

Atomic model validation using CCP-EM software suite

Agnel P Joseph
CCP-EM, STFC, Harwell, UK
EMBO course, 15th Sep 2021

This tutorial demonstrates the use of tools for atomic model validation currently implemented in the CCP-EM software suite. The “Validation:model” task in CCP-EM provides an interface to these tools. They cover aspects of model geometry and fit to the map. Multiple models can be used as inputs, for example models from different stages of refinement, set of fitted models *etc.*

Data for this tutorial: model_validation_data.tar.gz. Please download the dataset and extract the files from the archive.

The data consists of

1. the target map T0104 (EMD-0406) which is a single-particle cryo-EM reconstruction of horse liver alcohol dehydrogenase at a resolution of 2.9Å.
2. one of the models submitted (T0104EM0n2_2.pdb) to the EMDB model metrics challenge.
3. The other model (6nbb_model1.pdb, target structure) is the one deposited by the original authors with the target map in EMDB (PDB ID: 6nbb).

For faster computation, the map was cropped using Chimera to a smaller grid (emd_0406_cropped.mrc). The deposited pdb (6nbb) consists of 10 models that represent local conformational variations in the target map. At the moment, the tools implemented in the validation interface do not support multi-model pdbs. For the purpose of this tutorial, we use the first model (6nbb_model1.pdb) from the ensemble.

From the CCP-EM main window, open the ‘Validation: model’ task window (Figure 1):

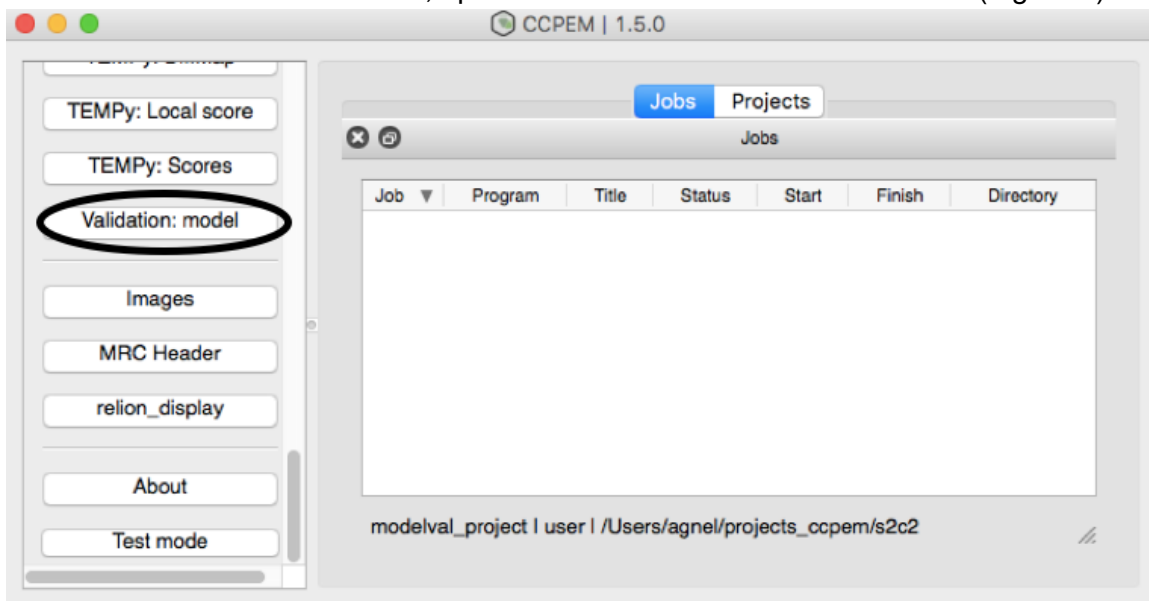


Figure 1: Validation: model task in CCP-EM

Input the job title, map (emd_0406_cropped.mrc) and models (6nbb_model1.pdb and T0104EM0n2_2.pdb) as below (click the “+” button to add a second input model, Figure 2). Select ‘Use Refmac to simulate map’. This uses Refmac to generate maps from atomic models using electron scattering factors and considers atomic B-factors in the map calculation. Hence, the synthetic maps are closer to the experimental maps than a simple atomic Gaussian map (default).

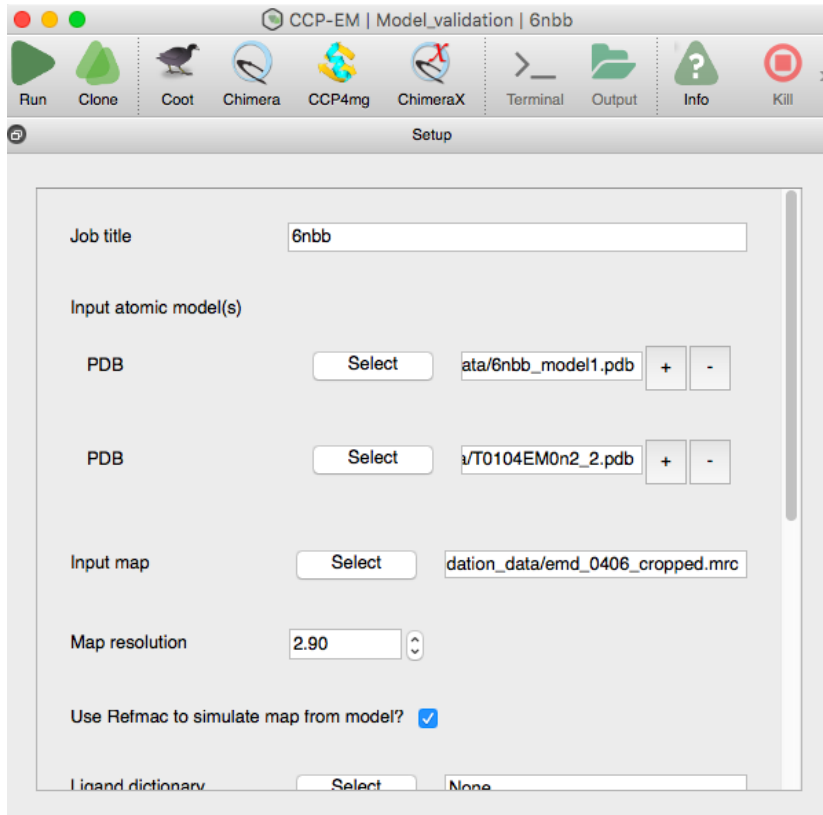


Figure 2: Input the map and models for assessment.

Select the validation metrics in Method Selection (Figure 3) and these are briefly explained below.

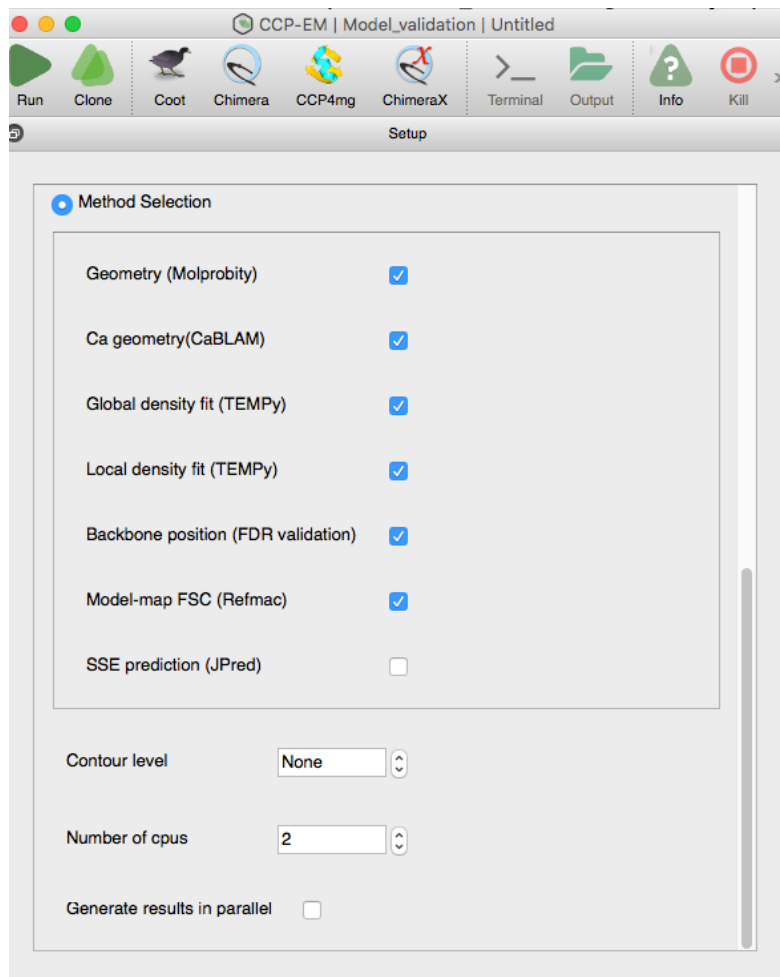


Figure 3: Select the methods to use for assessment

Leave the Contour level to “None” for now and the map is contoured automatically at 1.5 sigma for this calculation.

Selected tools: Molprobity¹ and CABLAM² for geometry checks; metrics from TEMPy for quantification of local³ and global map fit⁴; Confidence map based model validation⁵; REFMAC5⁶ for model-map FSC calculation.

Note: We haven’t selected JPred⁷ here: Jpred runs take some time at the moment as it connects to the Jpred4 server (<http://www.compbio.dundee.ac.uk/jpred4/index.html>) for the calculations. Jpred predicts secondary structure from sequence and is useful in some cases to compare the high confidence predictions of secondary structures vs modeled secondary structures to check if there are disagreements.

In the example screenshots we are using 2 cpus for the calculations but you should adjust the ‘Number of cpus’ according to the computer you are working on.

Click ‘Run’ to start the calculations. The job takes about 10 minutes for completion.

Once the job is finished, you will be able to see the Results tab with sub-tabs Global and Local (Figure 4).

On the Results (Global) tab, you can examine results under each section. Atomic B-factor distribution plots show that B-factors are perhaps refined for the deposited model (6nbb_model1.pdb, Figure 4) but not in the model T0104EM0n2_2.pdb where all atomic B-factors are zero.

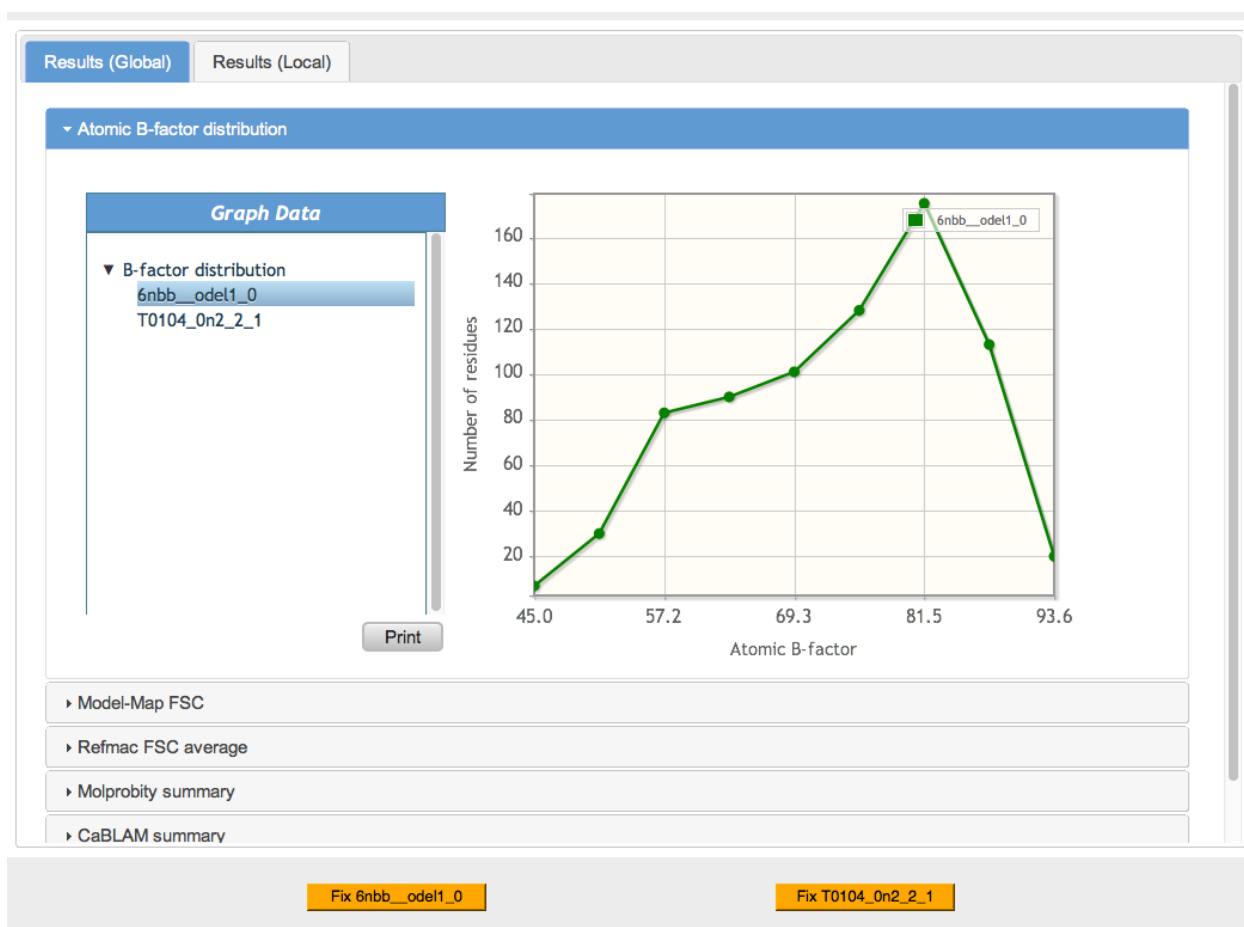


Figure 4: Results (Global): B-factor distribution

Expand the Model-Map FSC and select the map name from the legends to view the FSC curve. Look at the plots under 'Model-Map FSC'. Select 'emd_0_opped_0' (Figure 5, note that the model/map names are trimmed to 10 characters for ease of display) to compare model-map FSC plots of the two models. The plot clearly shows that the fit of model T0104EM0n2_2.pdb is worse than the deposited model 6nbb_model1.pdb.



Figure 5: Model-Map FSC

Under the next section 'Refmac FSC average', look at the table that compares the FSCavg values of the two models (Figure 6). The row 'FSCavg (FSC > 0.5)' gives FSCavg values calculated after ignoring resolution shells beyond model-map FSC 0.5. These figures confirm the results seen in the FSC curves, showing that 6nbb_model1.pdb fits the map better than T0104EM0n2_2.pdb.

Refmac FSC average

	emd_0_opped_0:6nbb_odel1_0	emd_0_opped_0:T0104_0n2_2_1
FSC average	0.5070	0.3939
FSC average (FSC > 0.5)	0.7456	0.6602

Figure 6: Refmac FSC average

'MolProbity summary' gives the comparison of Molprobity statistics (Figure 7). The expected range of values are in the column in the right. Lower 'Clashscore' indicates fewer serious clashes between atoms in the structure. Higher percentile reflects better quality or fewer clashes (among other structures solved at similar resolutions). The 'Molprobity score' is a combined score involving other measures, and smaller values indicate better quality. T0104EM0n2_2.pdb has more Ramachandran favored amino acids but has a large number of severe clashes (Clashscore: 66.28), and the Molprobity score is much worse (2.38), compared to the deposited model.

▼ Molprobability summary



	Outliers (6nbb__odel1_0)	Outliers (T0104_0n2_2_1)	Expected range
<i>Ramachandran outliers</i>	0.00 %	0.00 %	< 0.05%
<i>Ramachandran favored</i>	95.70 %	97.58 %	> 98%
<i>Rotamer outliers</i>	0.00 %	0.00 %	< 0.3%
<i>C-beta deviations</i>	0	0	0
<i>Clashscore</i>	5.17 (percentile:57.6)	66.28 (percentile:1.3)	
<i>Molprobability score</i>	1.57 (percentile:63.8)	2.38 (percentile:27.5)	
<i>Cis-proline</i>	5.00	5.00	0%
<i>Cis-general</i>	0.00	0.00	0%

Figure 7: MolProbability summary

'CaBLAM summary' (Figure 8) gives the percent of outlier types identified by CaBLAM. Generally, the expected range is <5% outliers.

▼ CaBLAM summary



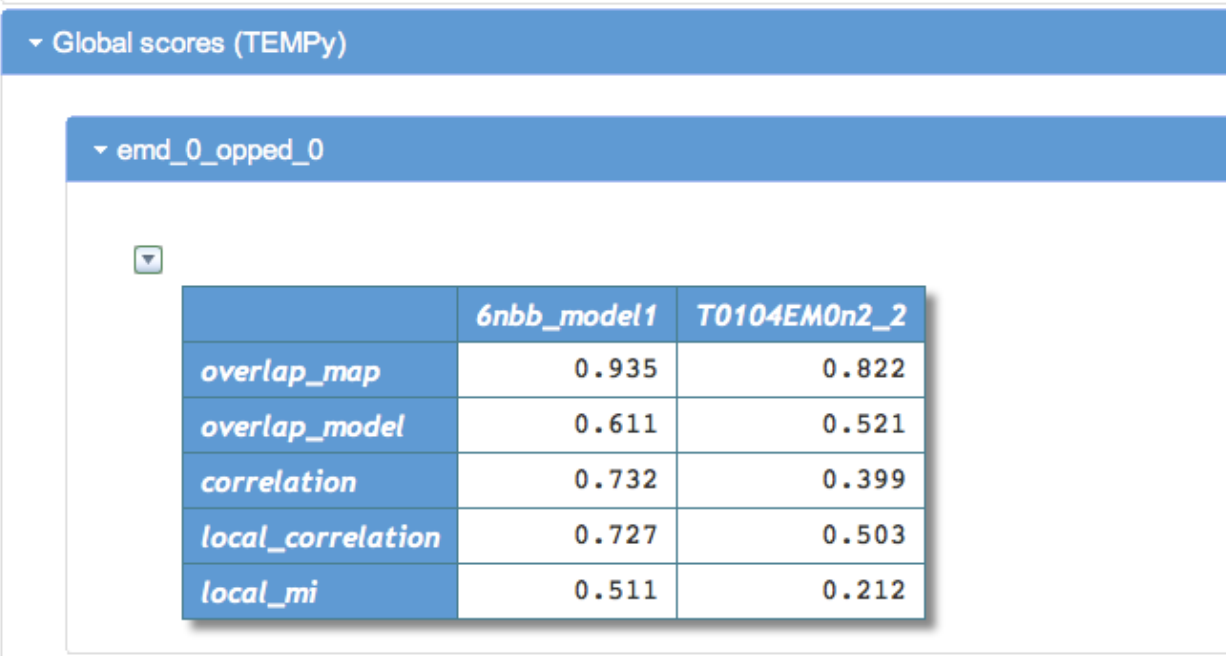
	Percent (6nbb__odel1_0)	Percent (T0104_0n2_2_1)
<i>CaBLAM Outlier</i>	2.297	2.162
<i>CaBLAM Disfavored</i>	5.541	4.054
<i>CA Geom Outlier</i>	0.811	1.081

Figure 8: CaBLAM summary

The Global TEMPy⁵ scores (Figure 9) are measures reflecting model-map agreement in real-space. 'Overlap_map' and 'overlap_model' give the fraction of map covered by the model and the fraction of model covered by the map, respectively⁶. The map is contoured automatically at 1.5 sigma for this calculation. In this case, the auto contour value is not ideal and hence the low overlap values. You can check the contour value chosen in the Pipeline log. Support for user input map contours will be added in the future releases.

Correlation gives global map-model correlation while local_correlation reflects the correlation within the area of overlap between the model and the map (useful for example when the model is not complete). The local_mi gives the Mutual Information⁶ score calculated within the area of

overlap. Values below 0.5 usually point to issues with the model fit. The absolute values of correlation and mutual information are affected by map resolution and post-processing.



The screenshot shows a web interface with a blue header bar containing the text "Global scores (TEMPy)". Below this is a sub-header bar with the text "emd_0_opped_0". Underneath the sub-header is a small square icon with a downward arrow. Below the icon is a table with three columns and six rows. The first column contains metric names, and the next two columns contain numerical values for two different models.

	<i>6nbb_model1</i>	<i>T0104EM0n2_2</i>
<i>overlap_map</i>	0.935	0.822
<i>overlap_model</i>	0.611	0.521
<i>correlation</i>	0.732	0.399
<i>local_correlation</i>	0.727	0.503
<i>local_mi</i>	0.511	0.212

Figure 9: Global scores TEMPy

The 'Results (Local)' tab gives the details of 'outliers' detected by different tools, per residue. The outlier details are divided by chain. Under Chain A, you will find residue outliers based on different metrics: Molprobity, CaBLAM and TEMPy SMOC⁷ and FDR backbone score⁵.

The per-residue score plot provides SMOC-segment-based Manders' overlap coefficient (SMOC) and FDR validation scores (Figure 10) for each residue in the chain. You can see that the deposited model has a much better fit. Note that lower SMOC scores need not necessarily mean residue mis-fit but can reflect low resolution or disorder in the map. SMOC score is explained briefly in additional notes.

The FDR validation score evaluates whether the position of the residue backbone (backbone trace) is inside the molecular contour or background. Scores lower than 0.9 usually require attention.

To visualize scores for the model: T0104EM0n2_2.pdb, toggle the models using the option at the bottom of the plot.



Figure 10: SMOC and FDR validation scores

The 'Summary table' (Figure 11) gives the summary of residue outliers clustered by proximity in space. Each cluster is a group of neighboring residues where outliers were detected by any of the metrics.

Results (Global)

Results (Local)

Summary table

Outlier summary table: 6nbb__odel1_0

Cluster	Residue	Molprobrity outliers	CaBLAM outliers	FDR outliers	SMOC outliers
1	109	backbone clash	-	-	-
	124	side-chain clash	-	-	Outlier
	125	-	-	-	Outlier
	126	-	-	Outlier	-
	128	-	-	Outlier	-
	129	side-chain clash	-	-	-
	153	-	-	-	Outlier
	154	-	-	-	Outlier
	155	side-chain clash	-	-	Outlier
	156	side-chain clash	-	-	-
	174	-	Outlier	-	-
	256	-	-	Outlier	-
	259	-	Outlier	-	-
	263	-	-	-	Outlier
	282	side-chain clash	-	-	Outlier
	284	-	-	-	Outlier
	287	side-chain clash	-	-	-
	288	side-chain clash	-	-	-
	313	side-chain clash	-	-	-
	318	backbone clash	-	-	Outlier

Fix 6nbb__odel1_0

Fix T0104_0n2_2_1

Figure 11: Summary Table

Click the button at the bottom of the window : 'Fix 6nbb__odel1_0' to open the model in Coot⁸ (Figure 12) along with a window listing all the outliers identified in this model (Molprobability to-do list). You can click on the list of outliers to navigate to them and try to fix manually them in Coot. Fixed residues can be flagged by checking the box under the 'Dealt with' column in the 'Molprobability to-do list' window. Try fixing some of the outliers using Coot refinement/regularize/rotamer tools. Use the Coot validation tools to guide the fixing process.

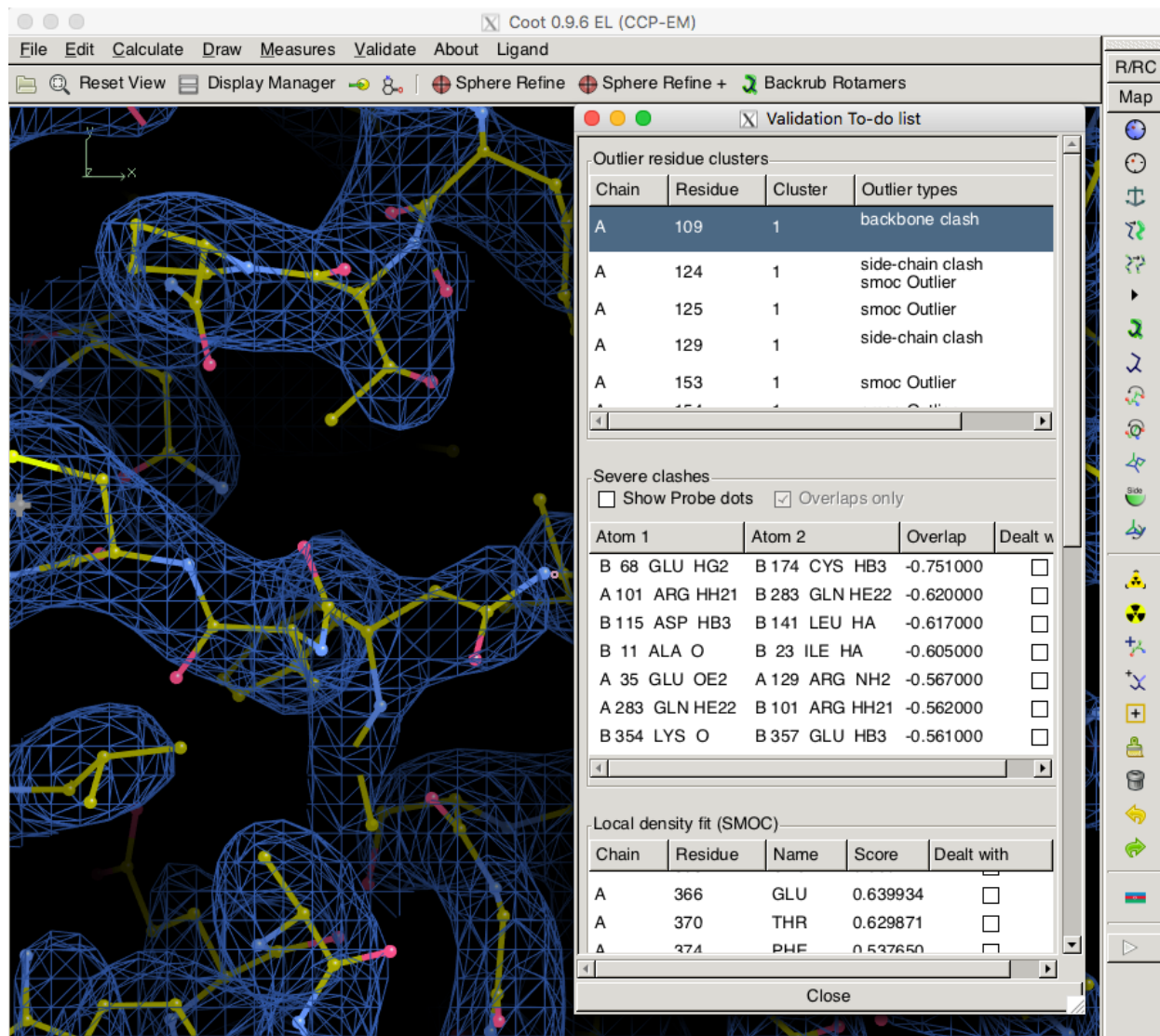


Figure 12: Molprobability to-do list in Coot

Additional Notes:

1. Refining B-factors may affect the metrics indicating fit-to-map significantly. You can test this later by refining B-factors for T0104EM0n2_2.pdb in Refmac by entering "refi bonly" in the Keywords input. Check how the fit-to-data metrics change with the refined model (refined.pdb).
2. CCP-EM validation task is currently not accessible directly from Coot, so if you want to re-calculate the CCP-EM validation scores after fixing outliers you will need to run the validation task again on the edited model. We hope to make this process smoother in the next CCP-EM release.
3. SMOC scores: Instead of the standard cross-correlation coefficient involving deviation from the mean, the score is calculated using the Manders' Overlap Coefficient, which is related to CCC⁷.

References:

1. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci. Publ. Protein Soc.* **27**, 293–315 (2018).
2. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink “waters,” and NGL Viewer to recapture online 3D graphics. *Protein Sci.* **29**, 315–329 (2020).
3. 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).
4. 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).
5. Olek, M. & Joseph, A. P. Cryo-EM Map–Based Model Validation Using the False Discovery Rate Approach. *Front. Mol. Biosci.* **8**, (2021).
6. Nicholls, R. A., Tykac, M., Kovalevskiy, O. & Murshudov, G. N. Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 492–505 (2018).
7. Drozdetskiy, A., Cole, C., Procter, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).
8. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).