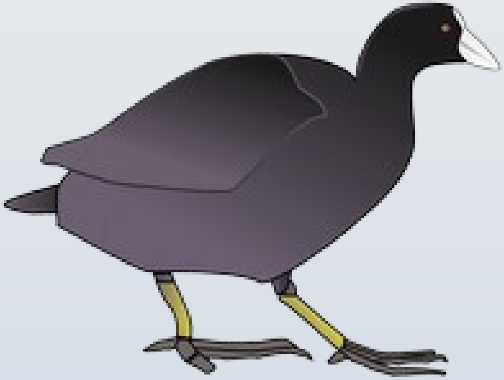


Coot Tools for Cryo-EM

Paul Emsley
MRC Laboratory of Molecular Biology

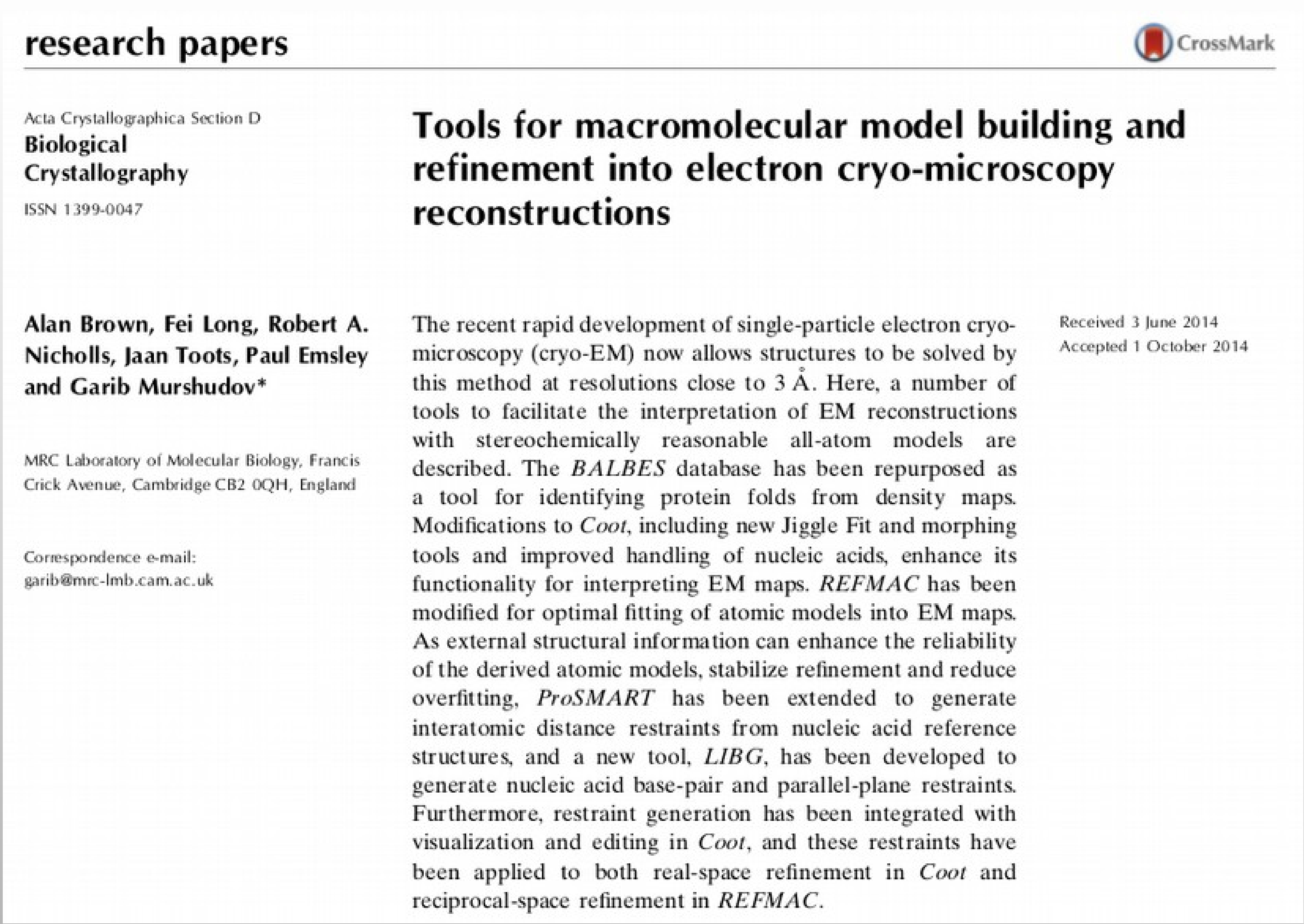
August 2018

Coot & Cryo-EM Model-building

- Coot was originally designed to build and edit protein molecule using x-ray data
  - interactive
  - refinement and validation built in
  - and became more sophisticated over the years
- Typically, we need to move more atoms that one does for crystallography
- the maps are lower resolution and the starting model is further from where you want them to be – usually systematically so
- addressing these needs has been the focus of my work extending/changing *Coot* for cryo-EM

Aside: A Note on Versions

- *Coot* is developed using Continuous Integration
  - so releases don't mean very much
- Latest Release: 0.8.9.1
- Latest Pre-release 0.8.9.2
- *Coot* Web Site
  - Download Binaries ⟶ pre-release ⟶ experimental



Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions

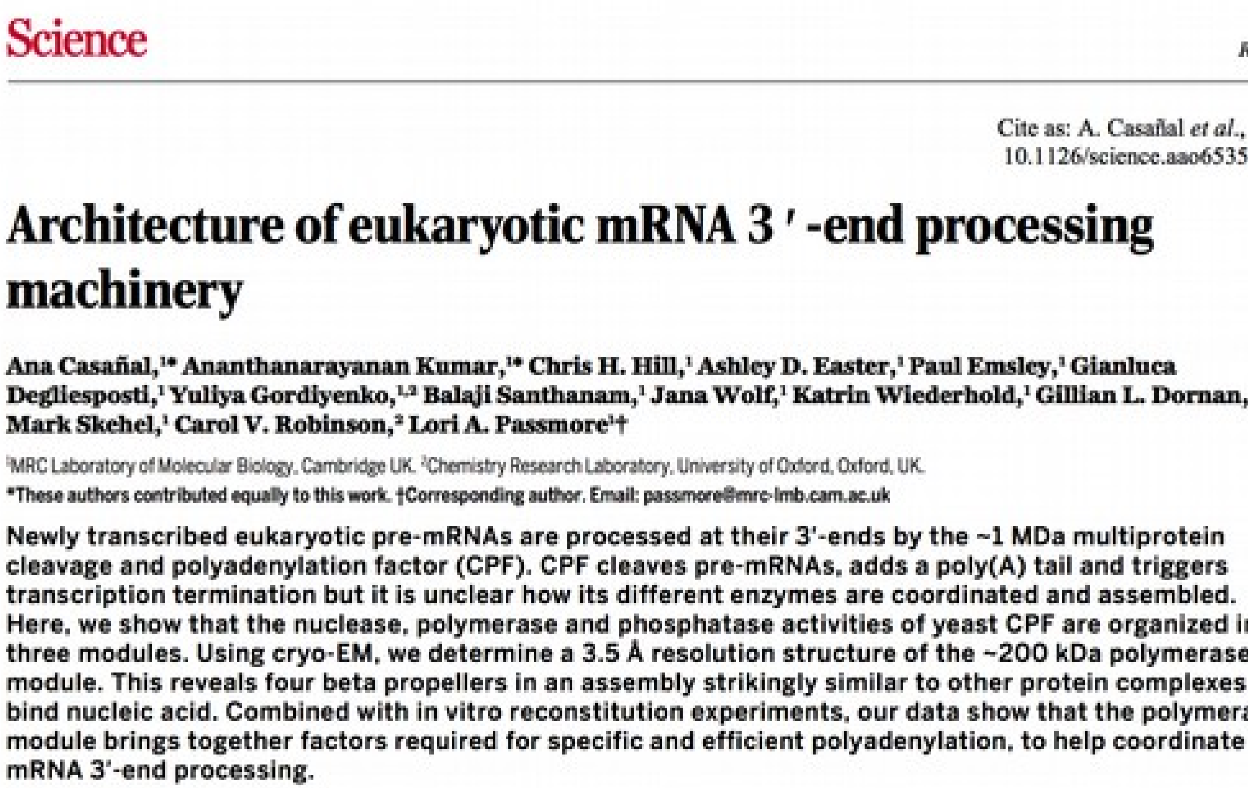
Alan Brown, Fei Long, Robert A. Nicholls, Jaan Toots, Paul Emsley and Garib Murshudov*

MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, England

Correspondence e-mail:
garib@mrc-lmb.cam.ac.uk

Received 3 June 2014
Accepted 1 October 2014

The recent rapid development of single-particle electron cryo-microscopy (cryo-EM) now allows structures to be solved by this method at resolutions close to 3 Å. Here, a number of tools to facilitate the interpretation of EM reconstructions with stereochemically reasonable all-atom models are described. The *BALBES* database has been repurposed as a tool for identifying protein folds from density maps. Modifications to *Coot*, including new Jiggle Fit and morphing tools and improved handling of nucleic acids, enhance its functionality for interpreting EM maps. *REFMAC* has been modified for optimal fitting of atomic models into EM maps. As external structural information can enhance the reliability of the derived atomic models, stabilize refinement and reduce overfitting, *ProSMART* has been extended to generate interatomic distance restraints from nucleic acid reference structures, and a new tool, *LIBG*, has been developed to generate nucleic acid base-pair and parallel-plane restraints. Furthermore, restraint generation has been integrated with visualization and editing in *Coot*, and these restraints have been applied to both real-space refinement in *Coot* and reciprocal-space refinement in *REFMAC*.



Science

REPORTS

Cite as: A. Casañal *et al*., *Science* 10.1126/science.aao6535 (2017).

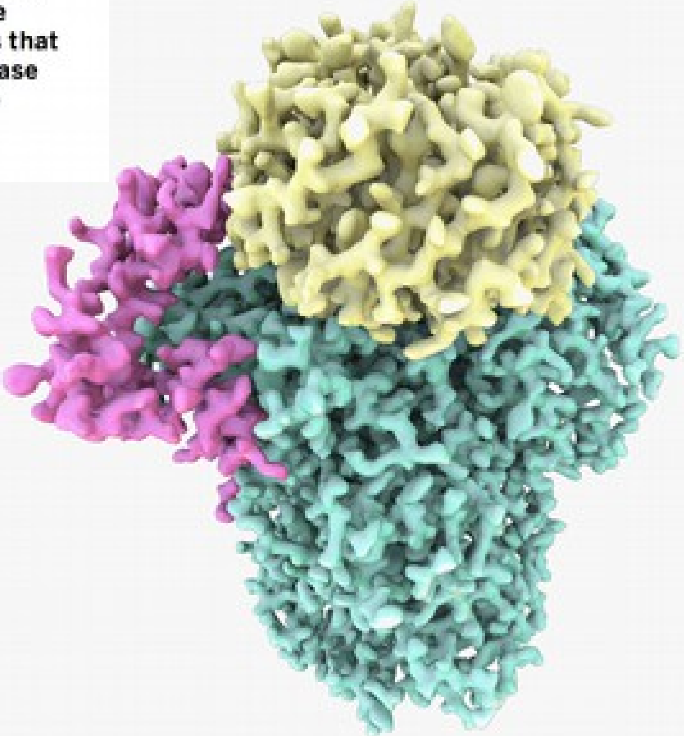
Architecture of eukaryotic mRNA 3′-end processing machinery

Ana Casañal,[1]* Ananthanarayanan Kumar,[1]* Chris H. Hill,[1] Ashley D. Easter,[1] Paul Emsley,[1] Gianluca Degliesposti,[1] Yuliya Gordiyenko,[1,2] Balaji Santhanam,[1] Jana Wolf,[1] Katrin Wiederhold,[1] Gillian L. Dornan,[1] Mark Skehel,[1] Carol V. Robinson,[2] Lori A. Passmore[1]†

[1]MRC Laboratory of Molecular Biology, Cambridge UK. [2]Chemistry Research Laboratory, University of Oxford, Oxford, UK.

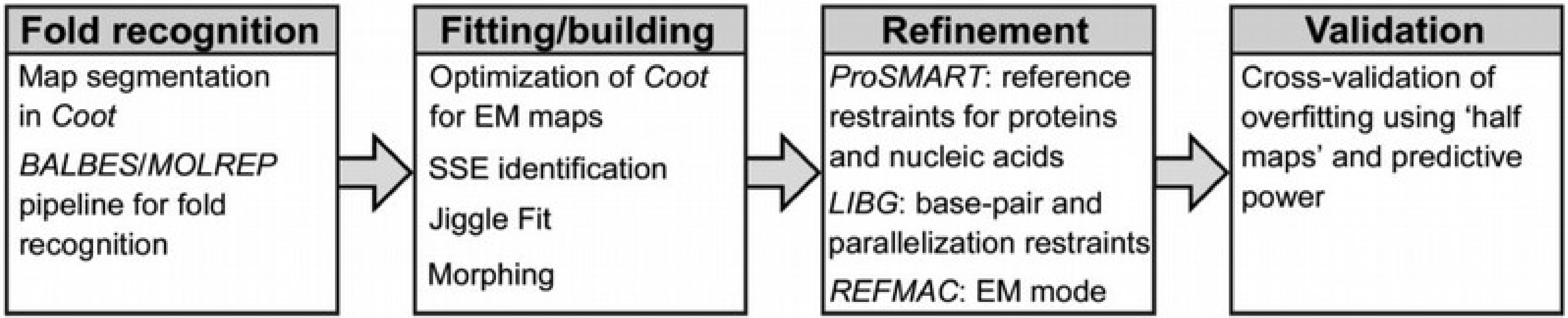
*These authors contributed equally to this work. †Corresponding author. Email: passmore@mrc-lmb.cam.ac.uk

Newly transcribed eukaryotic pre-mRNAs are processed at their 3′-ends by the ~1 MDa multiprotein cleavage and polyadenylation factor (CPF). CPF cleaves pre-mRNAs, adds a poly(A) tail and triggers transcription termination but it is unclear how its different enzymes are coordinated and assembled. Here, we show that the nuclease, polymerase and phosphatase activities of yeast CPF are organized into three modules. Using cryo-EM, we determine a 3.5 Å resolution structure of the ~200 kDa polymerase module. This reveals four beta propellers in an assembly strikingly similar to other protein complexes that bind nucleic acid. Combined with in vitro reconstitution experiments, our data show that the polymerase module brings together factors required for specific and efficient polyadenylation, to help coordinate mRNA 3′-end processing.



Model-Building Tools

Recent Developments



Fold recognition

Map segmentation in *Coot*

BALBES/*MOLREP* pipeline for fold recognition

Fitting/building

Optimization of *Coot* for EM maps

SSE identification

Jiggle Fit

Morphing

Refinement

ProSMART: reference restraints for proteins and nucleic acids

LIBG: base-pair and parallelization restraints

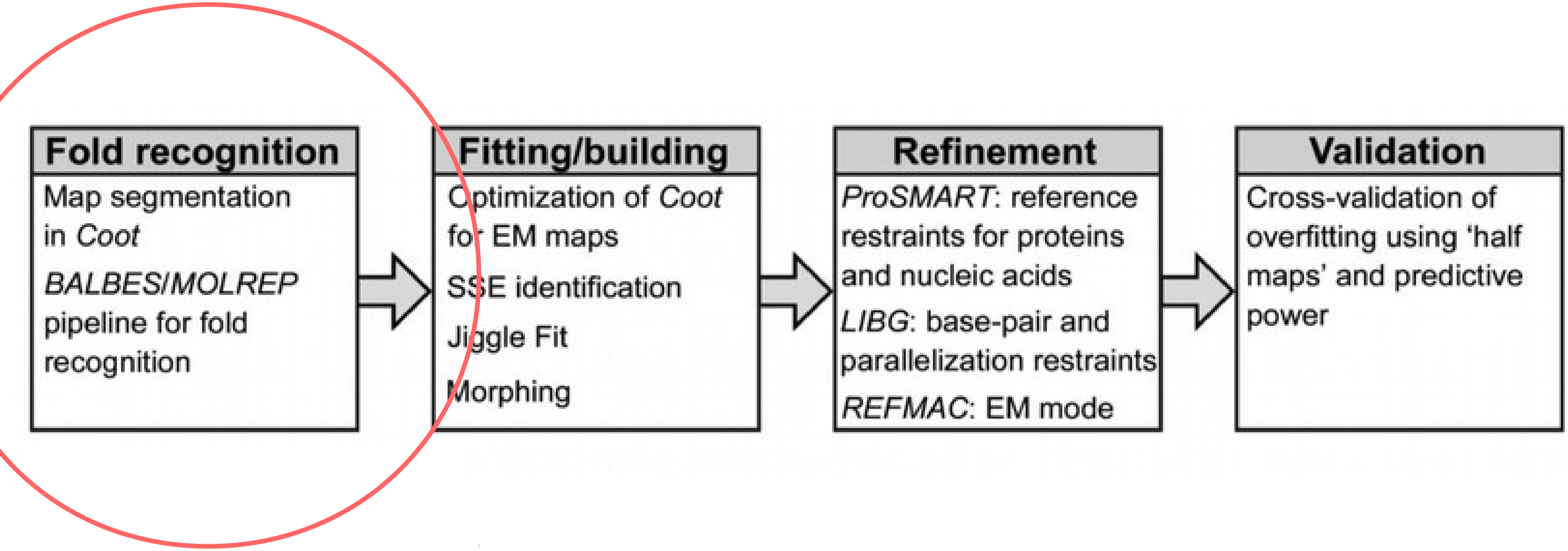
REFMAC: EM mode

Validation

Cross-validation of overfitting using 'half maps' and predictive power

Model-Building Tools

Recent Developments



Fold recognition

Map segmentation in *Coot*

BALBES/MOLREP pipeline for fold recognition

Fitting/building

Optimization of *Coot* for EM maps

SSE identification

Jiggle Fit

Morphing

Refinement

ProSMART: reference restraints for proteins and nucleic acids

LIBG: base-pair and parallelization restraints

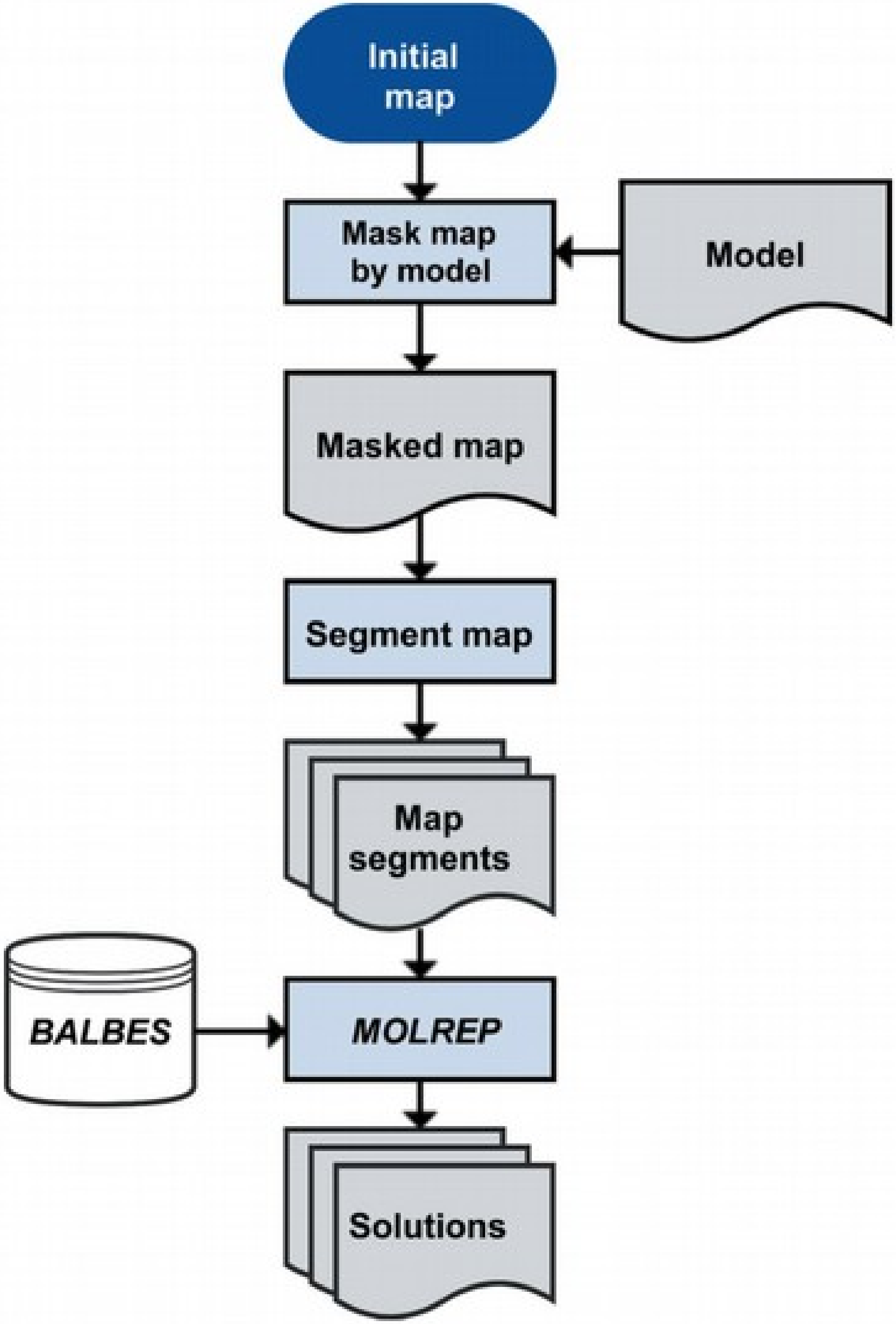
REFMAC: EM mode

Validation

Cross-validation of overfitting using 'half maps' and predictive power

Cryo-EM data

- “Sequence by Map”
- Alternatively, use fold recognition
- Using the BALBES database
  - (originally design for molecular replacement)
  - screen domains against unknown density


Initial map
Mask map by model
Model
Masked map
Segment map
Map segments
BALBES
MOLREP
Solutions

Fold Recognition

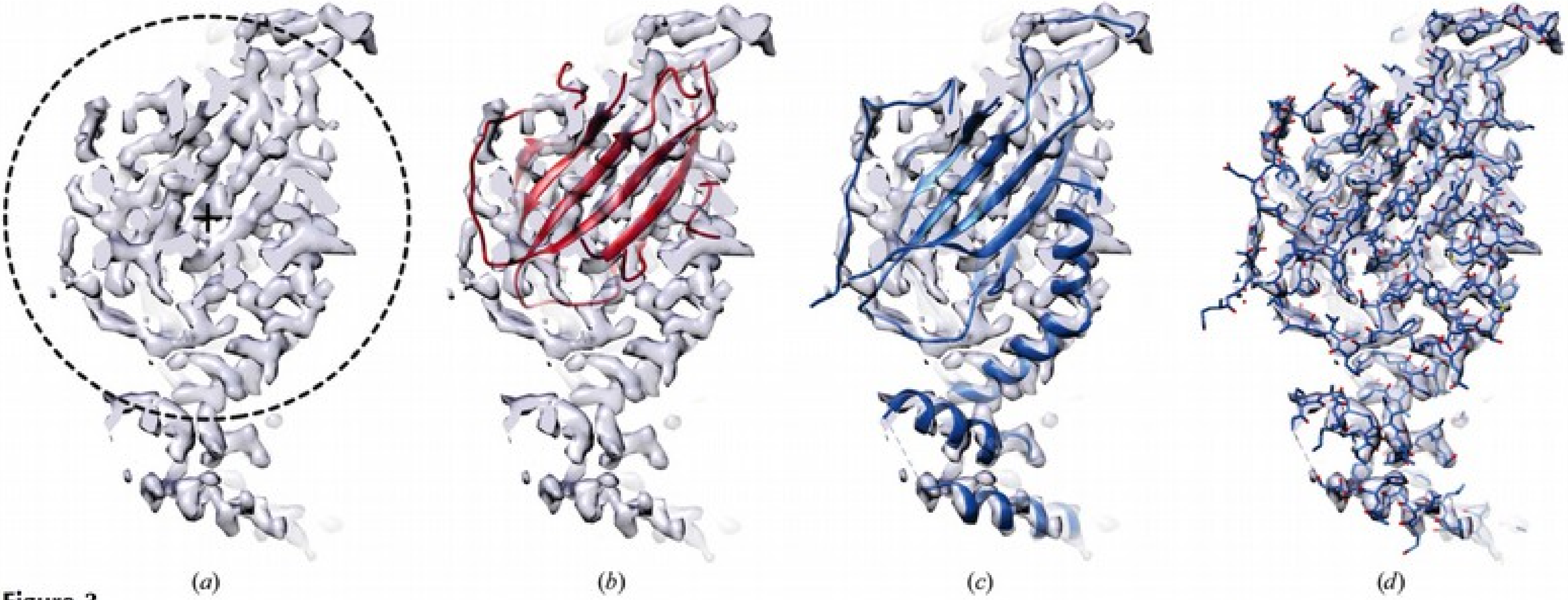
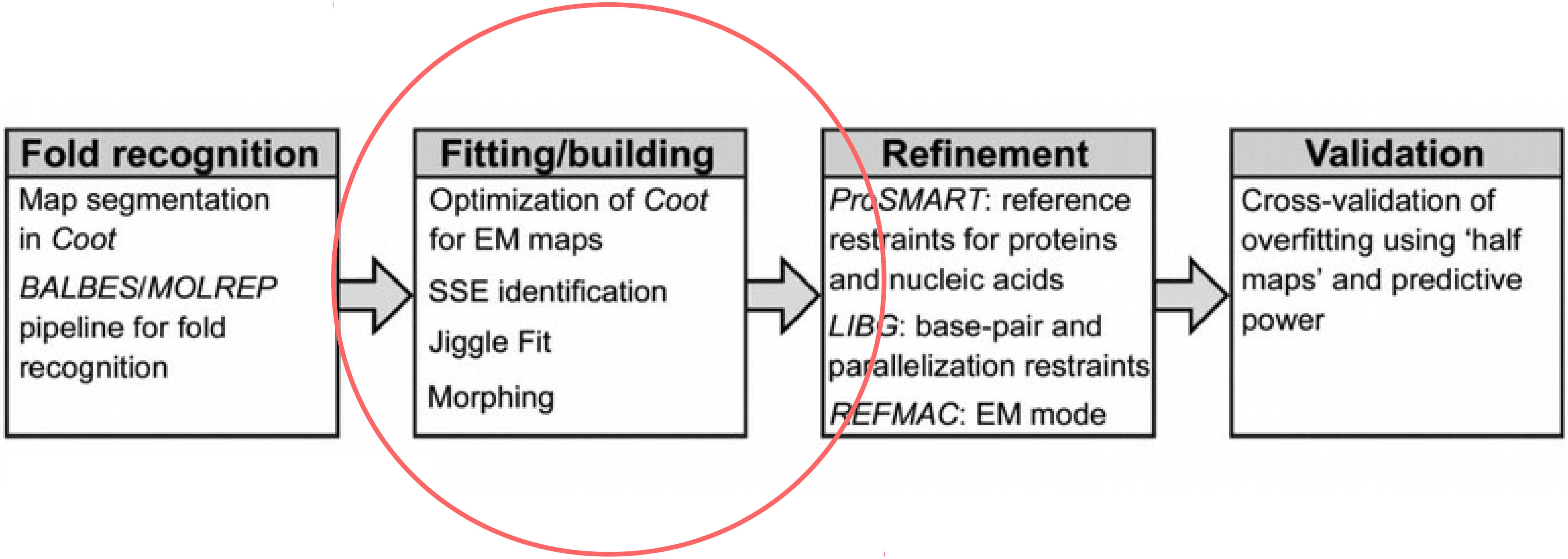

(*a*) (*b*) (*c*) (*d*)

Figure 3

Fold recognition can identify template molecules for model building. (*a*) Density map corresponding to the final model of the mitoribosomal protein mL38 with the segmented search map indicated. (*b*) Top solution from the *BALBES–MOLREP* pipeline. (*c*, *d*) Final refined model of mL38 in (*c*) cartoon and (*d*) full-atom representation.

Model-Building Tools

Recent Developments



Fold recognition

Map segmentation in *Coot*

BALBES/*MOLREP* pipeline for fold recognition

Fitting/building

Optimization of *Coot* for EM maps

SSE identification

Jiggle Fit

Morphing

Refinement

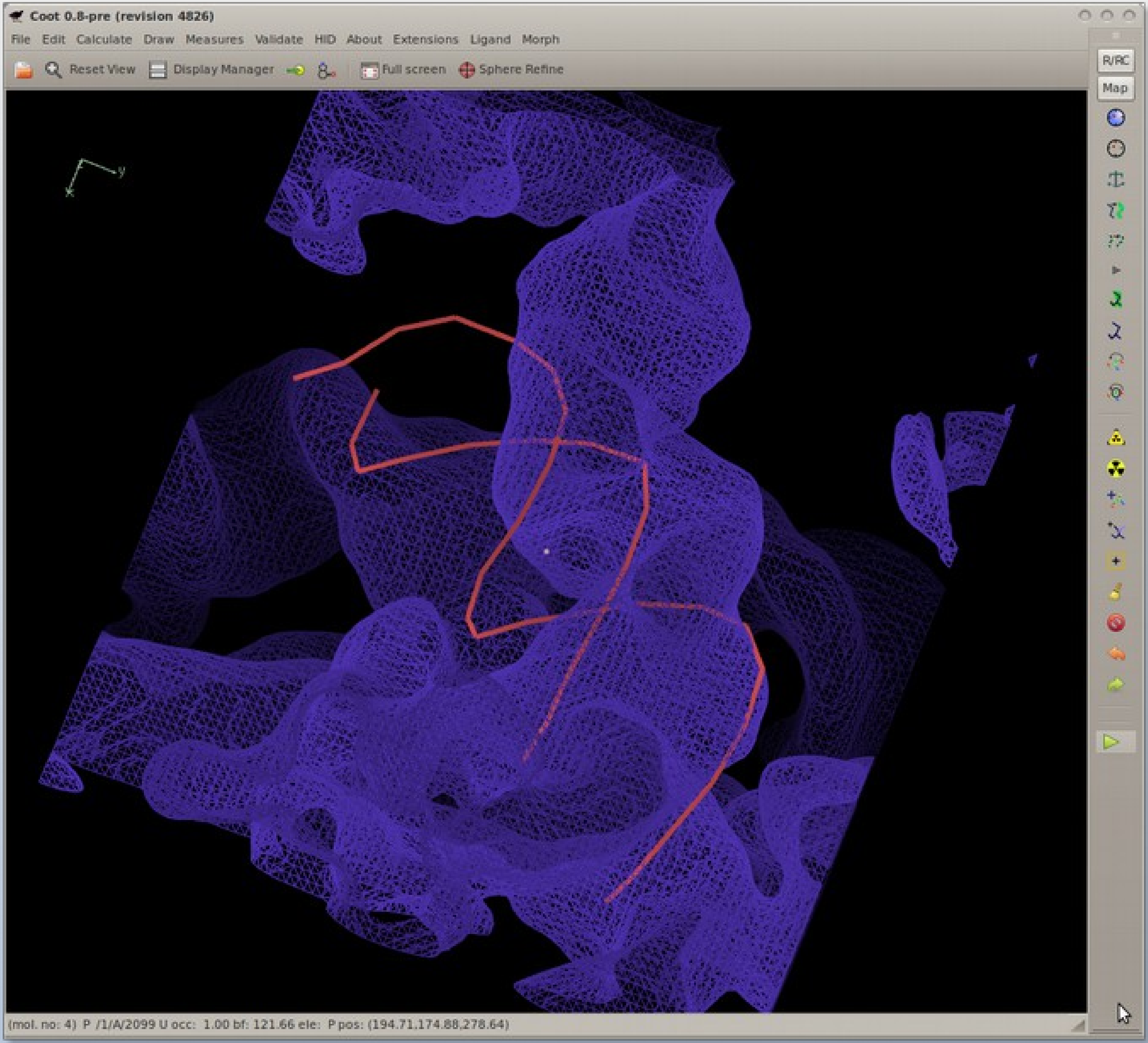
ProSMART: reference restraints for proteins and nucleic acids

LIBG: base-pair and parallelization restraints

REFMAC: EM mode

Validation

Cross-validation of overfitting using ‘half maps’ and predictive power



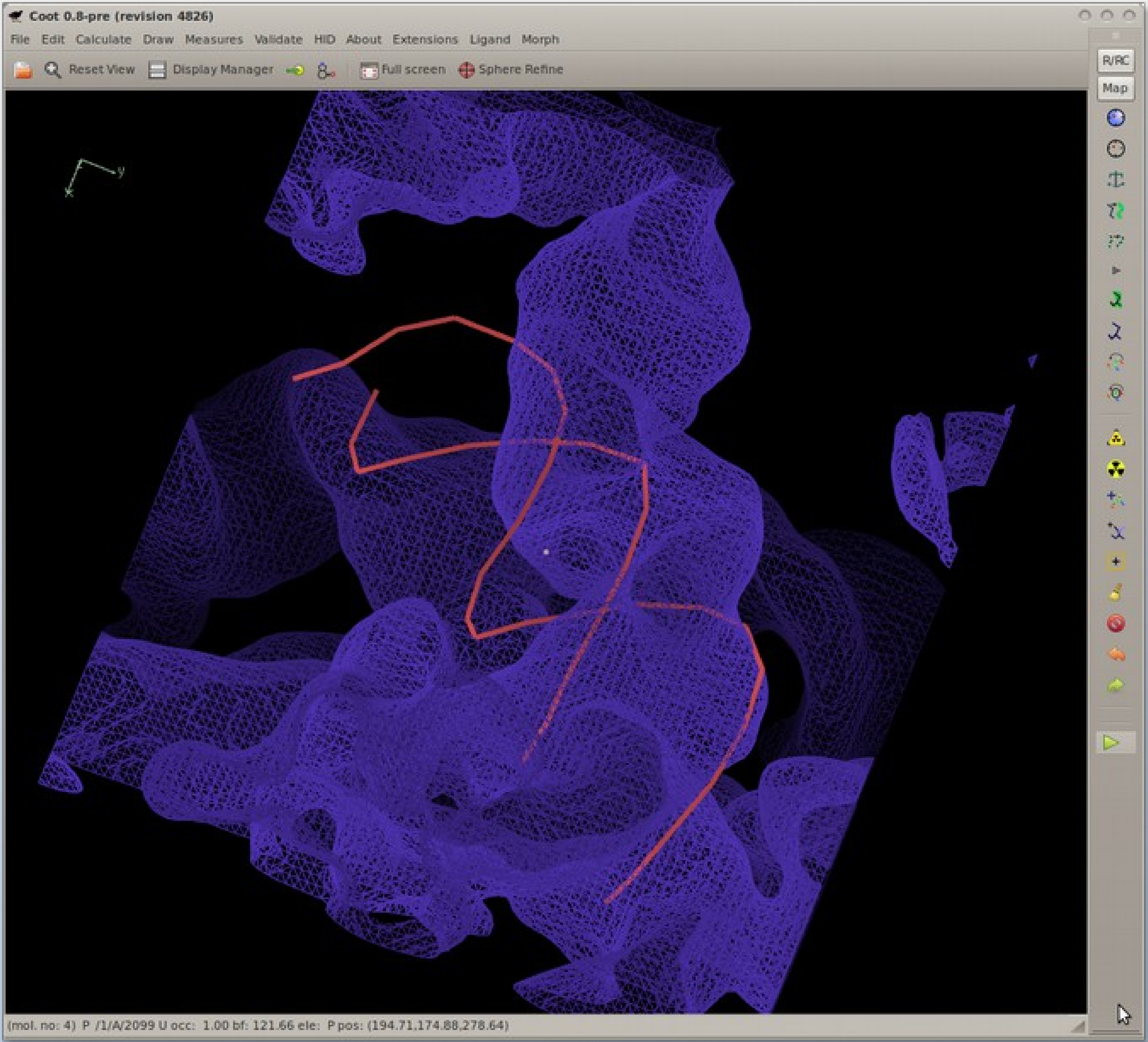
```
(mol. no: 4)  P  /1/A/2099 U occ:  1.00 bf: 121.66 ele:  P pos: (194.71,174.88,278.64)
```

Jiggle Fit

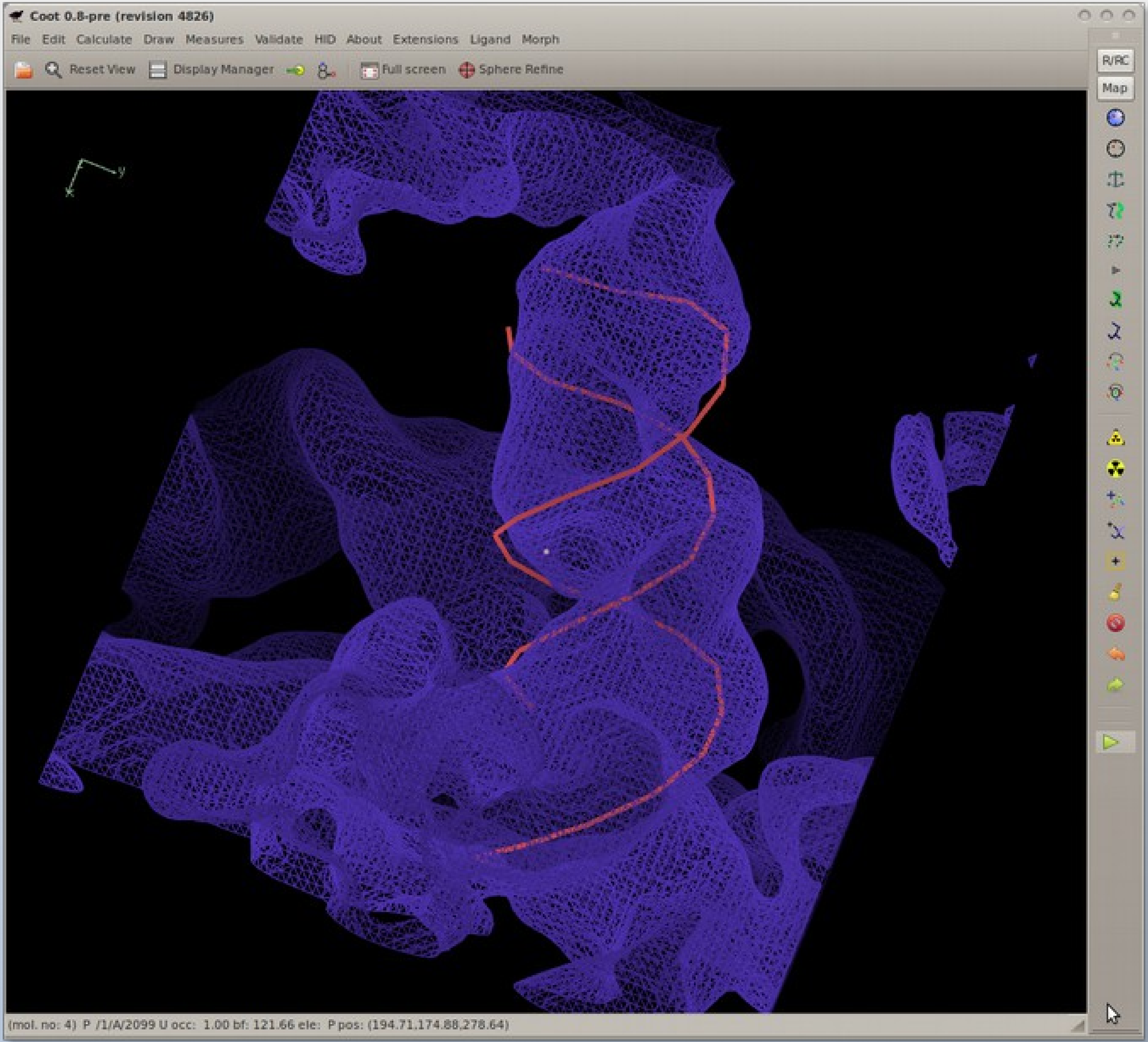
- How do I rotate and translate these atoms to fit the density?
  - 6-dimensional problem
- Originally used to fit simple ligands/solvent molecules to blobs of density
- Now extended to fit arbitrary atom selections
  - *e.g.* by Chain

Jiggle Fit: How it Works

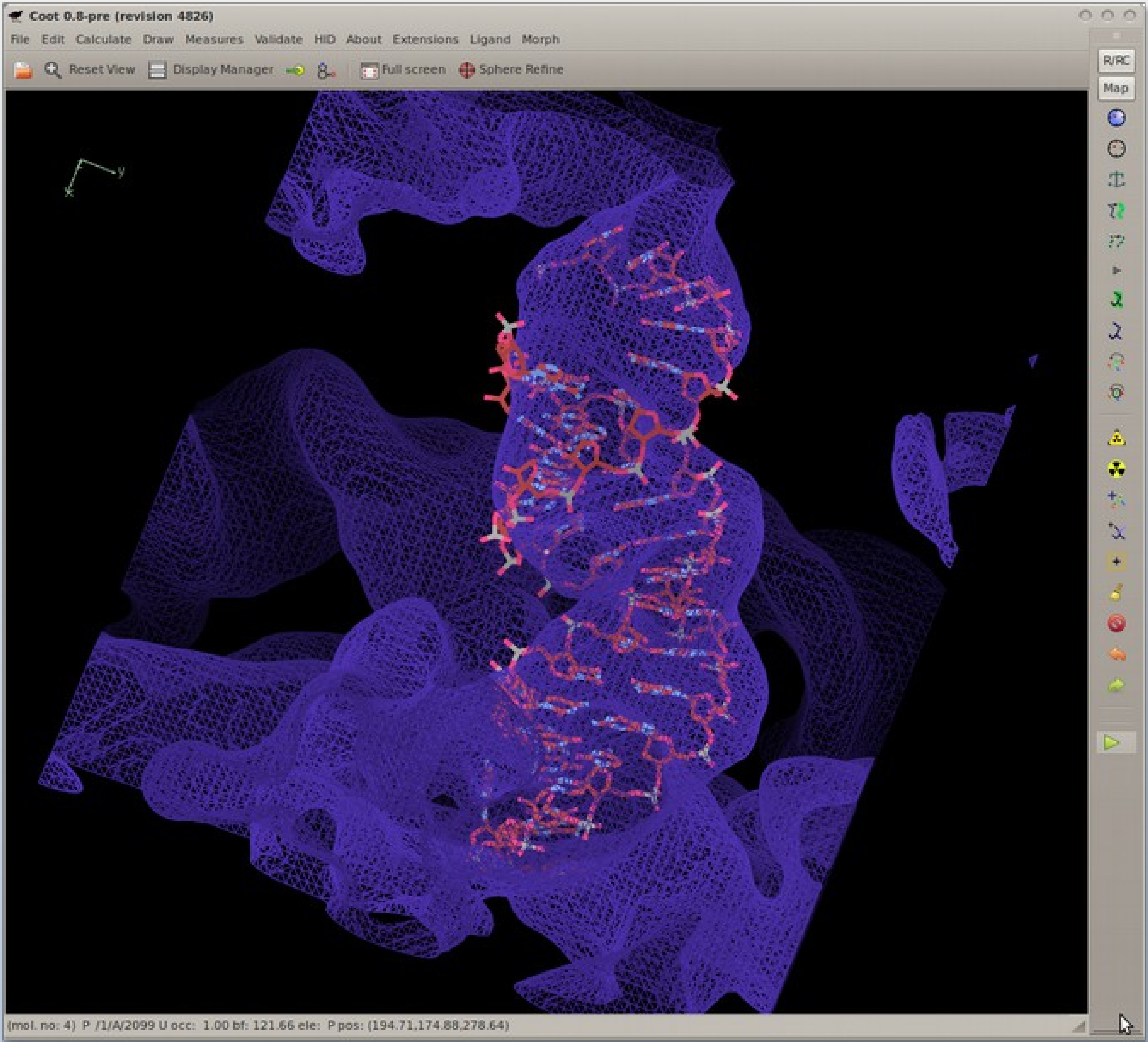
- Loop *n* (say 1000) times:
  - Generate sample angles and translations
  - Transform atom selection by these rotations and translation
  - Score and store the fit to density
- Rank density fit scores,
  - Pick top 10 solution, for each of them
    - Rigid body fit and score solutions
    - Pick the highest scoring solution if it's better than the starting model)
- Radius of Convergence is larger when using a low-pass map



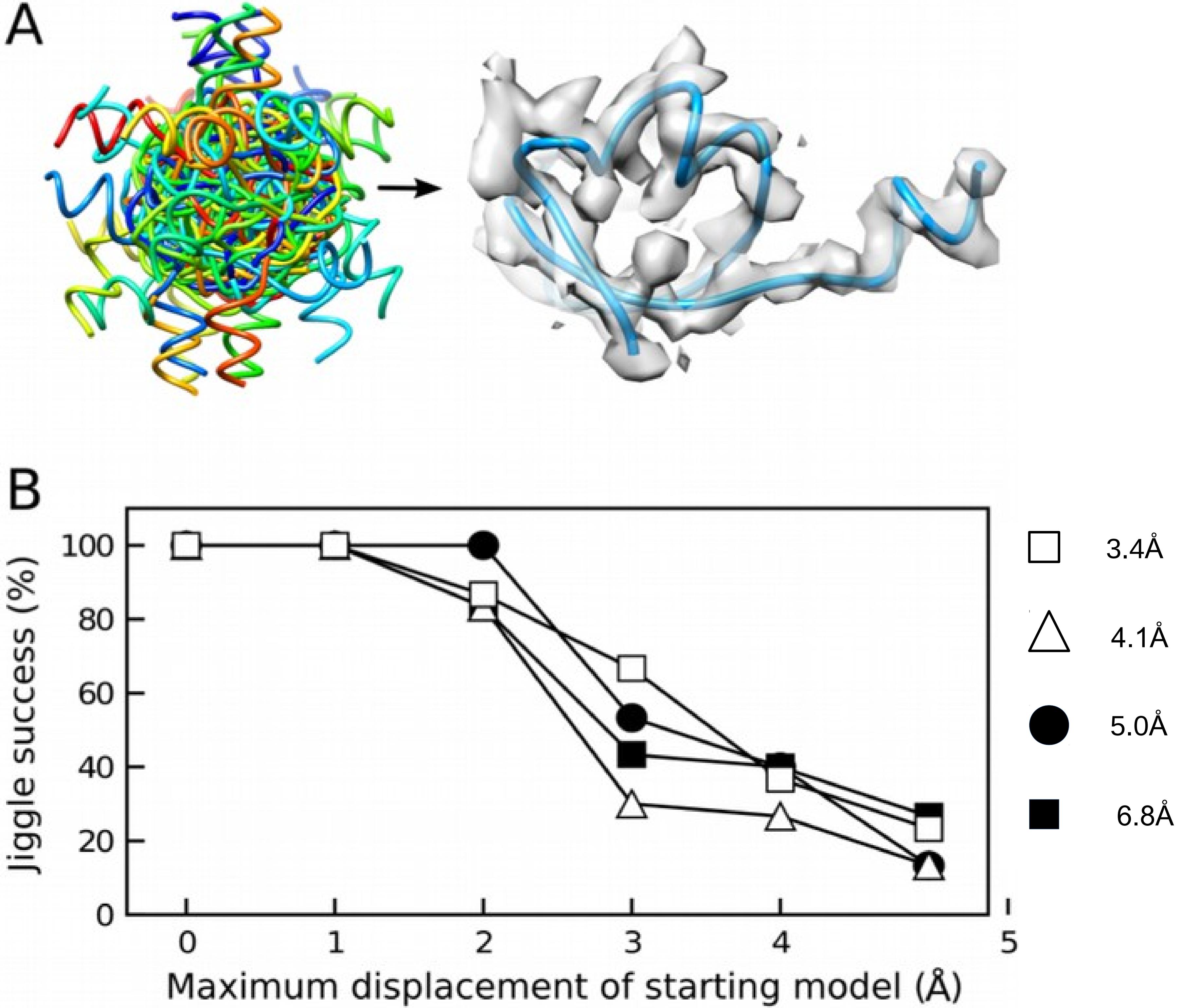
```
(mol. no: 4)  P  /1/A/2099 U occ:  1.00 bf: 121.66 ele:  P pos: (194.71,174.88,278.64)
```



(mol. no: 4) P /1/A/2099 U occ: 1.00 bf: 121.66 ele: P pos: (194.71,174.88,278.64)



```
(mol. no: 4)  P  /1/A/2099 U occ:  1.00 bf: 121.66 ele:  P pos: (194.71,174.88,278.64)
```


A
B
Jiggle success (%)
100
80
60
40
20
0
0
1
2
3
4
5
Maximum displacement of starting model (Å)
□ 3.4Å
△ 4.1Å
● 5.0Å
■ 6.8Å

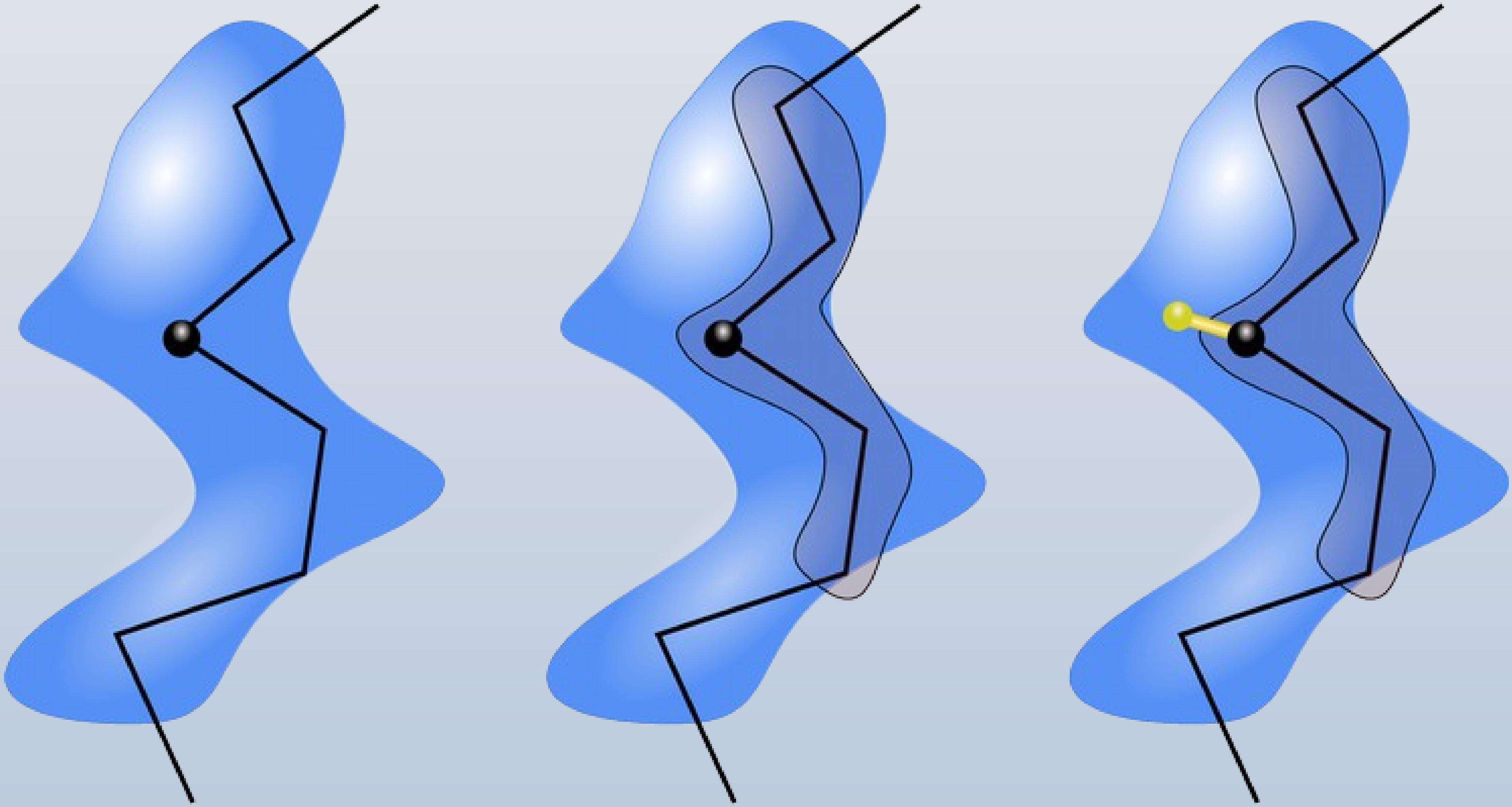
So we have our ideal RNA or homologous protein sitting roughly in the density

(not a great fit)

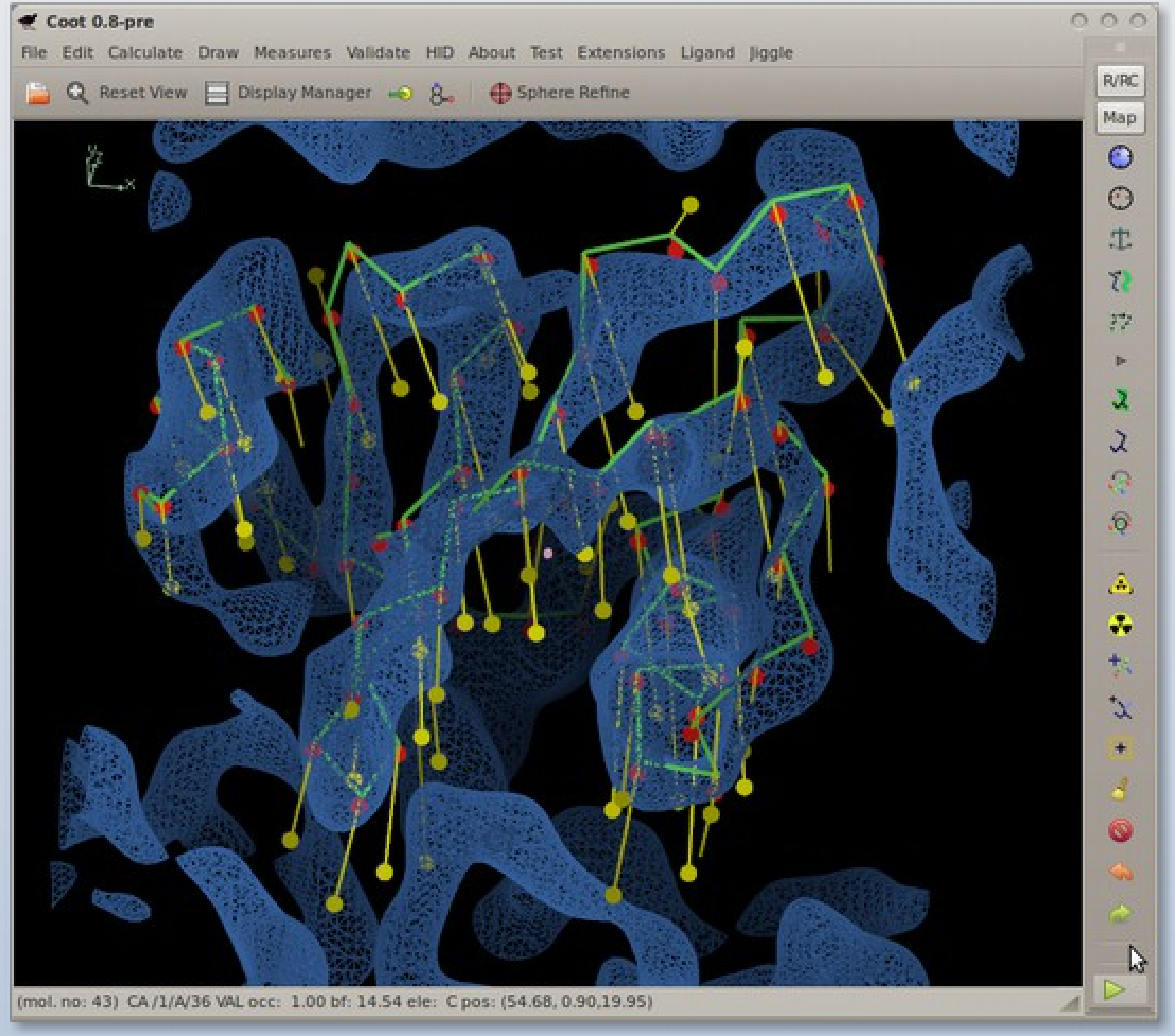
Model Morphing: How it Works

- For each residue in a chain, we ask:
  - where does a small fragment centred on this residue want to go?
  - (Robust) average the transformations and apply them on a per-residue basis
- Repeat

Model Morphing: Generating the Raw RTs



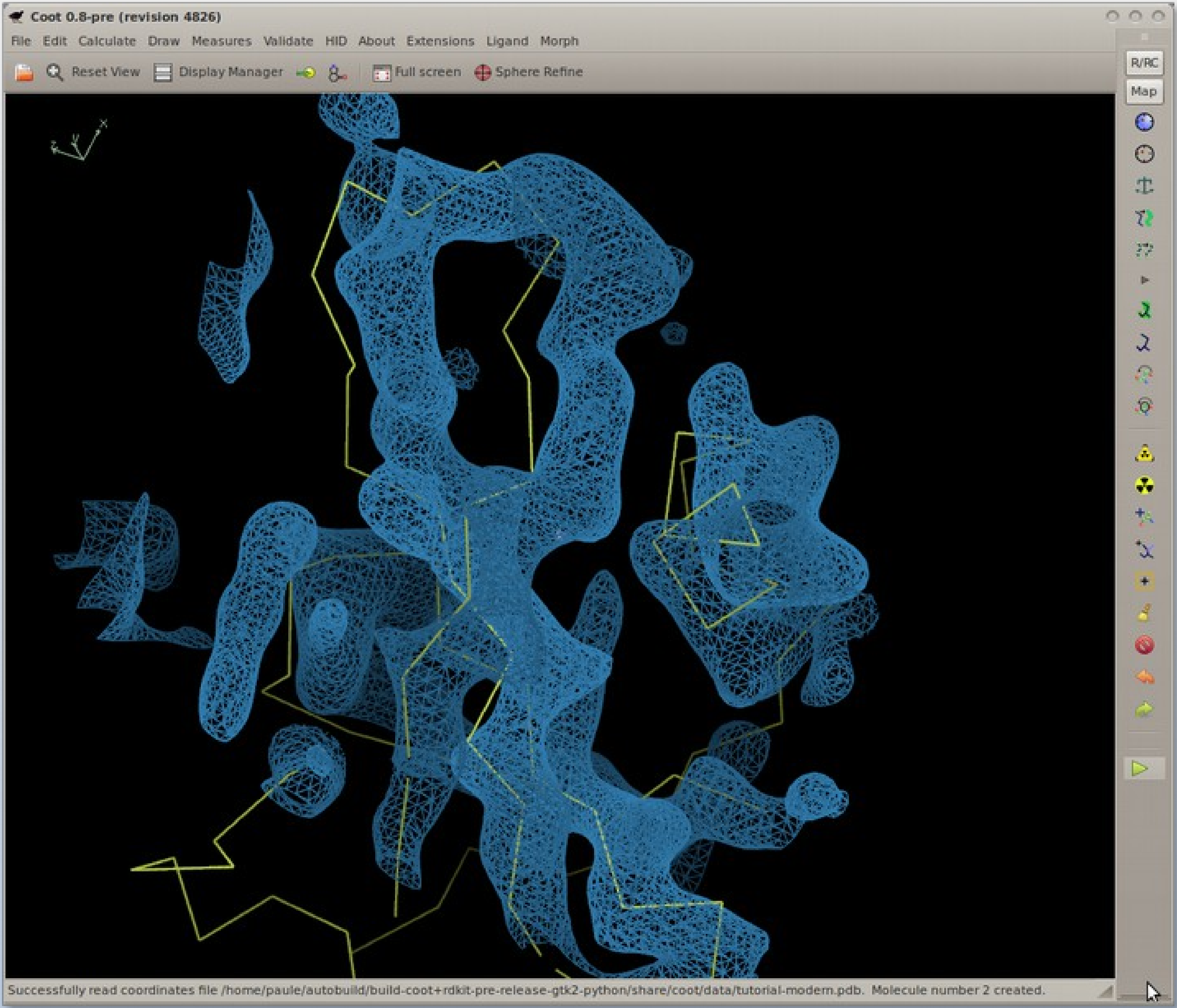
Model Morphing: Example



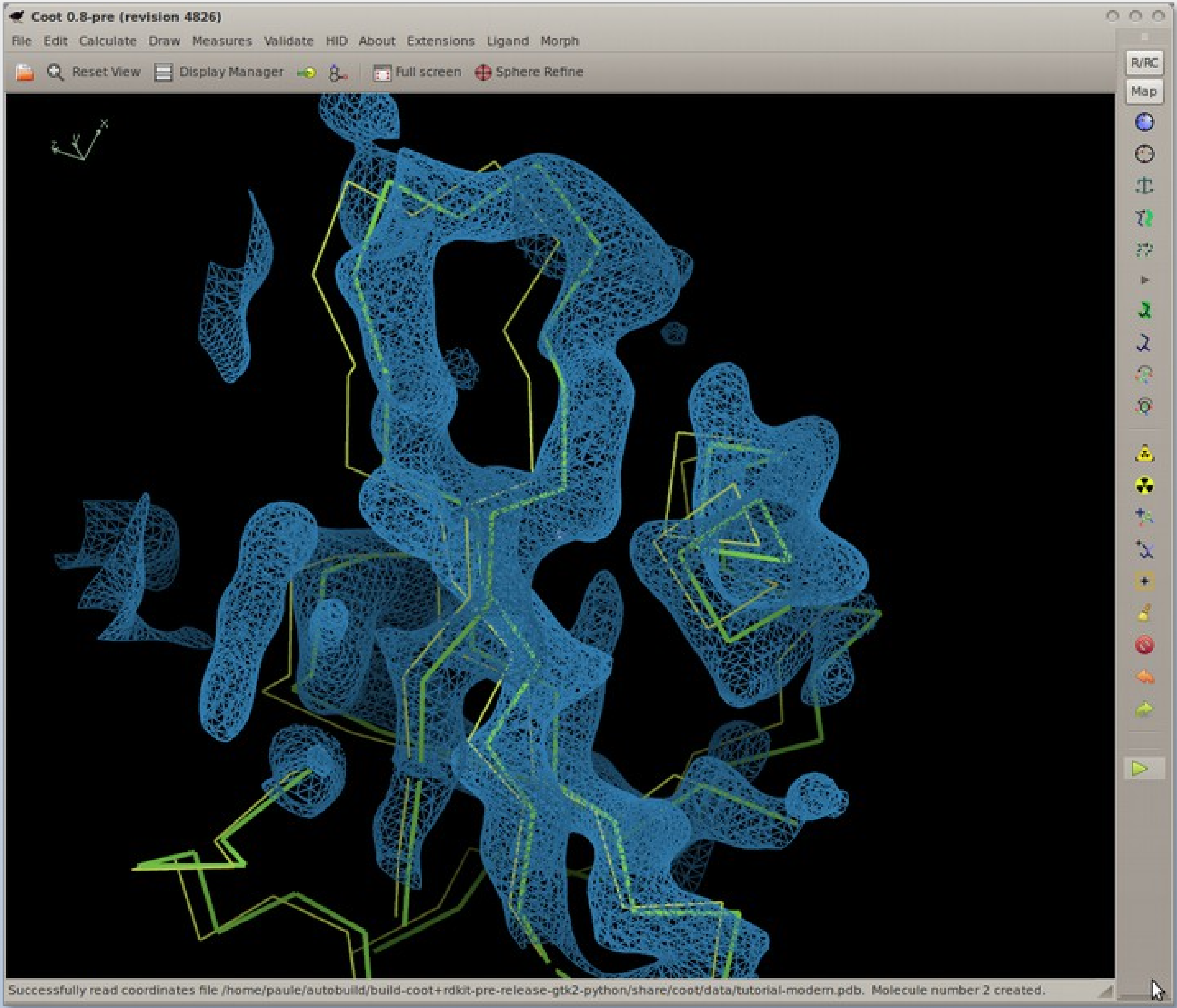
```
(mol. no: 43)  CA /1/A/36 VAL occ:   1.00 bf: 14.54 ele:   C pos: (54.68, 0.90,19.95)
```

Model Morphing: Robust Averaging

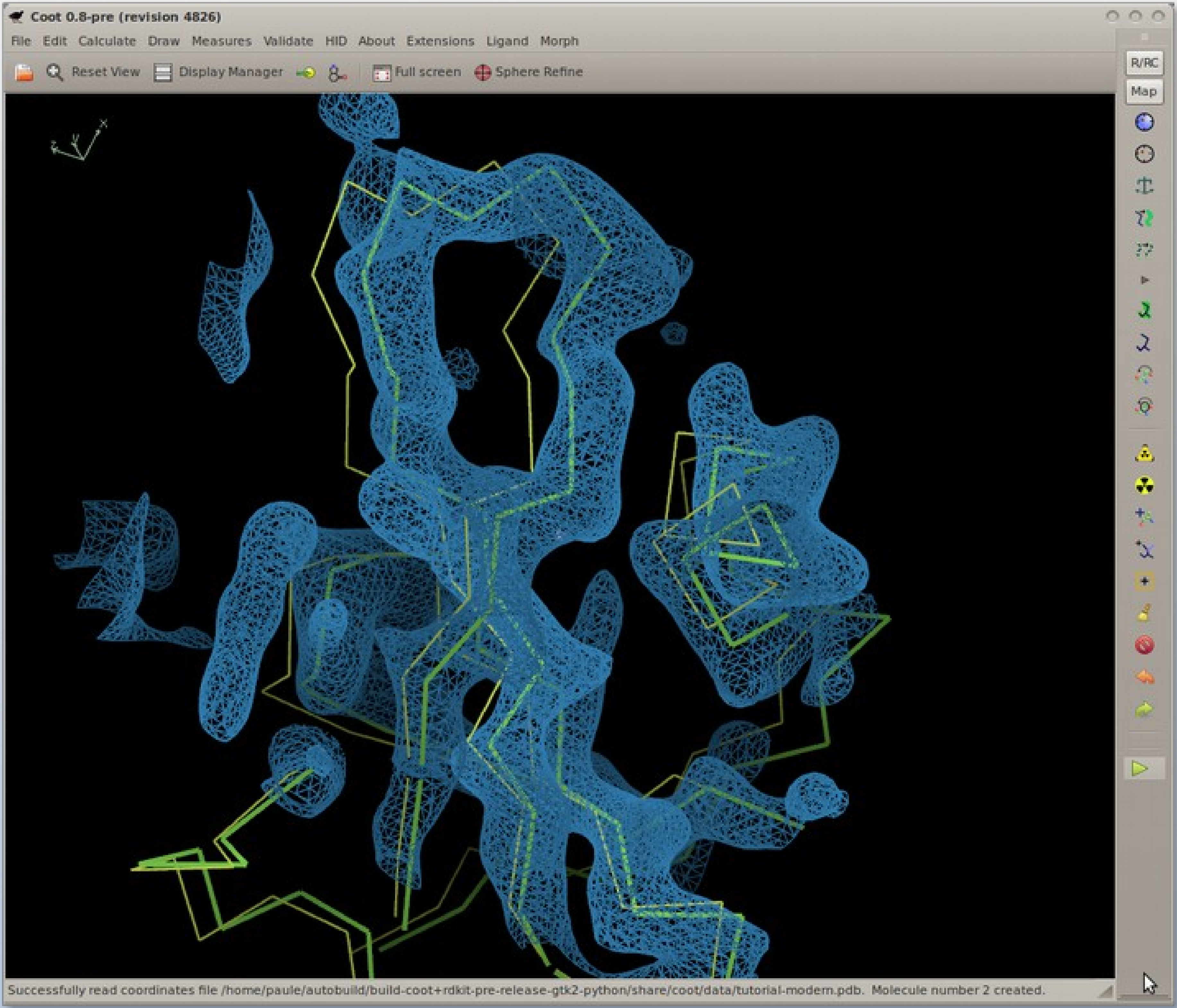
- What are the residues in the environment of a residue?
  - What are their RTs?
  - Create a metric 'distance', sort on that
  - Discard the top and bottom 25%
  - Use remaining RTs to generate average
  - ...which is then applied to central residue
- Repeat for all residues
- Larger environment radii make the shifts smaller/more conservative
  - More cycles needed



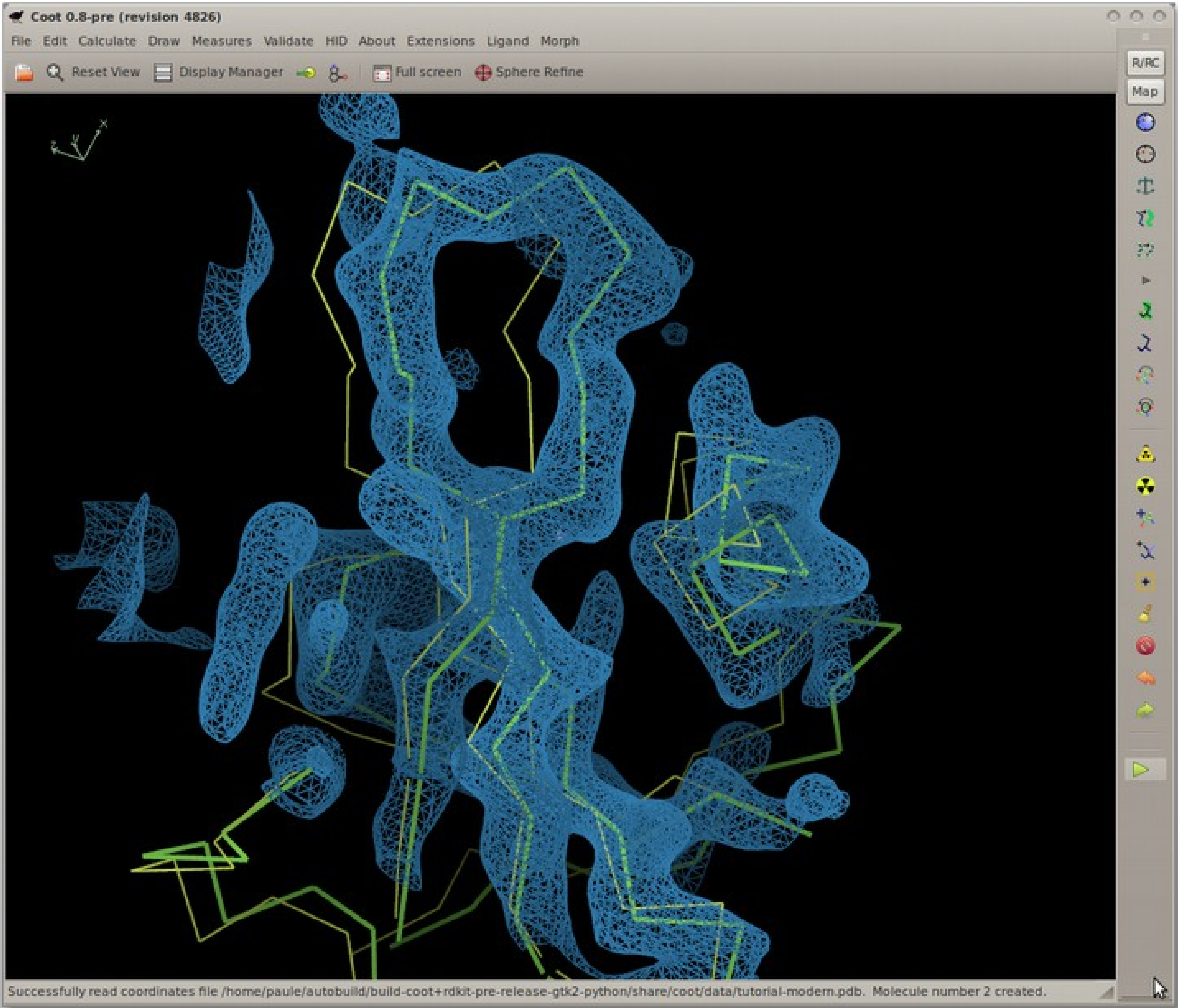
Successfully read coordinates file /home/paule/autobuild/build-coot+rdkit-pre-release-gtk2-python/share/coot/data/tutorial-modern.pdb. Molecule number 2 created.



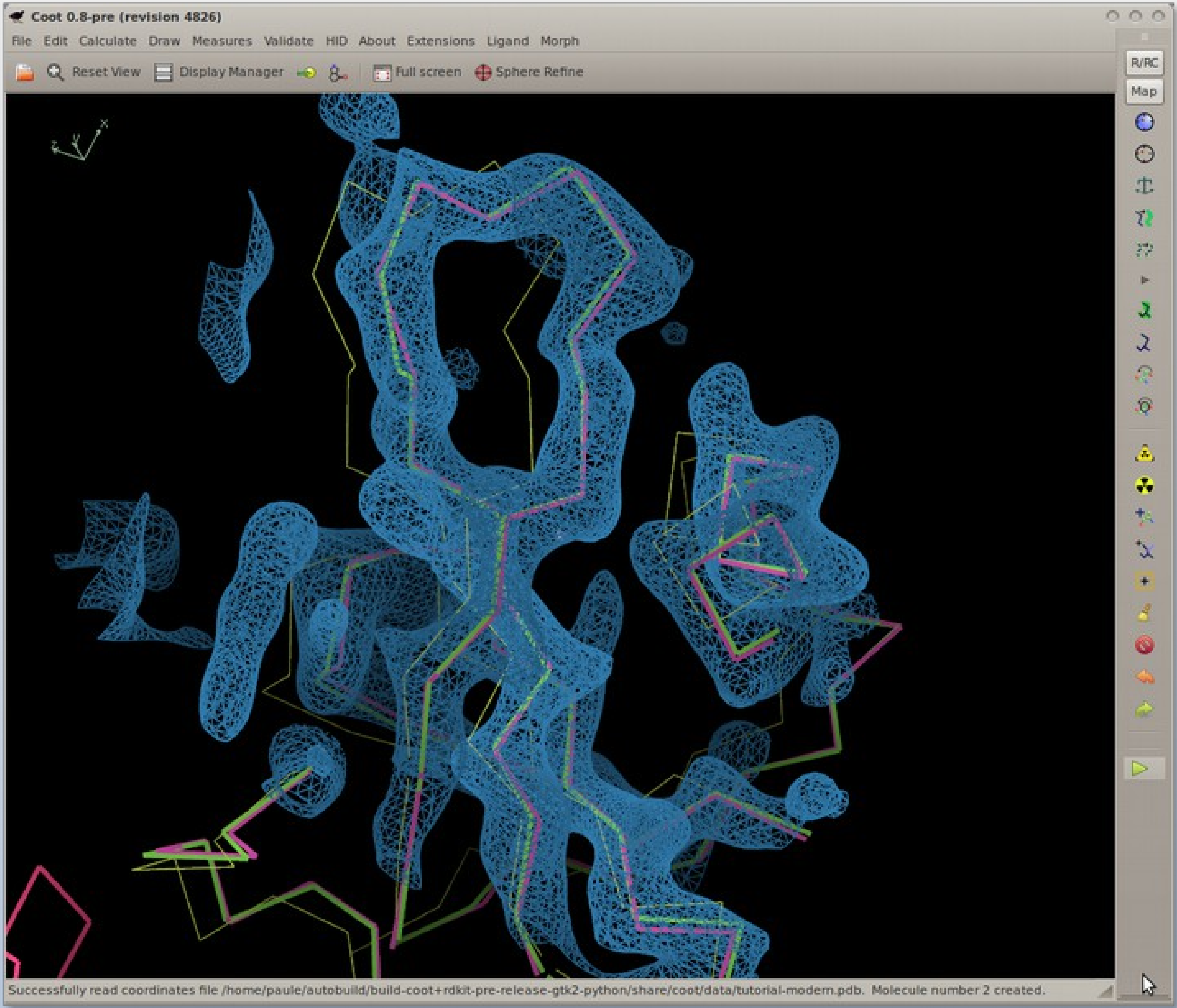
Successfully read coordinates file /home/paule/autobuild/build-coot+rdkit-pre-release-gtk2-python/share/coot/data/tutorial-modern.pdb. Molecule number 2 created.



Successfully read coordinates file /home/paule/autobuild/build-coot+rdkit-pre-release-gtk2-python/share/coot/data/tutorial-modern.pdb. Molecule number 2 created.

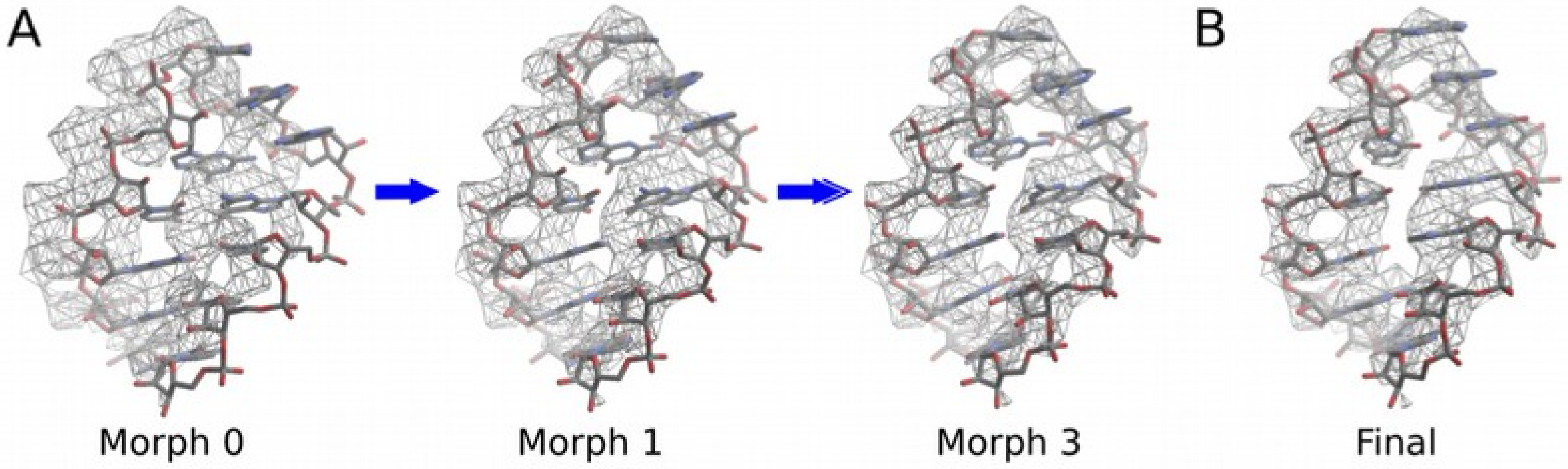


Successfully read coordinates file /home/paule/autobuild/build-coot+rdkit-pre-release-gtk2-python/share/coot/data/tutorial-modern.pdb. Molecule number 2 created.

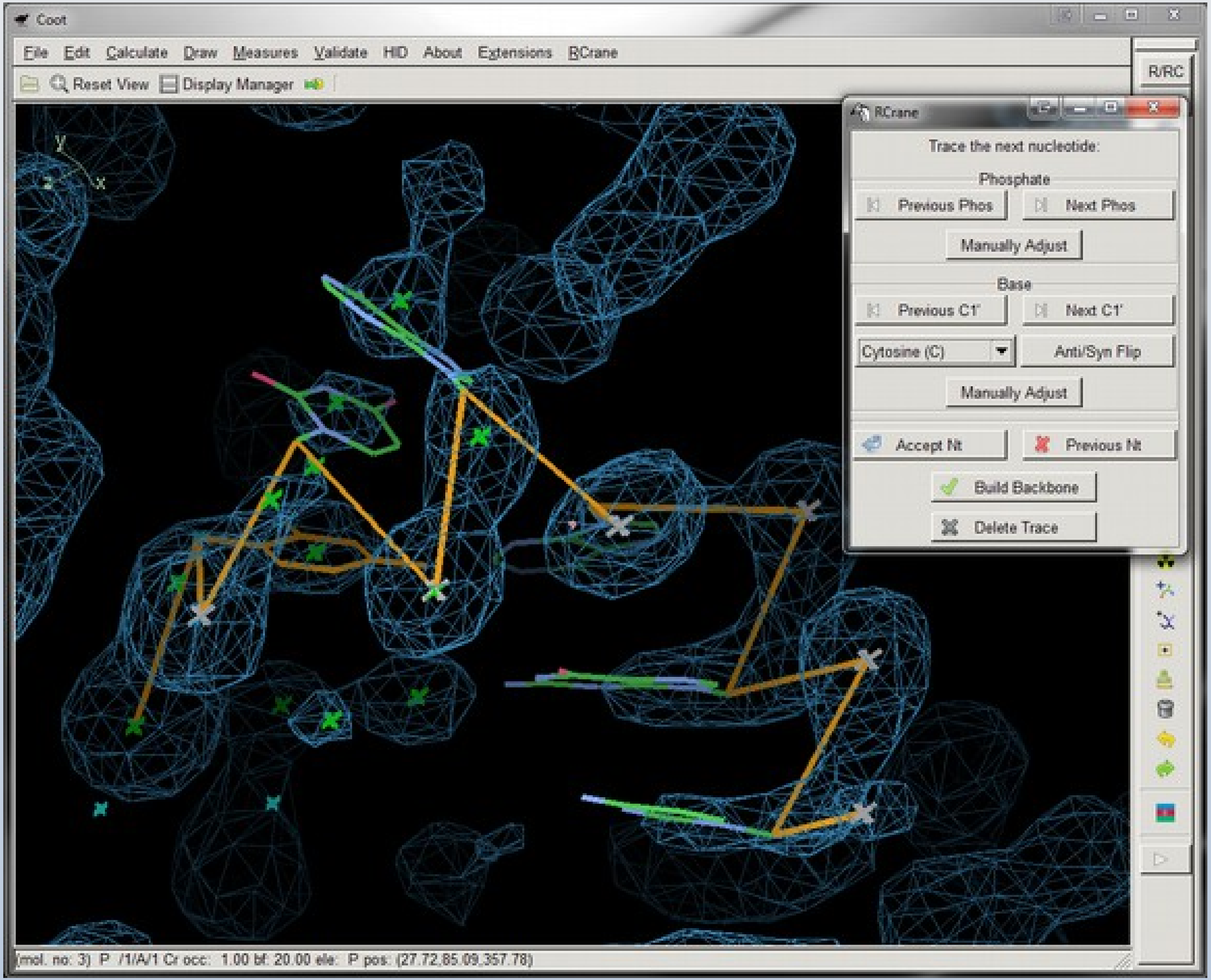


Successfully read coordinates file /home/paule/autobuild/build-coot+rdkit-pre-release-gtk2-python/share/coot/data/tutorial-modern.pdb. Molecule number 2 created.

Model Morphing



RCrane: Semi-Automated Building of RNA



Trace the next nucleotide:

```
(mol. no: 3) P /1/A/1 Cr occ: 1.00 bf: 20.00 ele: P pos: (27.72,85.09,357.78)
```

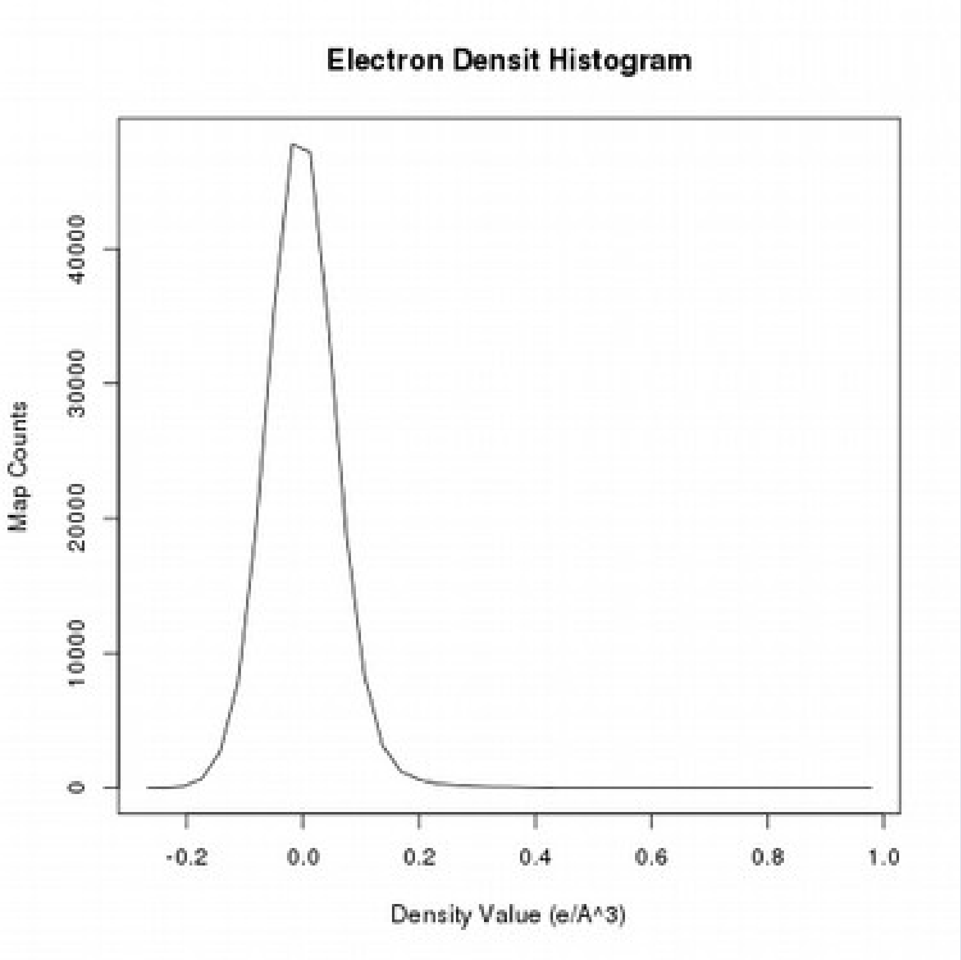
Helix Fitting

- The distribution of electron density is quite unlike that of x-ray maps
  - e.g. You don't see main-chain atoms at 4 rmsd in x-ray maps
    - regions of dense electron density contribute negatively to helix score in x-ray maps

Helix Fitting

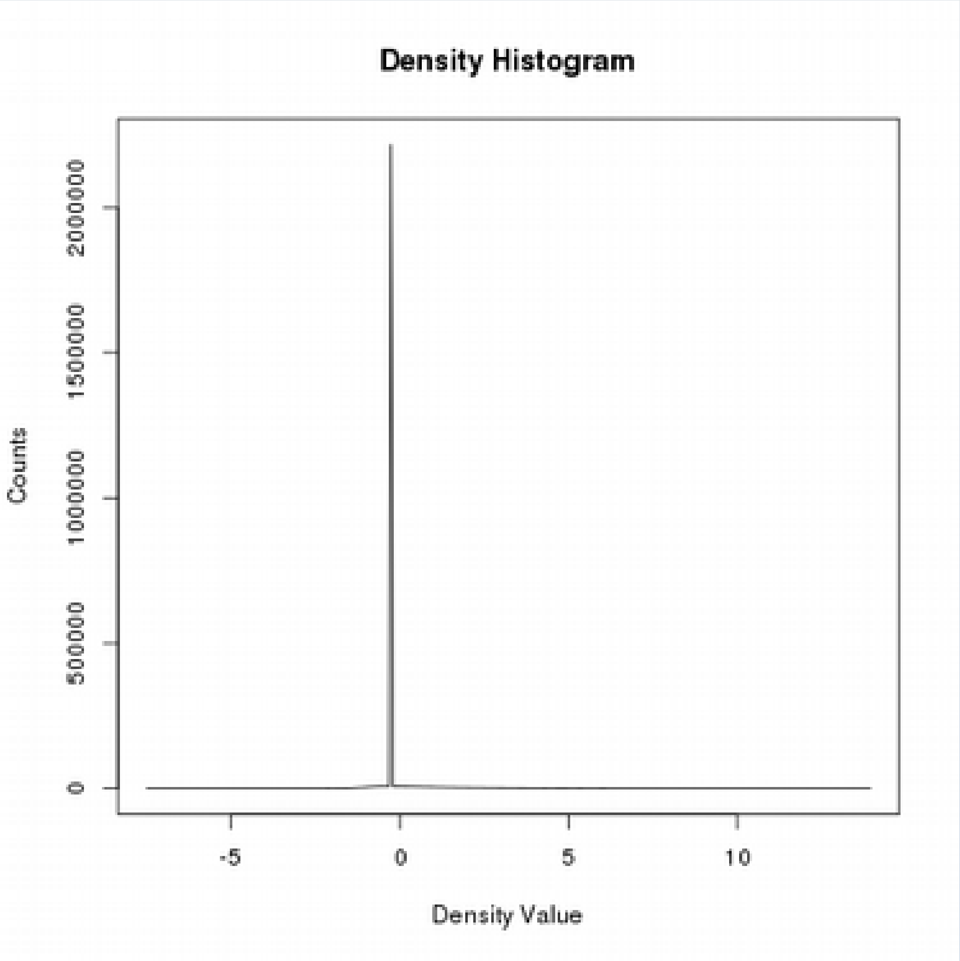
- The distribution of electron density is quite unlike that of x-ray maps

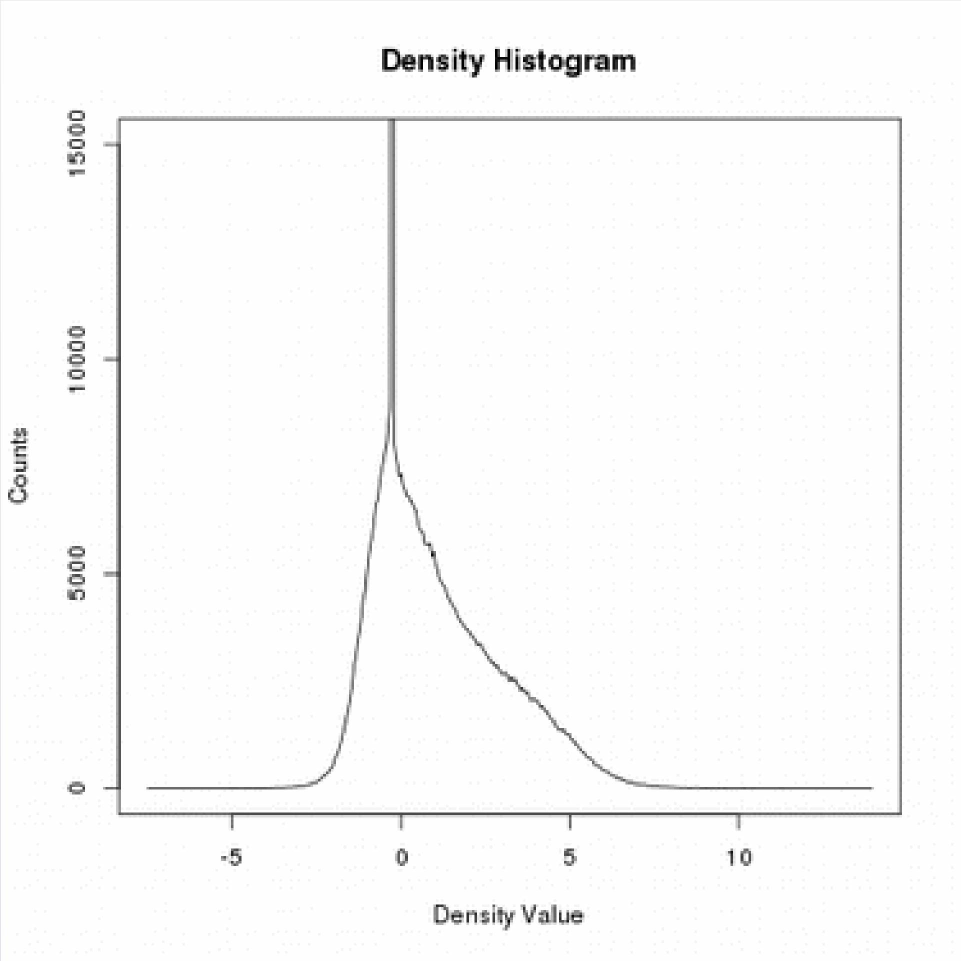
Typical Density Histogram from an X-ray map



Helix Fitting

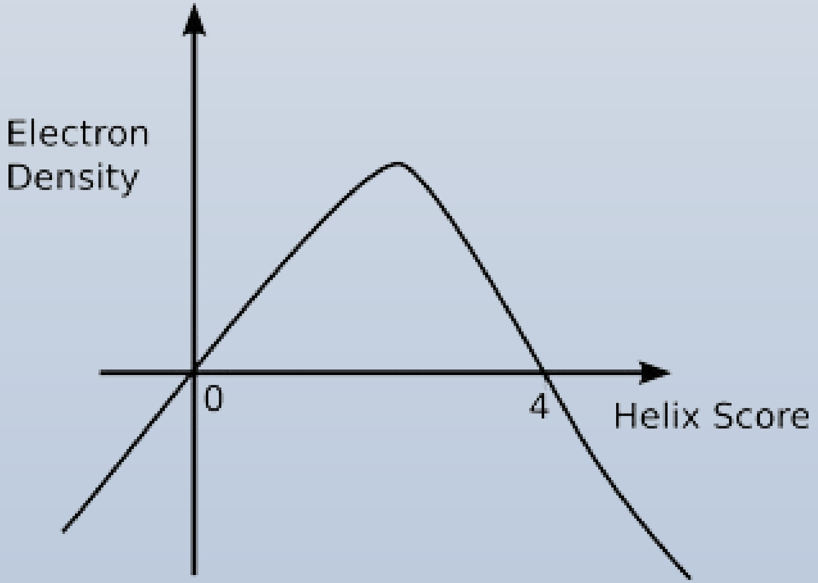
- The distribution of electron density is quite unlike that of x-ray maps





Helix Fitting

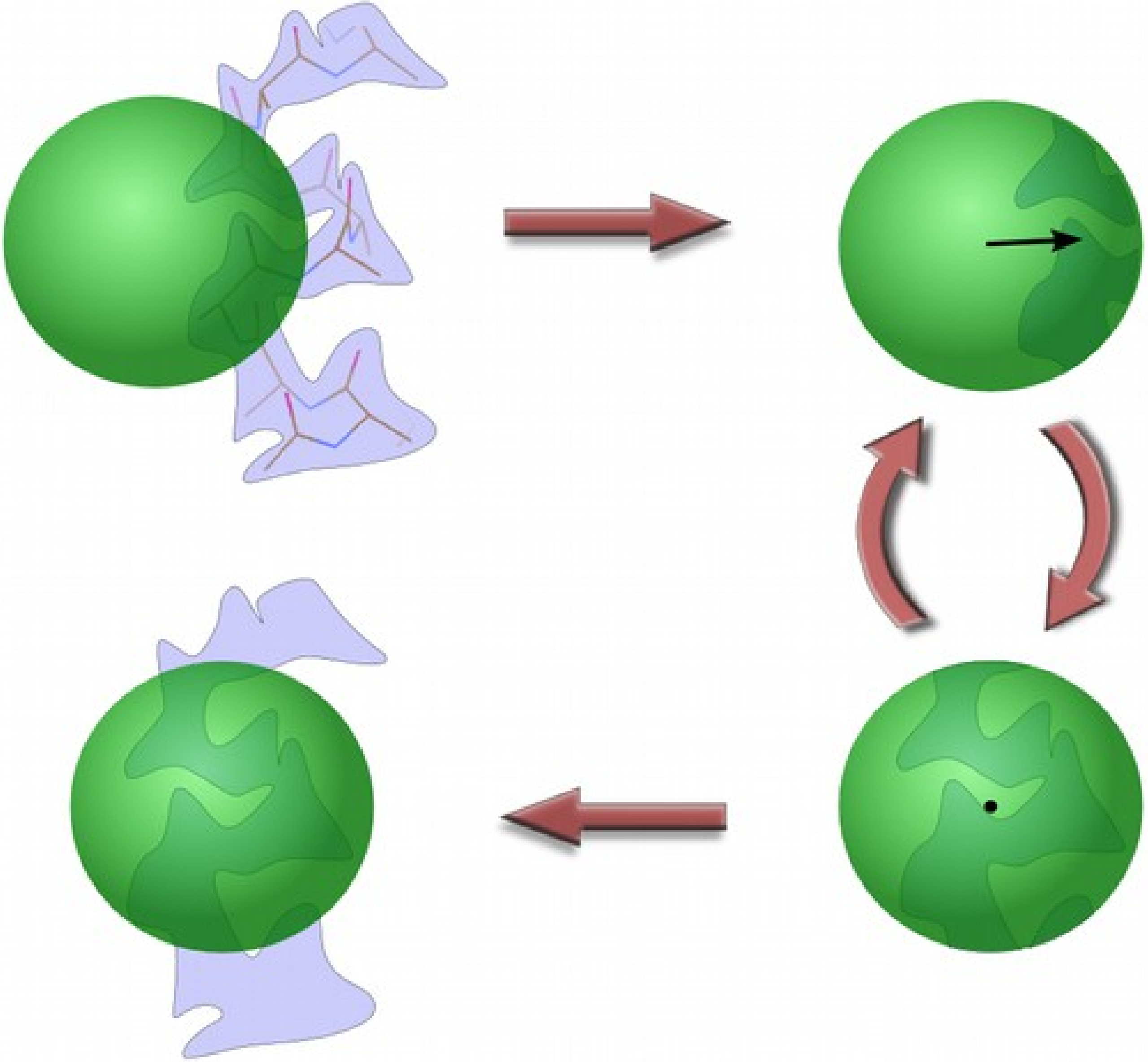
- The distribution of electron density is quite unlike that of x-ray maps
  - e.g. You don't see main-chain atoms at 4 rmsd in x-ray maps
    - regions of dense electron density contribute negatively to helix score
  - These EM maps were sharpened and in a big box of mostly nothing
  - Lots to see at 4 rmsd



Alpha Helix Placement

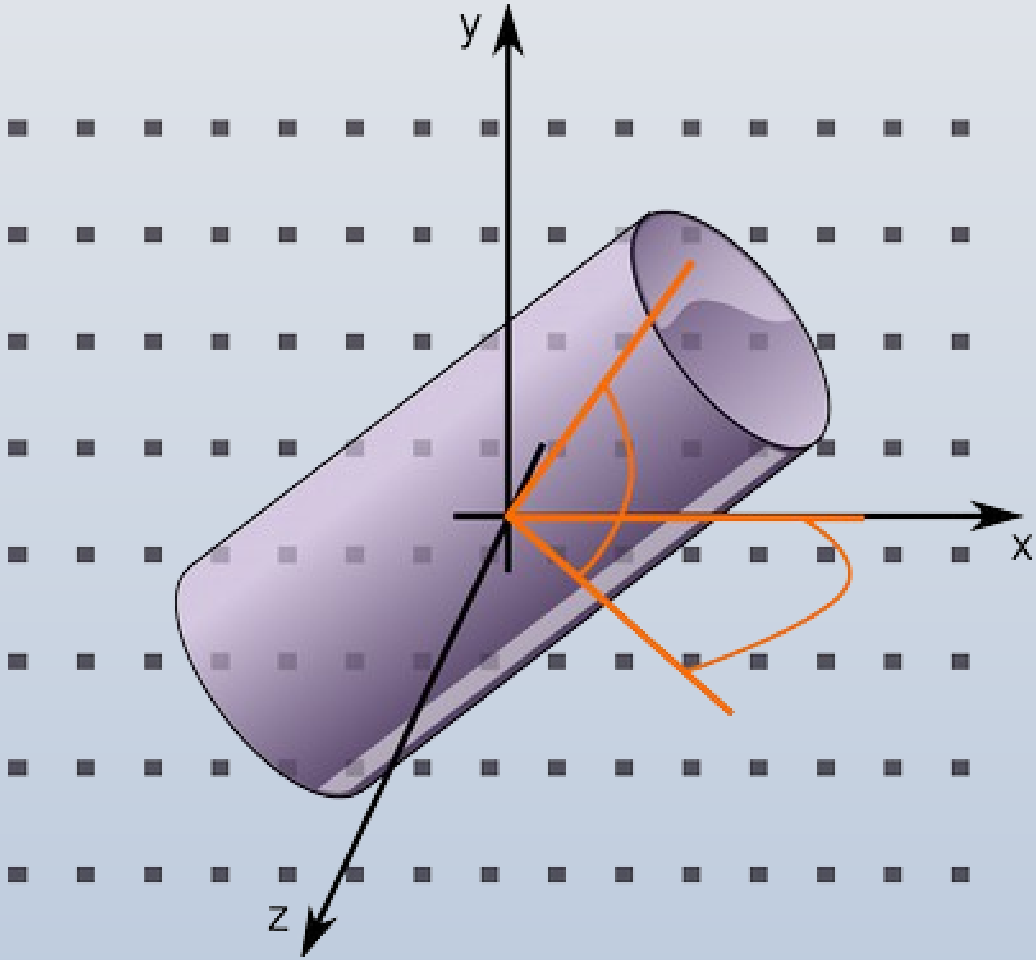
- Scenario: Looking at a new map, not built with automatic tools:
  - "I can see that there's a helix here - build it for me!"
- From a given point:
  - Move to local averaged maximum
  - Do a 2D MR-style orientation search on a cylinder of electron density
  - Build a helix (both directions)
  - 1D Rotation search to find best fit
  - Score based on density at CB positions
  - Trim 'n Grow

Centering the Rotation point

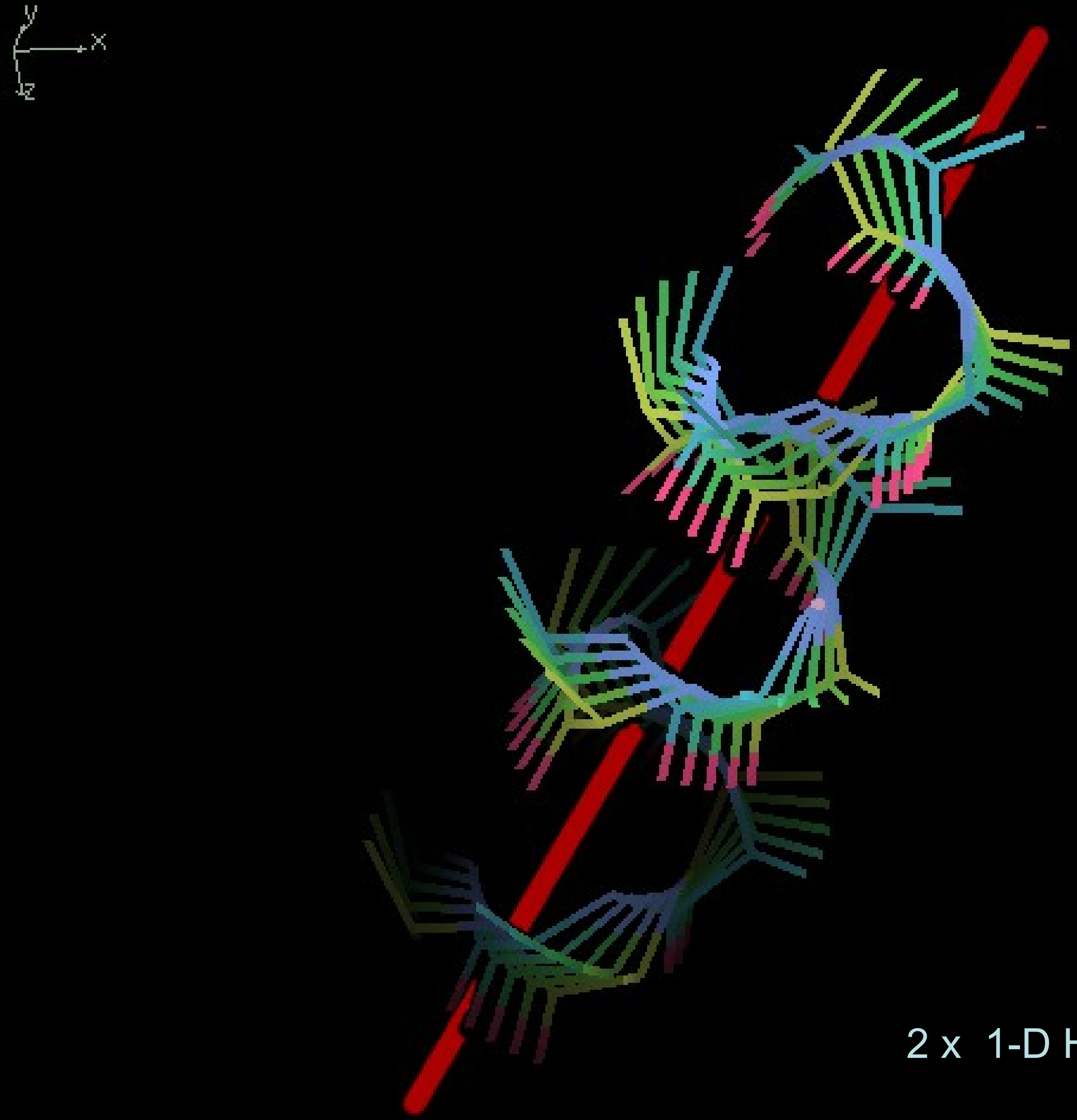


Helix Fitting:
Cylinder Search

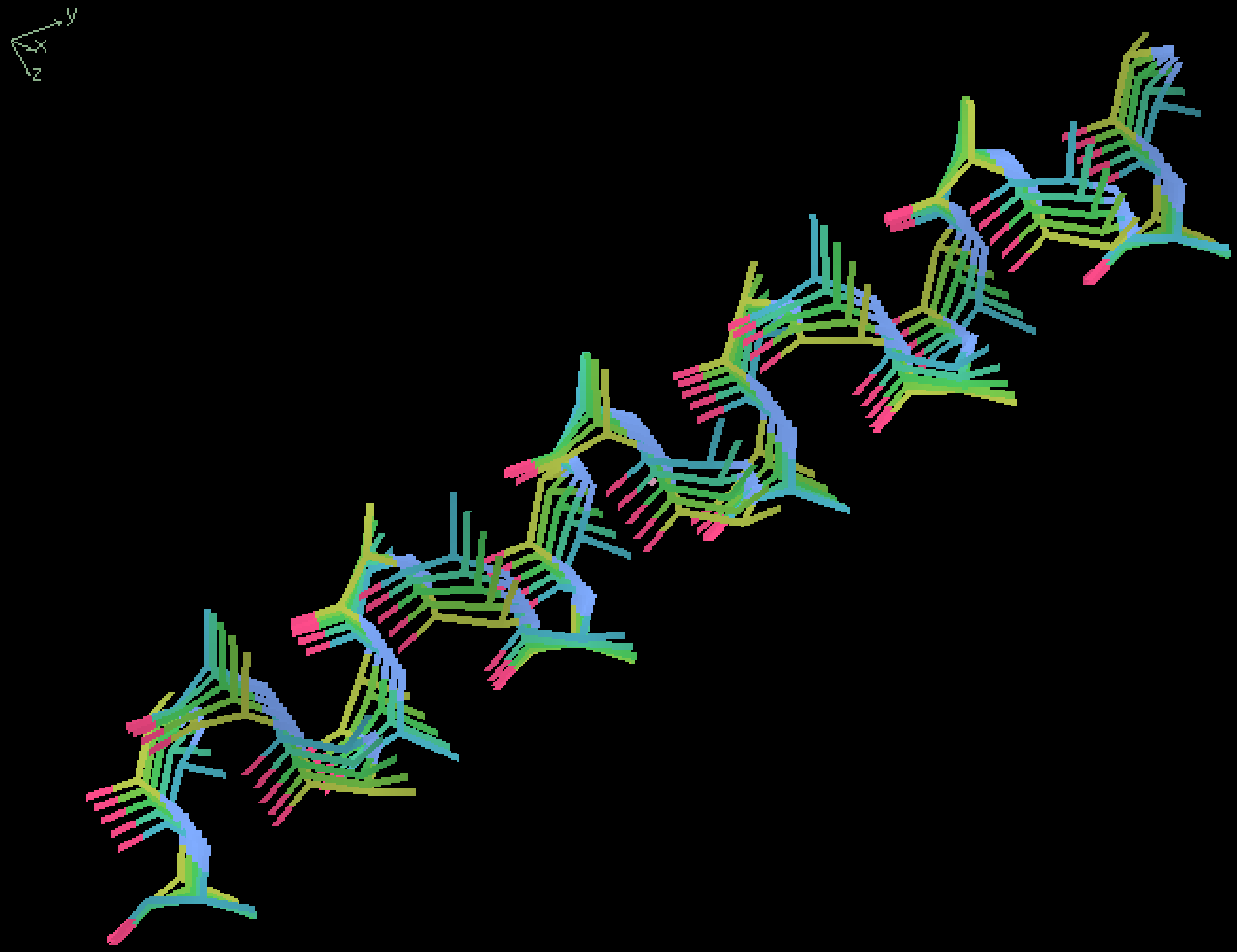
- Pick the orientation that encapsulates the most electron density

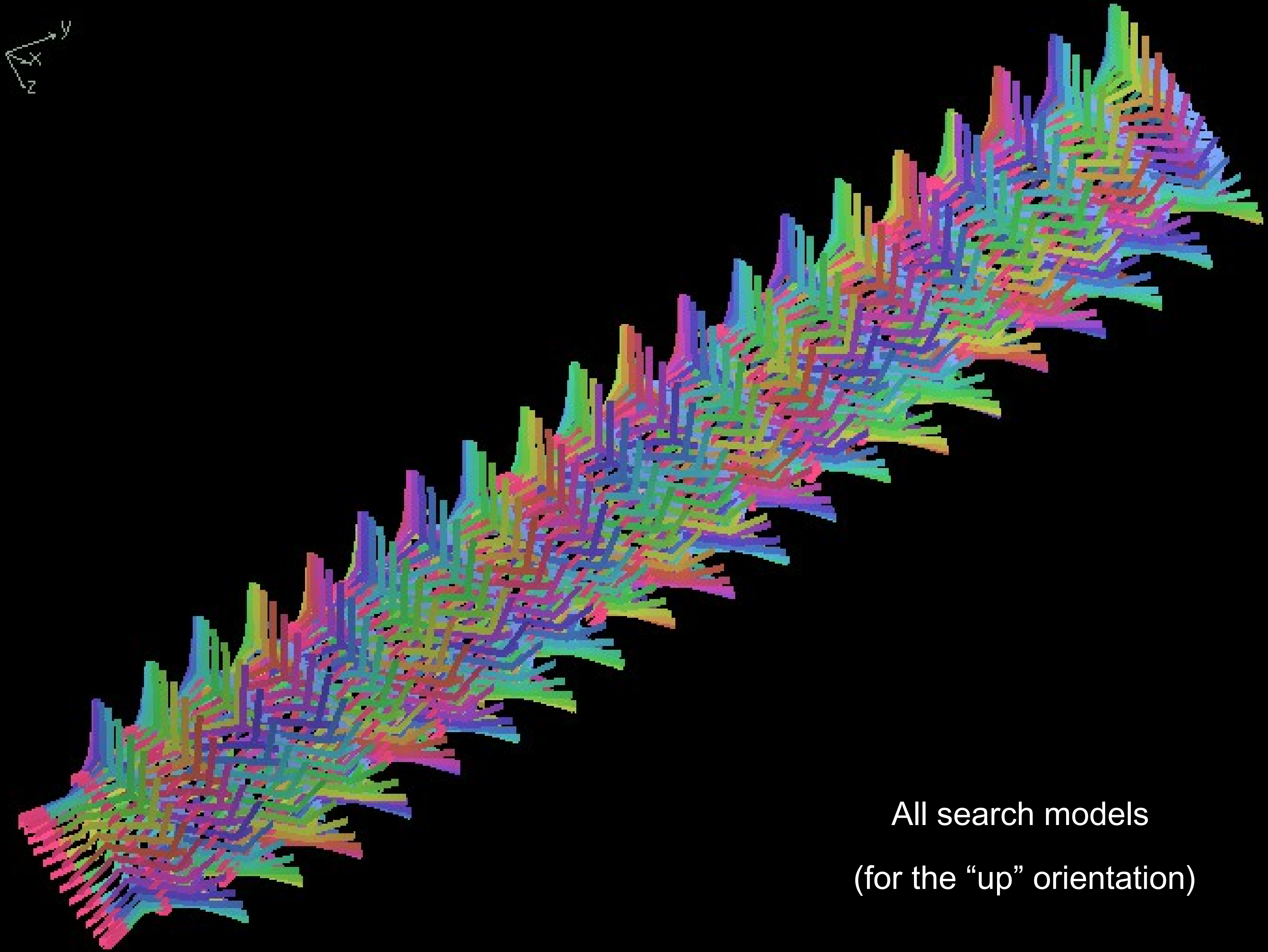


Using 2 rotation axes



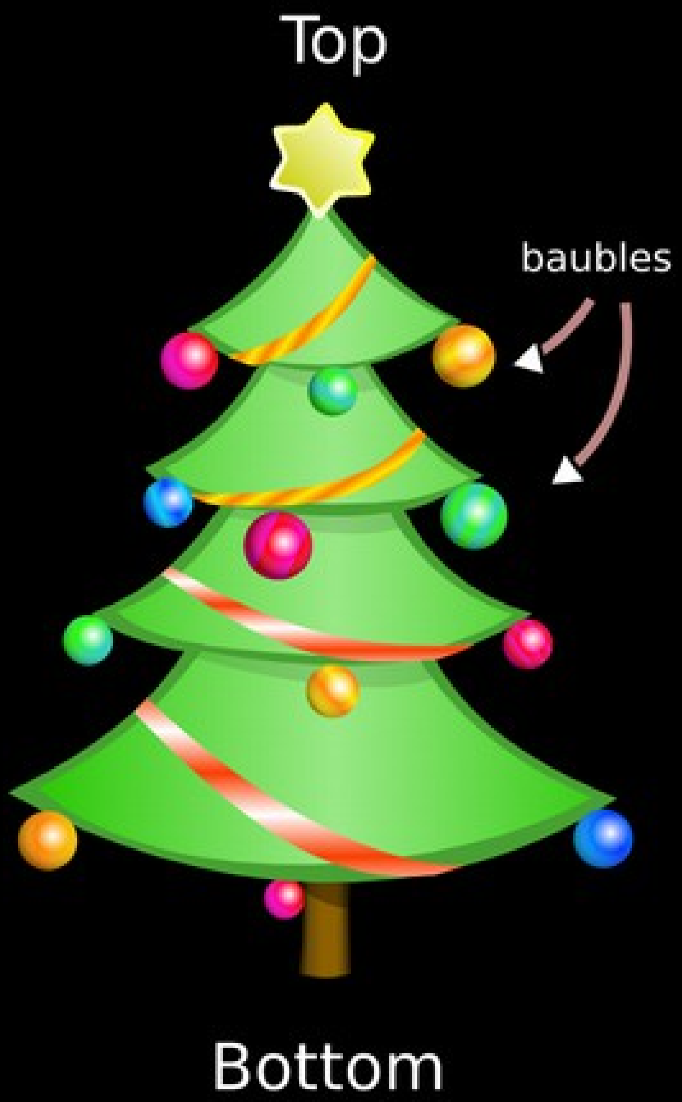
2 x 1-D Helix orientation searches


y
x
z

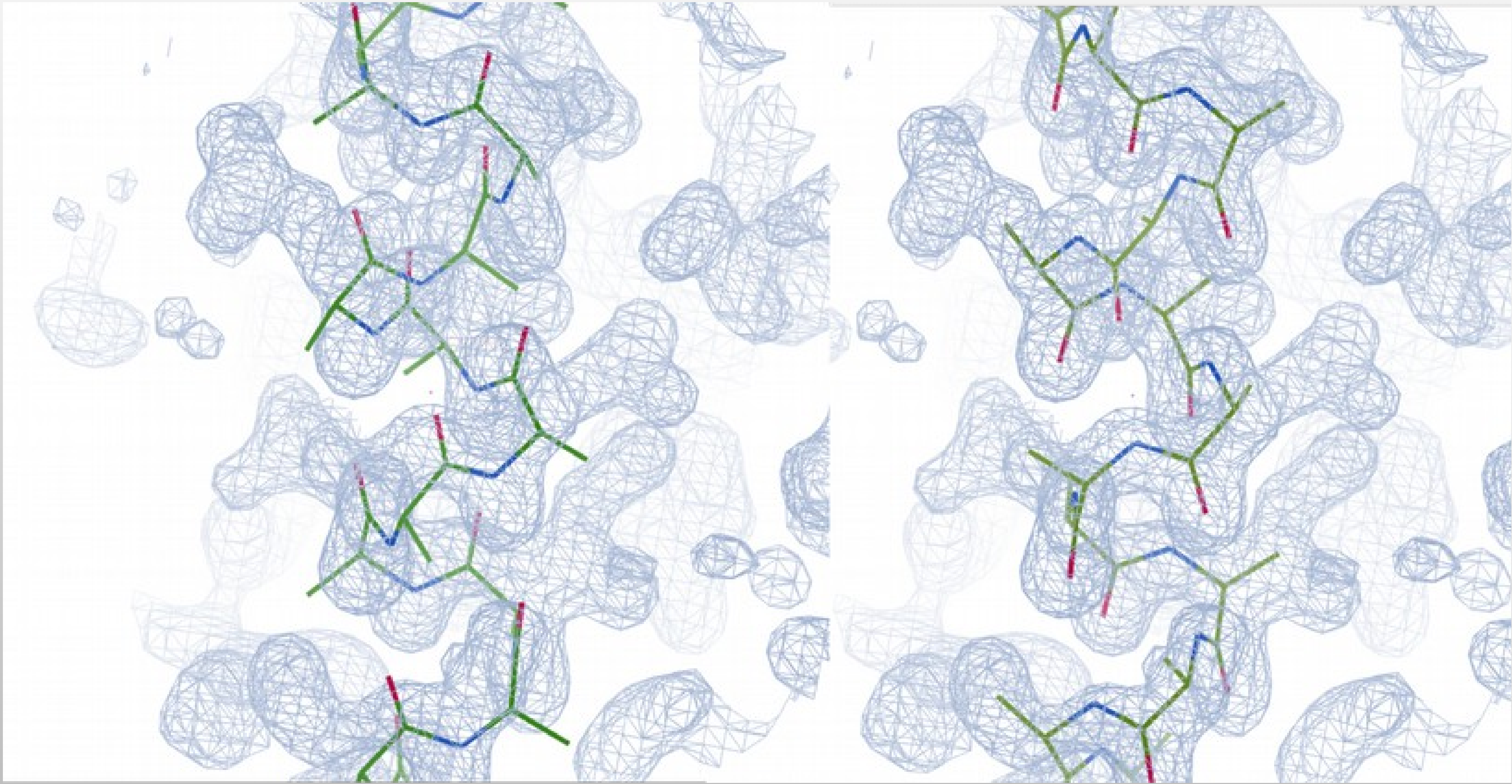


All search models

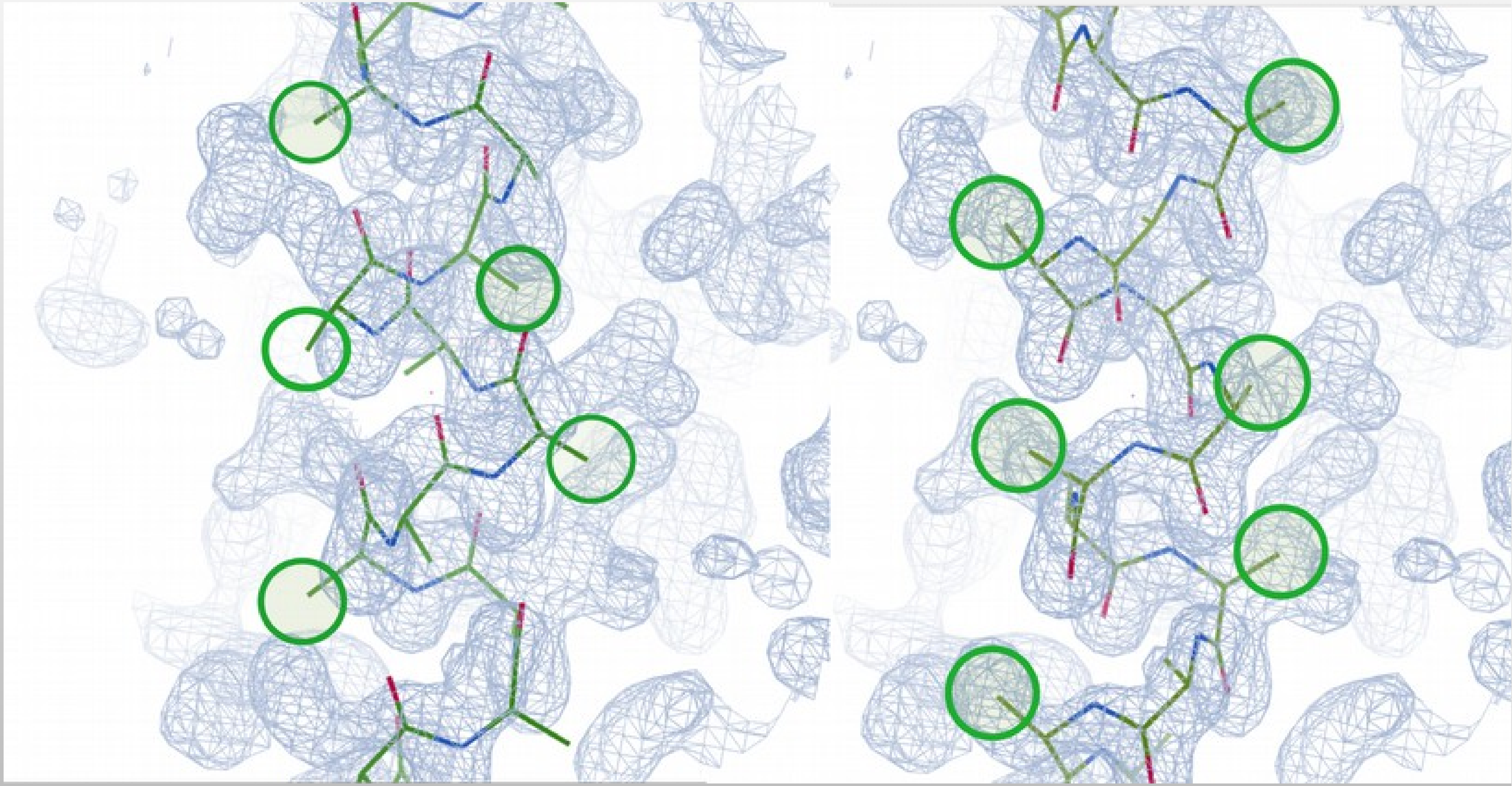
(for the “up” orientation)


Top
baubles
Bottom

Helix Fitting
Comparing orientation hypotheses



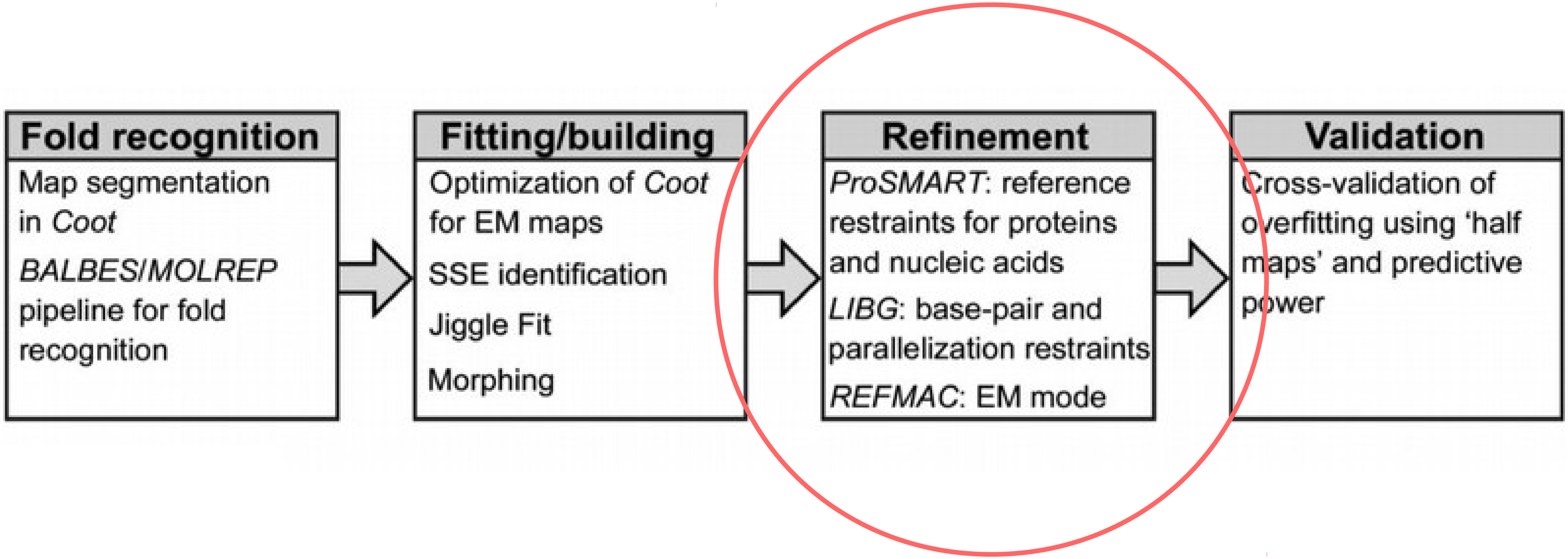
Helix Fitting
Comparing orientation hypotheses



c-betas are not fitted and are used for scoring

Model-Building Tools

Recent Developments



Fold recognition

Map segmentation in *Coot*

BALBES/MOLREP pipeline for fold recognition

Fitting/building

Optimization of *Coot* for EM maps

SSE identification

Jiggle Fit

Morphing

Refinement

ProSMART: reference restraints for proteins and nucleic acids

LIBG: base-pair and parallelization restraints

REFMAC: EM mode

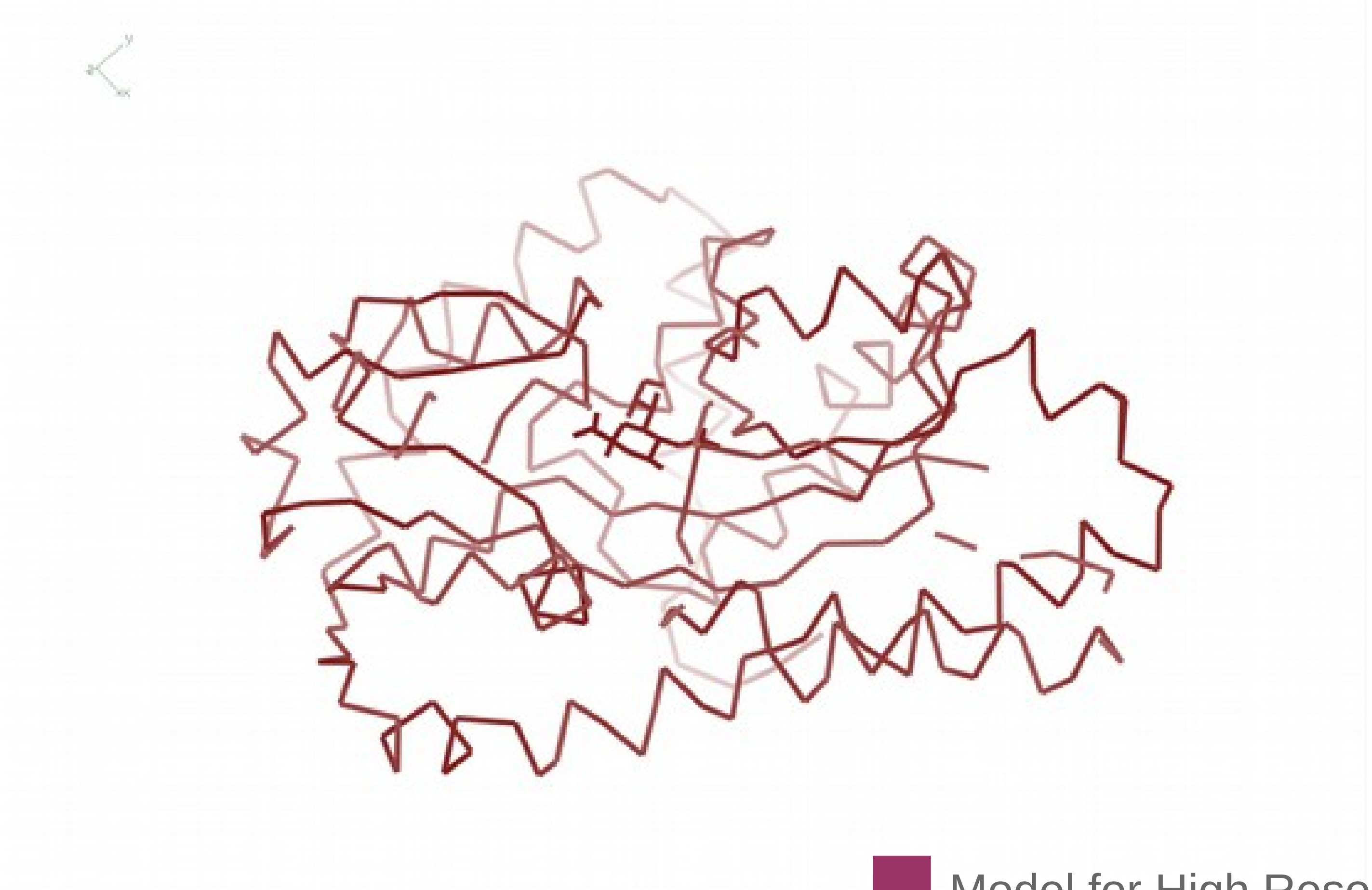
Validation

Cross-validation of overfitting using 'half maps' and predictive power

Local Distance Restraints

- ProSMART:
  - Use previous-solved "template" structures to inform the refinement of the (low resolution) target protein
  - by local-distance of atom pairs restraint generation

Local Distance Restraints: Prior Structure

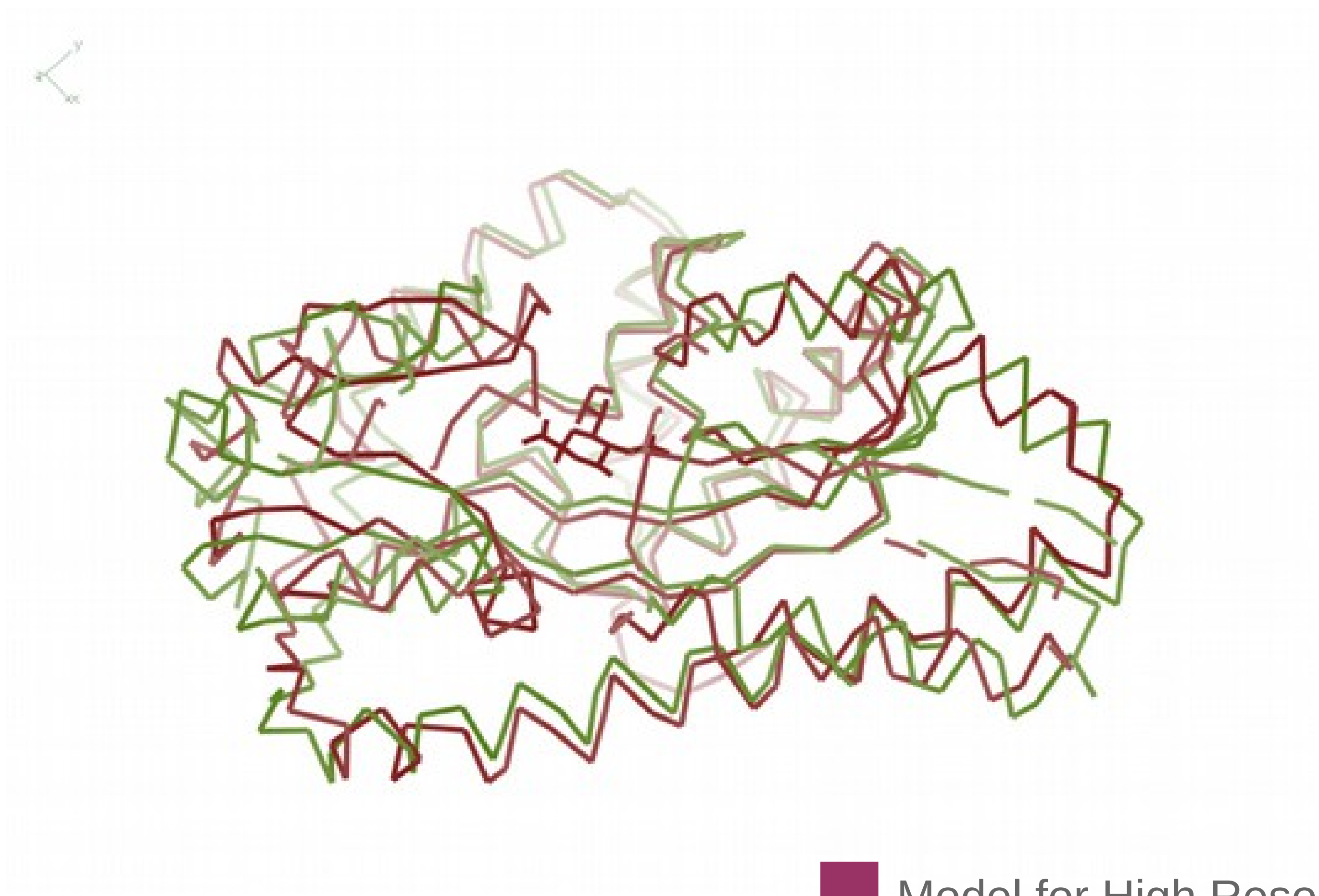


🟪 Model for High Resolution data

Local Distance Restraints: Today's Structure

🟩 Model for Low Resolution data

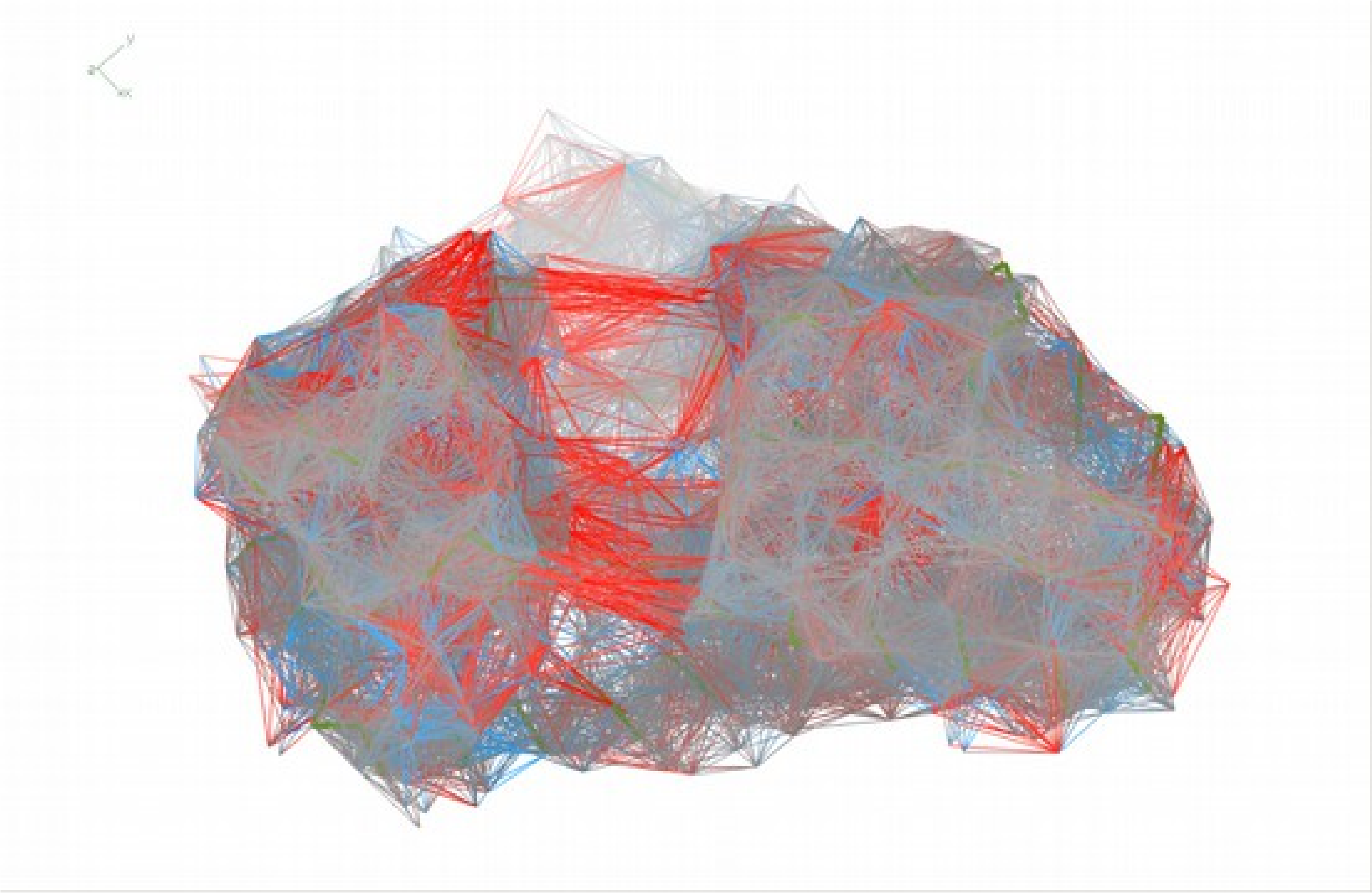
Local Distance Restraints: Structure Comparison



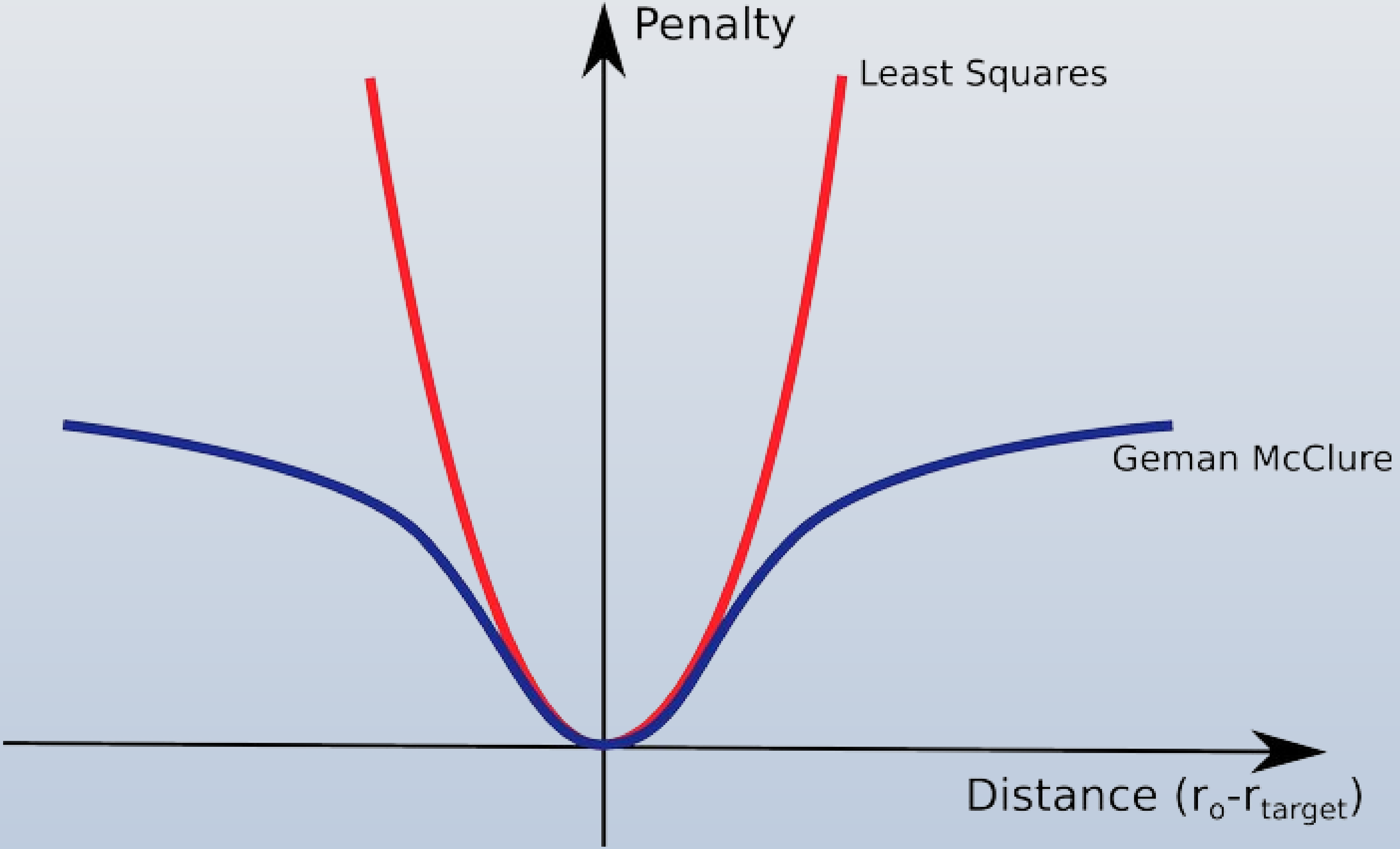
🟪 Model for High Resolution data

🟩 Model for Low Resolution data

Local Distance Restraints: ProSMART/Geman McClure Restraints



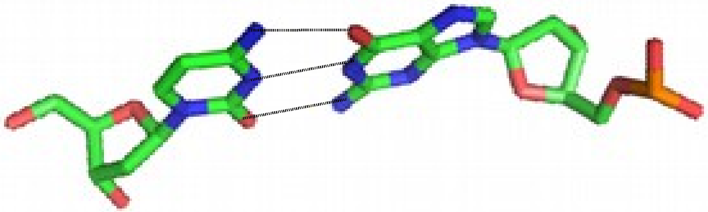
Modified Target Function



LIBG Restraints for DNA/RNA

LIBG – for the generation of nucleic acid restraints

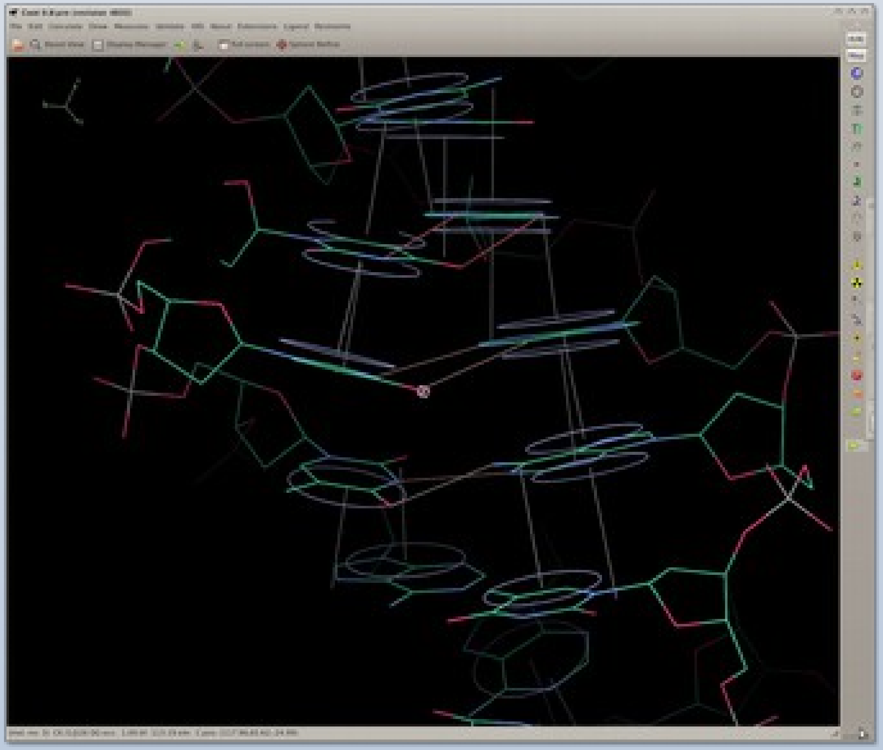
Base-pair restraints:

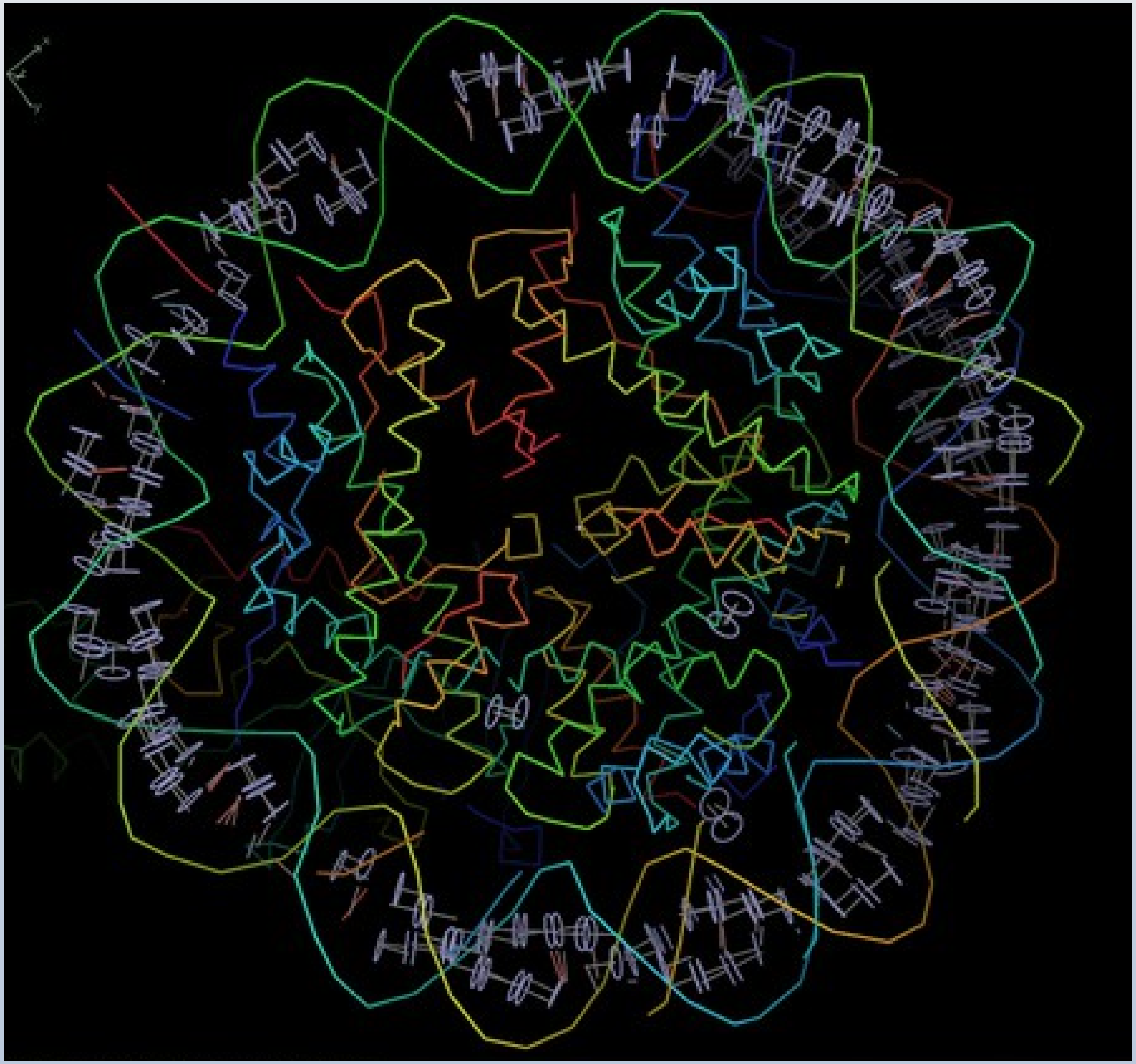


LIBG Restraints for DNA/RNA

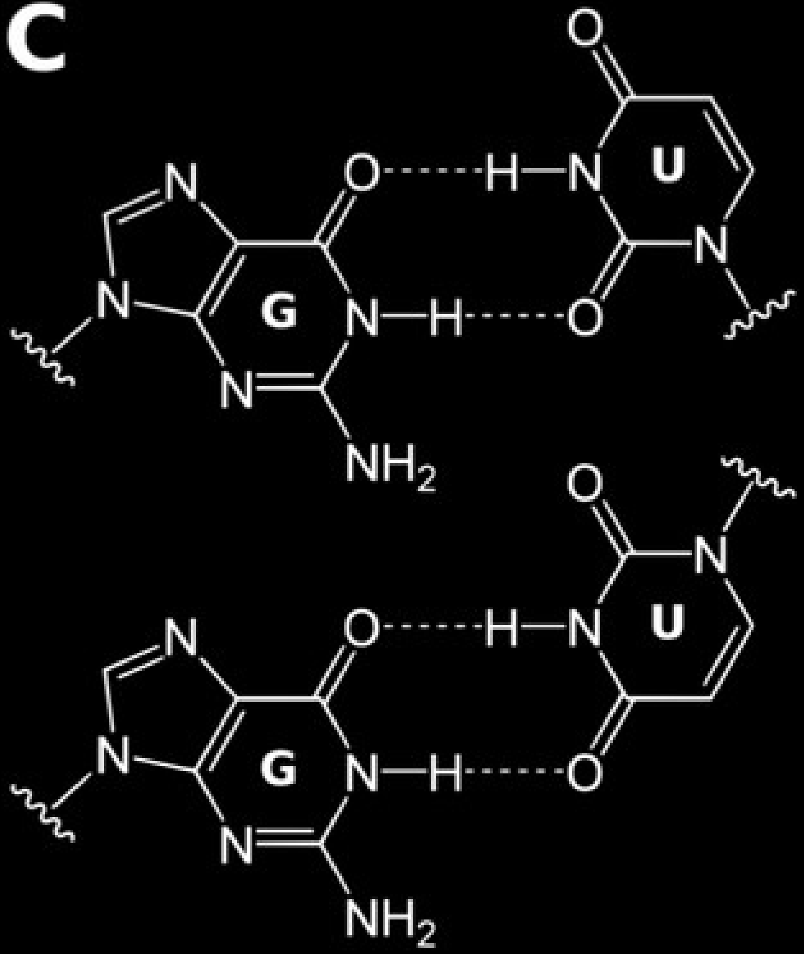
LIBG – for the generation of nucleic acid restraints

Base-stacking restraints:
(parallel plane restraints)

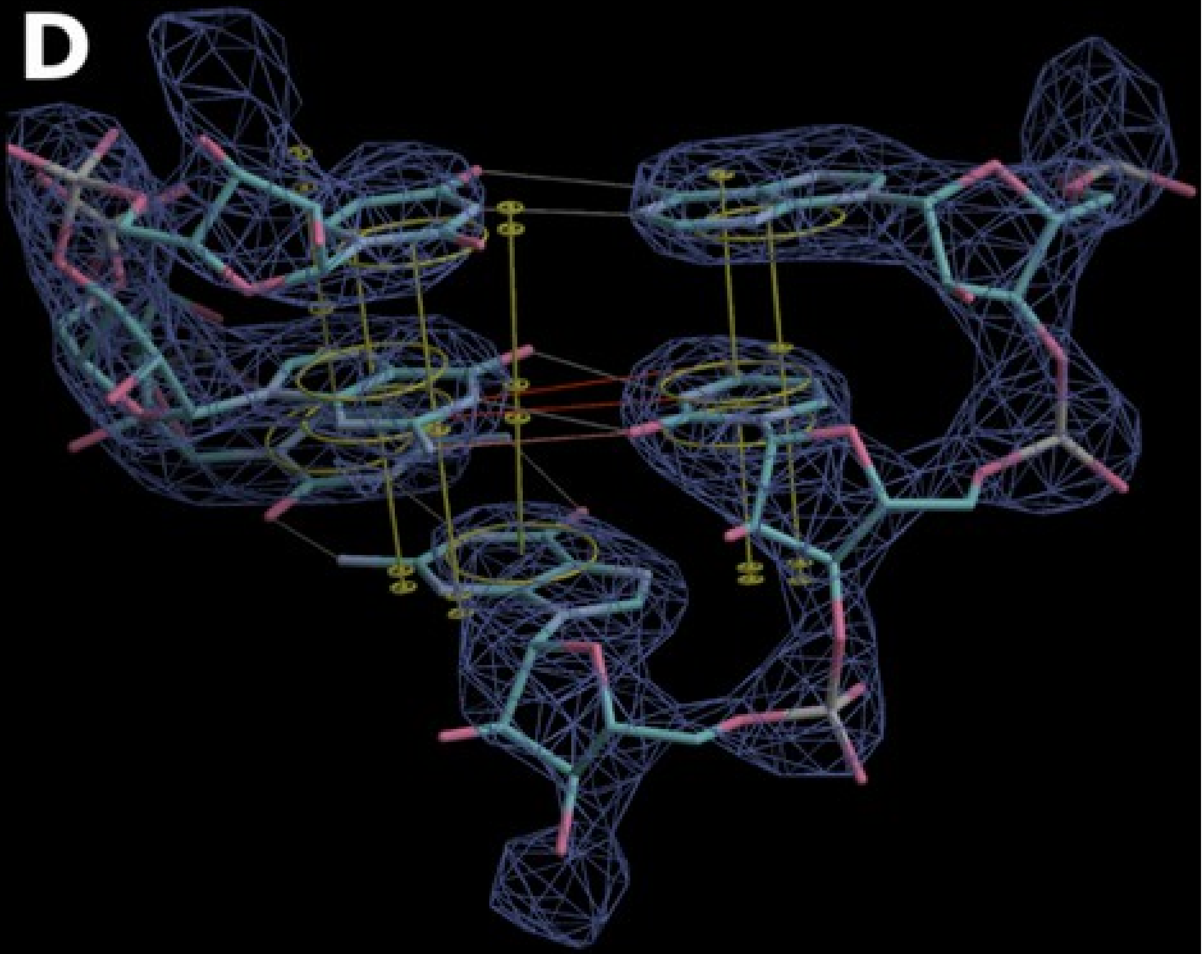




LIBG Restraints



(Watson Crick and)
Wobble, Reverse Wobble



Representation in Coot

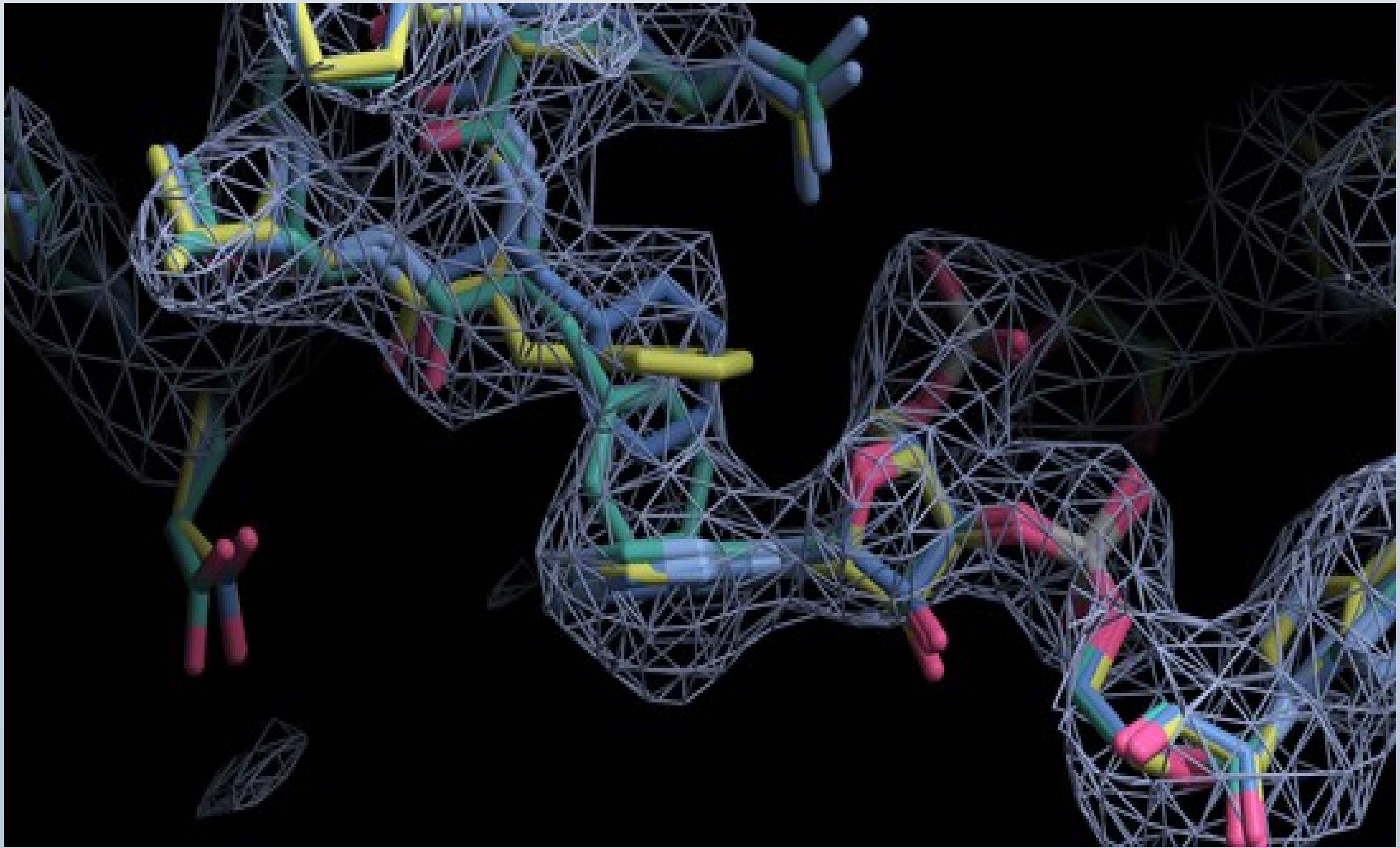
LIBG Restraints for DNA/RNA

LIBG – for the generation of nucleic acid restraints

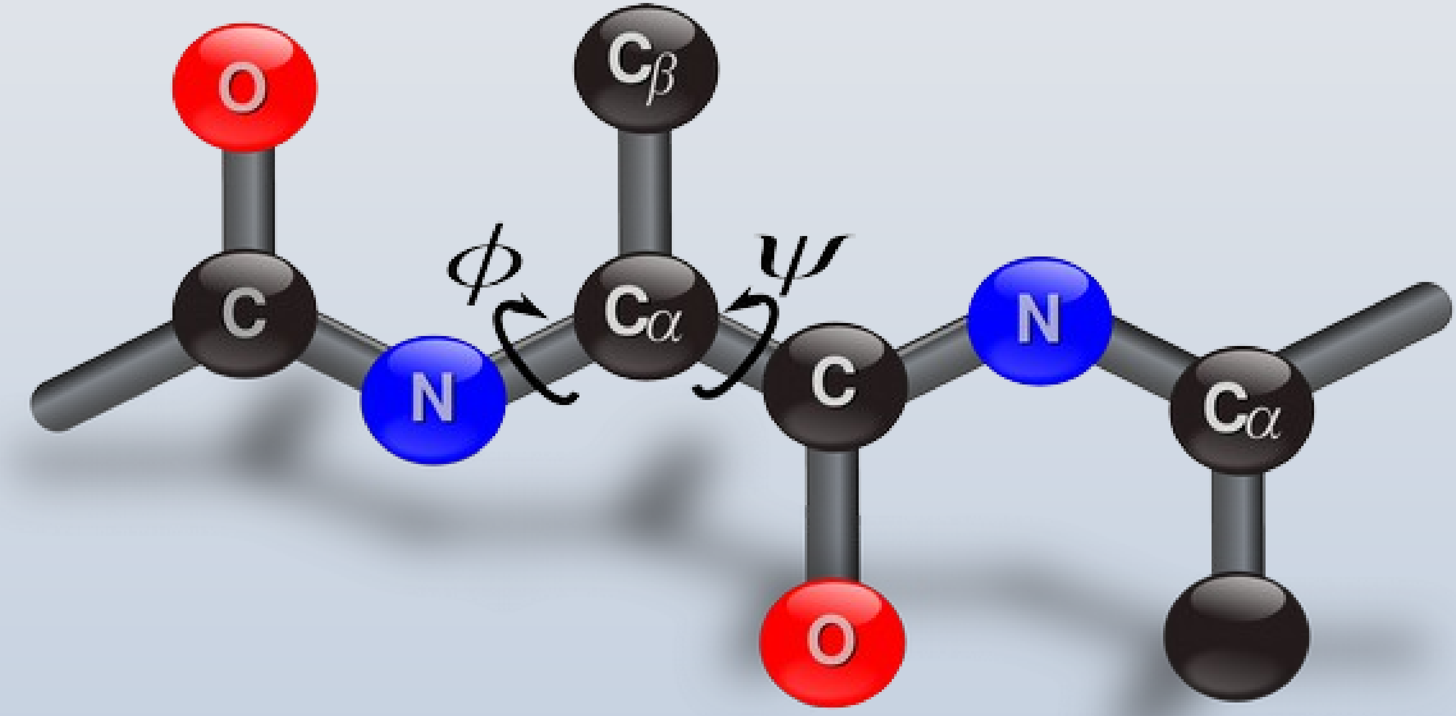
Base-stacking restraints:
(parallel plane restraints)

Green – before refinement
Blue – refined without LIBG
Yellow – refined with LIBG

Example:
3.2Å cryo-EM



Peptide Backbone Geometry



Low Resolution Model-Building

- “Backrub” rotamers

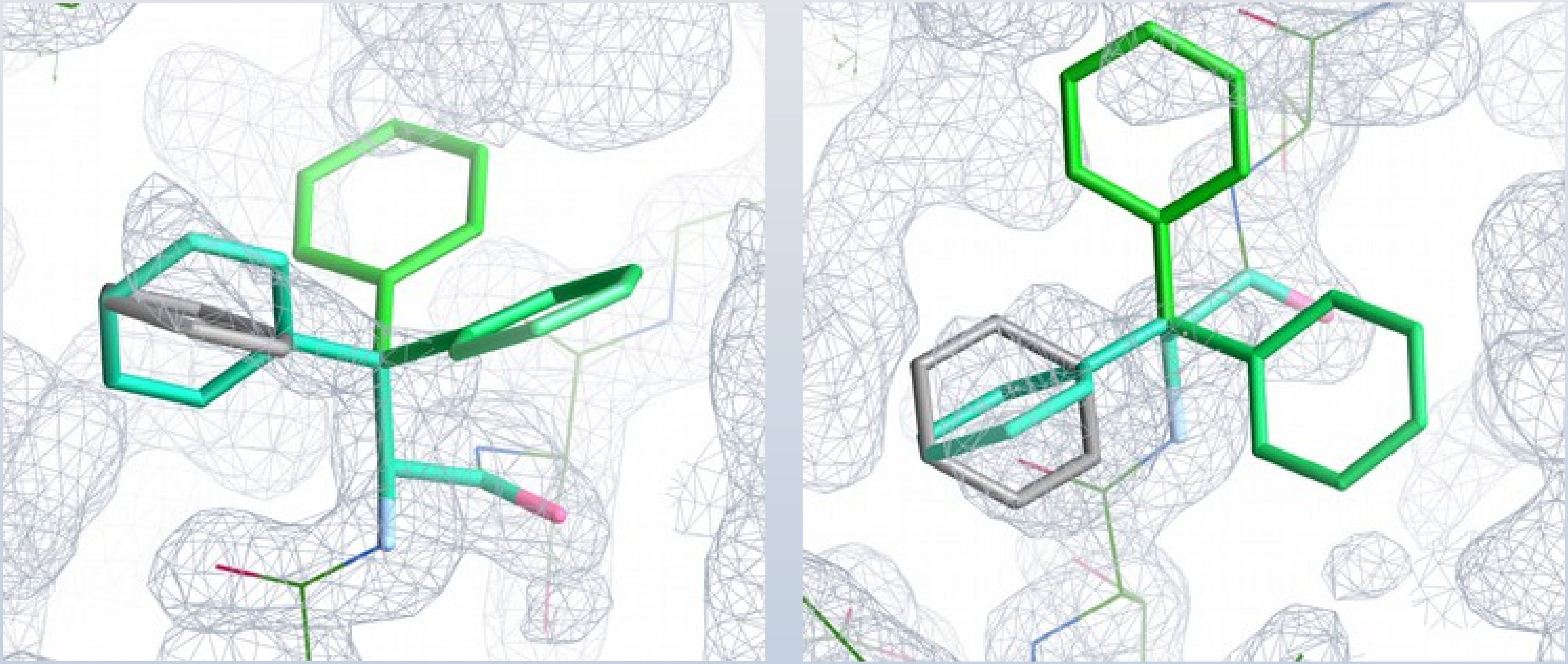
Rotamer Searching

- Two methods
    - Traditional
    - Backrub

Rotamers

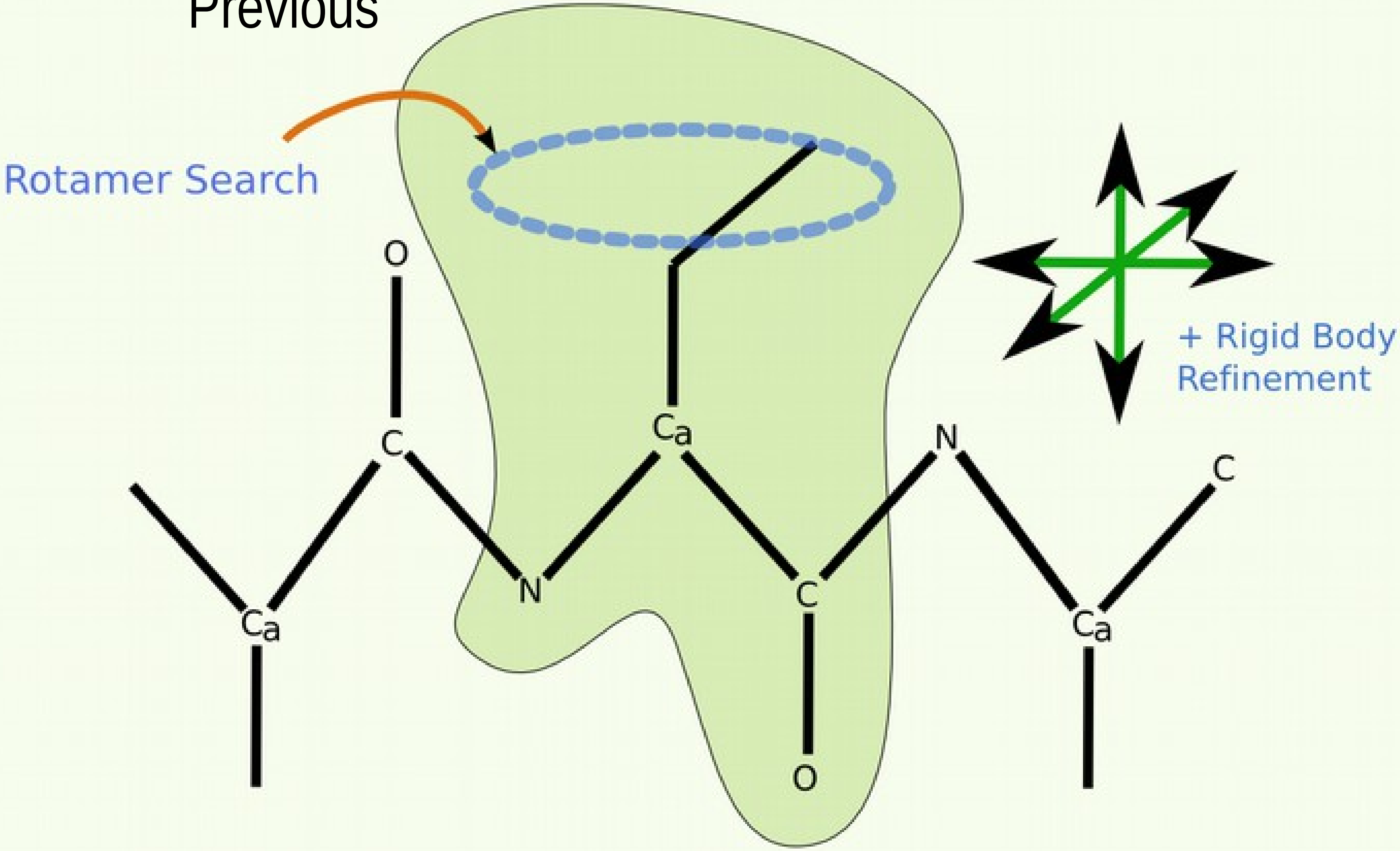
- Rotamers are preferred configurations of a side-chains rotatable bonds
  - where "preferred" means these configurations occur more frequently in a set of reference protein structures
  - "preferred" because they are low-energy conformations
- Several Rotamer "databases" exist
  - best: (Son of) Penultimate Rotamer Library

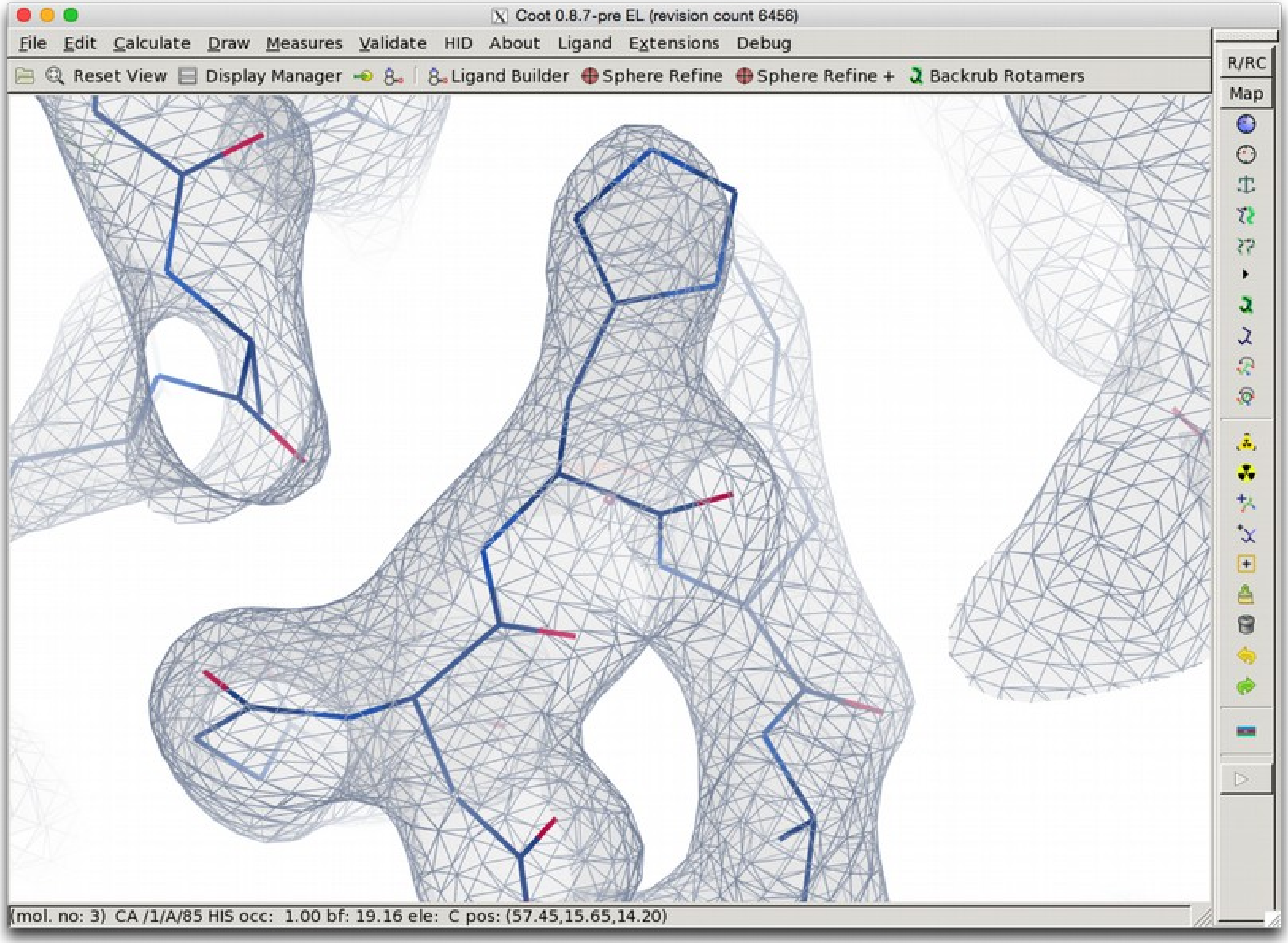
4 PHE Rotamers



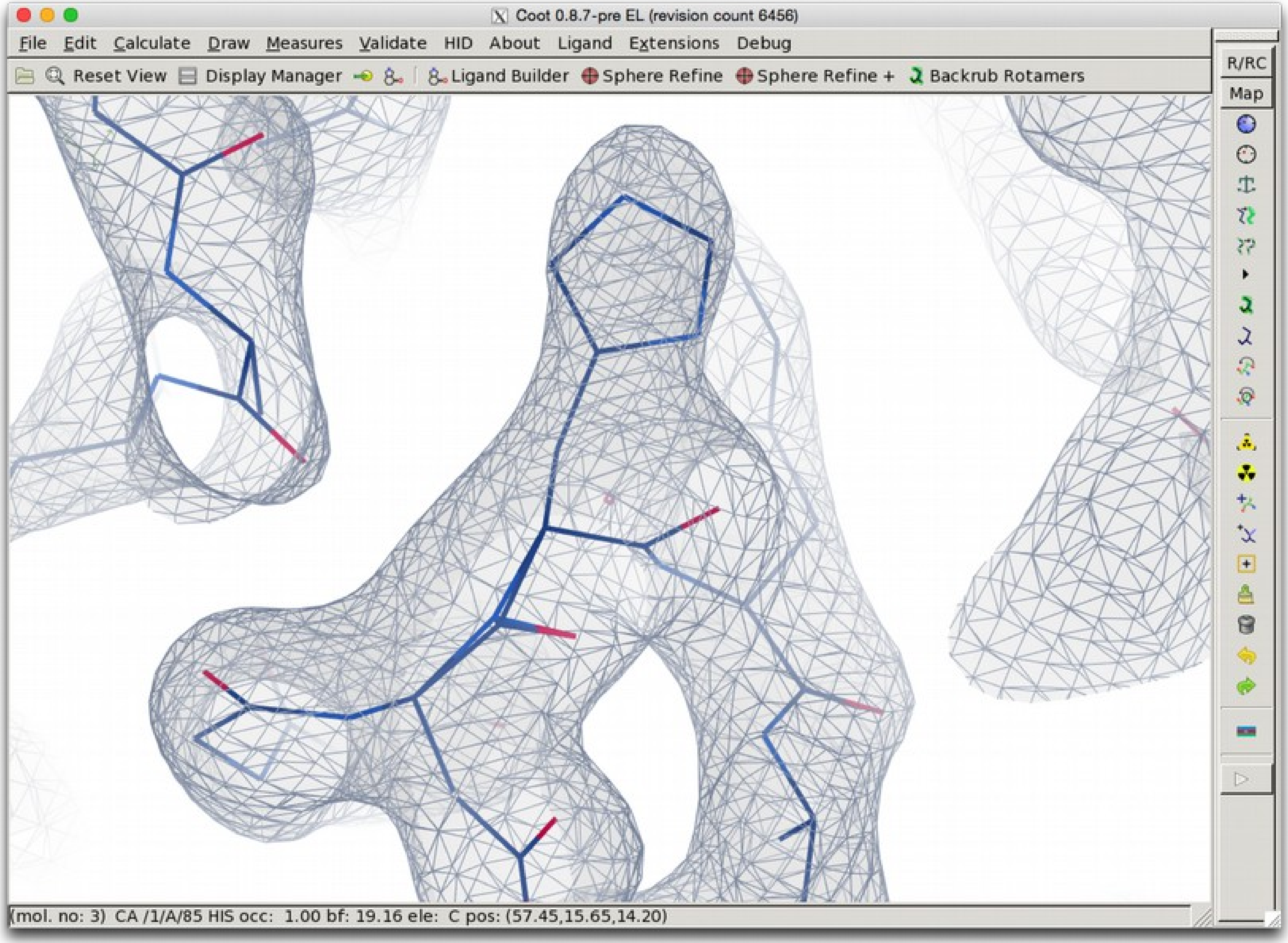
~~Current~~ Low Resolution Rotamer Search

Previous

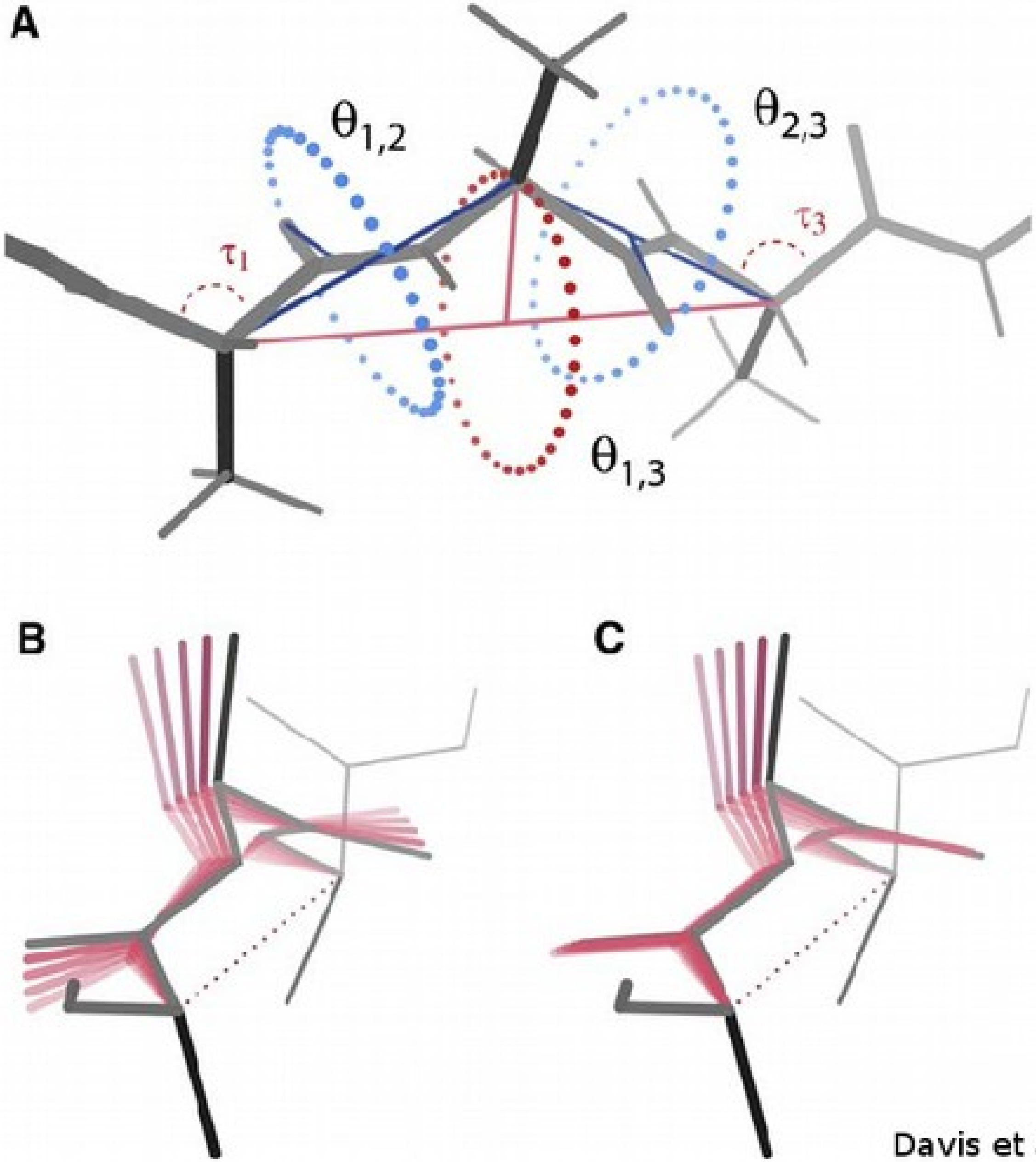




```
(mol. no: 3) CA /1/A/85 HIS occ:  1.00 bf: 19.16 ele:  C pos: (57.45,15.65,14.20)
```

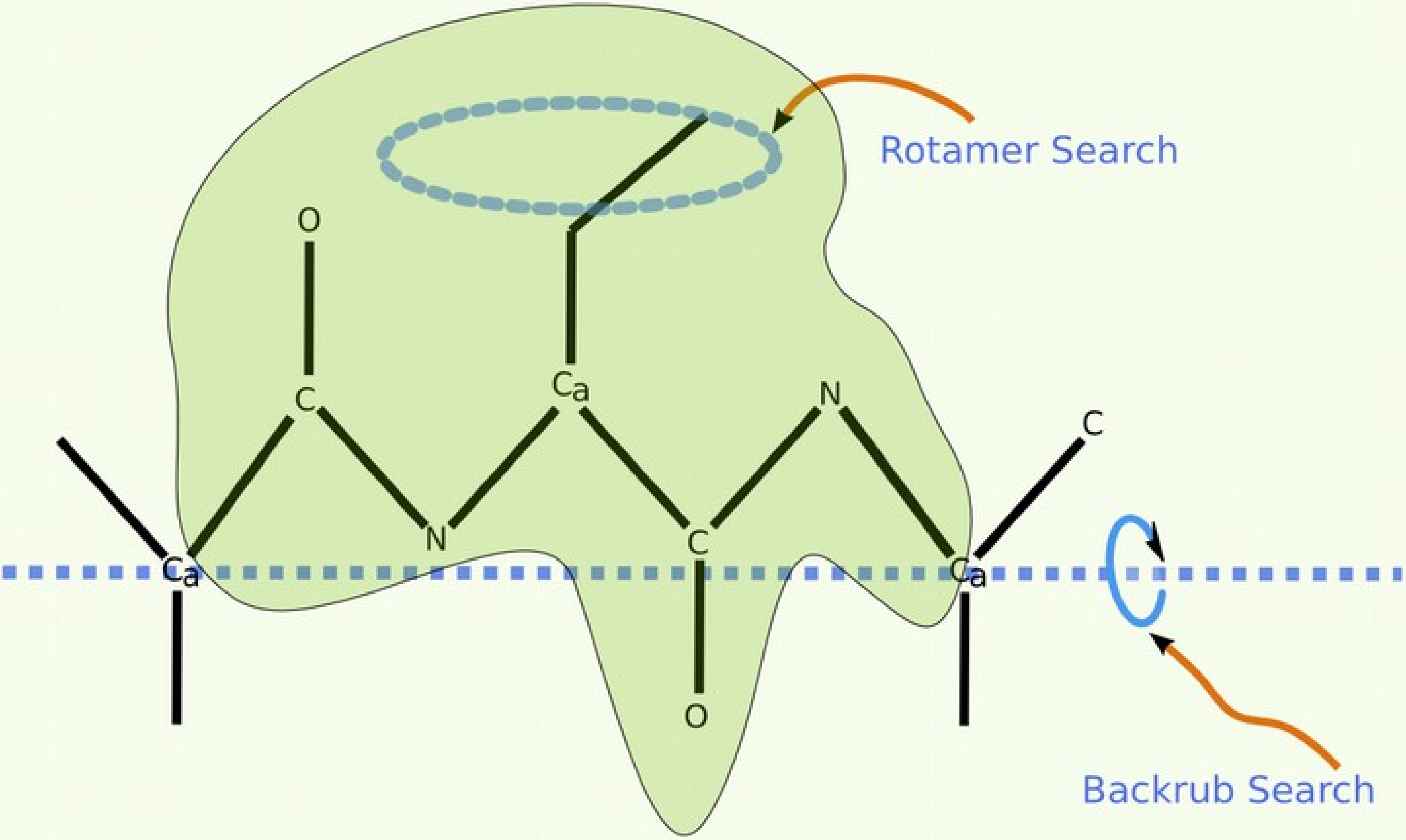


```
(mol. no: 3) CA /1/A/85 HIS occ:  1.00 bf: 19.16 ele:  C pos: (57.45,15.65,14.20)
```

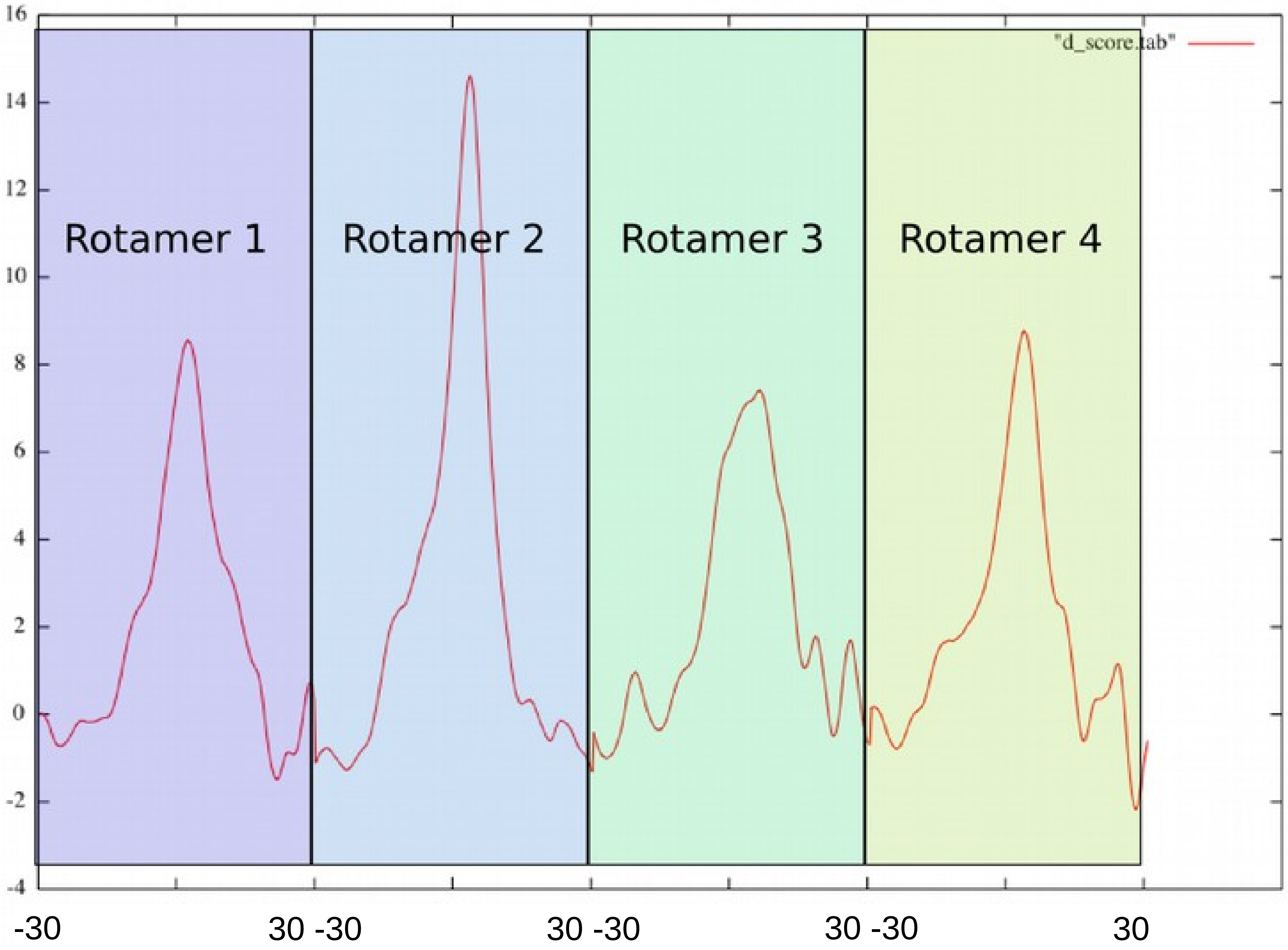


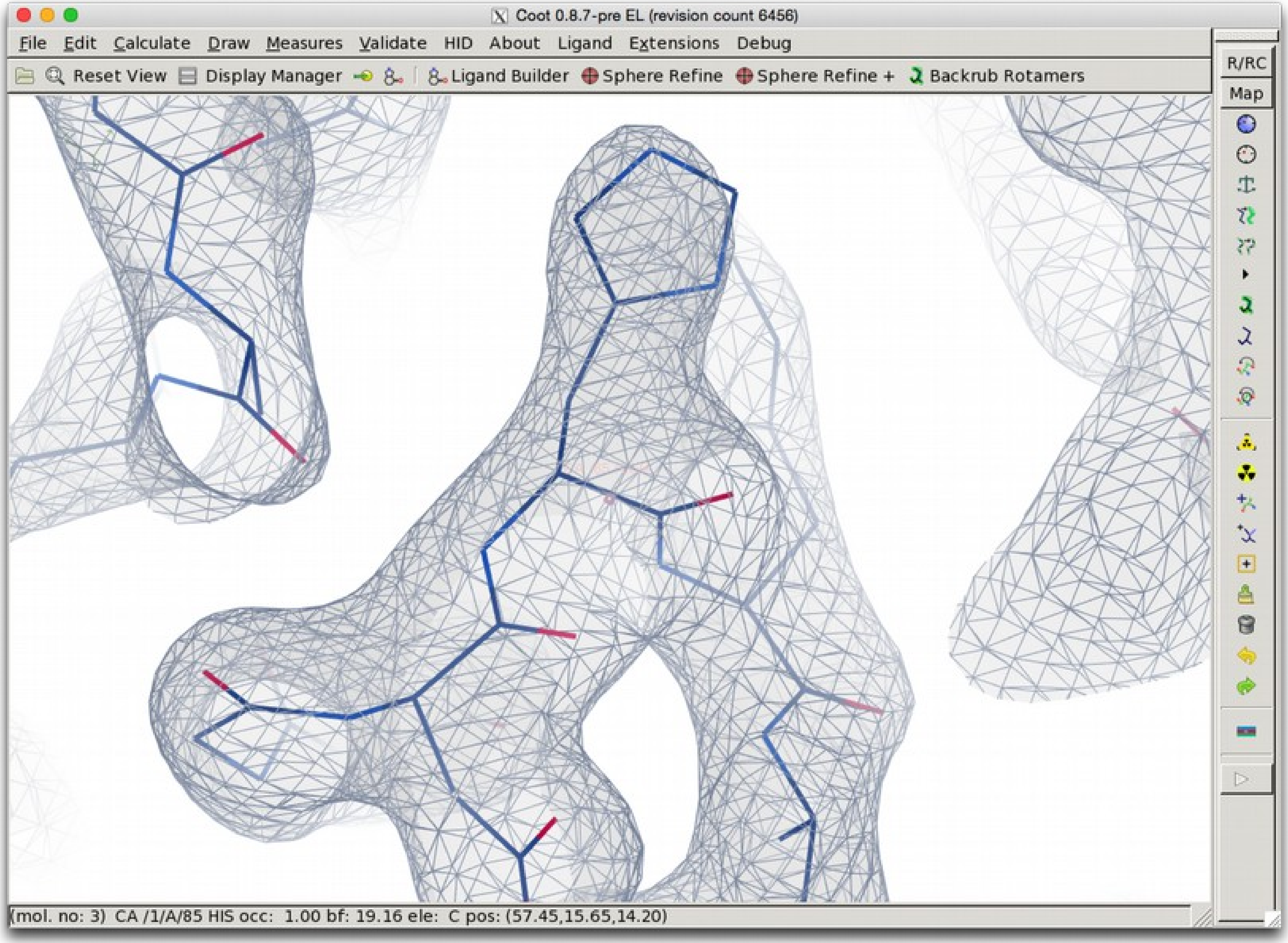
Davis et al. (2006) Structure

New Low Resolution Rotamer Search

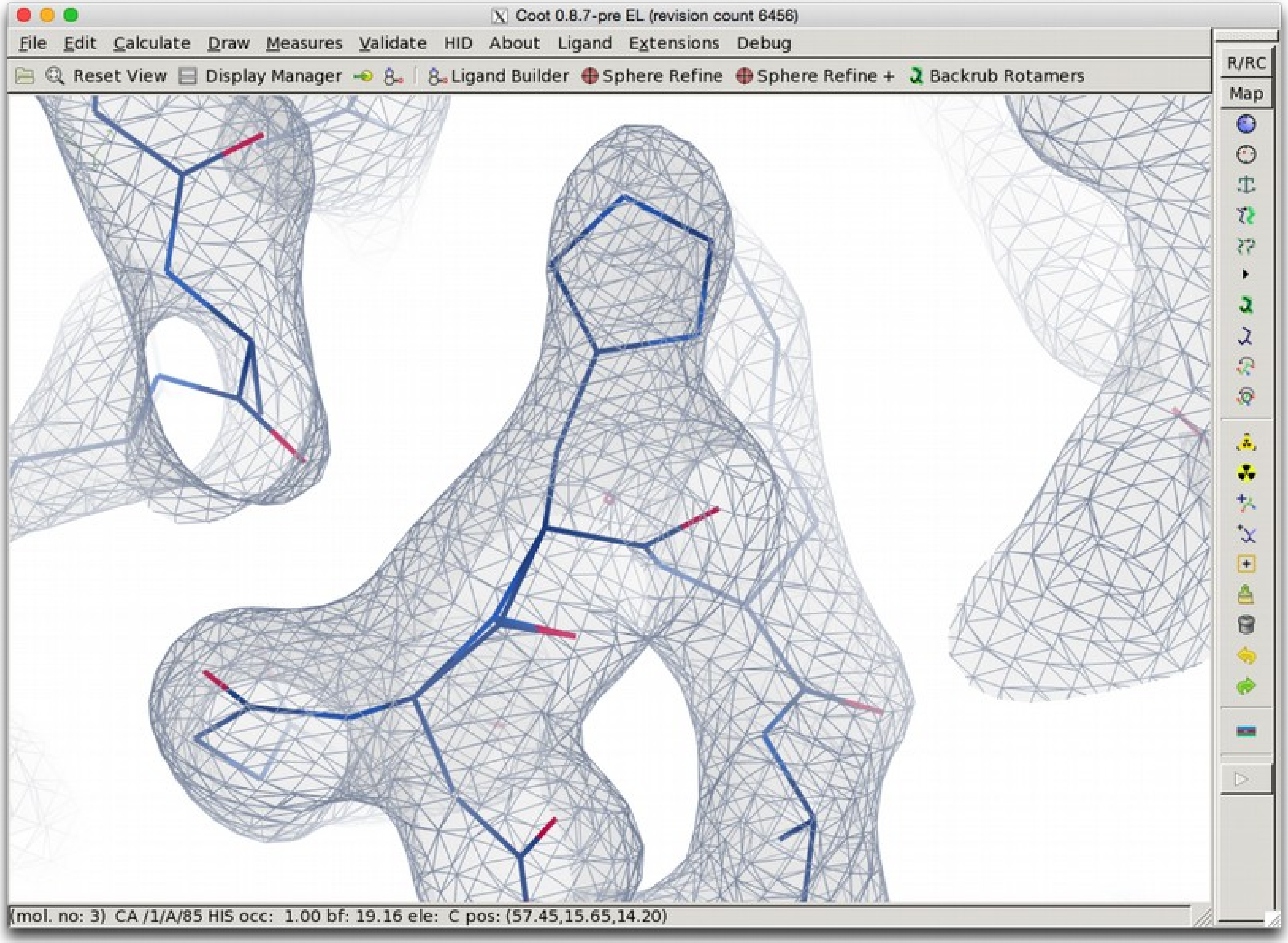


After Fitting Tools in KING/Molprobity

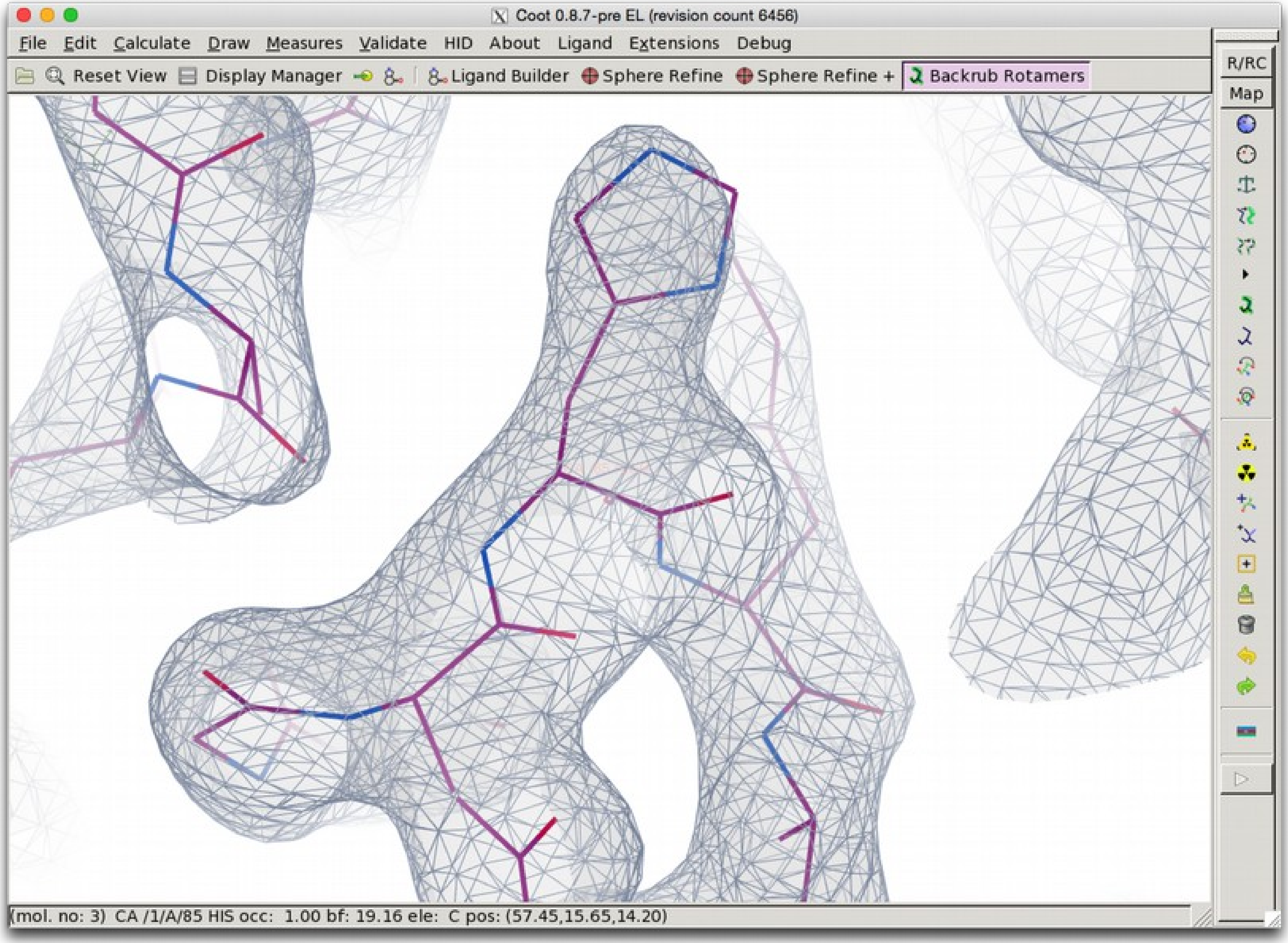

16
14
12
10
8
6
4
2
0
-2
-4
"d_score.tab"
Rotamer 1
Rotamer 2
Rotamer 3
Rotamer 4
-30
30 -30
30 -30
30 -30
30



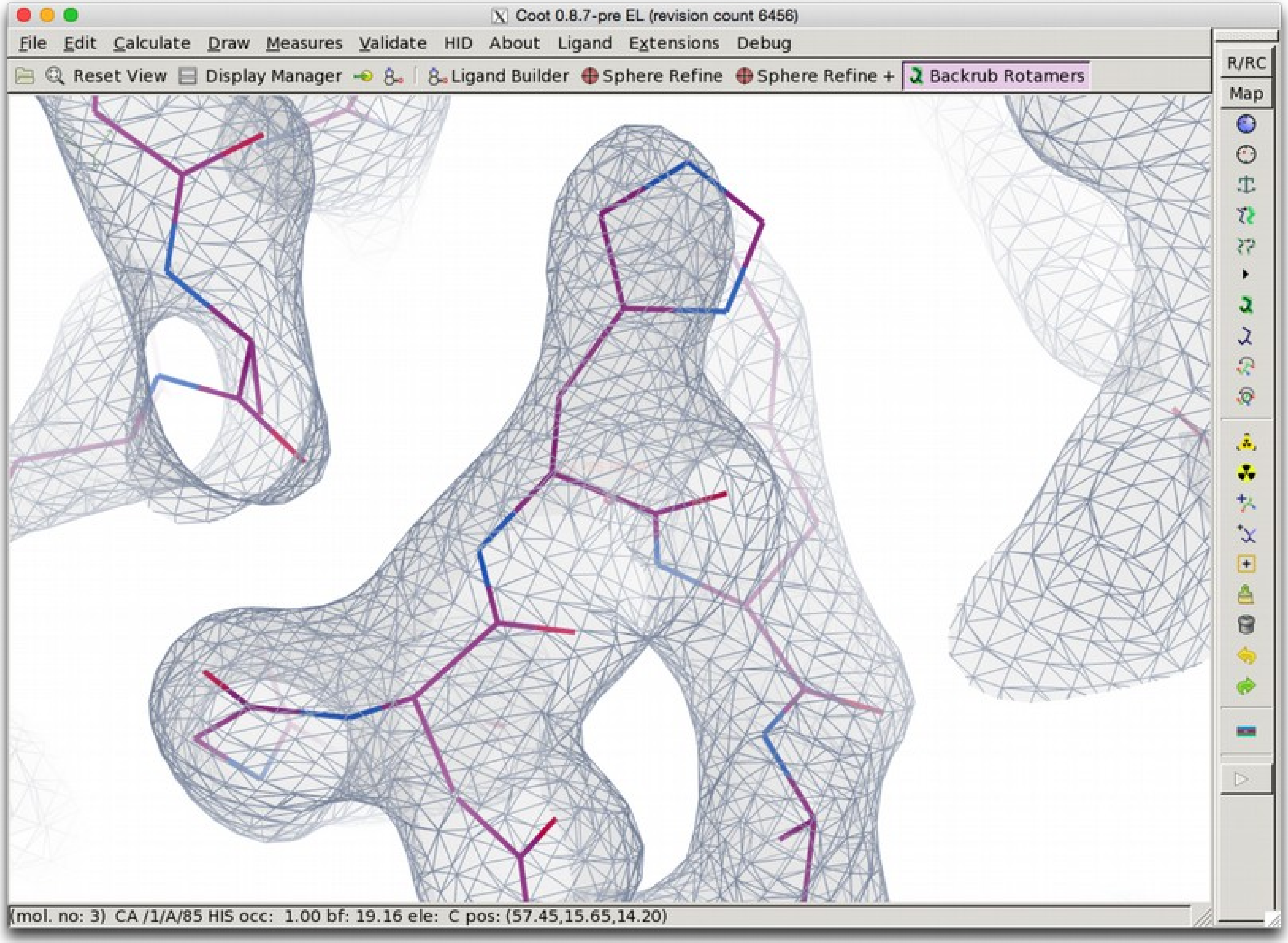
```
(mol. no: 3)  CA /1/A/85 HIS occ:  1.00 bf: 19.16 ele:  C pos: (57.45,15.65,14.20)
```



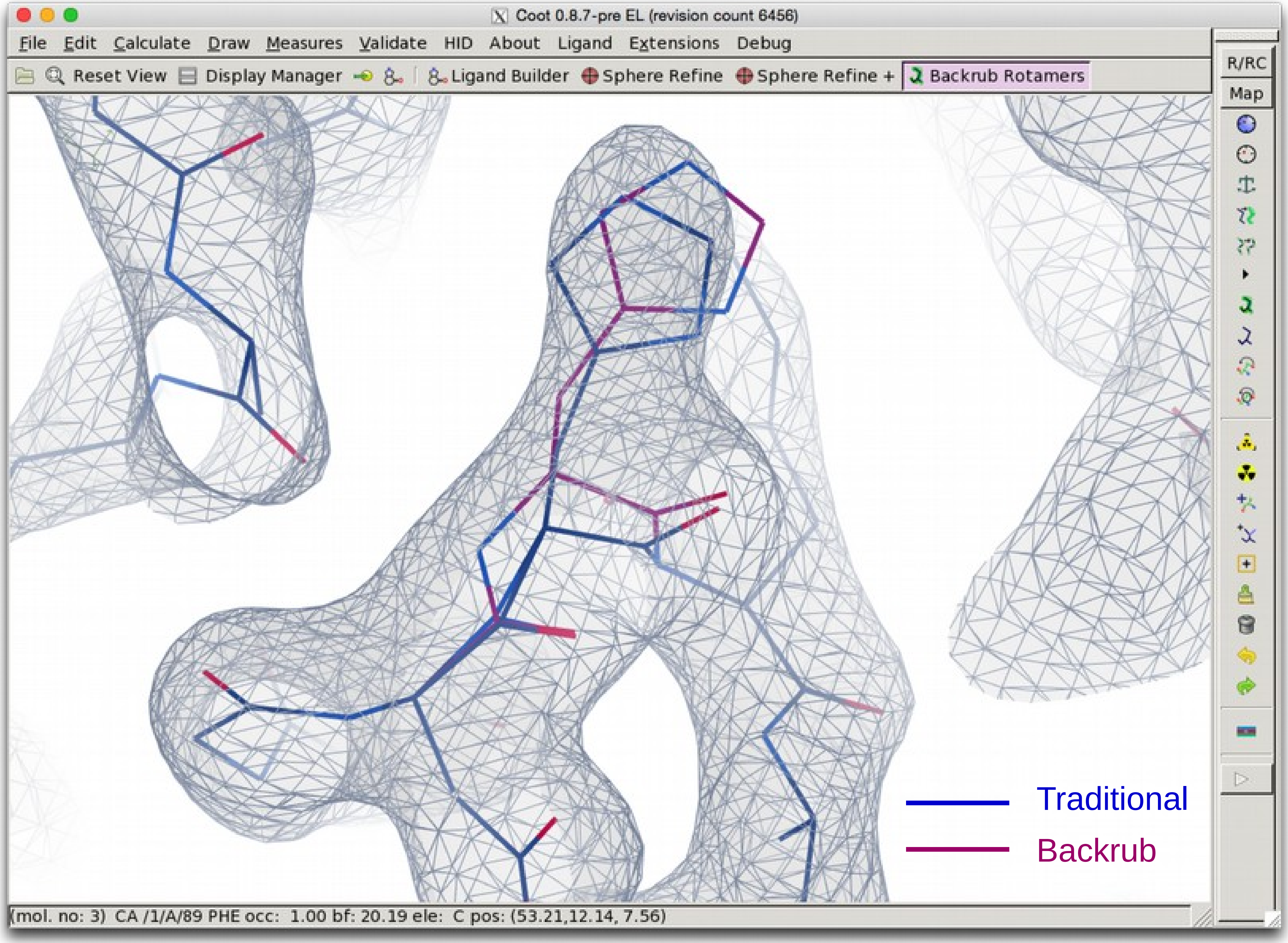
```
(mol. no: 3)  CA /1/A/85 HIS occ:  1.00 bf: 19.16 ele:  C pos: (57.45,15.65,14.20)
```



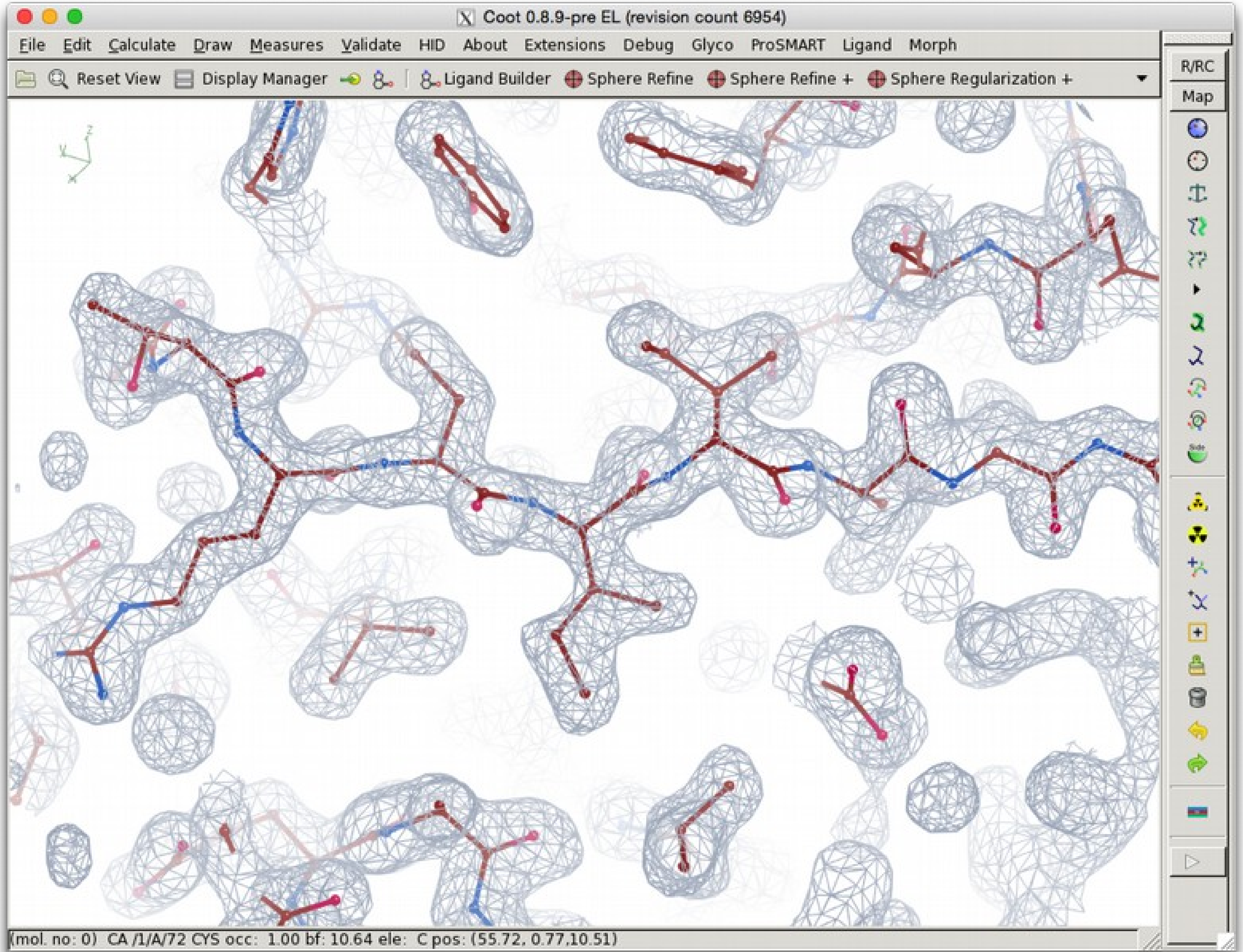
(mol. no: 3) CA /1/A/85 HIS occ: 1.00 bf: 19.16 ele: C pos: (57.45,15.65,14.20)



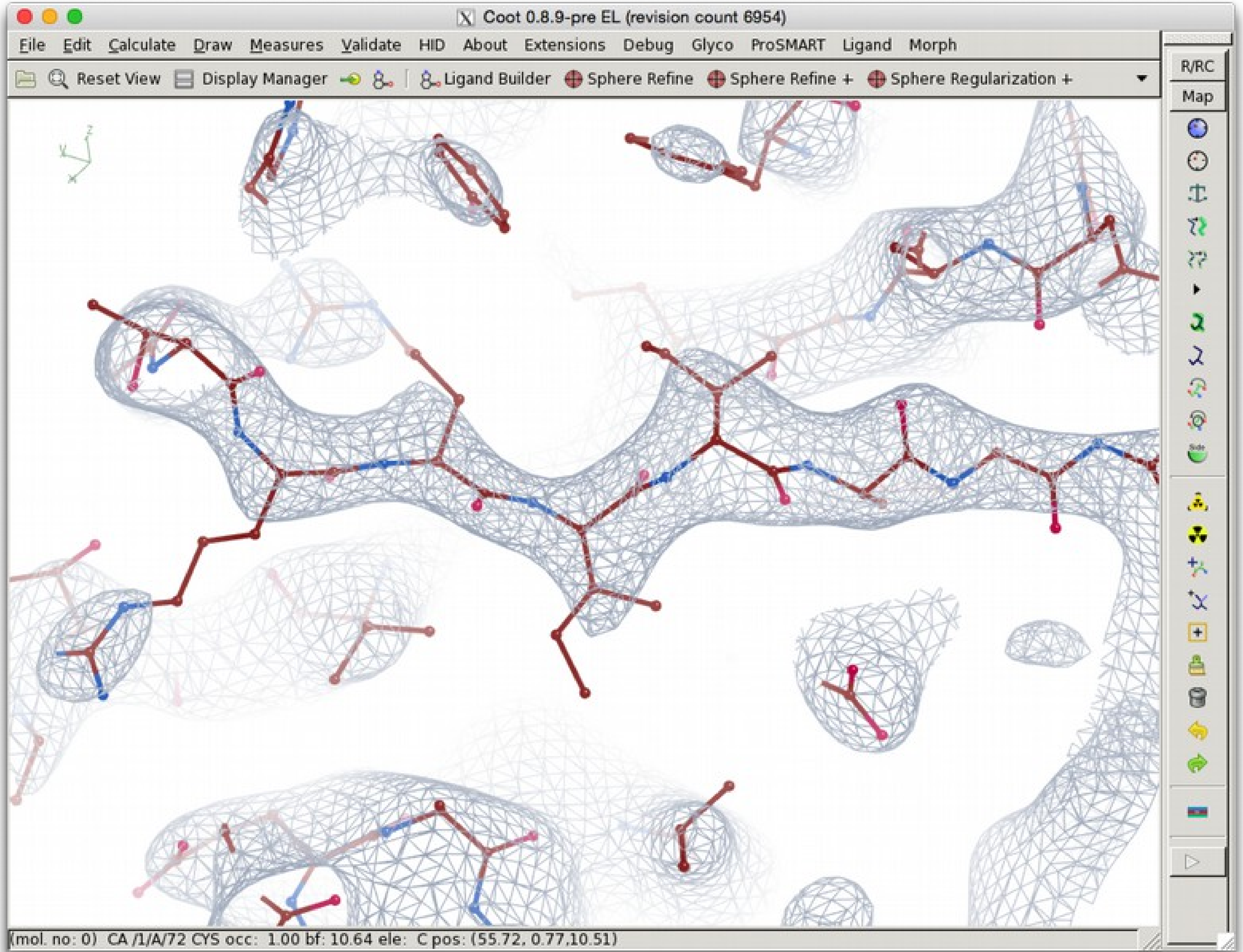
```
(mol. no: 3)  CA /1/A/85 HIS occ:  1.00 bf: 19.16 ele:  C pos: (57.45,15.65,14.20)
```



```
(mol. no: 3)  CA /1/A/89 PHE occ:  1.00 bf: 20.19 ele:  C pos: (53.21,12.14, 7.56)
```



```
(mol. no: 0)  CA /1/A/72 CYS occ:  1.00 bf: 10.64 ele:  C pos: (55.72, 0.77,10.51)
```



```
(mol. no: 0)  CA /1/A/72 CYS occ:  1.00 bf: 10.64 ele:  C pos: (55.72, 0.77,10.51)
```

Additional Reading

- “Atomic Resolution” a badly abused term in structural biology
  - Wlodawer & Dauter (2017)
- Responses are perhaps more interesting that the article itself

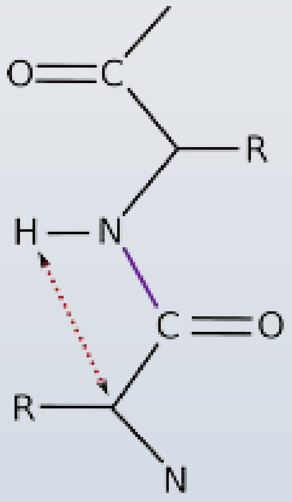
cis-Peptides

- What is a cis-peptide?
- Peptide restraints in Coot 2004-2015

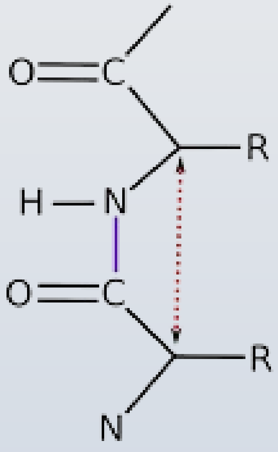
cis-Peptides

- A number of paper have been published recently highlighting the unusually large number of cis-peptides in some structures:
    - Croll: The rate of cis-trans conformation errors is increasing in low-resolution crystal structures *Acta Cryst.* (2015). **D**71, 706-709
    - Touw *et al.*: Detection of trans–cis flips and peptide-plane flips in protein structures *Acta Cryst.* (2015). **D**71, 1604-71614

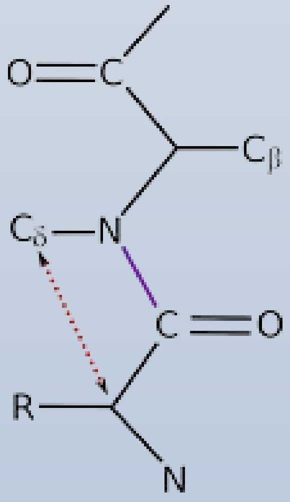
cis-Peptides



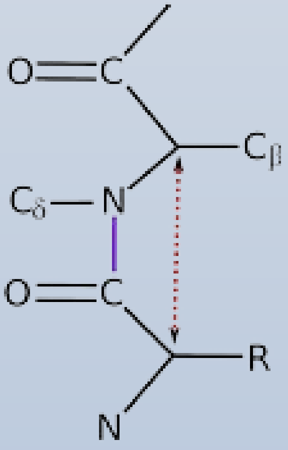
trans-peptide



cis-peptide

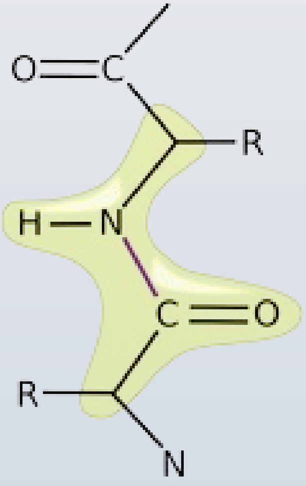


PRO *trans-peptide*

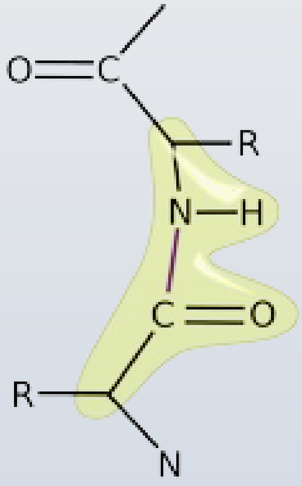


PRO *cis-peptide*

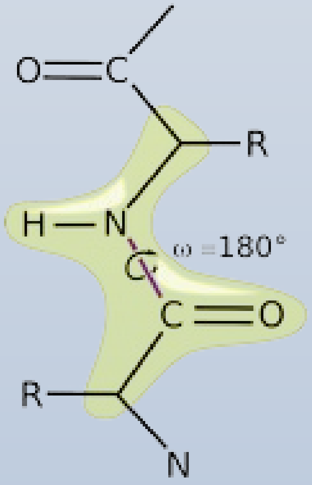
cis-Peptides



trans-peptide
with plane restraints

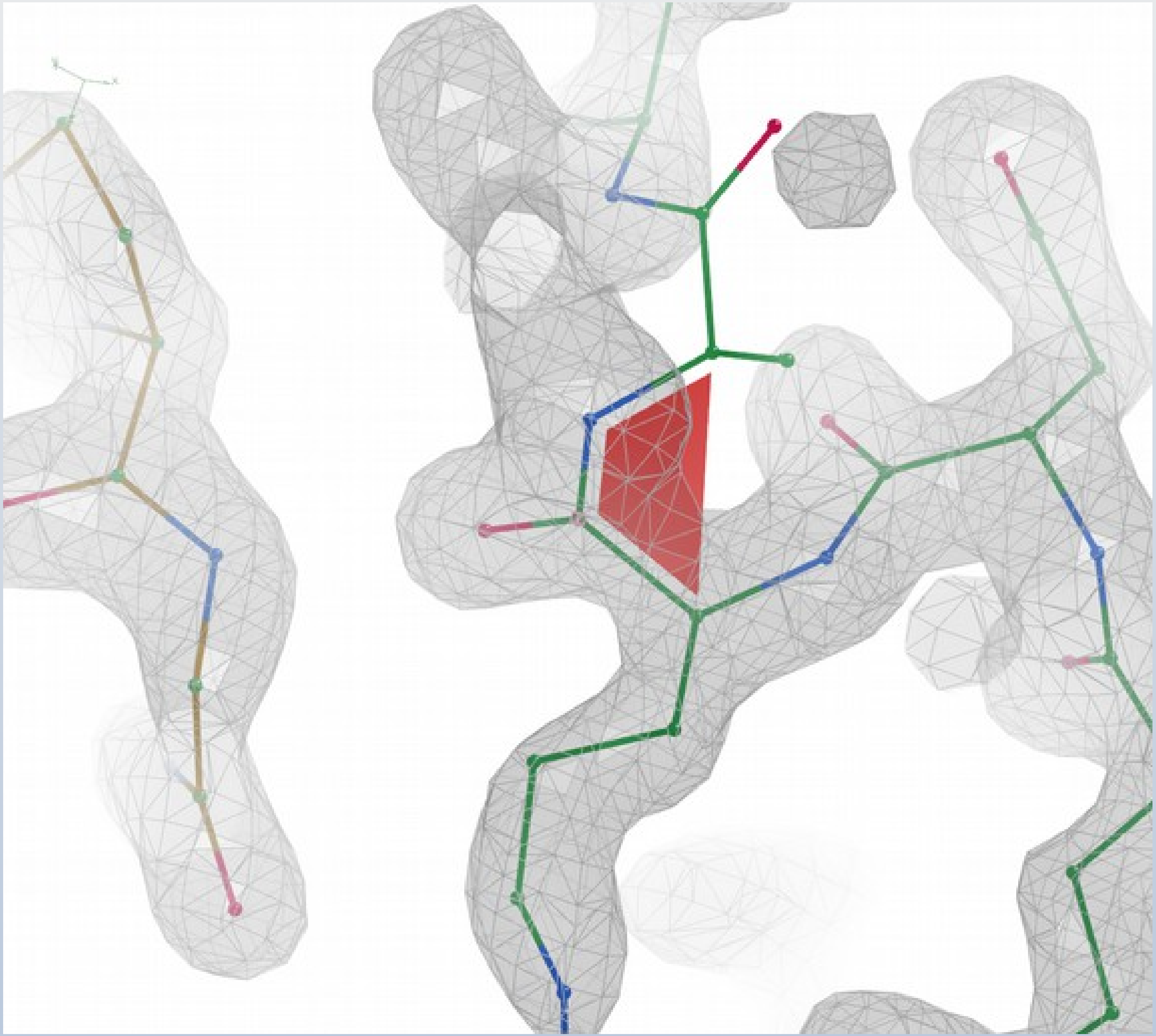


cis-peptide
with plane restraints



trans-peptide
with plane and trans restraints

cis-peptide Representation

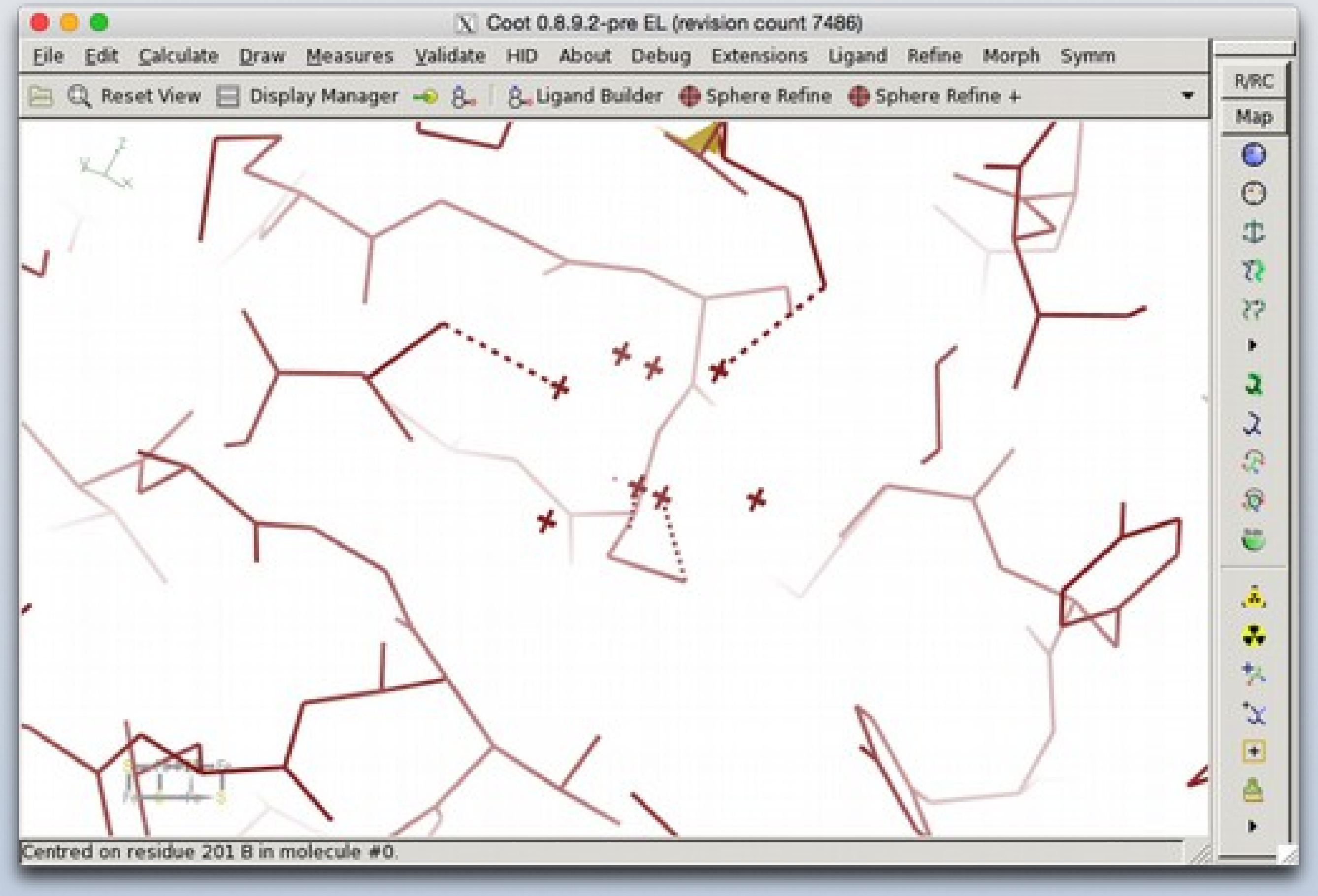


Pre-PRO 🟩

Twisted-trans 🟨

Non-pre-PRO 🟥

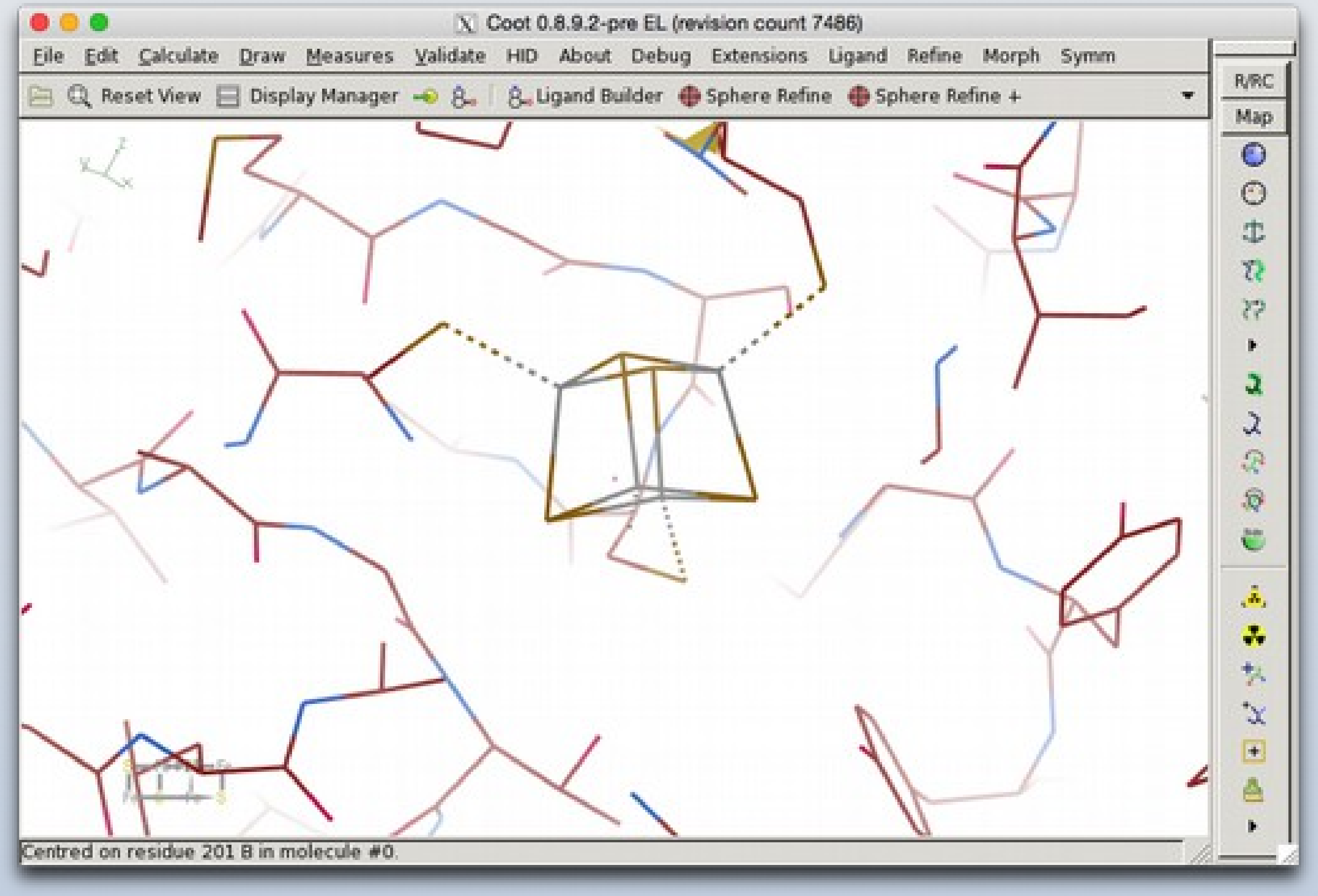
Iron-Sulfur Complexes



Centred on residue 201 B in molecule #0.

2017

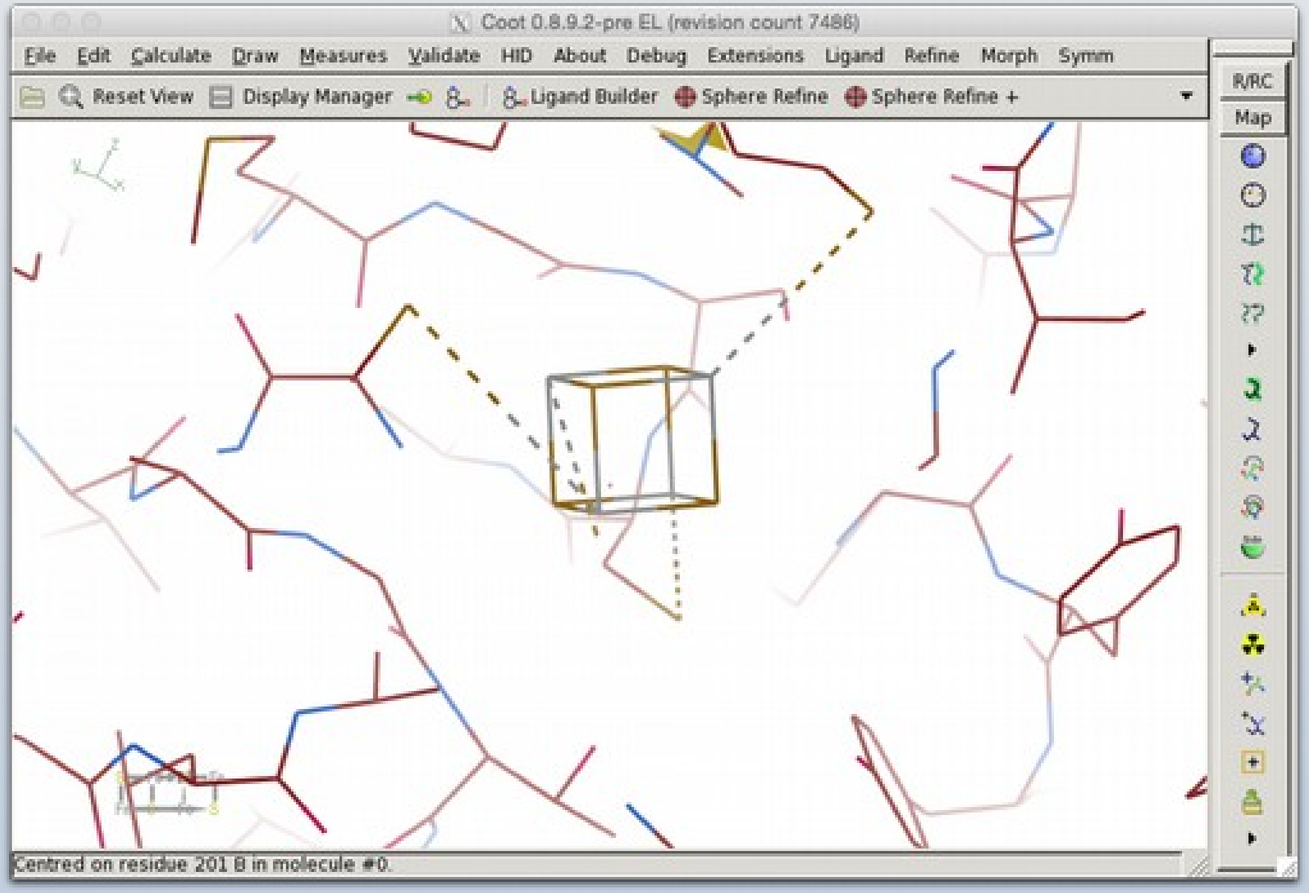
Iron-Sulfur Complexes



Centred on residue 201 B in molecule #0.

2018

Iron-Sulfur Complexes

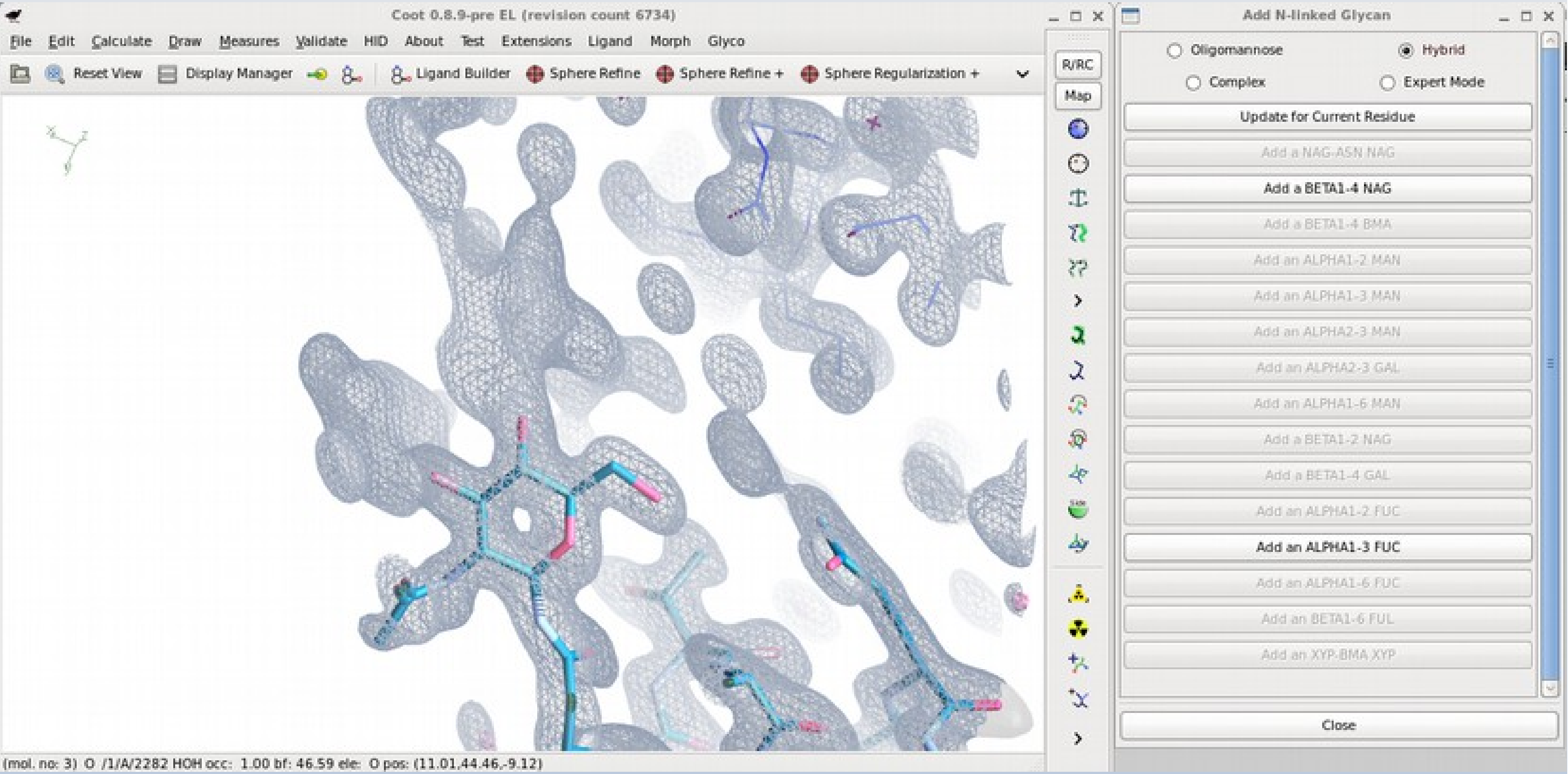


Centred on residue 201 B in molecule #0.

2018

N-linked Carbohydrates

- Improved algorithm and re-worked GUI

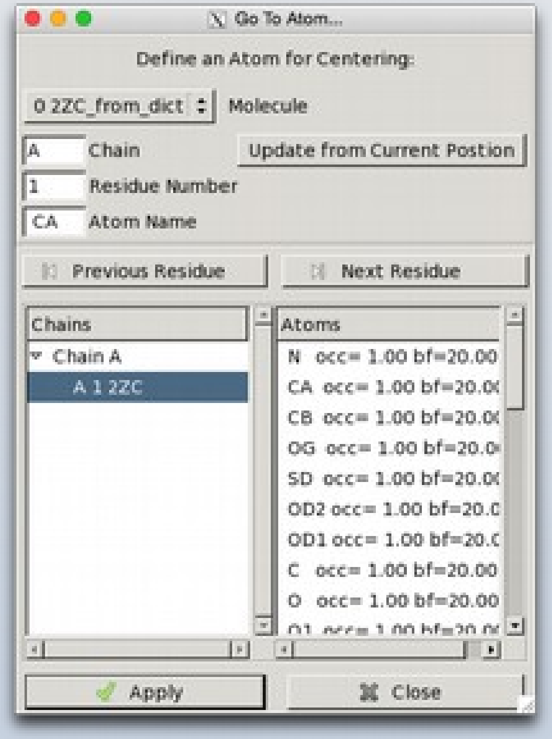


○ Oligomannose ● Hybrid

○ Complex ○ Expert Mode

(mol. no: 3) O /1/A/2282 HOH occ: 1.00 bf: 46.59 ele: O pos: (11.01,44.46,-9.12)

IISTDTIDIW



Molecule: 0 2ZC_from_dict

Chain: A

Residue Number: 1

Atom Name: CA

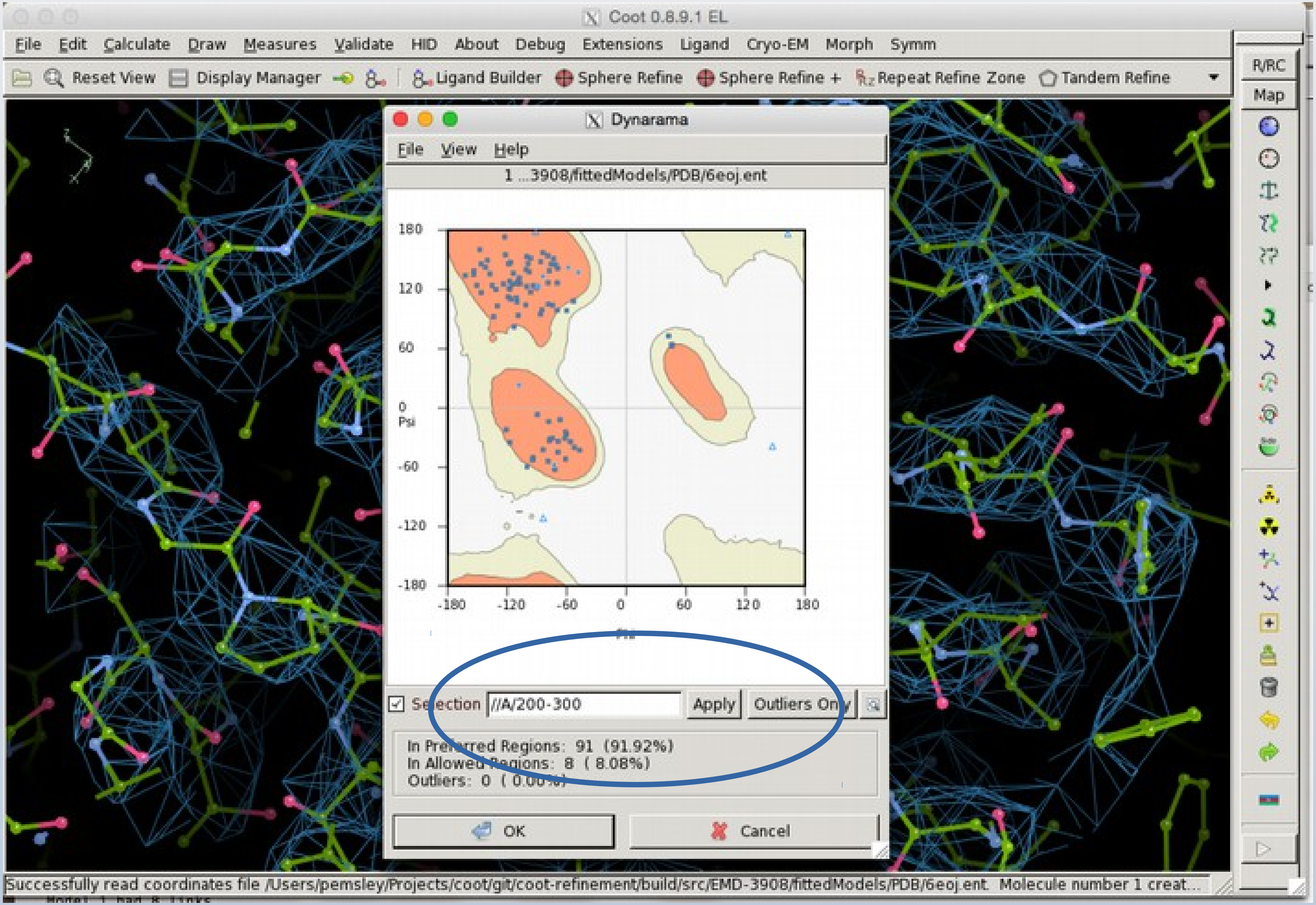
Chain A
A 1 2ZC

```
N   occ= 1.00 bf=20.00
CA  occ= 1.00 bf=20.0
CB  occ= 1.00 bf=20.0
OG  occ= 1.00 bf=20.0
SD  occ= 1.00 bf=20.0
OD2 occ= 1.00 bf=20.0
OD1 occ= 1.00 bf=20.0
C   occ= 1.00 bf=20.00
O   occ= 1.00 bf=20.00
```

- If I See This Dialog Then I'm Doing It Wrong

New/Bleeding Edge/Still-in-Development

The New Ramachandran Plot

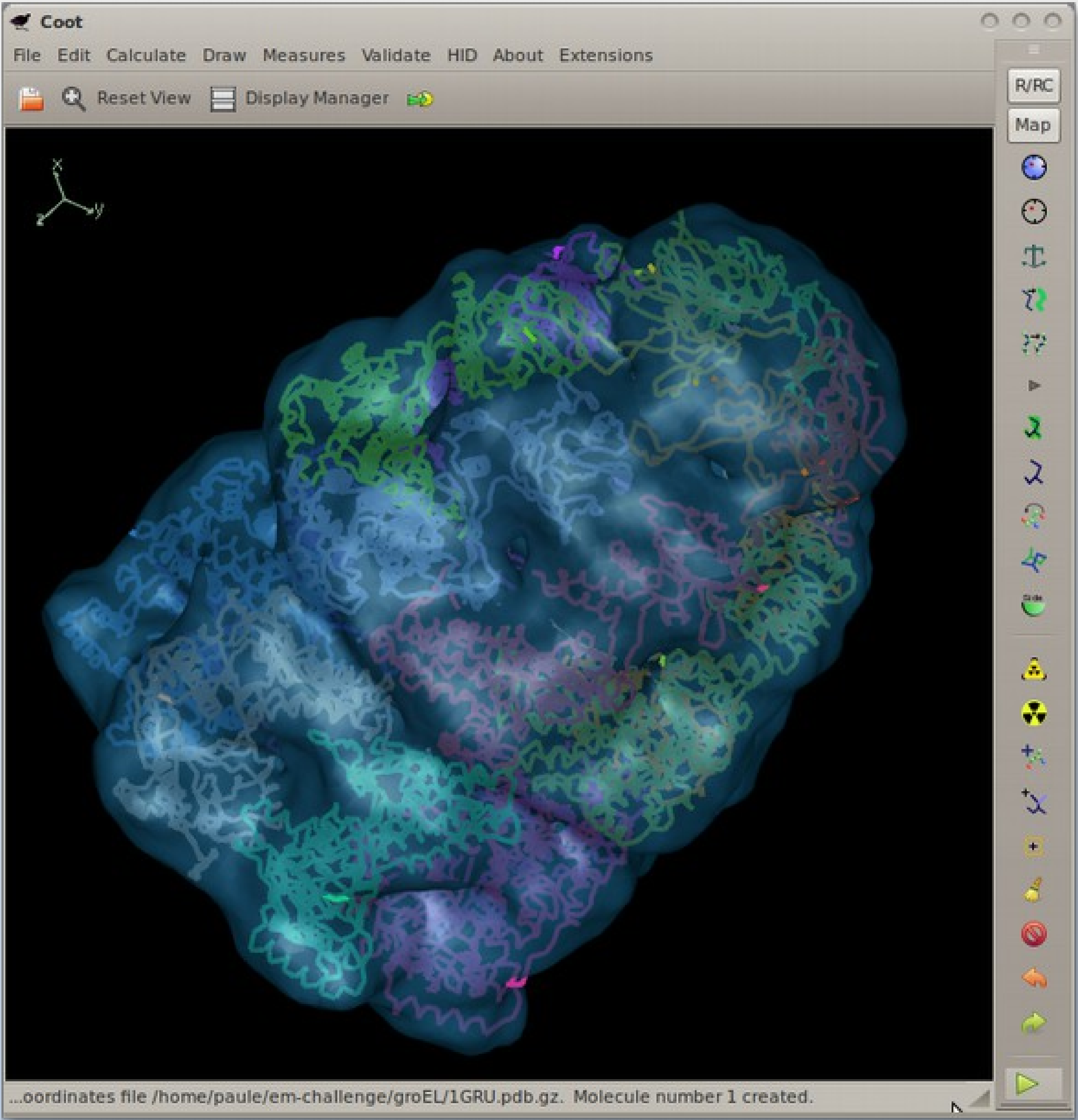


1 ...3908/fittedModels/PDB/6eoj.ent

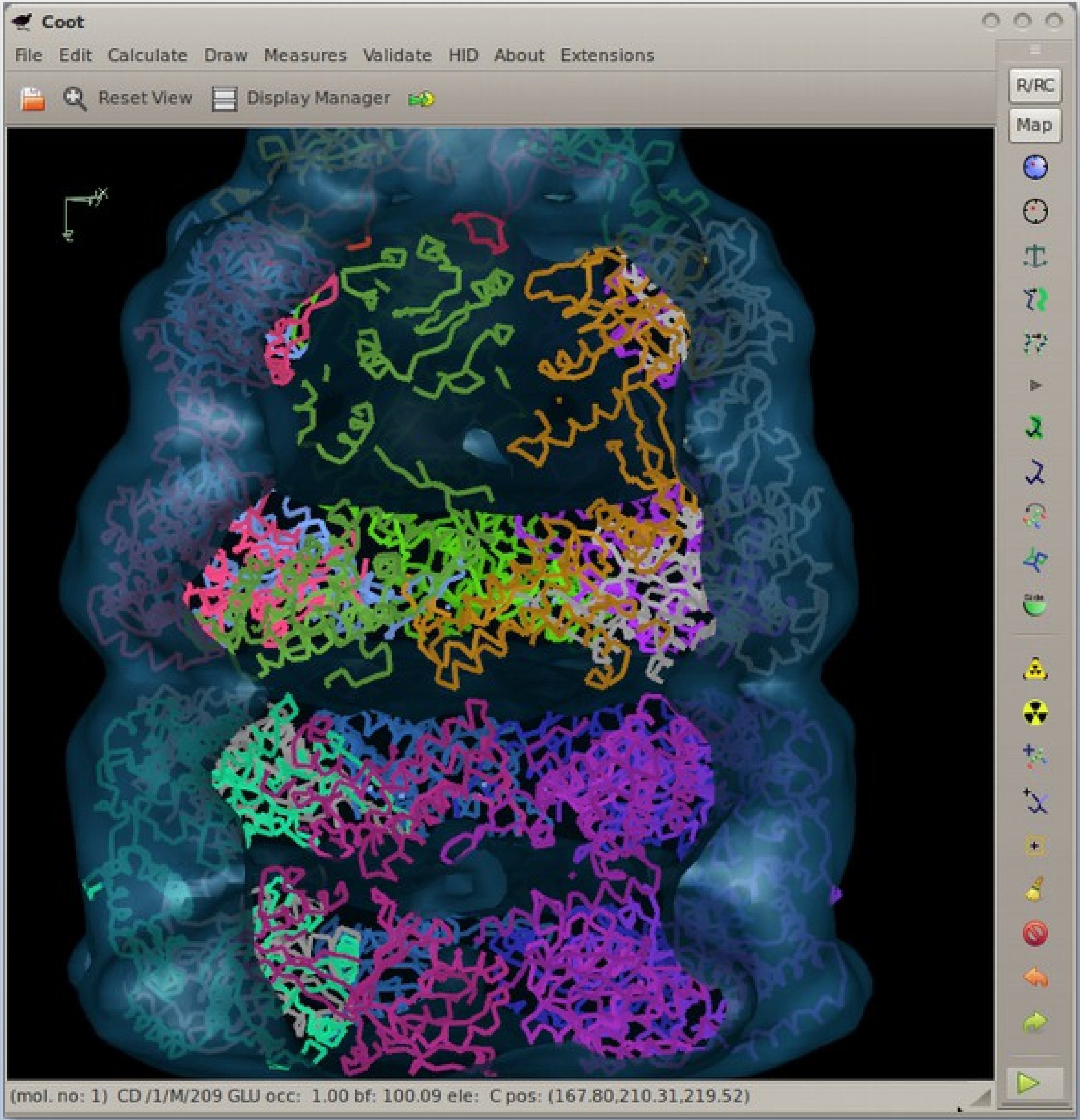
☑ Selection //A/200-300

In Preferred Regions: 91 (91.92%)
In Allowed Regions: 8 (8.08%)
Outliers: 0 (0.00%)

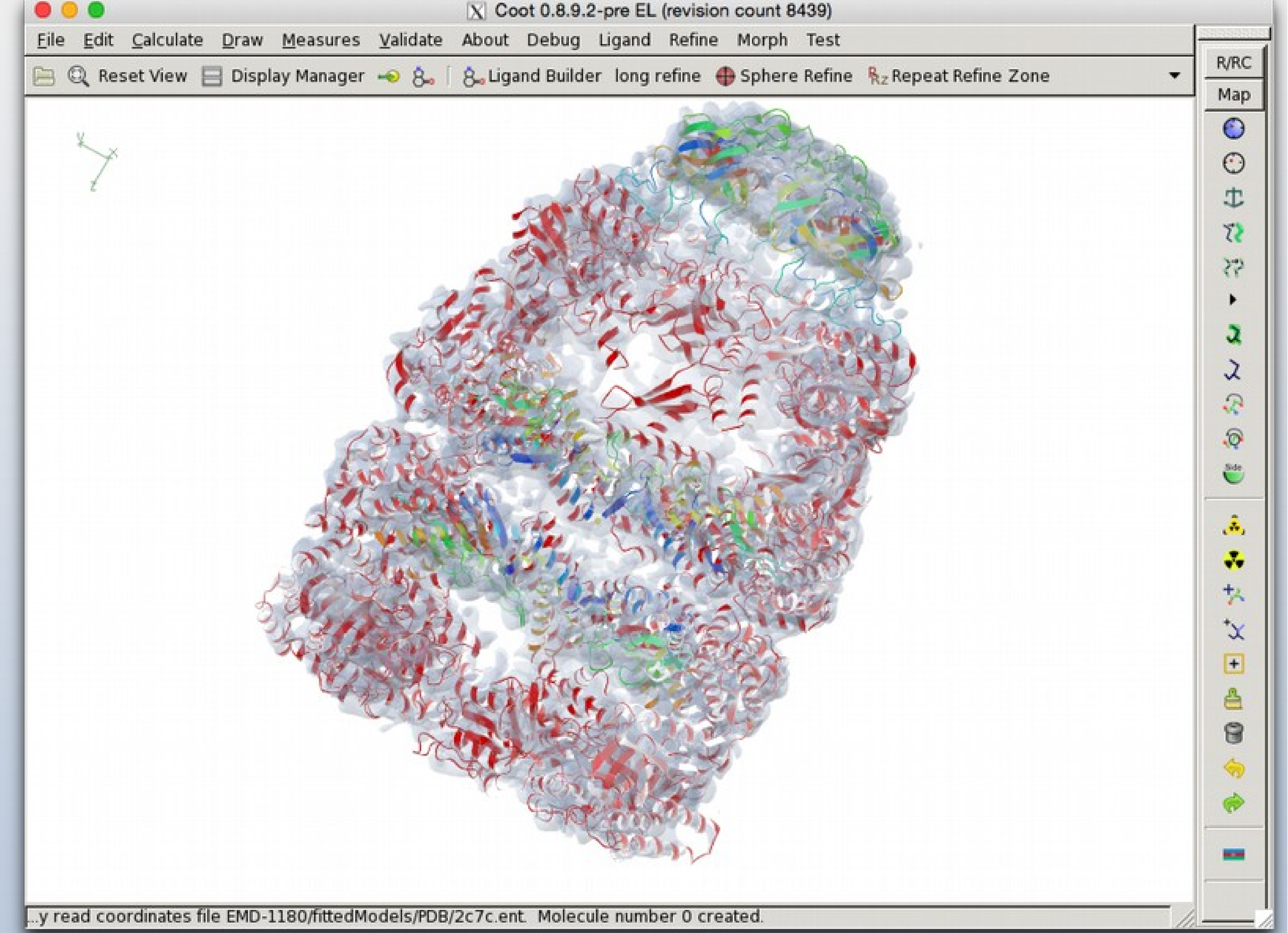
Successfully read coordinates file /Users/pemsley/Projects/coot/git/coot-refinement/build/src/EMD-3908/fittedModels/PDB/6eoj.ent. Molecule number 1 creat...



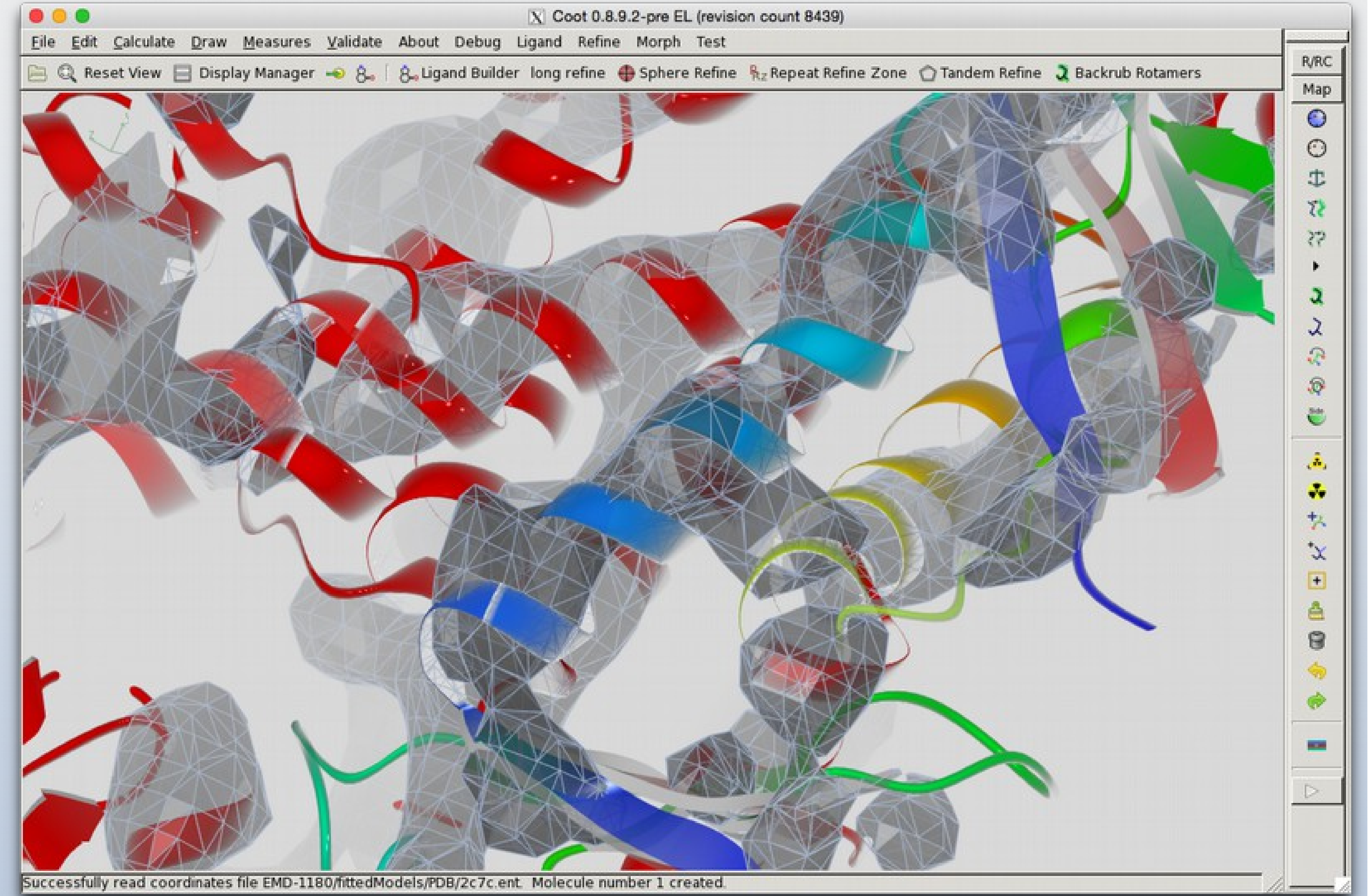
...oordinates file /home/paule/em-challenge/groEL/1GRU.pdb.gz. Molecule number 1 created.



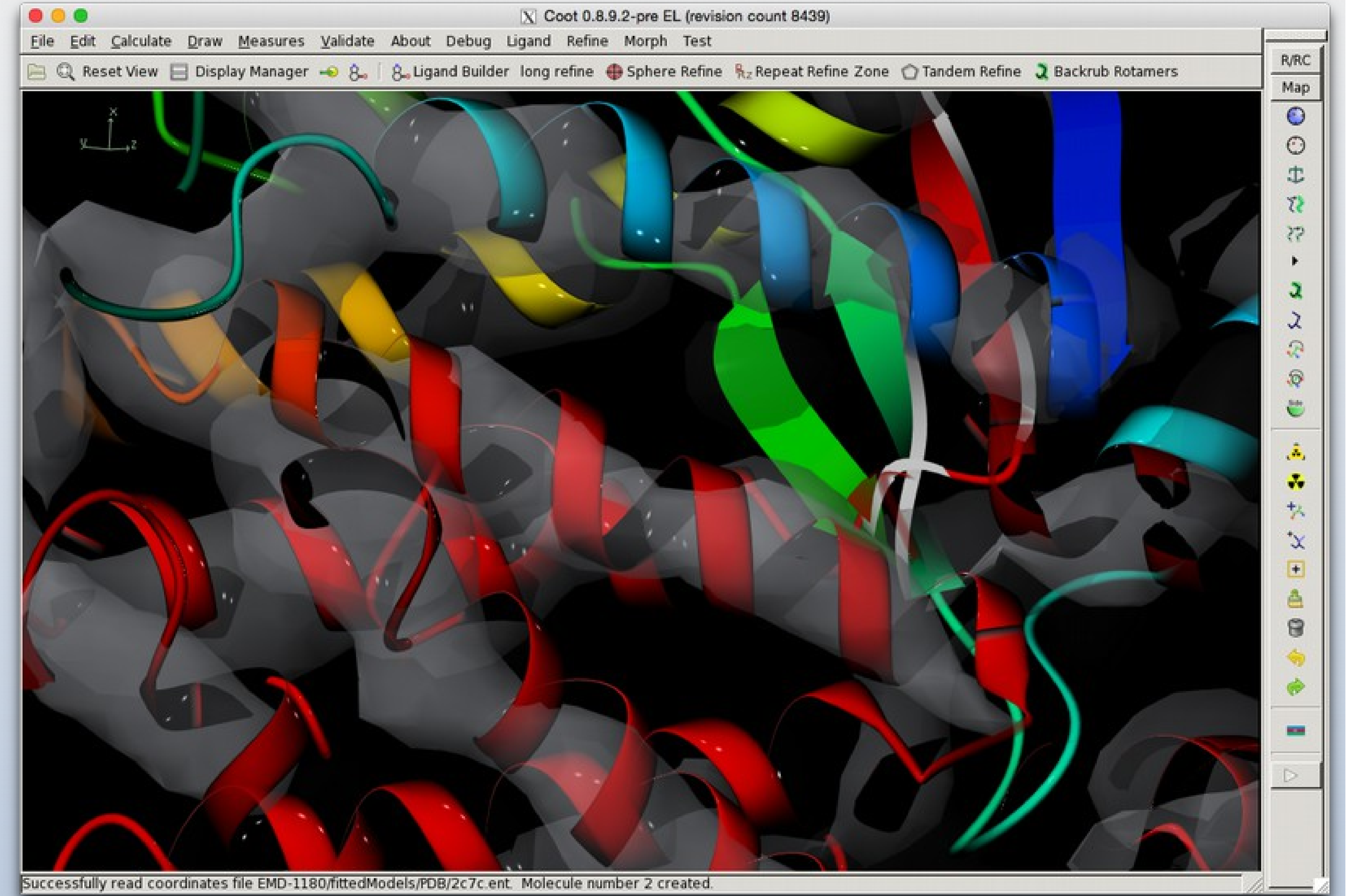
```
(mol. no: 1)  CD /1/M/209 GLU occ:  1.00 bf: 100.09 ele:  C pos: (167.80,210.31,219.52)
```



```
..y read coordinates file EMD-1180/fittedModels/PDB/2c7c.ent.  Molecule number 0 created.
```

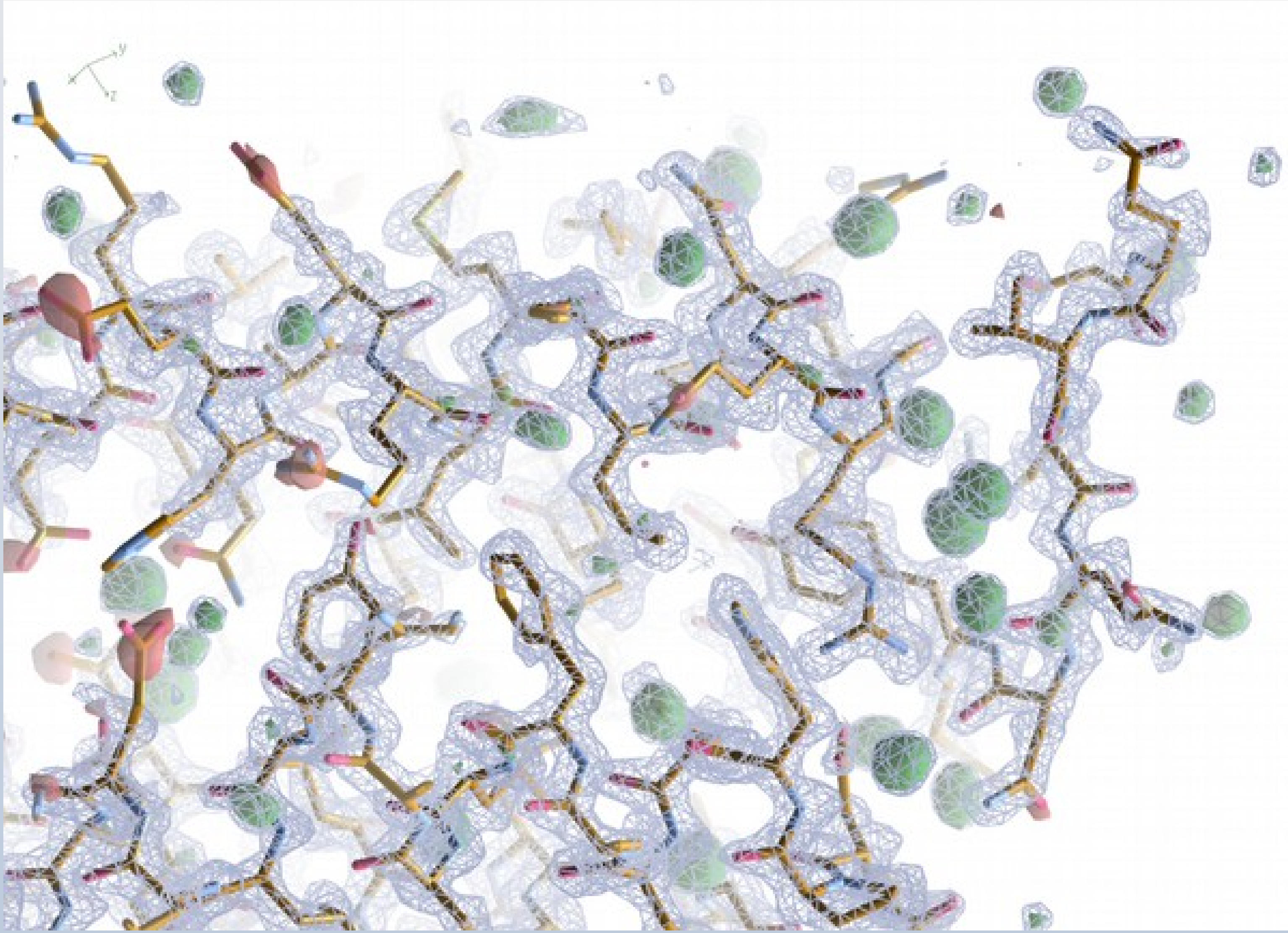


Successfully read coordinates file EMD-1180/fittedModels/PDB/2c7c.ent. Molecule number 1 created.

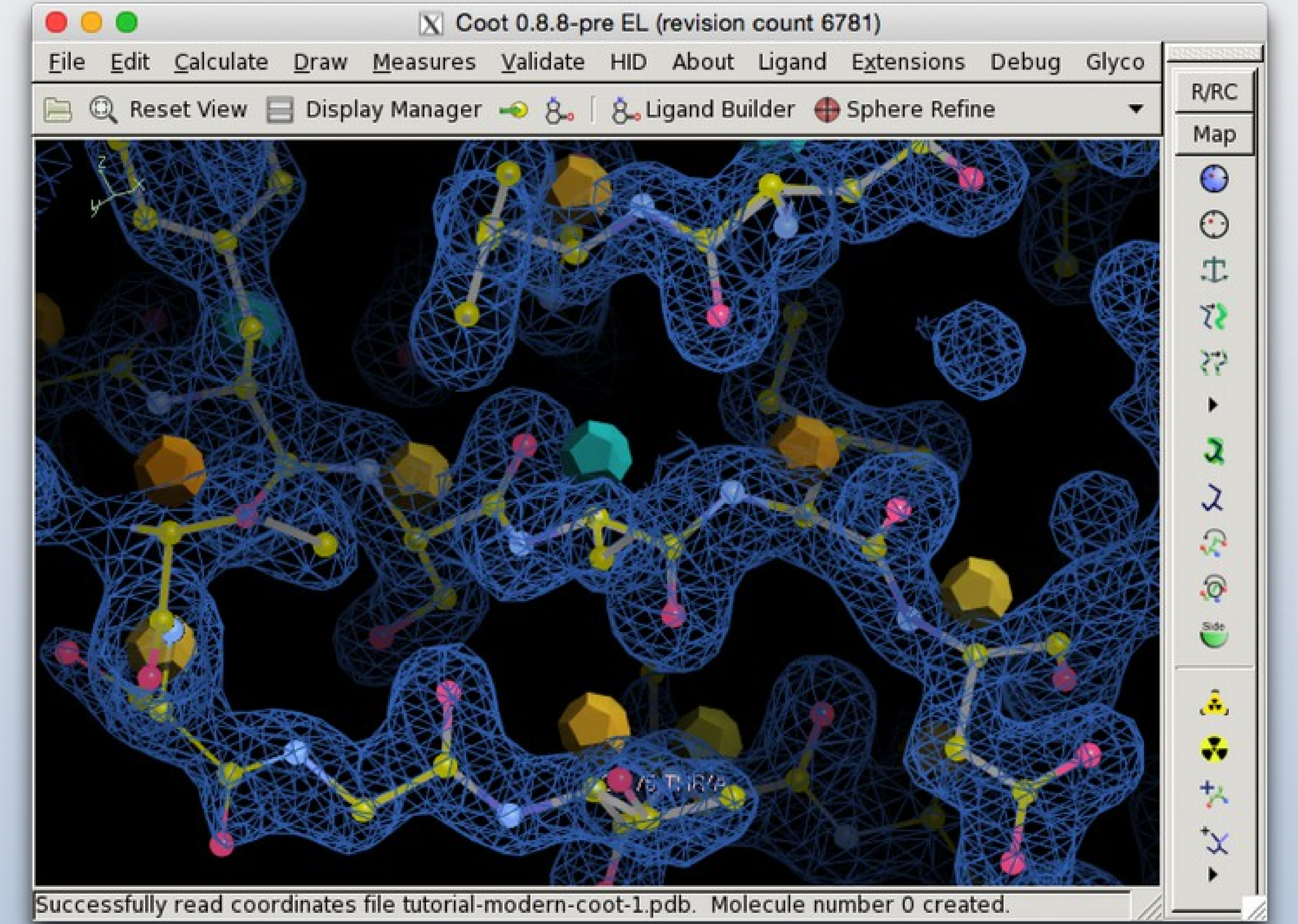


Successfully read coordinates file EMD-1180/fittedModels/PDB/2c7c.ent. Molecule number 2 created.

EMD-0144:
1.65 Å Human Apoferritin Relion-3

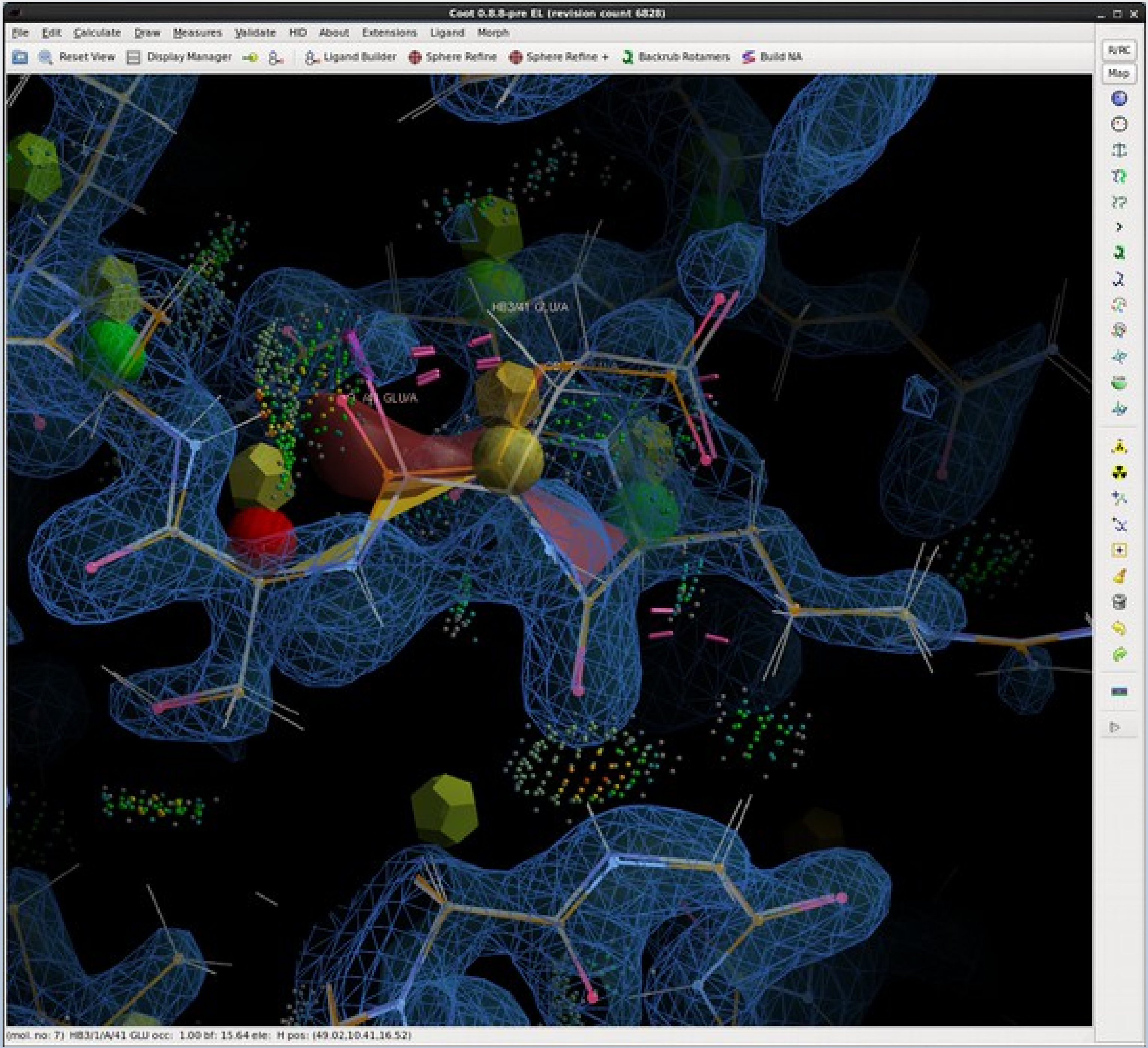


Interactive Rotamer Goodness



Successfully read coordinates file tutorial-modern-coot-1.pdb. Molecule number 0 created.

Multi-Criteria Markup



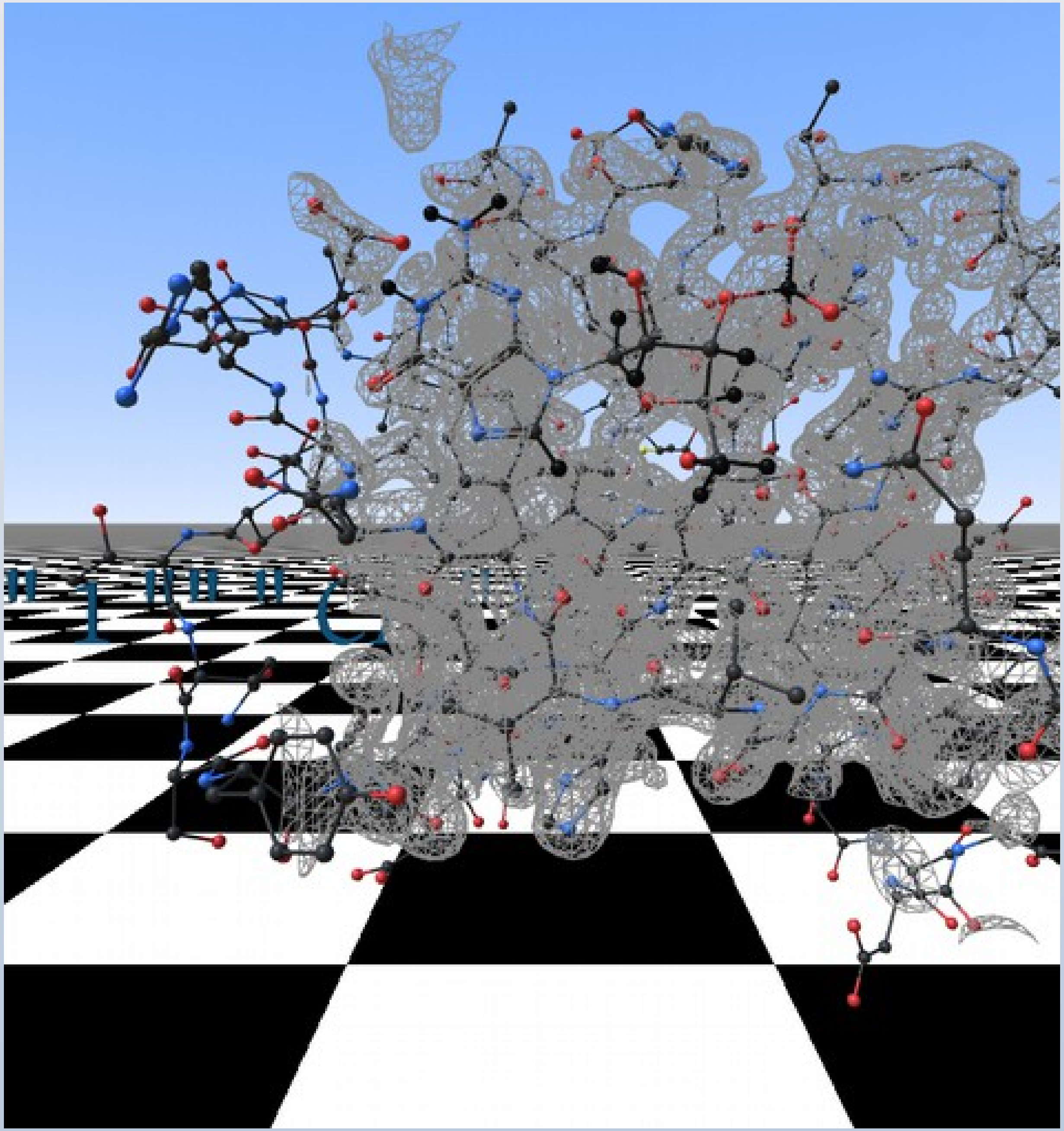
```
(mol. no: 7) HB3/1/A/41 GLU occ:  1.00 bf: 15.64 ele:  H pos: (49.02,10.41,16.52)
```

Rotamer and Ramachandran Markup

- `set_show_intermediate_atoms_rota_markup(1)`
- `set_show_intermediate_atoms_rama_markup(1)`

Coot Futures: Virtual Reality

Hamish Todd



- **An Intuitive Interface:**
- Stereoscopic Representation
- Greater Field of View
- 2 Hands with Articulation
- However:
  - current tools are not immediately transferable
  - because: nausea

Acknowledgements

- LMB:
  - Garib Murshudov, Rob Nicholls

- Kevin Cowan, Bernhard Lohkamp

- Libraries & Dictionaries:
  - Jane & Dave Richardson
  - Alexei Vagin
  - Eugene Krissinel