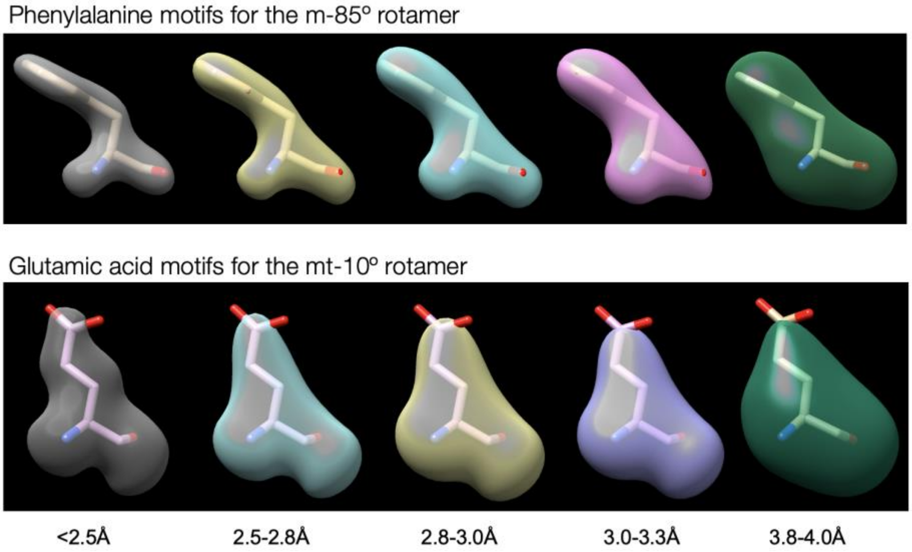
Introduction to map-model validation with Strudel Score

Strudel Score is a map-model validation method which entails calculating the linear correlation coefficient between an amino-acid map-motif from the Strudel library and the cryo-EM map values around a target residue.

The Strudel amino-acid motif library (see picture below) contains 3D motifs mined from EMDB maps by averaging large numbers of map fragments that correspond to amino-acid residues rotamers. The library covers seven resolution bands (0.0-2.3, 2.3-2.5, 2.5-2.8, 2.9-3.0, 3.0-3.5 and 3.5-4.0Å). Each resolution band contains up to 155 map motifs which correspond to distinct amino-acid rotamers as defined by the penultimate rotamer library (Lovell et all 2000). (In the higher resolution bands some motifs may be missing if there were not enough (<10) observations).

This tutorial consists of two parts: calculation of Strudel Scores with the **strudel_mapMotifValidation.py** program (part of the threed_strudel Python package) and inspection and visualisation of the validation results with the Strudel Score plugin for ChimeraX.

In this tutorial we will validate the models T0102EM054_1 and T0102EM010_1 submitted to the Model Challenge 2019 . Both models are an interpretation of the apoferritin target map T0102 (EMD-20027).


Phenylalanine motifs for the m-85° rotamer
Glutamic acid motifs for the mt-10° rotamer
<2.5Å
2.5-2.8Å
2.8-3.0Å
3.0-3.3Å
3.8-4.0Å

Examples of Strudel motifs for two particular rotamers, in different resolution bands.

Requirements

Mac or Linux computer.

Preparation

1. Install ChimeraX version >= 1.2
2. Install the Strudel Score plugin for ChimeraX:
   ChimeraX --exit --nogui --cmd 'toolshed install strudelscore'

Or using the ChimeraX toolshed menu: Tools/More Tools…/StrudelScore

3. Create a directory for this tutorial, e.g *strudel* in your home directory:
   ```
   mkdir ~/strudel ; cd ~/strudel
   ```
4. Download and unpack the latest version of the Strudel motif library (file ending voxel-0.5.tar) in the directory you just created from:
   https://ftp.ebi.ac.uk/pub/databases/emdb_vault/strudel_libs/
   now you should have a new directory called ~/strudel/strudel-libs_ver-2.0_voxel-0.5 containing the Strudel motif libraries
5. In your strudel directory create another directory for the tutorial data:
   ```
   mkdir data
   ```
   Download the tutorial data from https://drive.google.com/drive/folders/1Yn12qo0xgw_4tm1dAEUQFImobH_LZ0Hx?usp=sharing to the data directory.

threed_strudel installation

```
pip install threed_strudel
```

Set ChimeraX path

```
strudel_setChimeraX.py path_to_chimeraX_executable
```

Run the Strudel validation

This step requires approximatively 30 seconds per residue on a single compute core (2.9 GHz intel core i7). For a 175-residue protein the calculations will require ~20 minutes in total using 4 cores or ~5 minutes using 16 cores. For faster results we recommend the use of a compute cluster. The bulk of the compute time is required to align and calculate correlations between each residue and each of the 155 motifs in the library (175*155 times in this case).

If you used the suggested folder naming the command for running the validation step is:

```
strudel_mapMotifValidation.py -np 4 -p data/T0102EM054_1_err.cif -m data/T0102.map -l strudel-libs_ver-2.0_voxel-0.5/motifs_0.0-2.3 -o T0102EM054_out
```

where
-np – number of cores

for more options run: *strudel_mapMotifValidation.py -h*

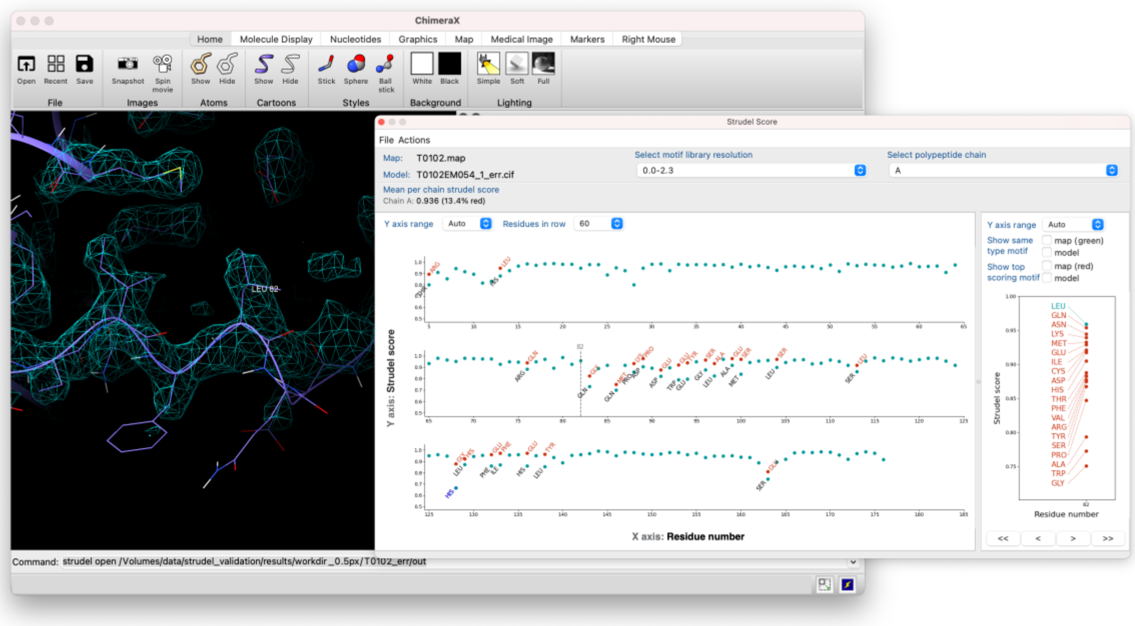
If the program execution was interrupted for any reason run the same command again; the program execution will continue from where it stopped.

After the program finishes, you can inspect the validation results using the Strudel Score plugin in ChimeraX. This can be done in several ways:

1.
   ```
   ChimeraX   --cmd 'strudel open ~/strudel/T0102EM054_out'
   ```

2. 2.1 Open ChimeraX
   2.2 On the ChimeraX command line type*:*
   strudel open ~/strudel/T0102EM054_out
3. 3.1 Open ChimeraX
   3.2 Select: Tools/General/Strudel Score
   3.3 In the Strudel Score window: select File/Open Project and select the validation output folder (~/strudel/T0102EM054_out)

The plugin looks like this:



Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score
Chain A: 0.936 (13.4% red)

Y axis range: Auto

Residues in row: 60

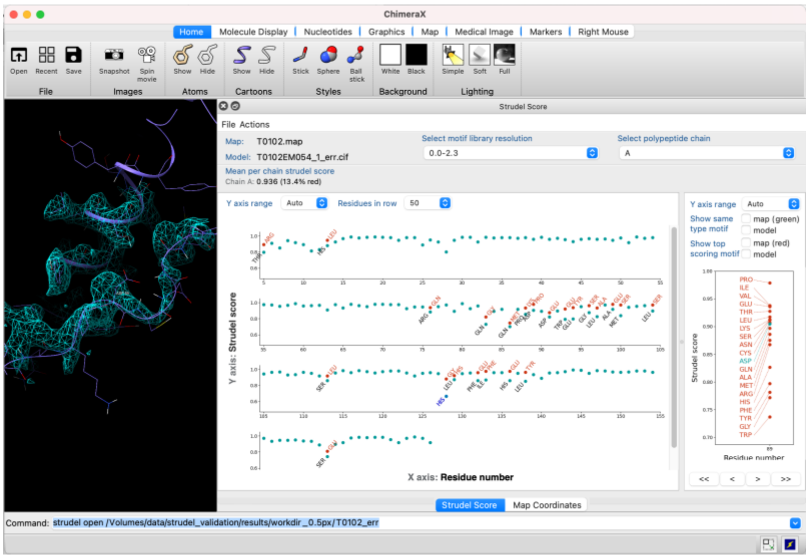
Y axis range: Auto

Show same type motif: □ map (green) □ model

Show top scoring motif: □ map (red) □ model

Command: strudel open /Volumes/data/strudel_validation/results/workdir_0.5px/T0102_err/out

You can keep Strudel Score as a separate window or fit it into the ChimeraX main window by dragging it into the desired position.



Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score
Chain A: 0.936 (13.4% red)

Y axis range: Auto

Residues in row: 50

Y axis range: Auto

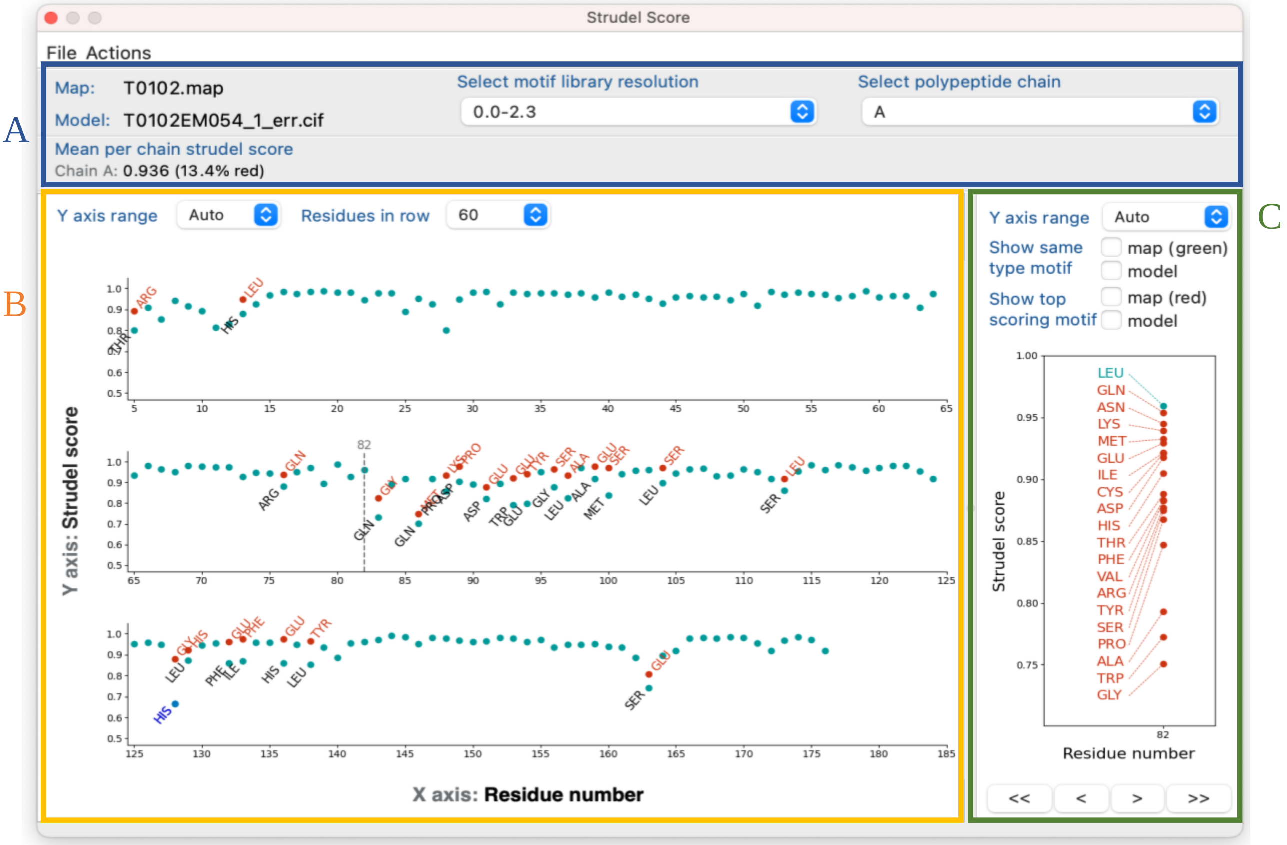
Show same type motif: □ map (green) □ model

Show top scoring motif: □ map (red) □ model

Command: strudel open /Volumes/data/strudel_validation/results/workdir_0.5px/T0102_err

The plugin window consists of three panels:
A – the info and selection panel

B – the plot panel
C – the residue-specific panel



Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

Y axis range: Auto

Residues in row: 60

Y axis range: Auto

Show same type motif: □ map (green) □ model

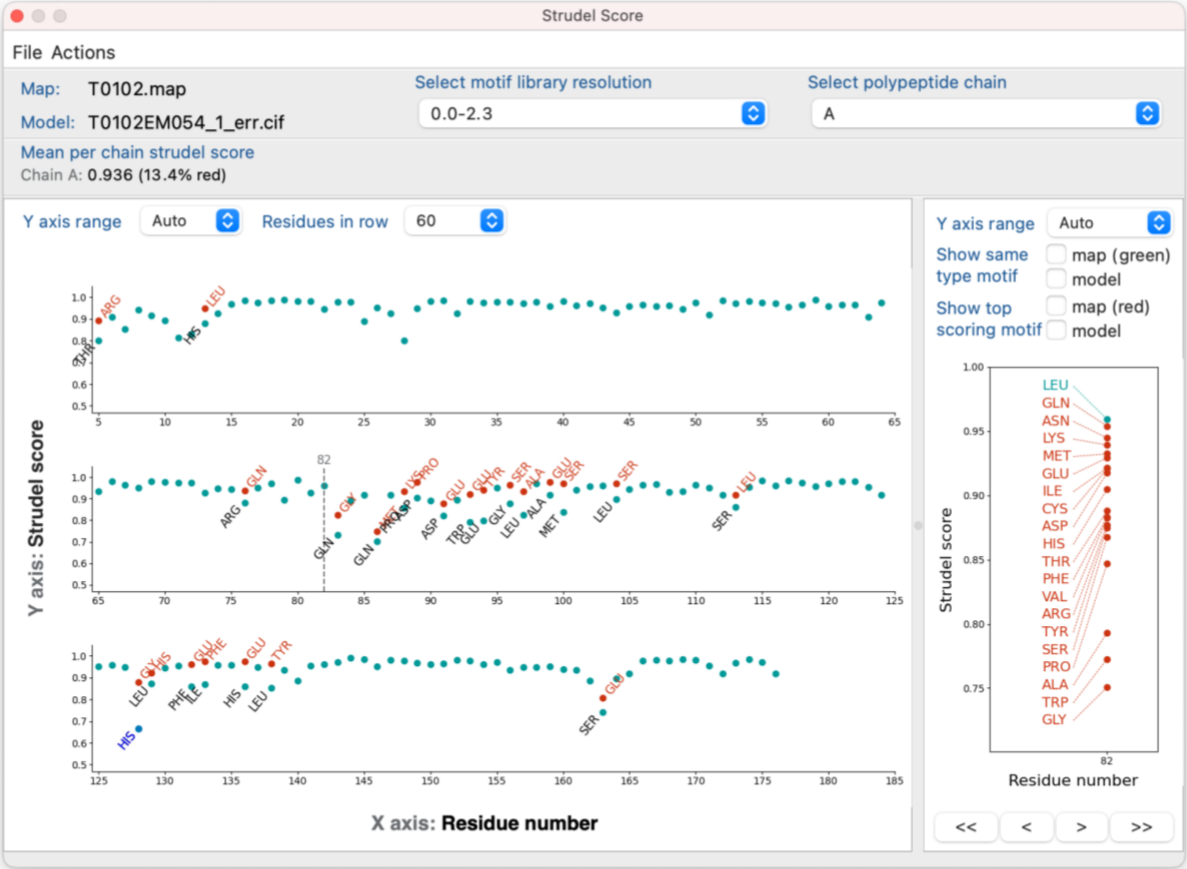
Show top scoring motif: □ map (red) □ model

The info and selection panel contains (from left to right):
a) The map and model names.
b) Average strudel score and the percentage of outliers for each chain.
c) A drop-down menu showing the resolution range of the Strudel library against which the validation was performed. The user can switch between the data for different resolution ranges if the results are available.

The plot panel contains:
a) Scatter plots showing the Strudel score as a function of residue number. A teal dot represents a residue which scores best (within 5%) with a library motif of the same residue type. If this is not the case, the residue is considered an outlier and a red dot is shown for the top-scoring residue type. Upon clicking on a residue in the graph the residue-specific panel the panel on the right will show the scores for the best-scoring rotamer motif of each residue type.
b) Drop-down menus which control the Y axis range and the maximum number of residues in each plot.

The residue-specific panel
a) For a selected residue, the scatter plot shows the scores for the best-scoring rotamer motif of each residue type. The motif of the same type as the modelled residue is shown in teal, all others in red. The labels in the right panel can be clicked to show the motifs aligned with the target residue in the map.

b) The arrow buttons at the bottom allow navigation to the next/previous residue in the sequence (single arrows) and next/previous outlier residue (double arrows) in the sequence.
c) Tick boxes control showing/hiding of the same type and top scoring motifs model and map.



Strudel Score

Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score
Chain A: 0.936 (13.4% red)

Y axis range: Auto

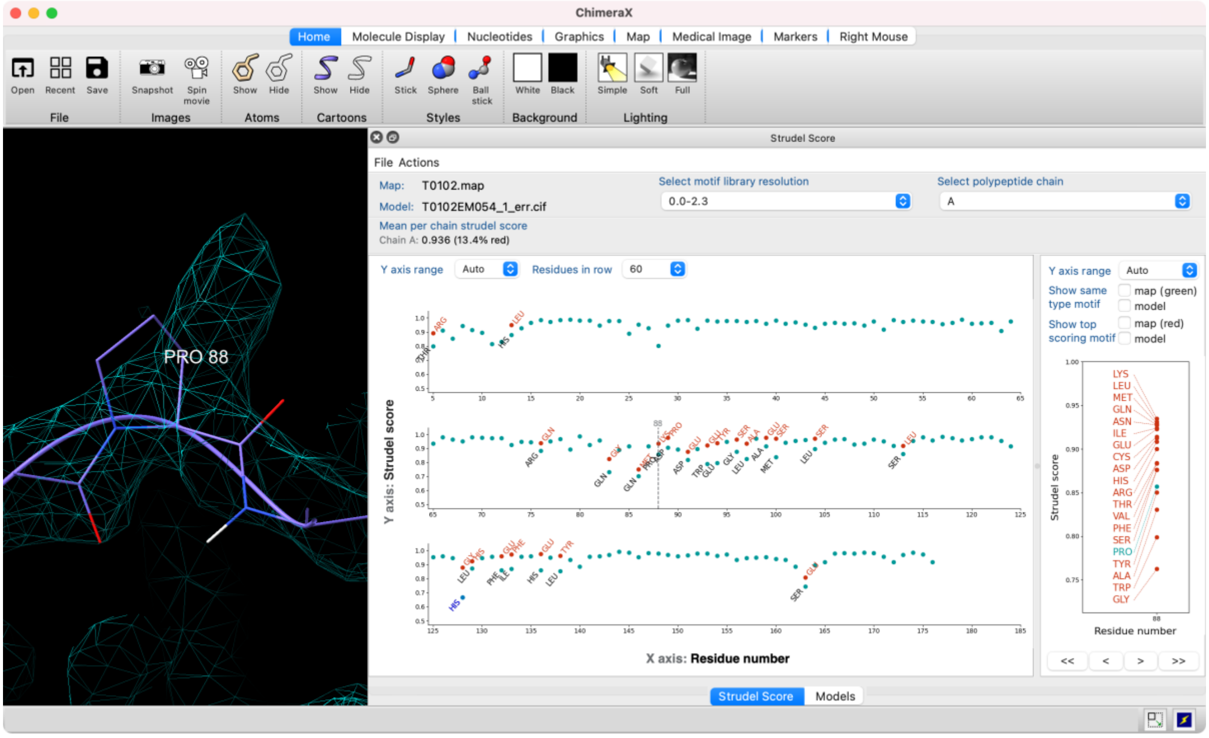
Residues in row: 60

Y axis range: Auto

Show same type motif: □ map (green) □ model

Show top scoring motif: □ map (red) □ model

After a quick glance at the Strudel Score plot, you can see that the region 76-113 in chain A has many outliers. To inspect an outlier, click on the corresponding dot in the Strudel Score plot. If you look at residue A88, you will notice that it was modelled as a proline. However, it scores poorly with the proline library motifs, ranking only 16th out of 20 residue types (see residue-specific panel). In contrast, the motifs of several residue types with longer side chains such as Lys, Leu and Met score significantly better.



Map: T0102.map

Model: T0102EM054_1_err.cif

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Y axis range: Auto

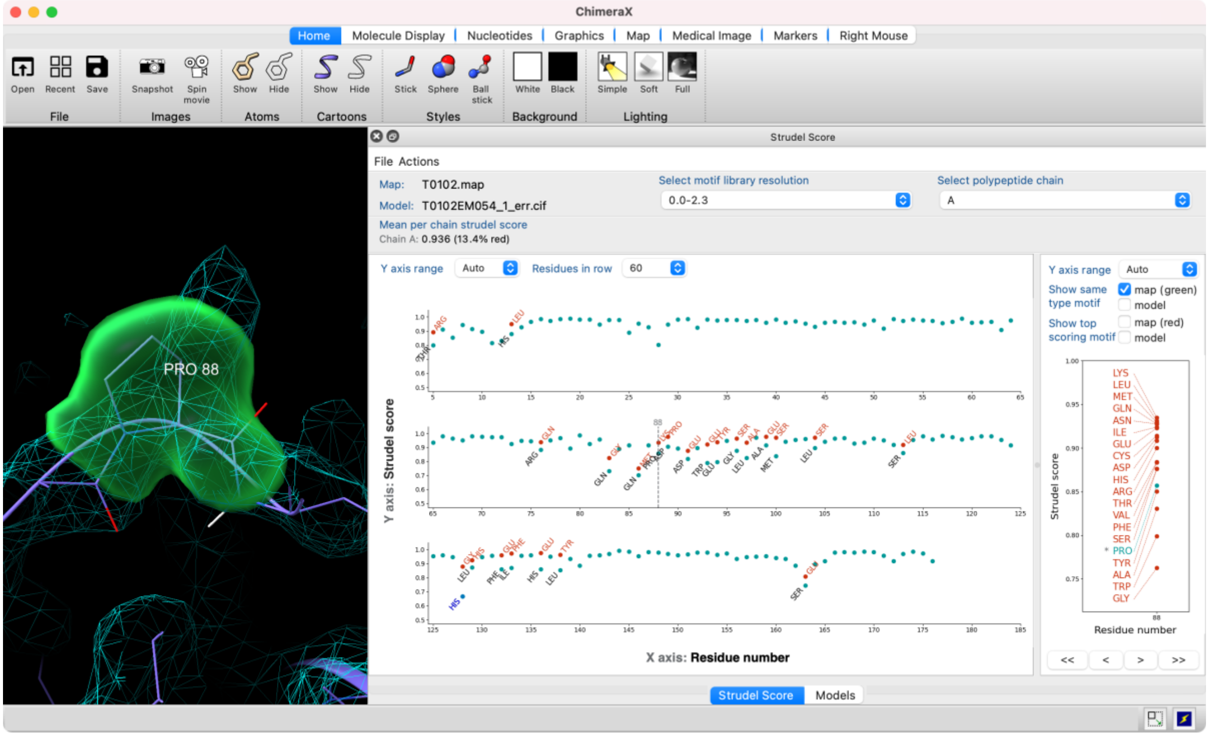
Residues in row: 60

Y axis range: Auto

Show same type motif: ☐ map (green) ☐ model

Show top scoring motif: ☐ map (red) ☐ model

If you click on a residue name in the residue-specific panel (right panel) the corresponding map motif for the best-matching rotamer of that residue type will be shown. If we click on proline (PRO), the best-fitting proline motif clearly matches poorly with the map for this residue. (It is shown in green as this is the residue type that was modelled here.)



Map: T0102.map

Model: T0102EM054_1_err.cif

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Y axis range: Auto

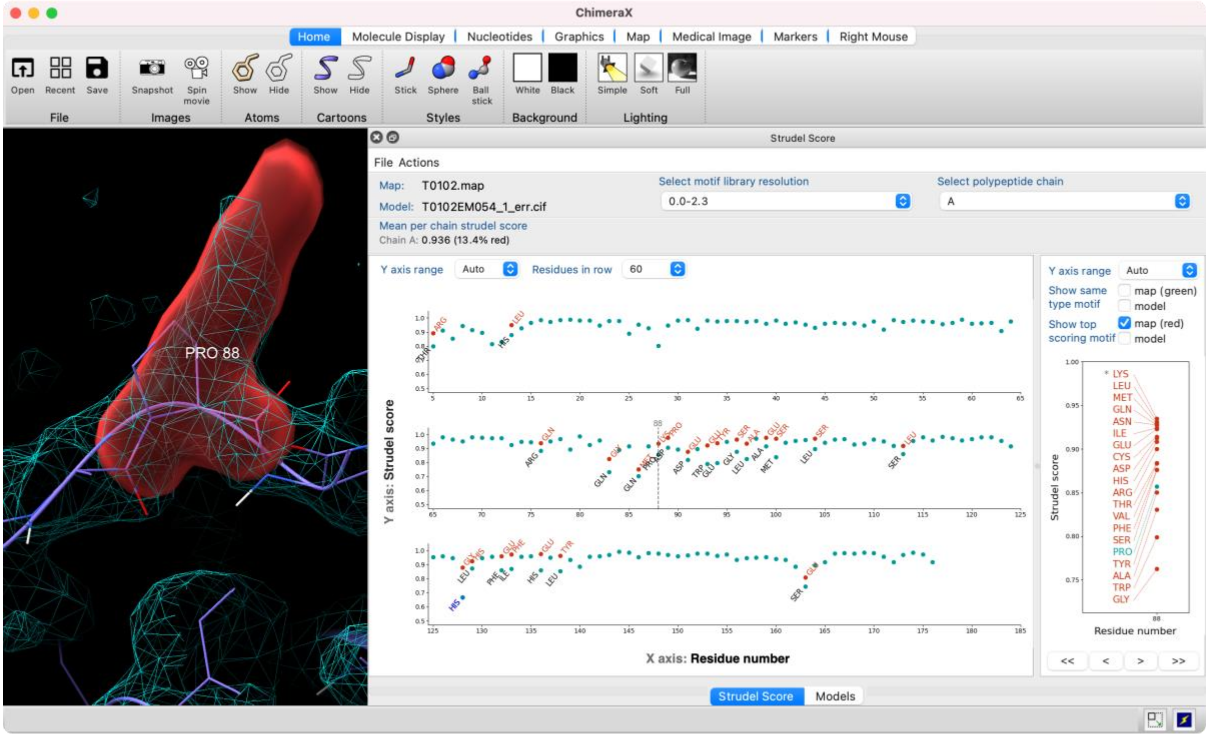
Residues in row: 60

Y axis range: Auto

Show same type motif: ☑ map (green) ☐ model

Show top scoring motif: ☐ map (red) ☐ model

Top-scoring motifs such as those for Lys, Leu and Met (shown in red) are a better match to the map for this residue.



Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

Y axis range: Auto

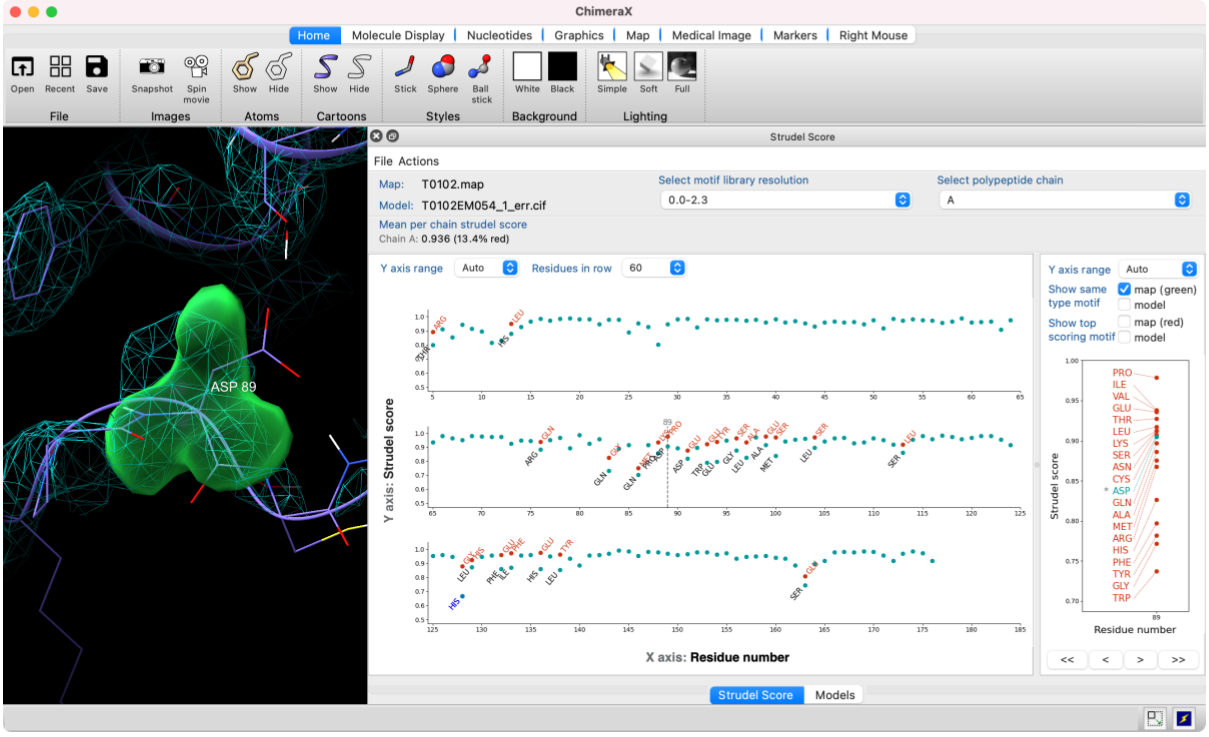
Residues in row: 60

Y axis range: Auto

Show same type motif: □ map (green) □ model

Show top scoring motif: ☑ map (red) □ model

For the next outlier (residue A89) you will see a similar situation. The residue was modelled as an aspartic acid (Asp), but it scores much better with a proline motif. And indeed, upon inspection, the proline motif matches much to the region of the residue A89 in the experimental map.



Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

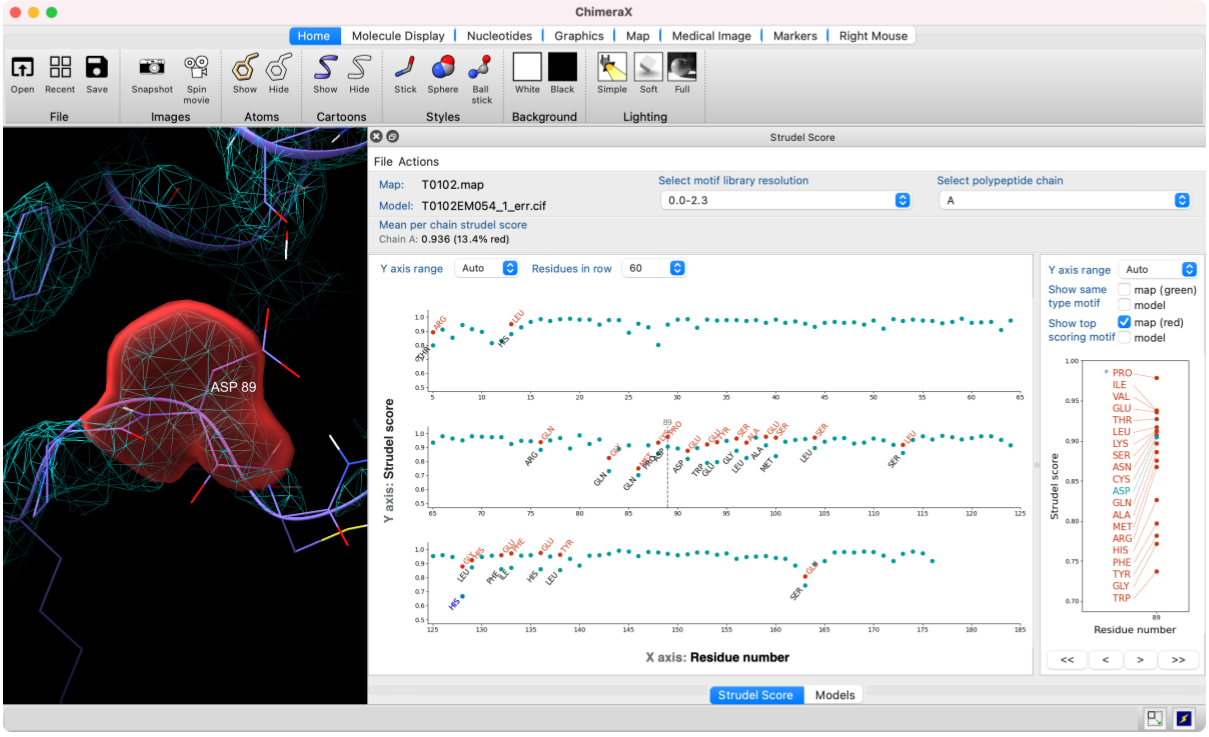
Y axis range: Auto

Residues in row: 60

Y axis range: Auto

Show same type motif: ☑ map (green) □ model

Show top scoring motif: □ map (red) □ model



Map: T0102.map

Model: T0102EM054_1_err.cif

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Y axis range: Auto

Residues in row: 60

Y axis range: Auto

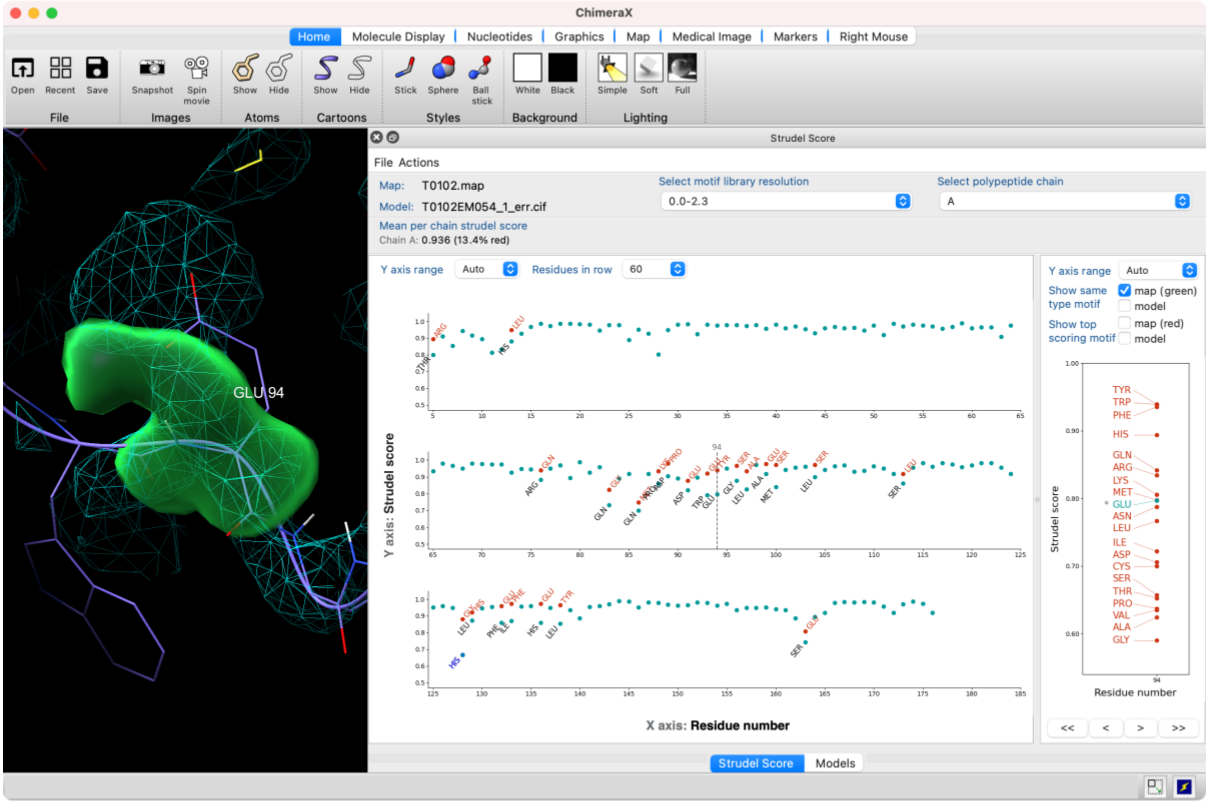
Show same type motif: □ map (green) □ model

Show top scoring motif: ☑ map (red) □ model

There are three proline residues in this protein sequence (marked in red below). Only one of them is in the examined region (Pro A88, underlined in the sequence) and you observed that this residue was also an outlier. On the other hand, the map for the next residue, A89, looks very much like you would expect for a proline. Thus, it can be hypothesised that there is one-residue register error (or register shift) in this part of the model. To confirm or refute this hypothesis, you will need to inspect several more residues in this region.

```
TTASTSQVRQNYHQDSEAAINRQINLELYASYVYLSMSYYFDRDDVALKNFAKYFLHQSHEEREHAEKLM
KLQNQRGGRIFLQDIKKPDCDDWESGLNAMECALHLEKNVNQSLLELHKLATDKNDPHLCDFIETHYLN
EQVKAIKELGDHVTNLRKMGAPESGLAEYLFDKHTLGDSDNES
```

Have a closer look at residue 94. This residue was modelled as a glutamic acid (Glu). However, the Glu motif does not fit well to this region of the map and scores much lower than those of several aromatic residues, Tyr, Trp and Phe. If you refer back to the sequence, you will notice that the previous residue is actually a tryptophane. This lends further support to your register-error hypothesis. Confirm by careful inspection that residues 96, 100, 104 also appear to have been modelled in map features belonging to a neighbouring residue.



File Actions

Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

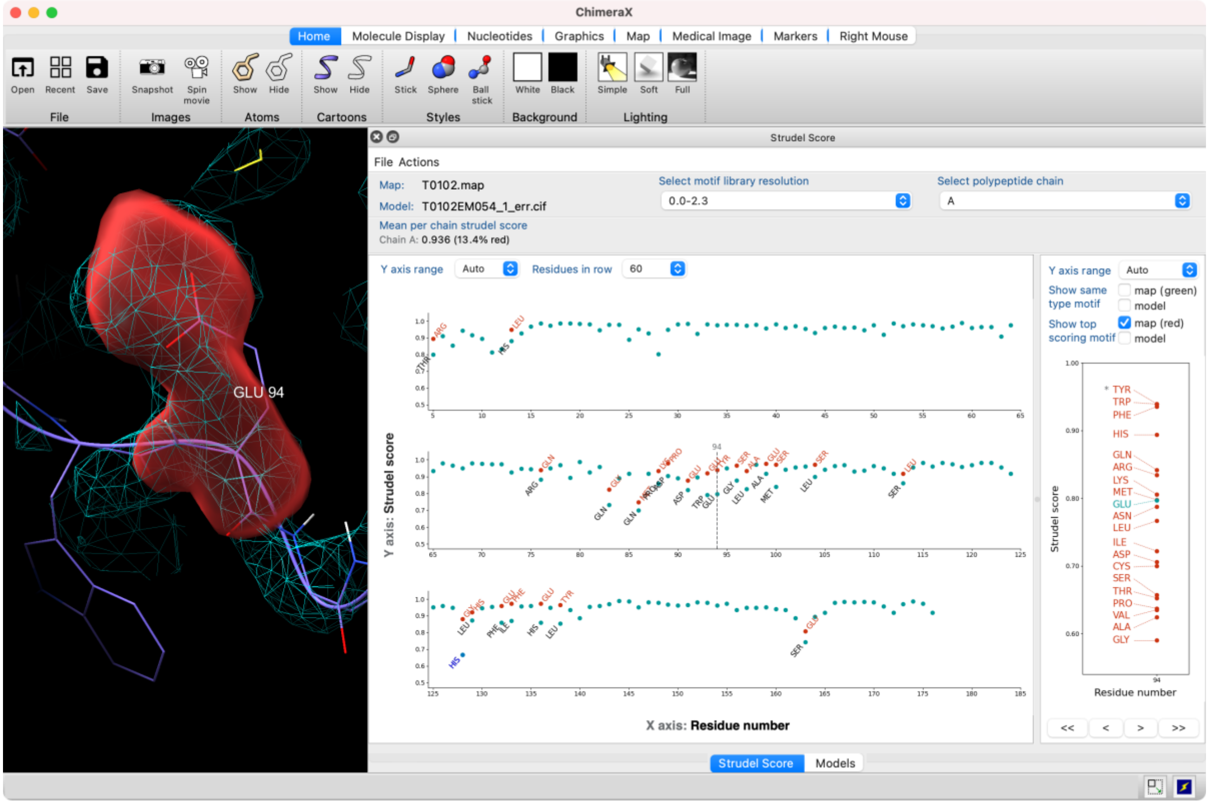
Y axis range: Auto

Residues in row: 60

Y axis range: Auto

Show same type motif: ☑ map (green) □ model

Show top scoring motif: □ map (red) □ model



File Actions

Map: T0102.map

Model: T0102EM054_1_err.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score

Chain A: 0.936 (13.4% red)

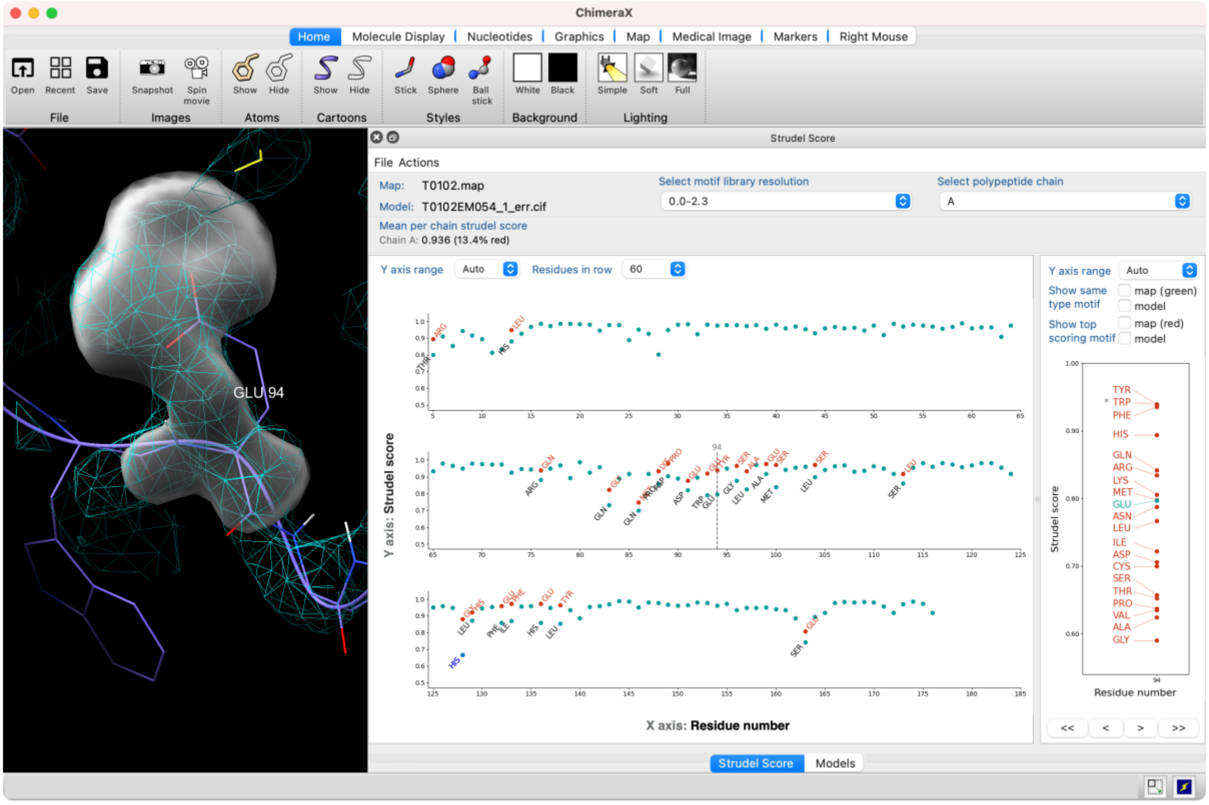
Y axis range: Auto

Residues in row: 60

Y axis range: Auto

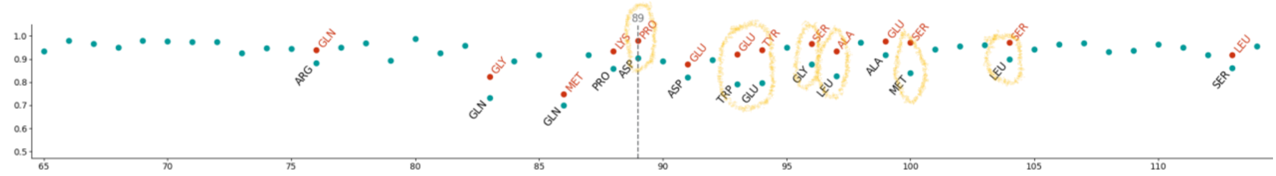
Show same type motif: □ map (green) □ model

Show top scoring motif: ☑ map (red) □ model

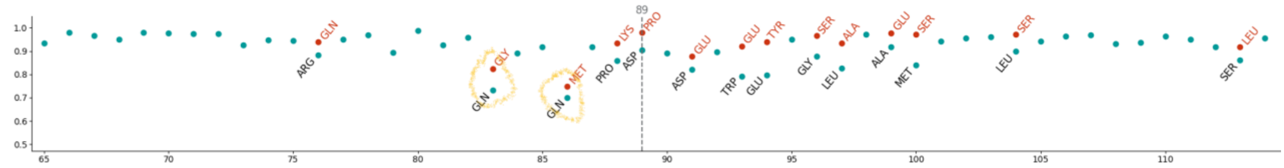


Map: T0102.map
Model: T0102EM054_1_err.cif
Mean per chain strudel score
Chain A: 0.936 (13.4% red)
Select motif library resolution: 0.0-2.3
Select polypeptide chain: A
Y axis range: Auto
Residues in row: 60
Y axis range: Auto
Show same type motif: □ map (green) □ model
Show top scoring motif: □ map (red) □ model

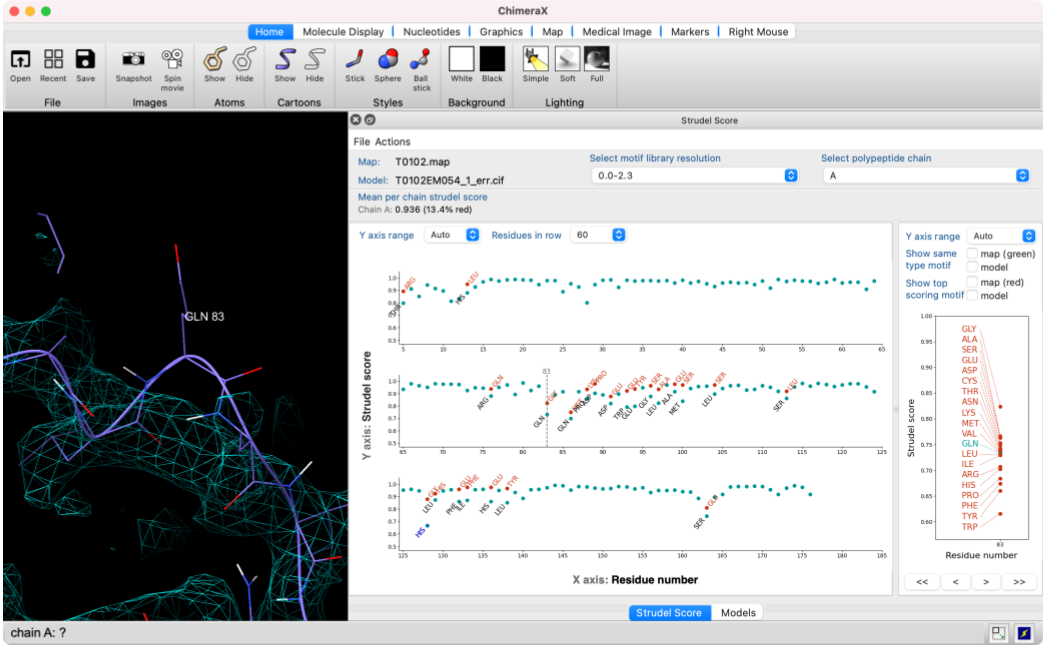
Referring to the Strudel Score plot, you may notice that in all the cases examined above the score values for the top-scoring motifs are relatively high (circled in yellow in the figure below), comparable to those of residues that are not outliers (e.g., in the region from residue A35 to A55). This strongly suggests that in each of these cases, the correct residue type is likely to be among the top 2-3 suggested alternatives.



The figure below highlights two outliers where both the modelled residue type and the best-matching one have scores that are significantly lower than those of non-outlier residues. In such cases, it is likely that the map in that region is poor, or that the residue was built (mostly) outside the map.



An example of this is shown below for Gln A83 which has somehow been placed entirely outside the map.

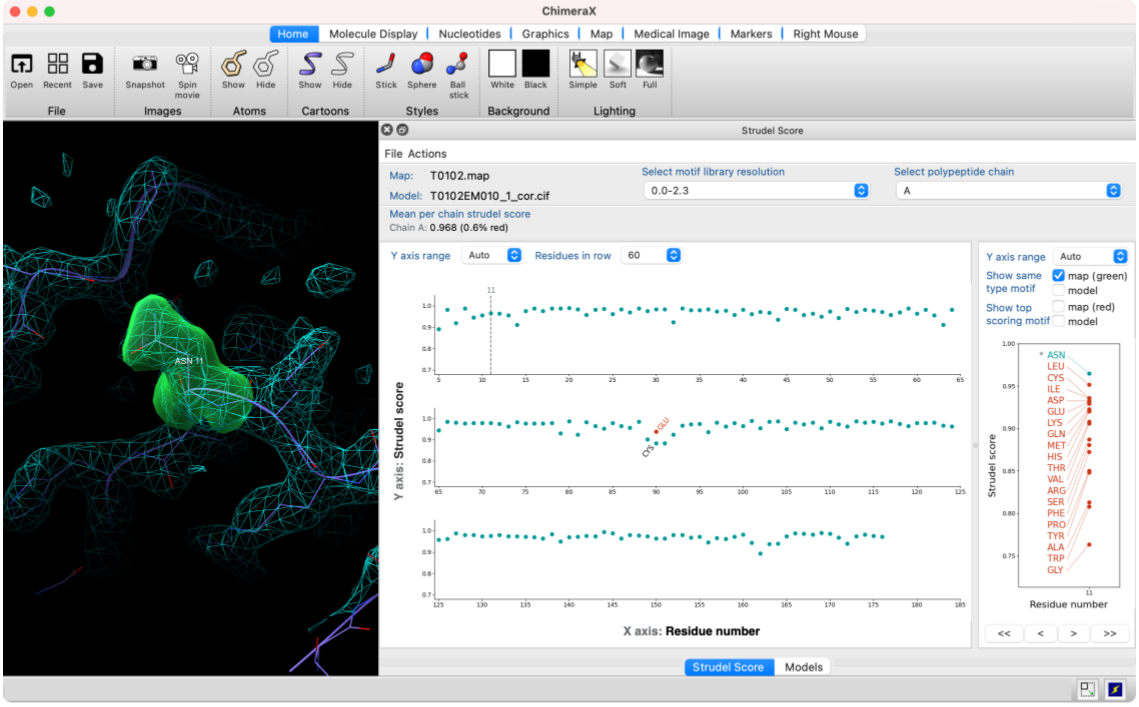


Map: T0102.map
Model: T0102EM054_1_err.cif
Select motif library resolution: 0.0-2.3
Select polypeptide chain: A
Mean per chain strudel score
Chain A: 0.936 (13.4% red)
Y axis range: Auto
Residues in row: 60
Y axis range: Auto
Show same type motif: □ map (green) □ model
Show top scoring motif: □ map (red) □ model

Next run the validation for model T0102EM010_1 which was built into the same map as the one you just examined, but which is of considerably better quality.

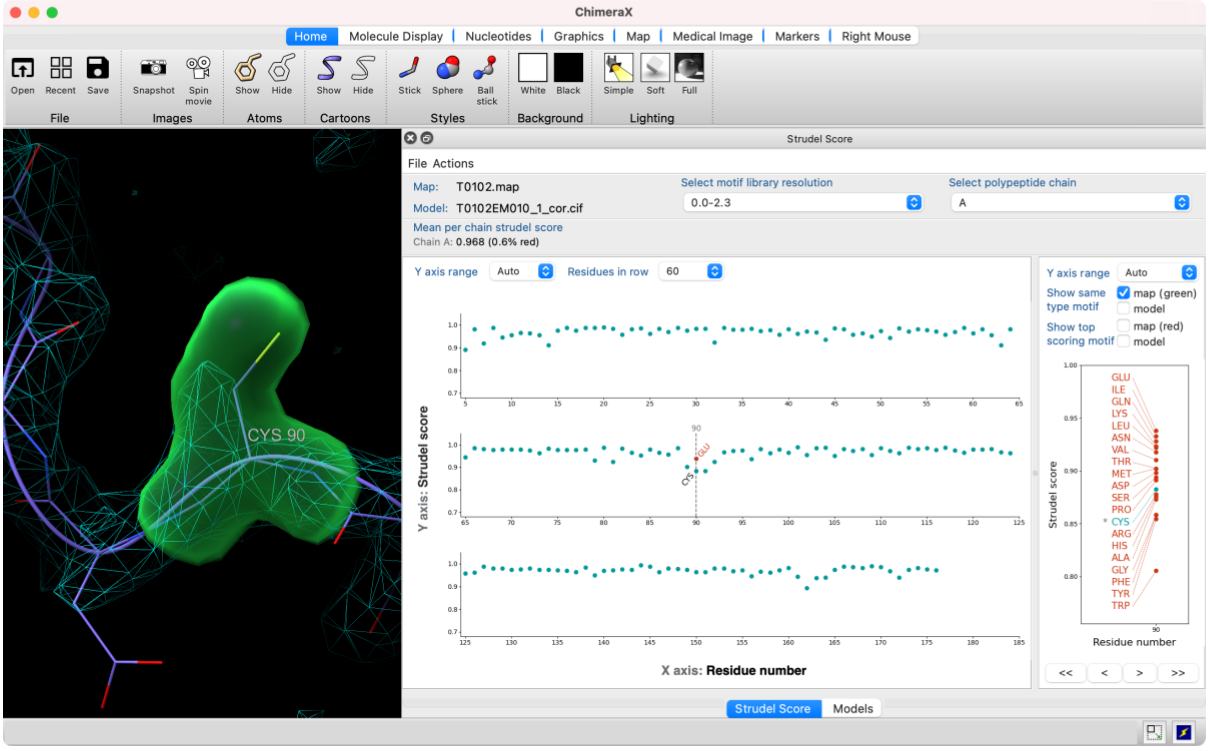
In strudel folder run:
strudel_mapMotifValidation.py -np 4 -p data/T0102EM010_1_corr.cif -m data/T0102.map -l strudel-libs_ver-2.0_voxel-0.5/motifs_0.0-2.3 *-o T0102EM010_out*

To inspect the results, type the following in the ChimeraX command line*:*
strudel open ~/strudel/T0102EM010_out



Map: T0102.map
Model: T0102EM010_1_cor.cif
Select motif library resolution: 0.0-2.3
Select polypeptide chain: A
Mean per chain strudel score
Chain A: 0.968 (0.6% red)
Y axis range: Auto
Residues in row: 60
Y axis range: Auto
Show same type motif: ☑ map (green) □ model
Show top scoring motif: □ map (red) □ model

The Strudel Score plot reveals that this model scores very well. There is only one outlier due to a poorly defined map for residue Cys A80. Check yourself if your hypothesis regarding a one-residue register error is supported by this model.



Map: T0102.map

Model: T0102EM010_1_cor.cif

Select motif library resolution: 0.0-2.3

Select polypeptide chain: A

Mean per chain strudel score

Chain A: 0.968 (0.6% red)

Y axis range: Auto

Residues in row: 60

Y axis range: Auto

Show same type motif: ☑ map (green) □ model

Show top scoring motif: □ map (red) □ model