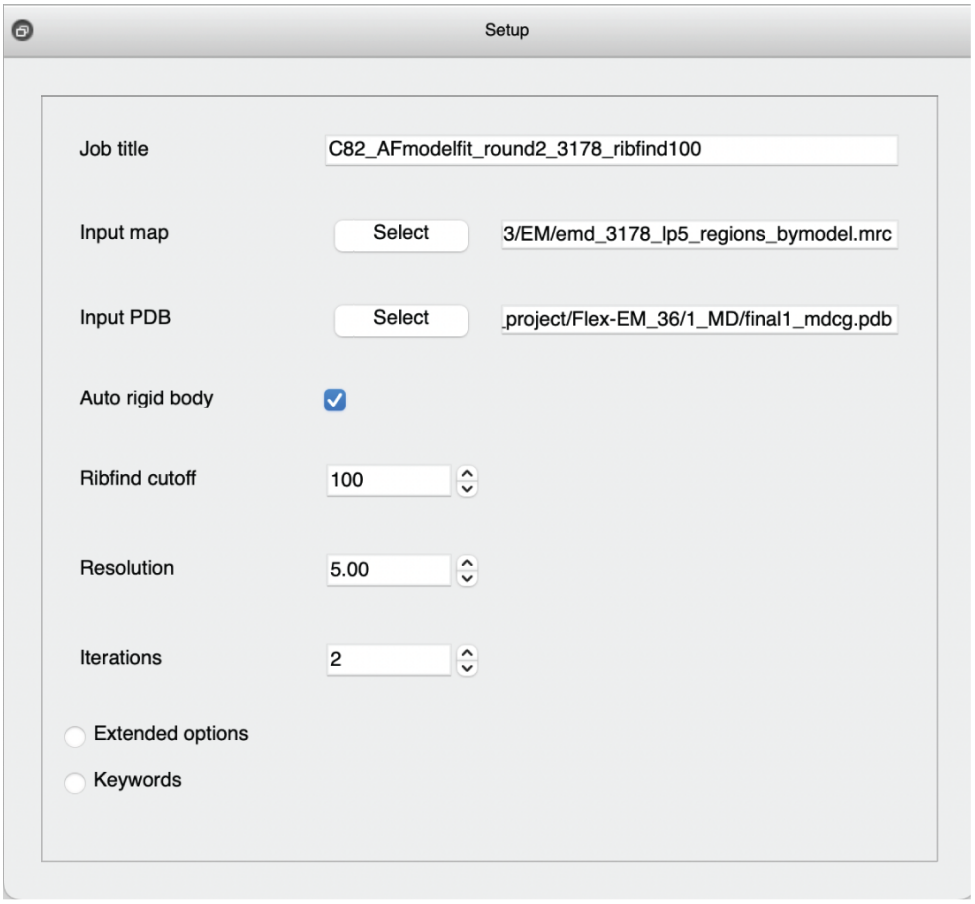
We can use the rigid body file corresponding to the cluster cut-off of 40 for the Flex-EM run. Open Flex-EM task in the CCP-EM GUI and fill in the input parameters. We need the RIBFIND file ending with “denclust_40.txt”.

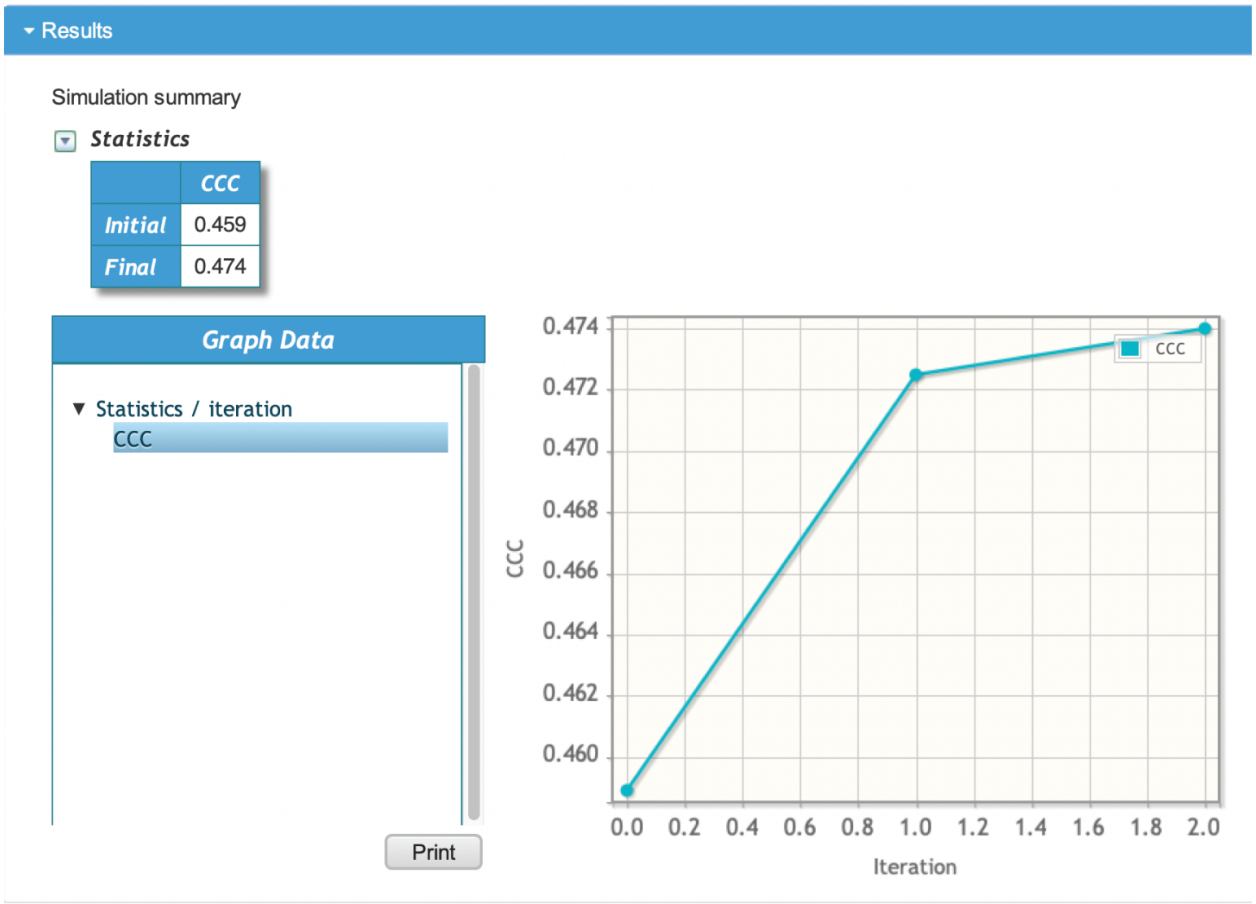


‘Run’ the Flex-EM job for 2 iterations to start the flexible fitting process. Each iteration will take about ~20 mins in this case.
Once the Flex-EM run is finished, the cross-correlation scores for models from each iteration is plotted in the ‘Results’ tab. The cross-correlation with density appears to settle between 0.458 and 0.460, after the first iteration.



To further fit the model in the density, the secondary structures can be treated as rigid bodies in the next Flex-EM run (cluster cutoff 100). We can use the output from the first iteration of the above run (*md1_2.pdb* in the Flex-EM data folder) as input for the next Flex-EM run.


Setup
Job title
C82_AFmodelfit_round2_3178_ribfind100
Input map
Select
3/EM/emd_3178_lp5_regions_bymodel.mrc
Input PDB
Select
project/Flex-EM_36/1_MD/final1_mdcg.pdb
Auto rigid body
Ribfind cutoff
100
Resolution
5.00
Iterations
2
Extended options
Keywords



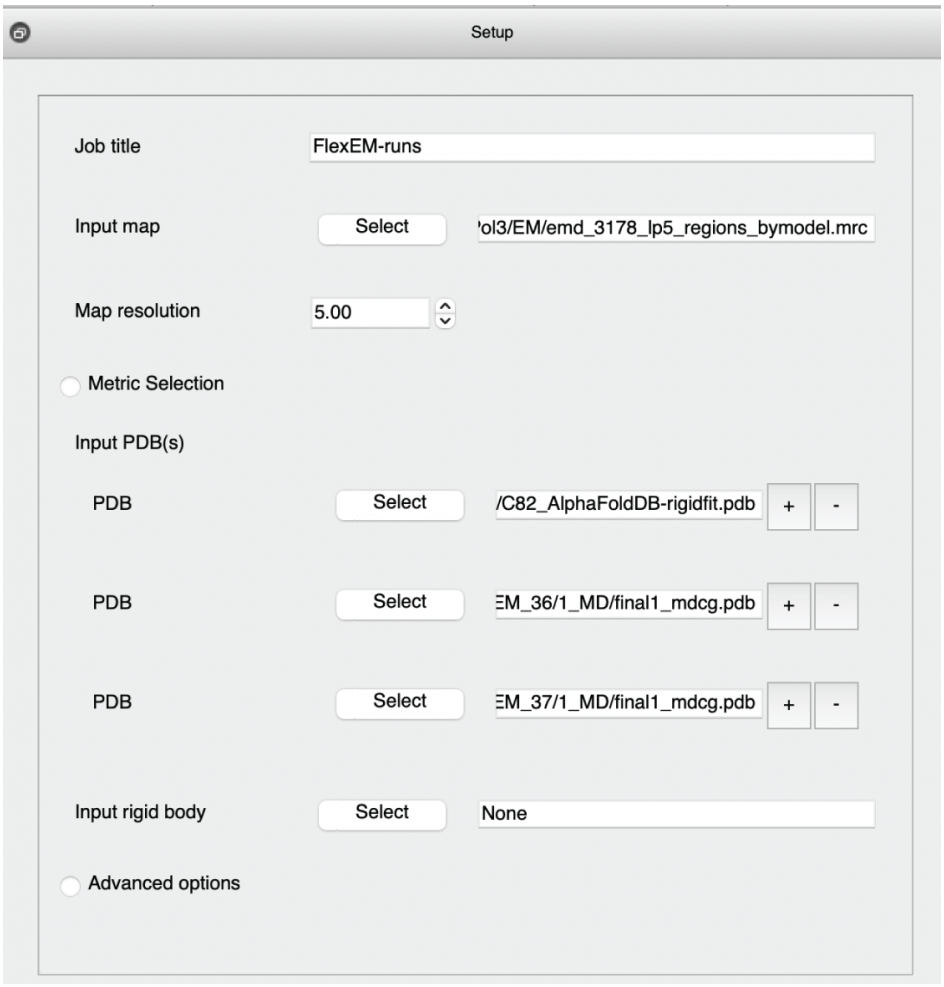
The output from the second Flex-EM run shows further improvement in the cross-correlation of the model with density.

Open the fitted model from the second iteration (*final1_mdcg.pdb*) in Chimera from the launcher along with the map to view the model fit in density. This model has been copied to the *precomputed_results* data folder as flexem_round2.pdb. Open the starting model (**C82_AlphaFoldDB-rigidfit.pdb**) as well to compare.

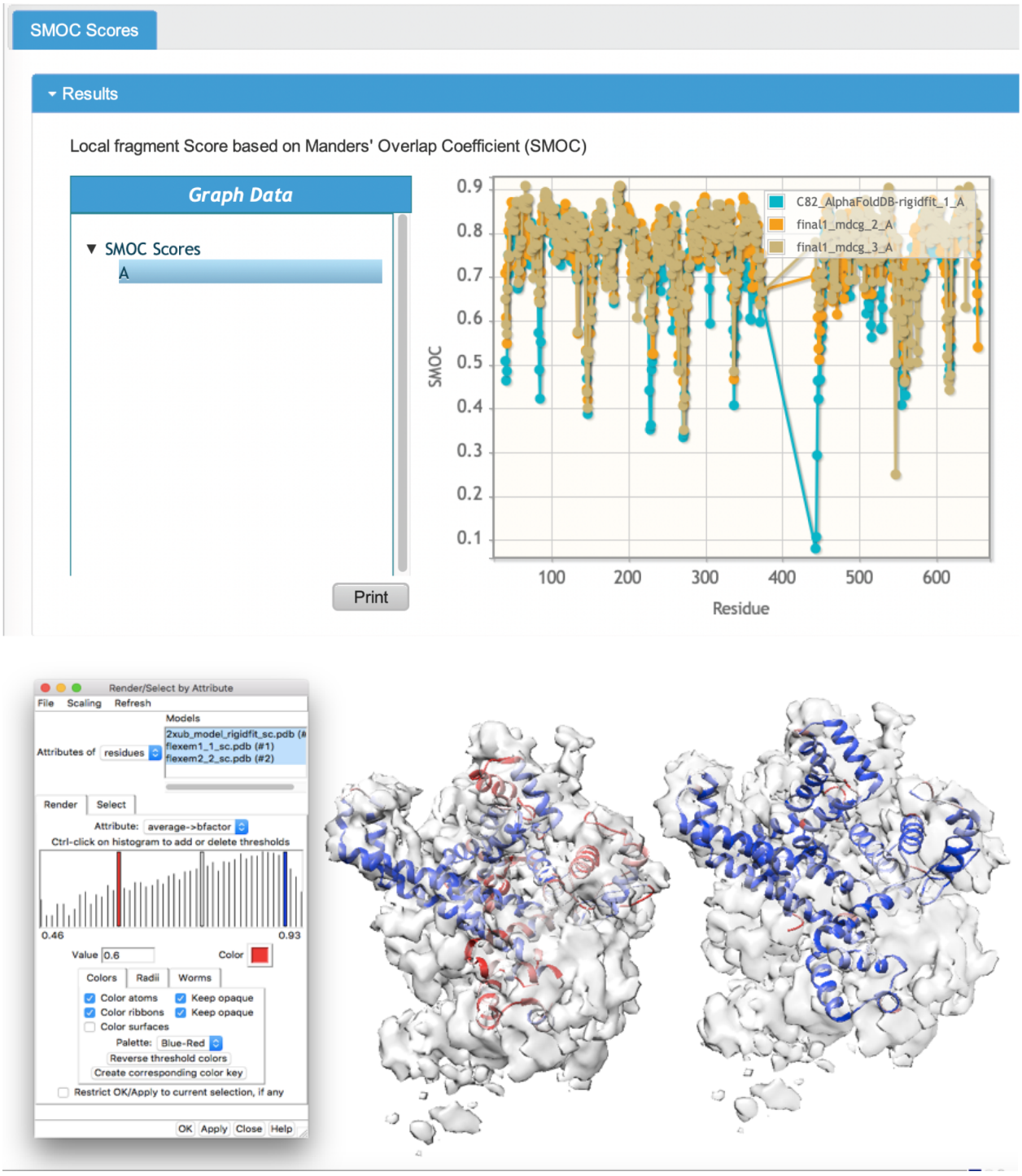
5. Assessment of improvement in local density fit

To compare the fitted models and evaluate the improvement in local density fit, we will use TEMPy SMOC (segment-based Manders' overlap coefficient) score2,7. A local overlap coefficient is calculated over voxels covered by each residue.

1. Select TEMPy Local scores task in the CCP-EM GUI interface
2. Input the map. Input pdb files to be scored : initial model **C82_AlphaFoldDB-rigidfit.pdb**, md1_2.pdb (from your first flexEM run) and final1_mdcg.pdb (from your second flexEM run)



Click Run and wait for the Results to be displayed.

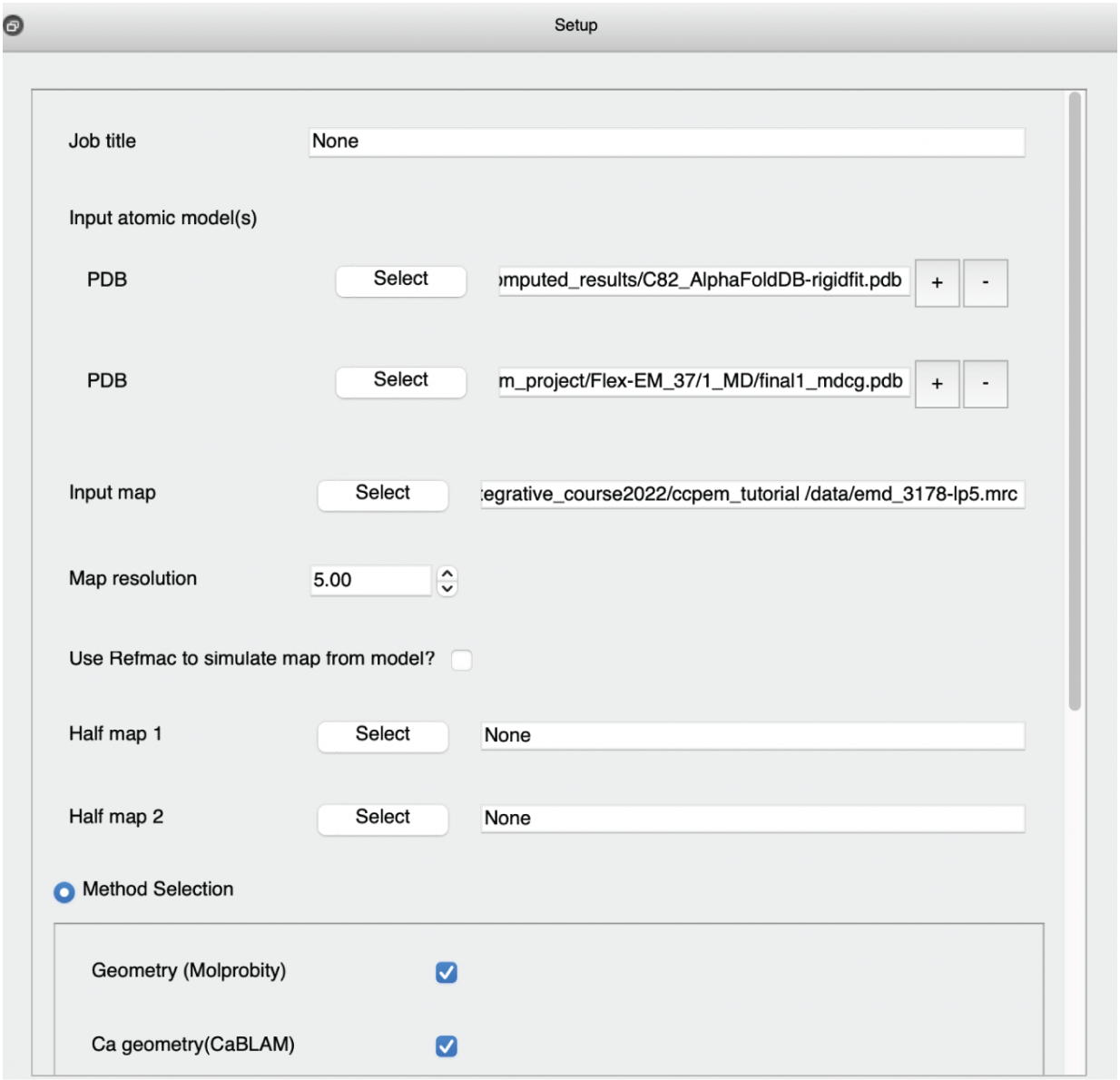


We can see that the overall local fit to density improves along the chain. When you click ‘View in Chimera’ the models open colored based on the SMOC scores. The range of values for coloring can be adjusted in the Structure Analysis window (Tools->Render by Attribute->Structure Analysis, choose residues/average bfactor)

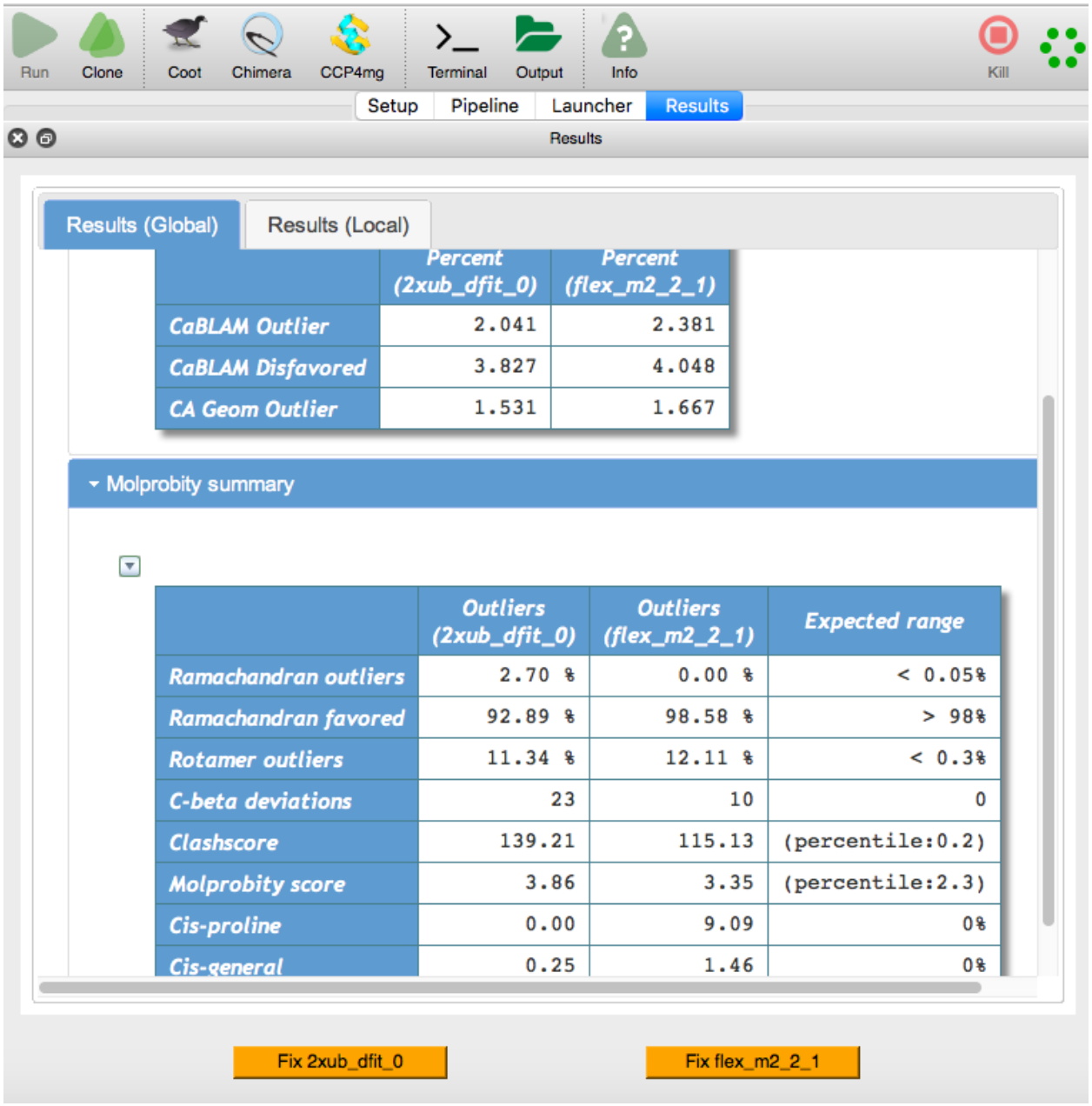
There are a few segments in the chain that are low scoring and require further inspection. Any low scoring areas can be further fixed interactively in Coot[8], followed by a round of Refmac[9] refinement to further improve the fit. Note that the residue at the chain terminus of the chain and chain breaks have relatively low scores. The geometry

6. Validating model geometry

The *Validation:Model* task in CCP-EM interface enables use of model quality evaluators like Molprobity (model geometry) and CaBLAM (backbone quality).



On the *Results (global)* tab, the summary of outlier statistics from Molprobity and CaBLAM appear.



Clearly, the model quality can be improved further by fixing clashes, Rotamer and Ramachandran outliers. Any severe issues can be fixed in Coot and the model can be refined further using Refmac wherein the quality usually improves as well.

References:

1. Topf, M. *et al.* Protein Structure Fitting and Refinement Guided by Cryo-EM Density. *Structure* **16**, 295–307 (2008).
2. Joseph, A. P. *et al.* Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods San Diego Calif* **100**, 42–49 (2016).
3. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
4. Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* **170**, 427–438 (2010).
5. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
6. Pandurangan, A. P. & Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. *J. Struct. Biol.* **177**, 520–531 (2012).
7. Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J. Struct. Biol.* **199**, 12–26 (2017).
8. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
9. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).