

Model Refinement and Validation

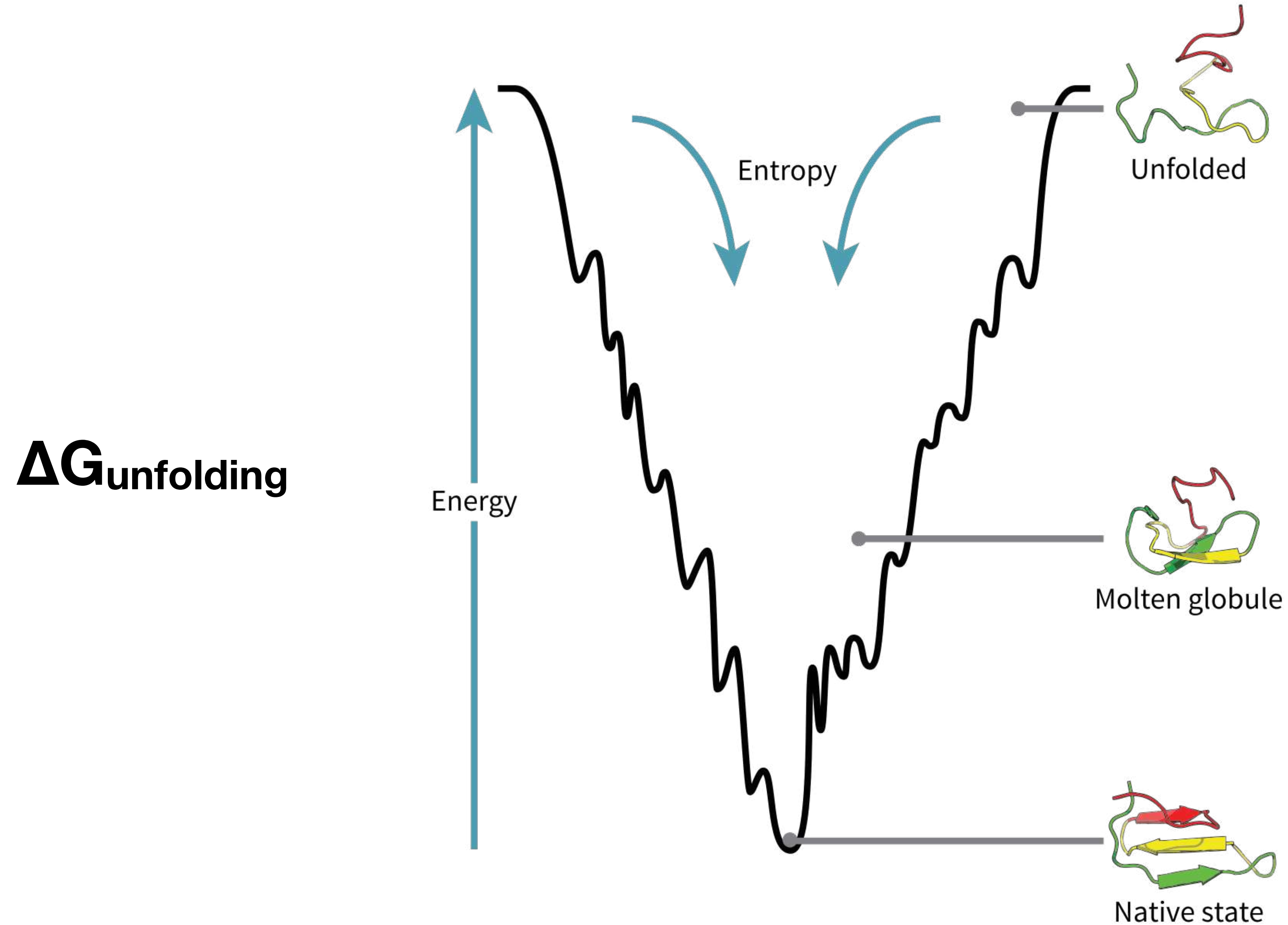
Damian Ekiert and Gira Bhabha
Johns Hopkins University

Goal of model refinement:

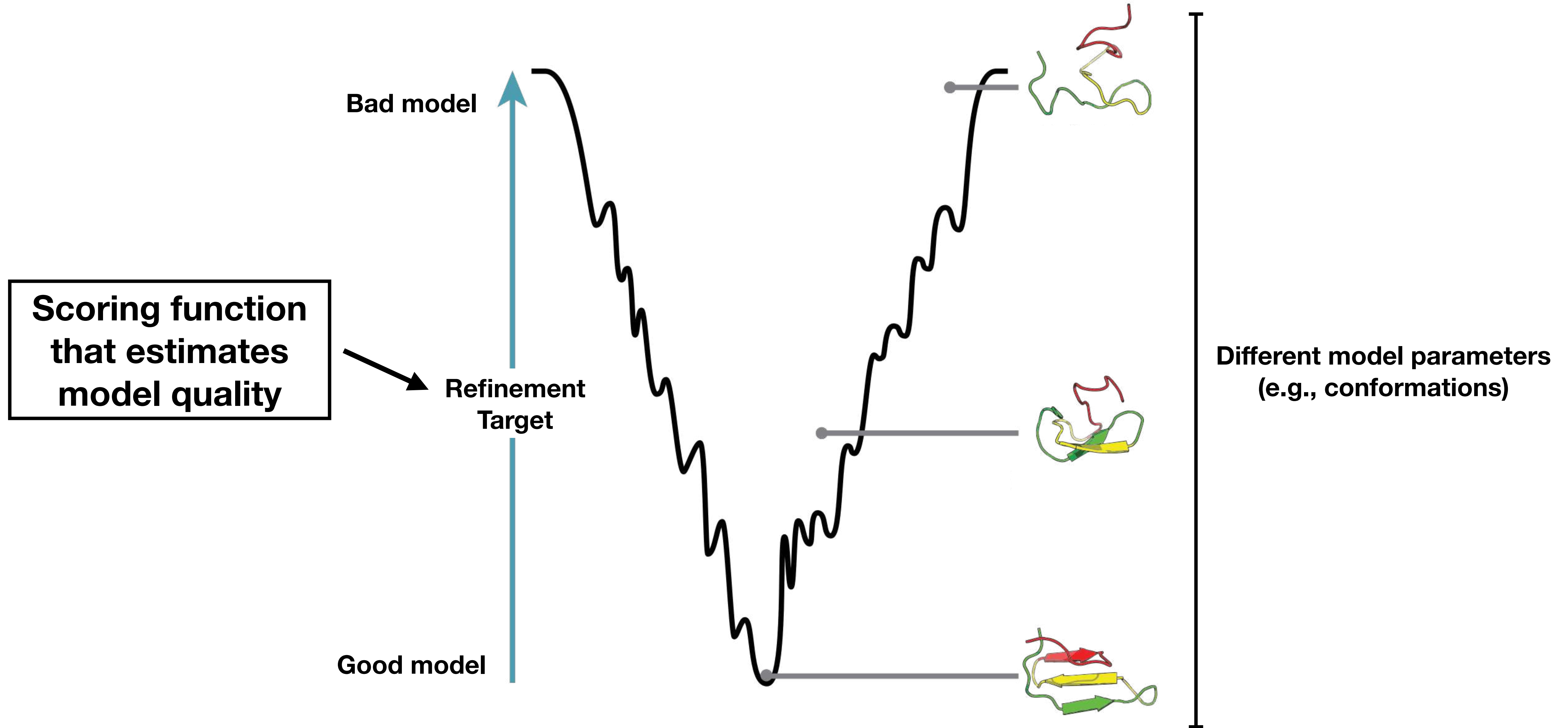
To create a set of coordinates that

- 1) explains the data as best we can, but
- 2) also conforms with what we know about proteins in general

Model Refinement vs Protein Folding Funnel



Refinement target describes differences between model and data

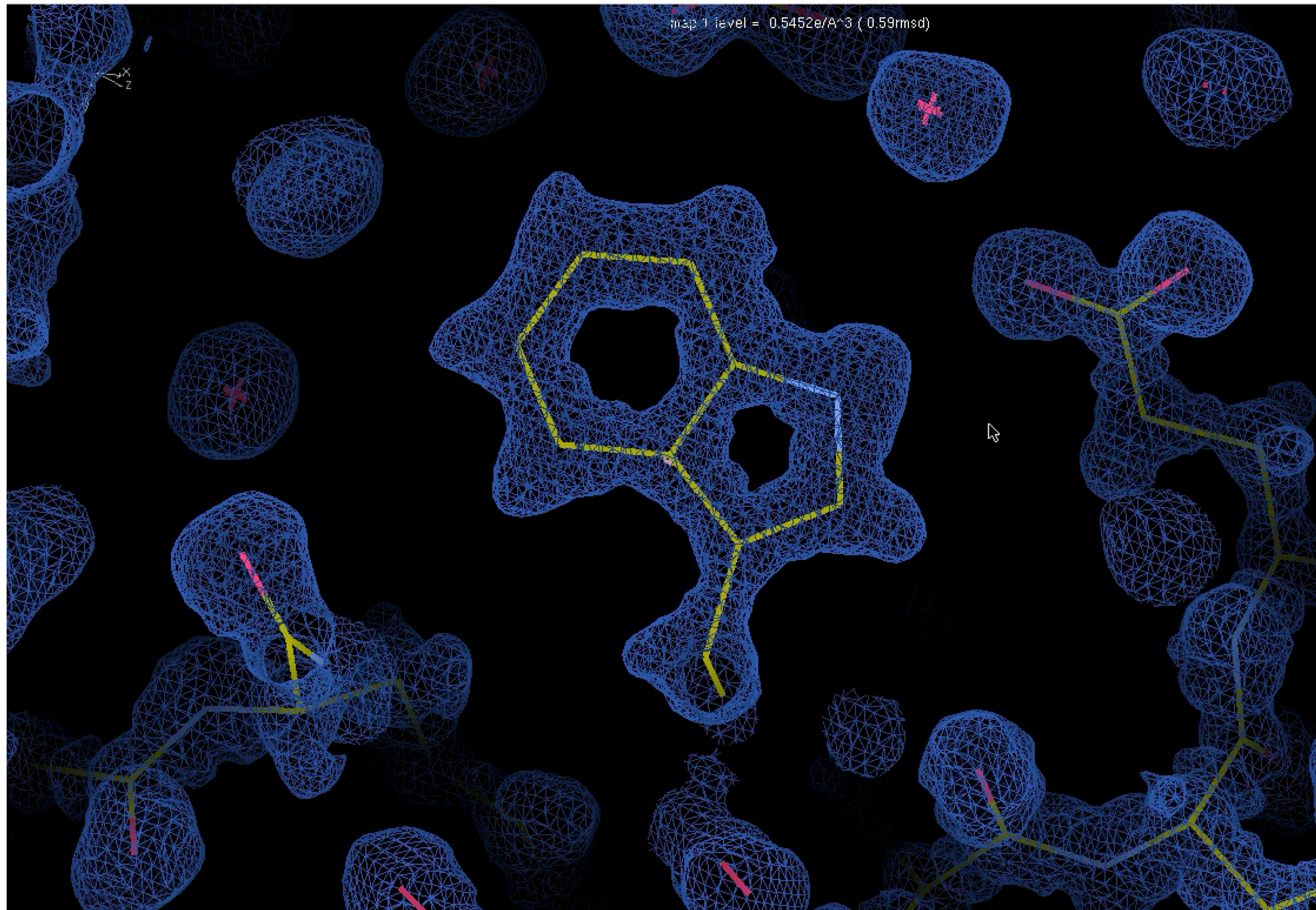


Simplest: Refinement Target = (Model vs Data)

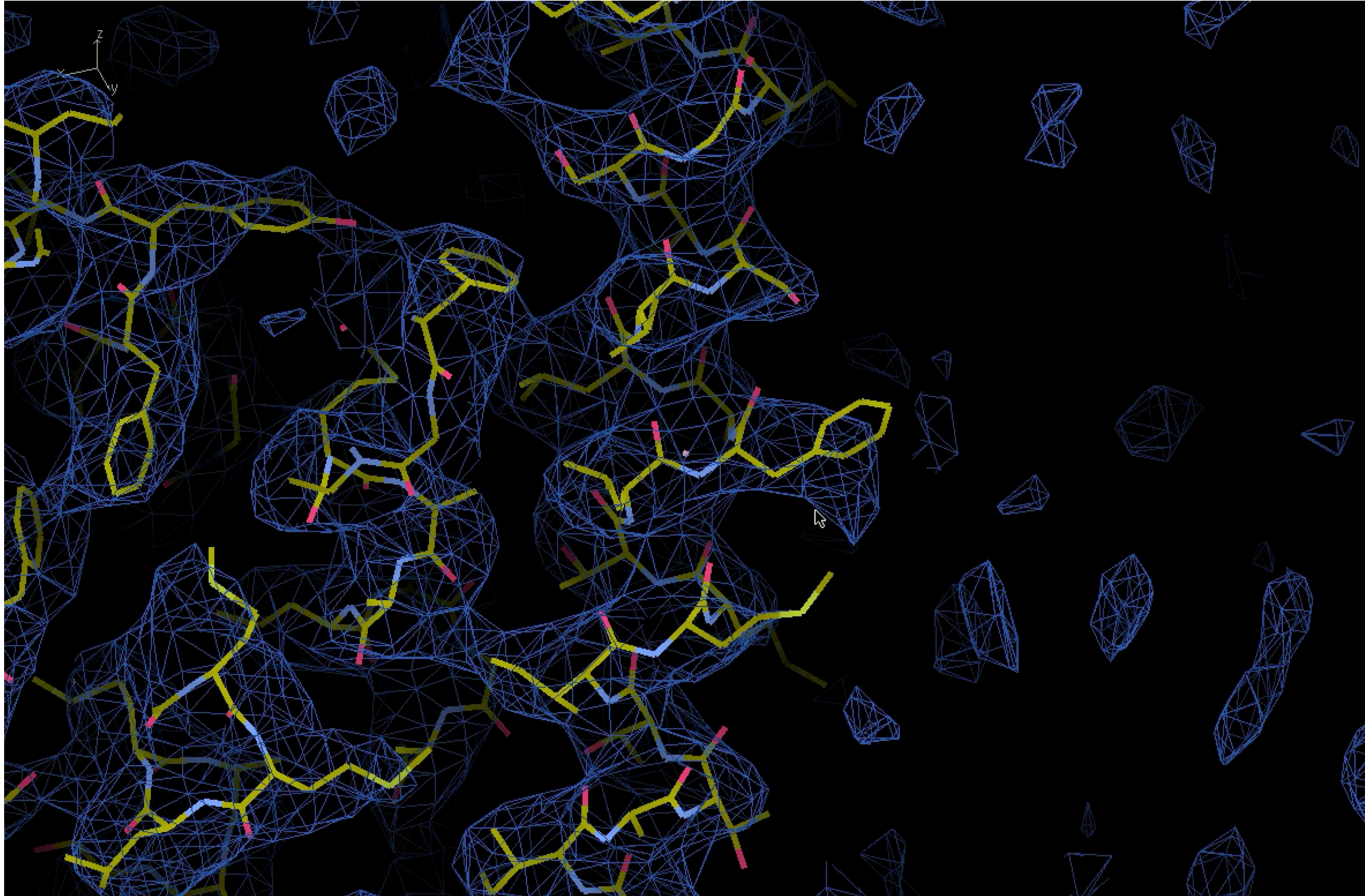
Compare and quantify the differences



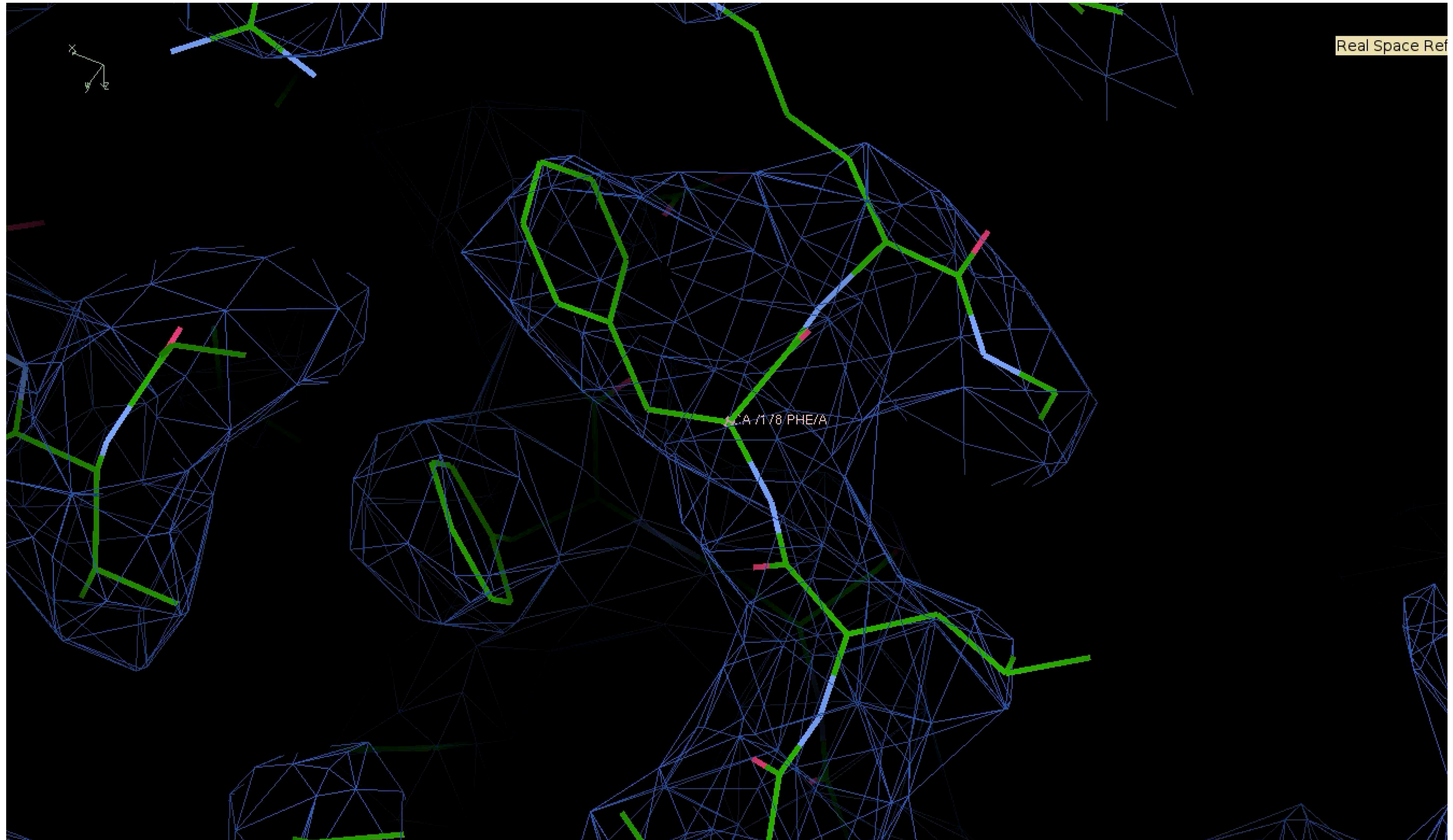
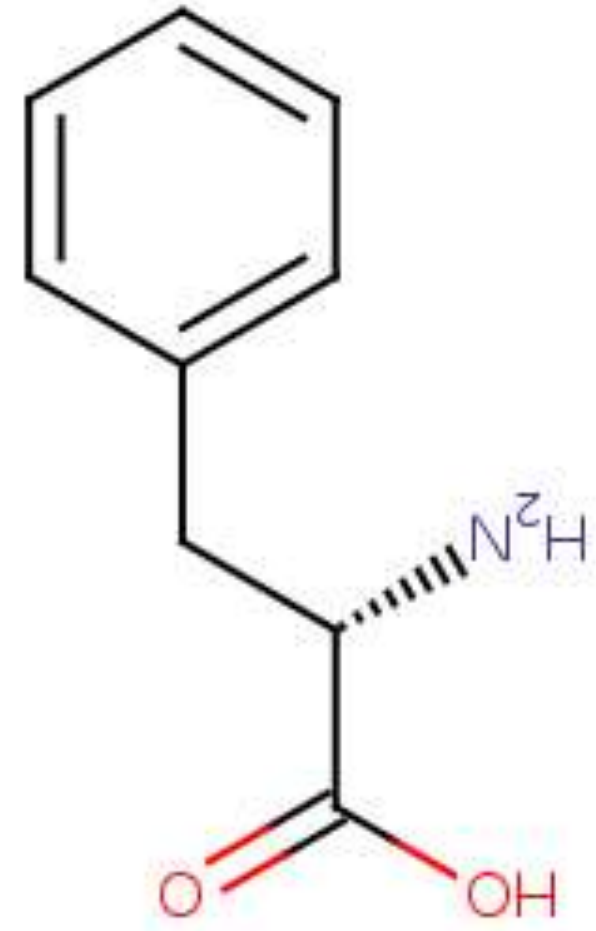
At atomic resolution, position of individual atoms is well-defined



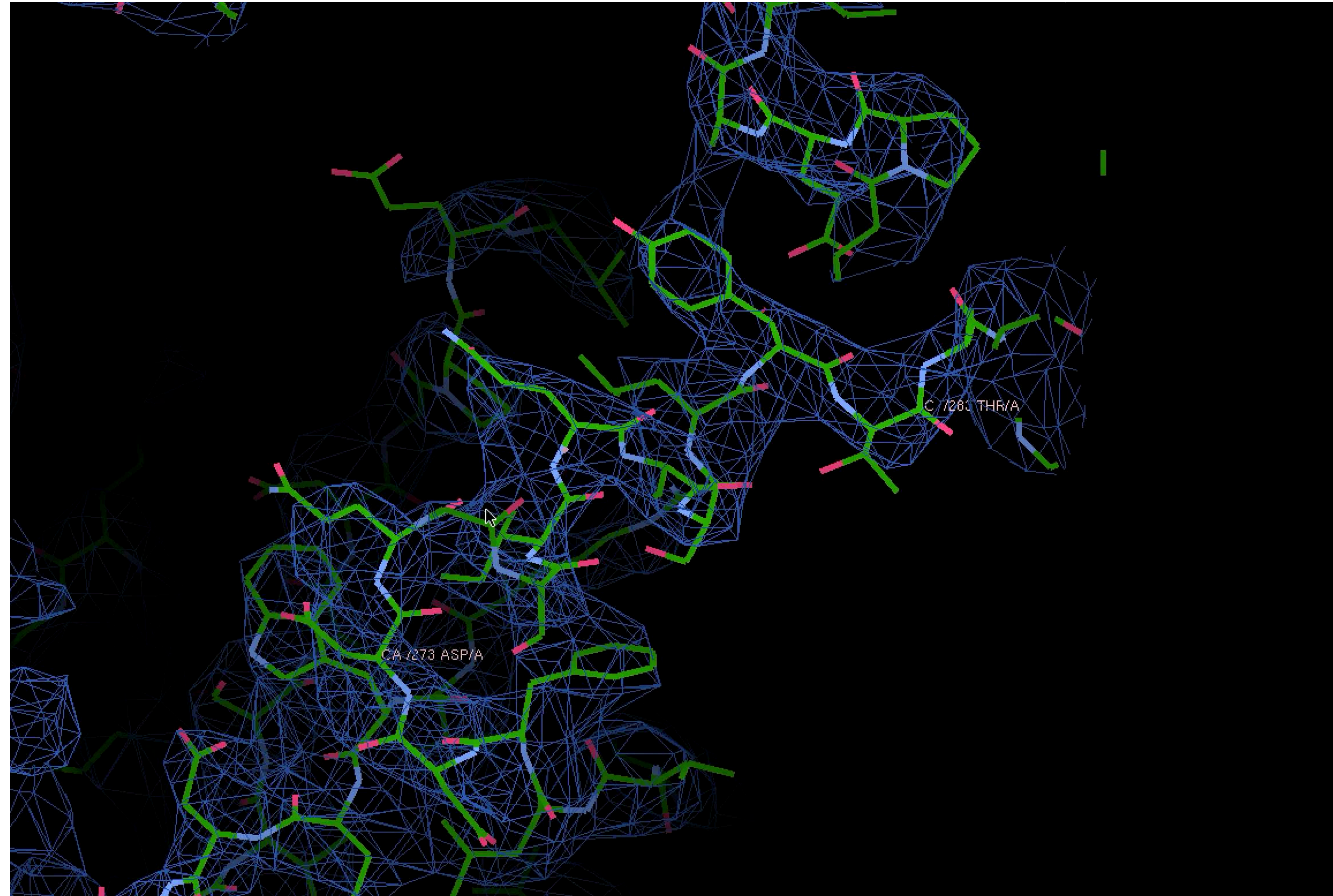
But at “near-atomic” resolution, the position of residues and side chains is not always clear



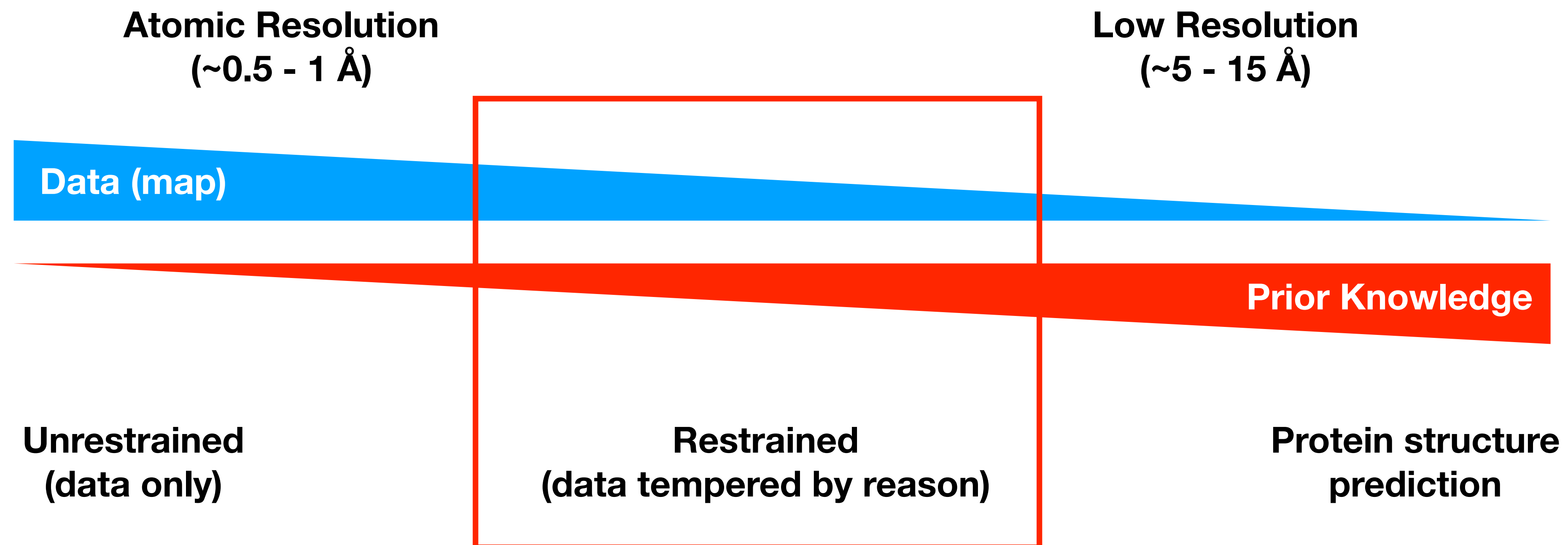
Refinement using only data



Refinement using only data

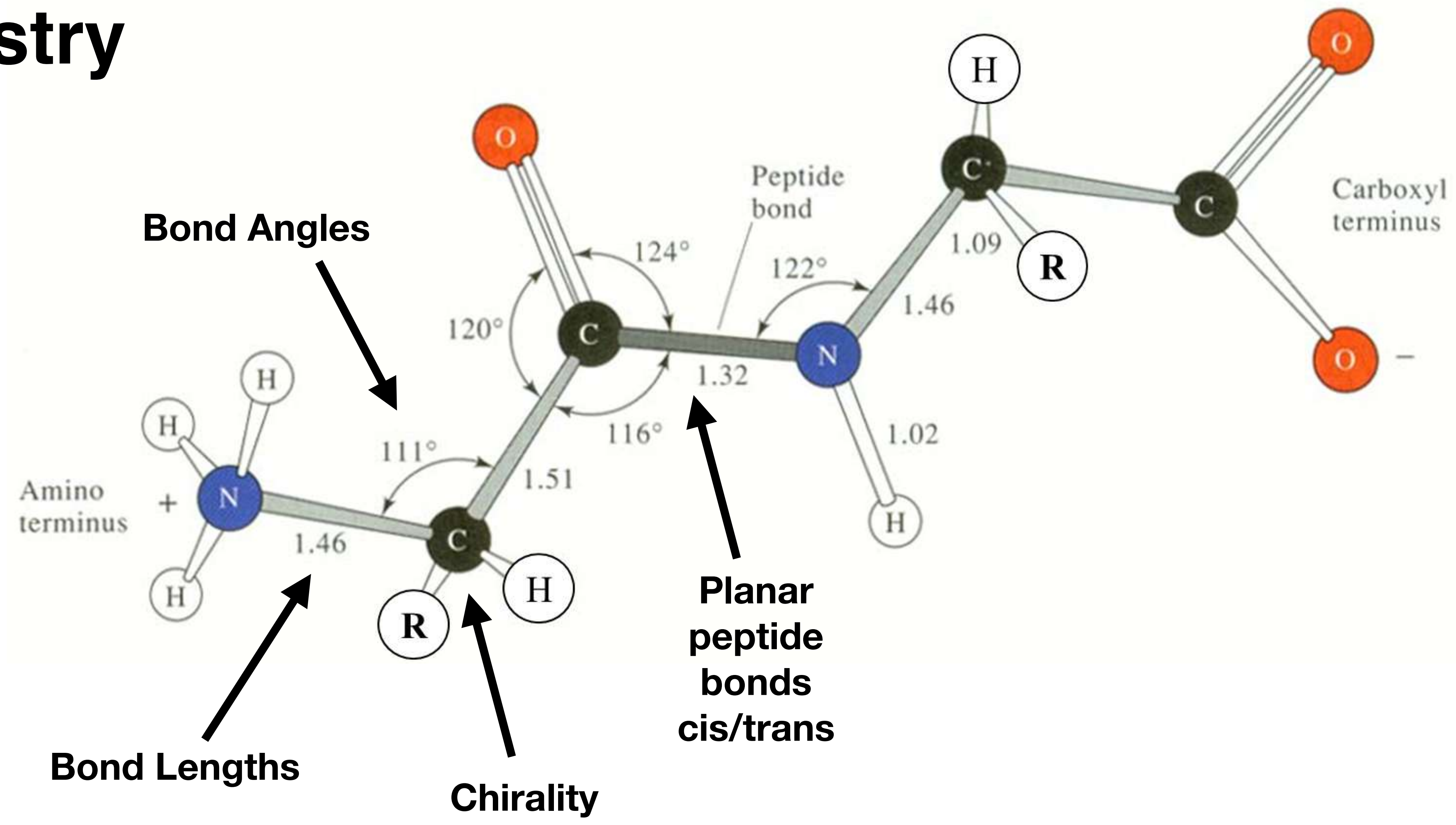


Harnessing prior knowledge of protein structure to bridge the gap

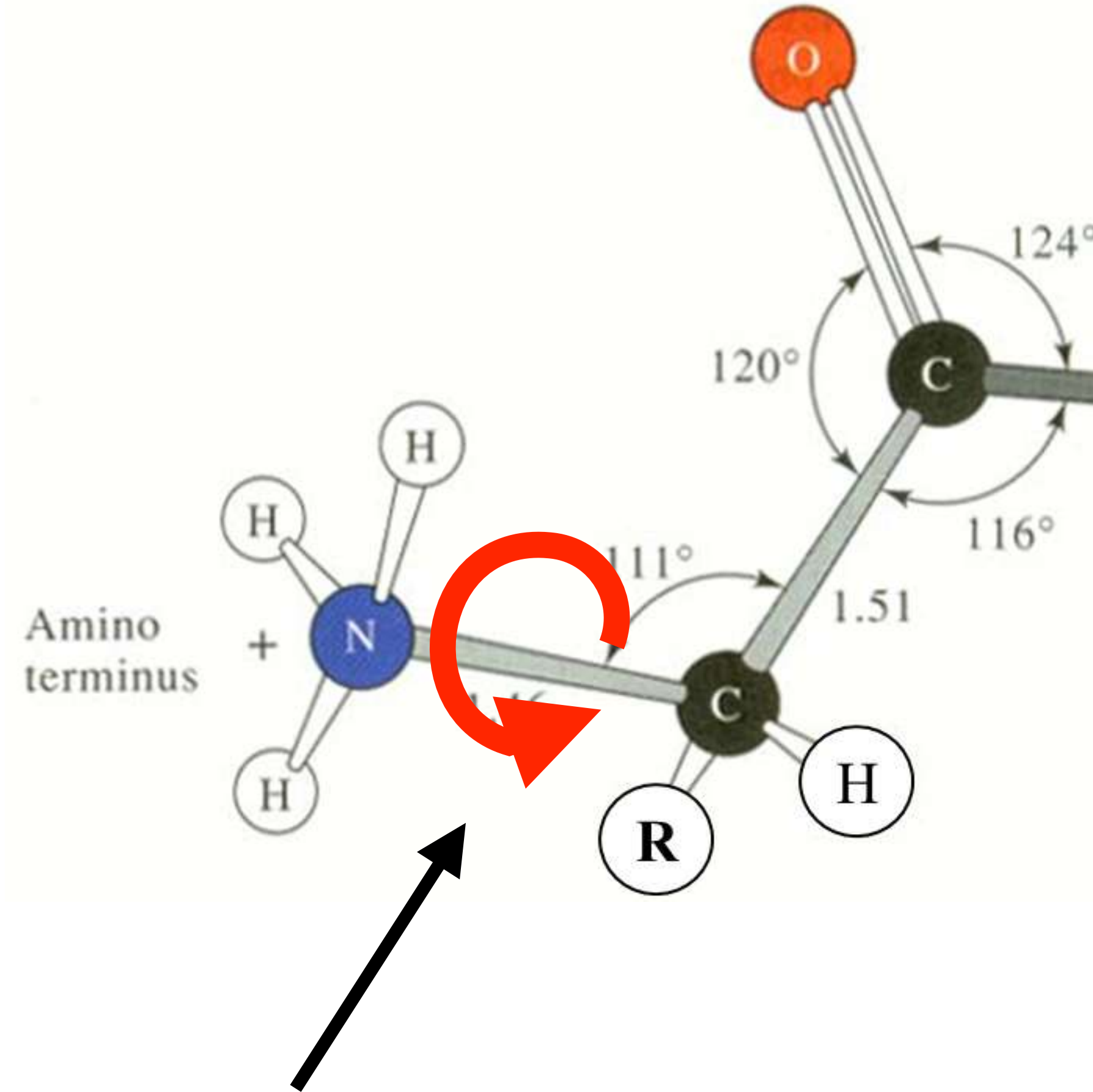


$$\text{Refinement Target} = (\text{Model vs Data}) + w_1 (\text{Model vs Prior Knowledge})$$

Stereochemistry



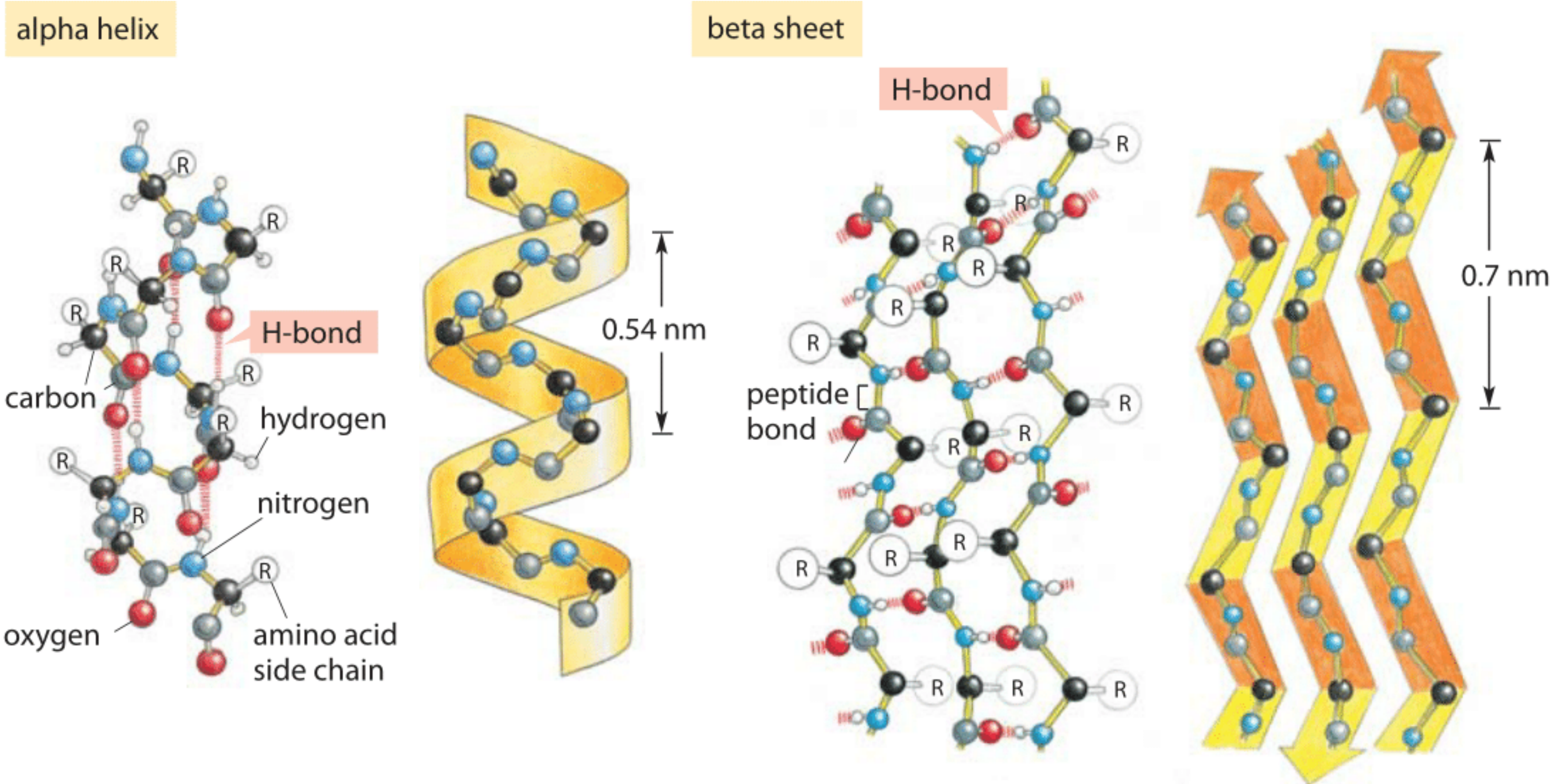
Stereochemistry



Torsion/dihedral Angles

Constraints backbone conformations as well as side chain rotameric states

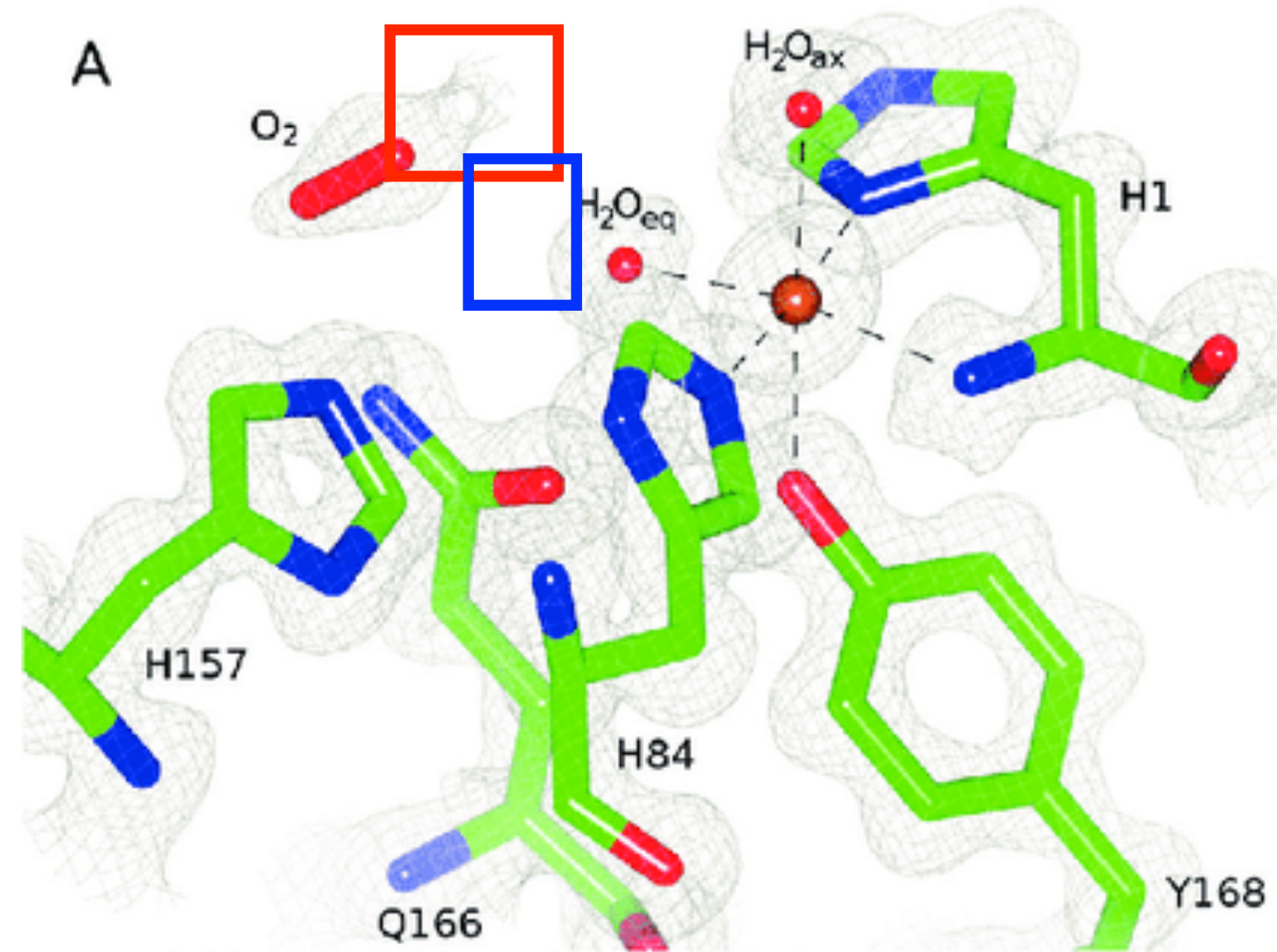
Secondary Structure and Hydrogen Bonds



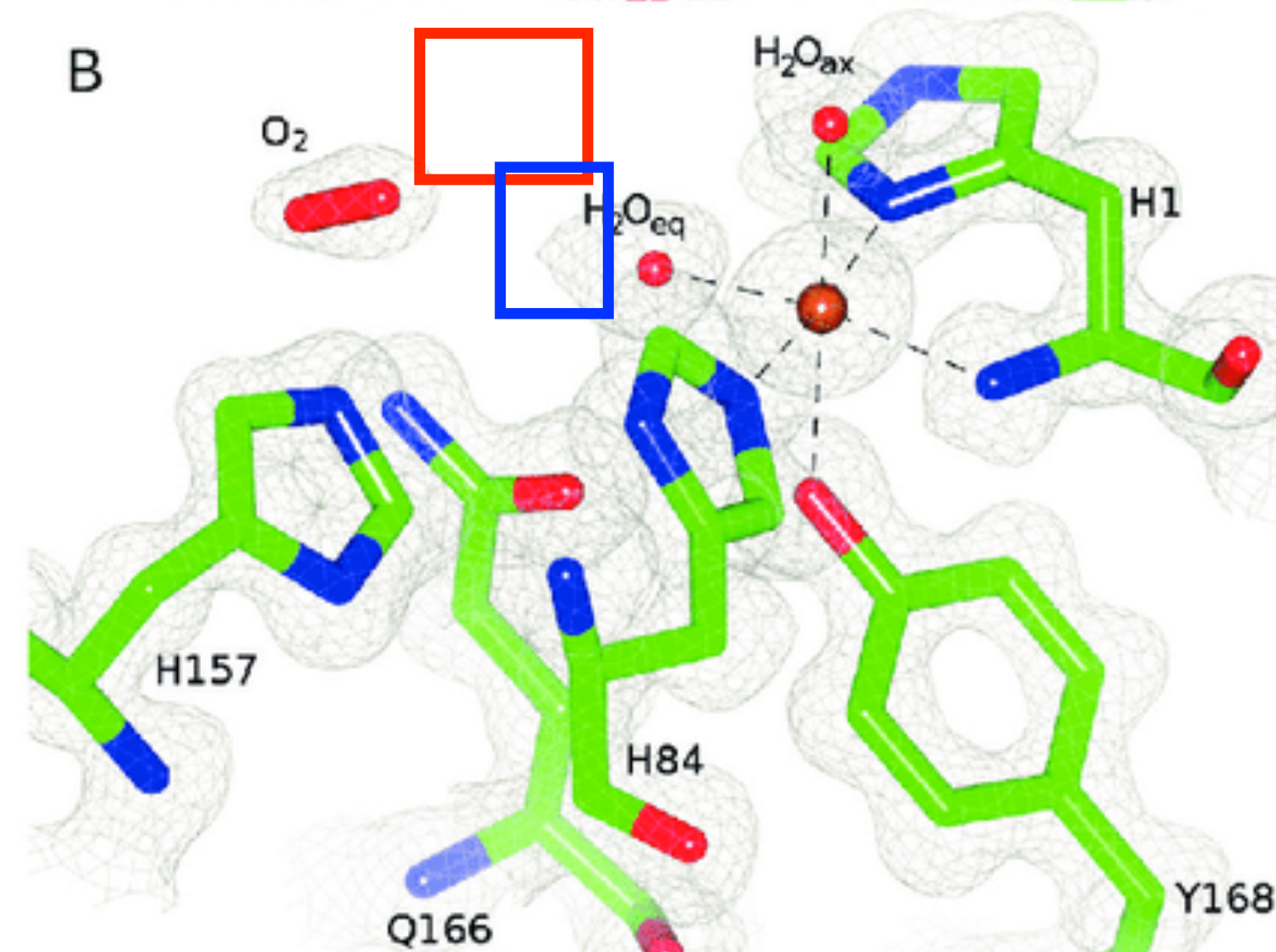
- Torsion angle restraints to maintain appropriate backbone conformation
- Distance restraints between H-bonding atoms

“Non-crystallographic symmetry” (NCS), reference model

Active Site #1



Active Site #2



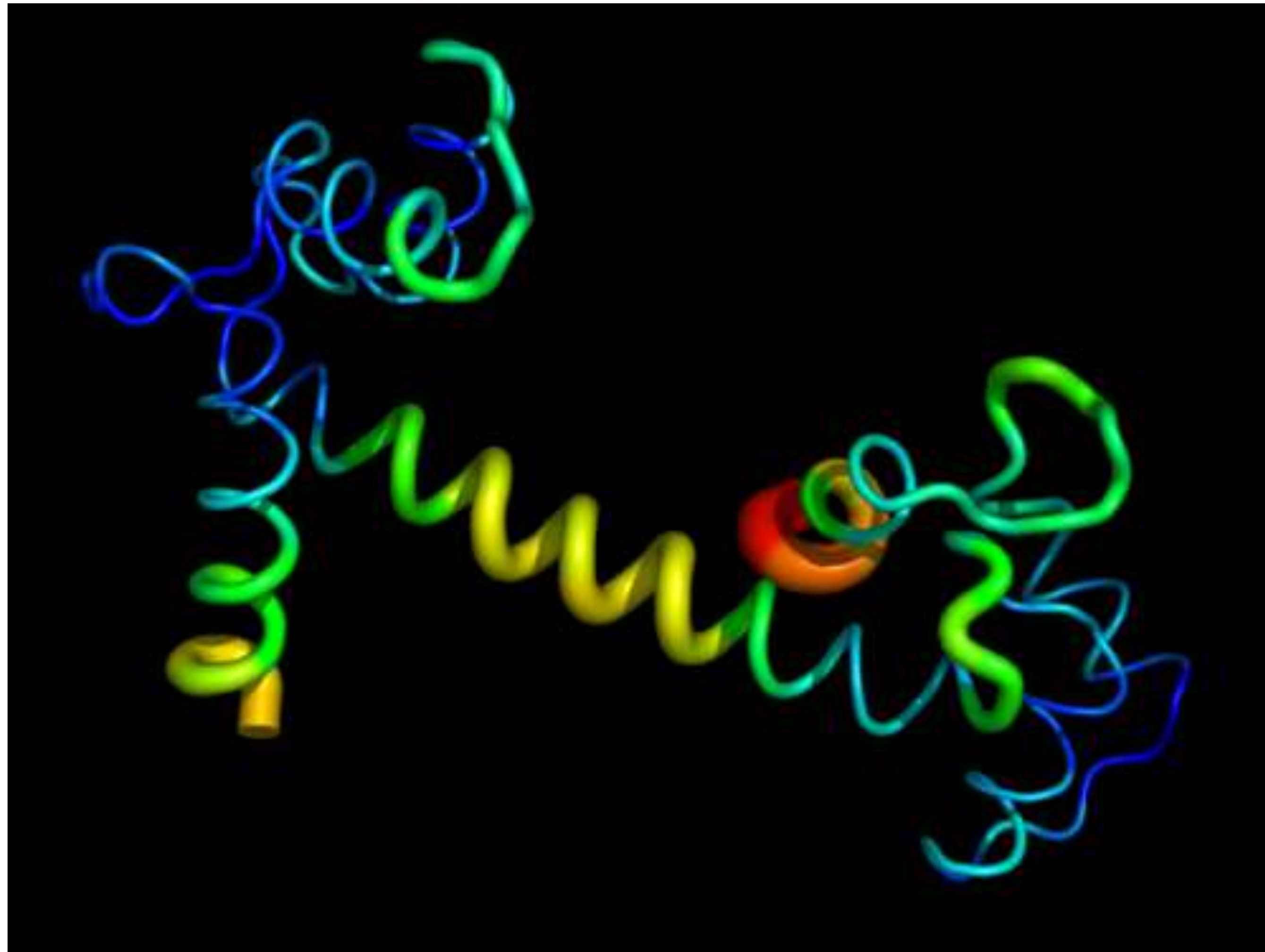
Restrain to be similar
Constrain to be identical

Especially helpful at
lower resolution with
non-symmetrized maps

Chains can be restrained to
be similar to other chains in
structure, or a “reference”,
higher resolution structure
(or the starting model)

B factor / ADP restraints

Higher B factor = **fatter ribbon**, **warmer color**

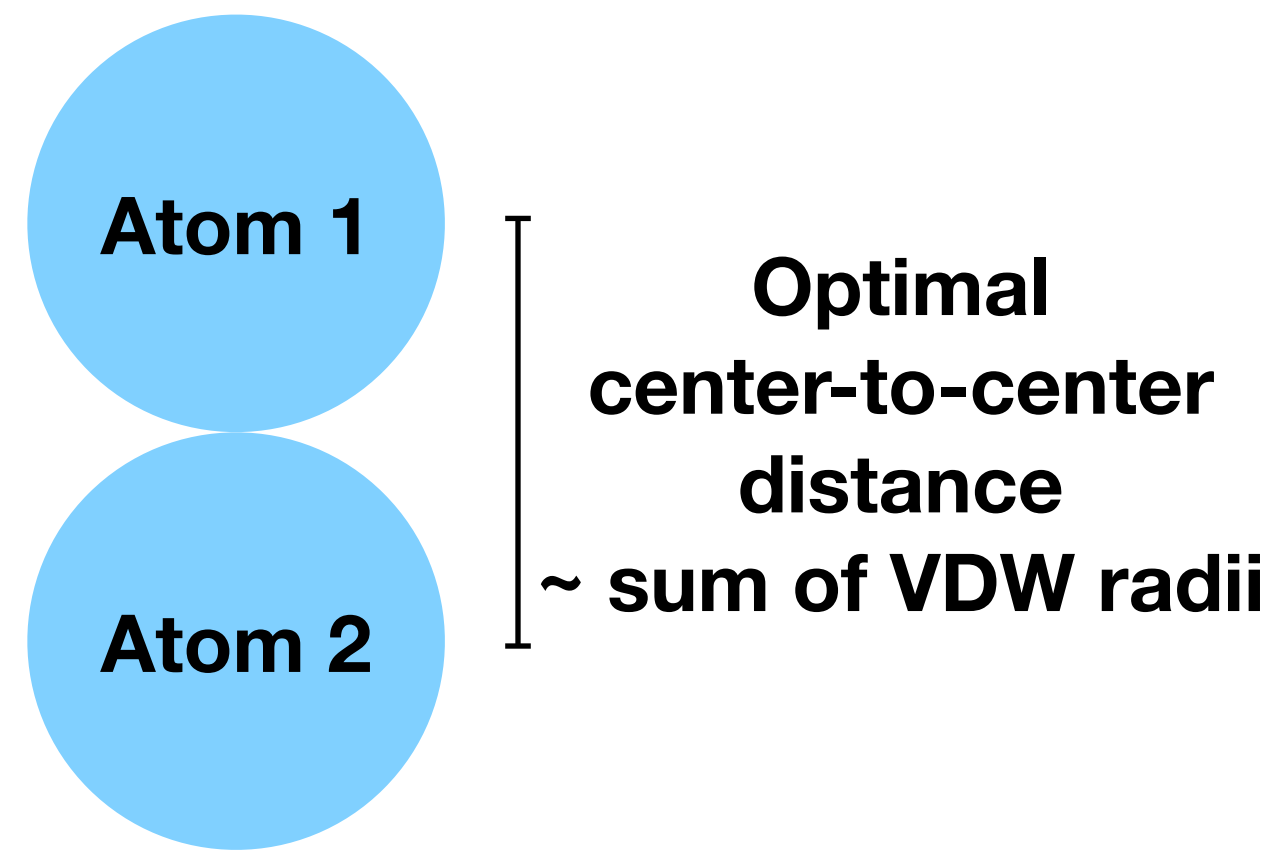


B factors are not randomly distributed

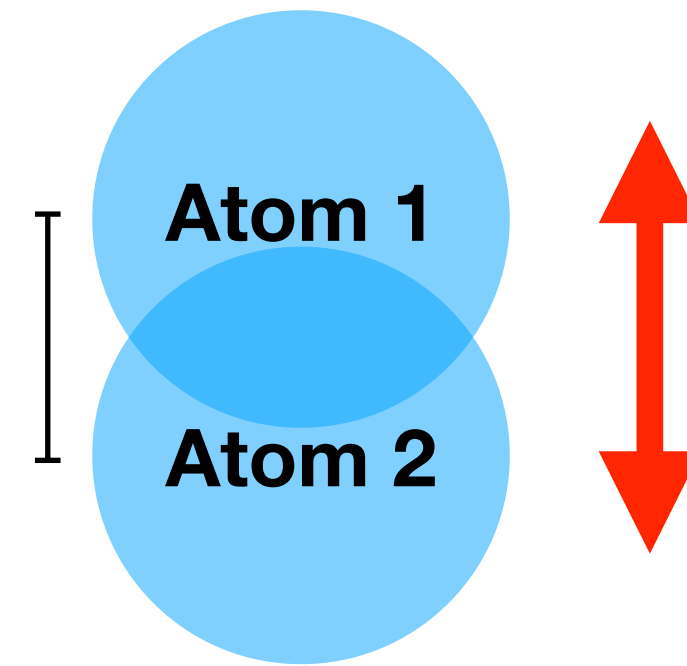
B factor of a particular residue is a good predictor of the residue just before and after

Therefore, we can retrain B factors such that connected atoms/residues must have similar B factors

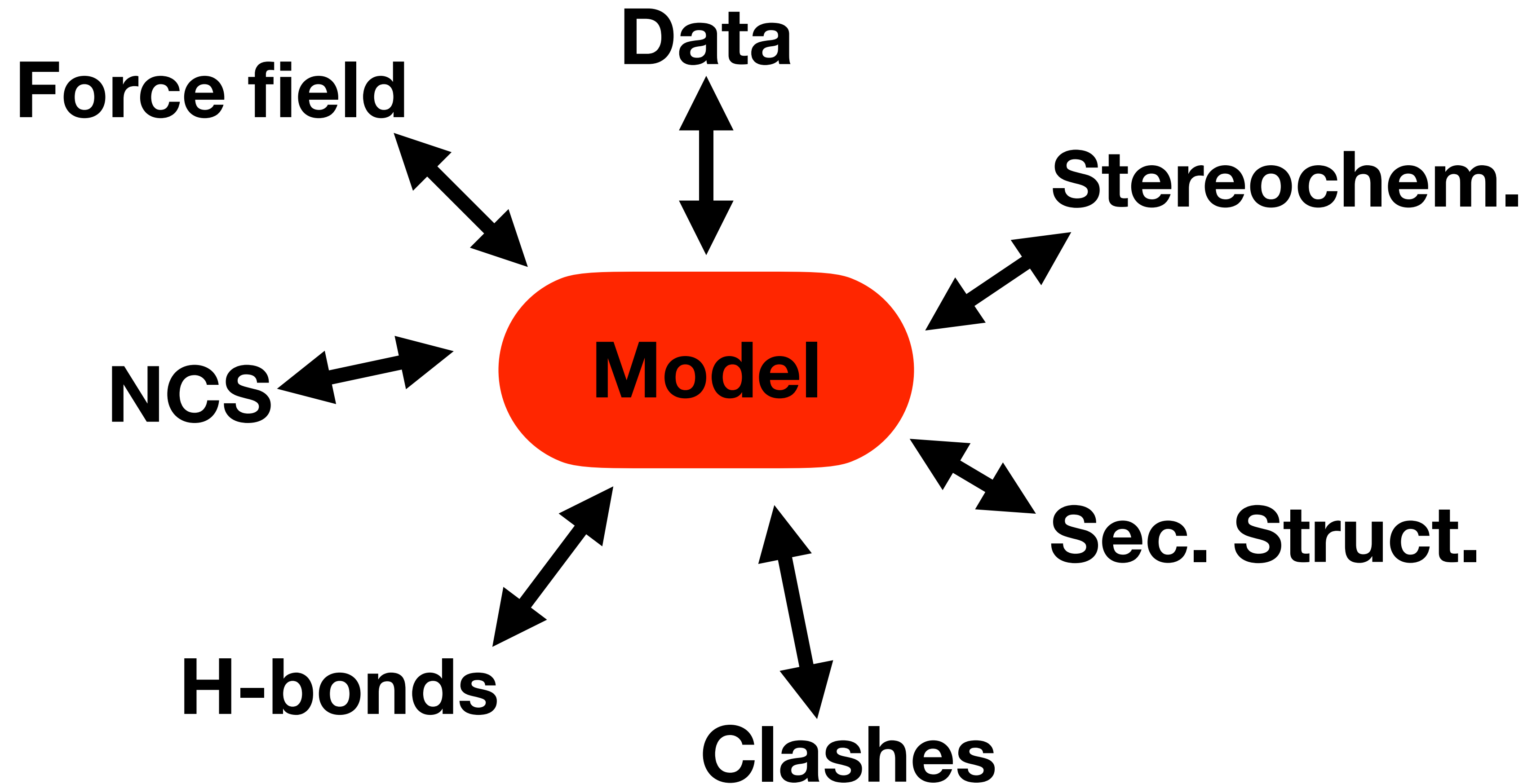
Steric repulsion



If atoms get too close together, need a force to push them apart

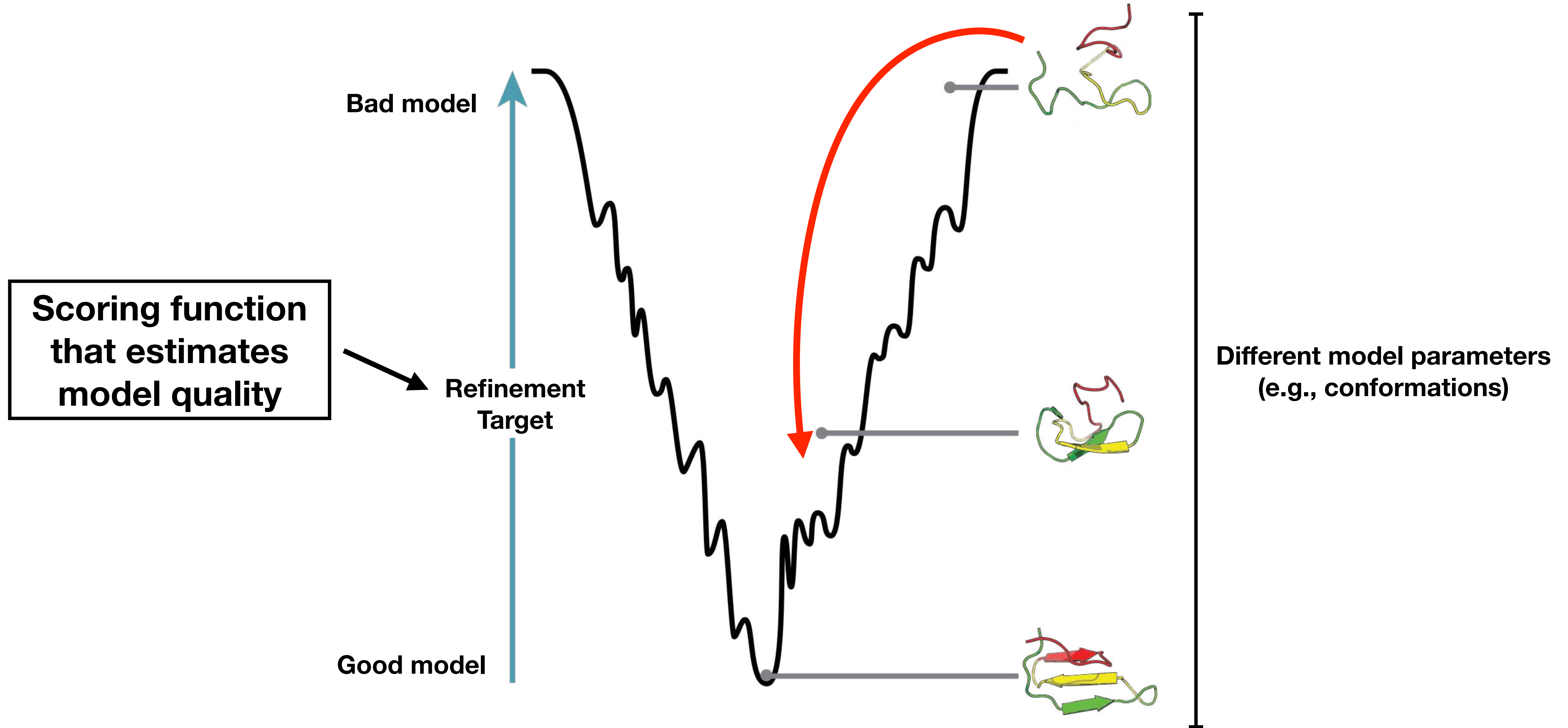


More complete refinement target includes many terms

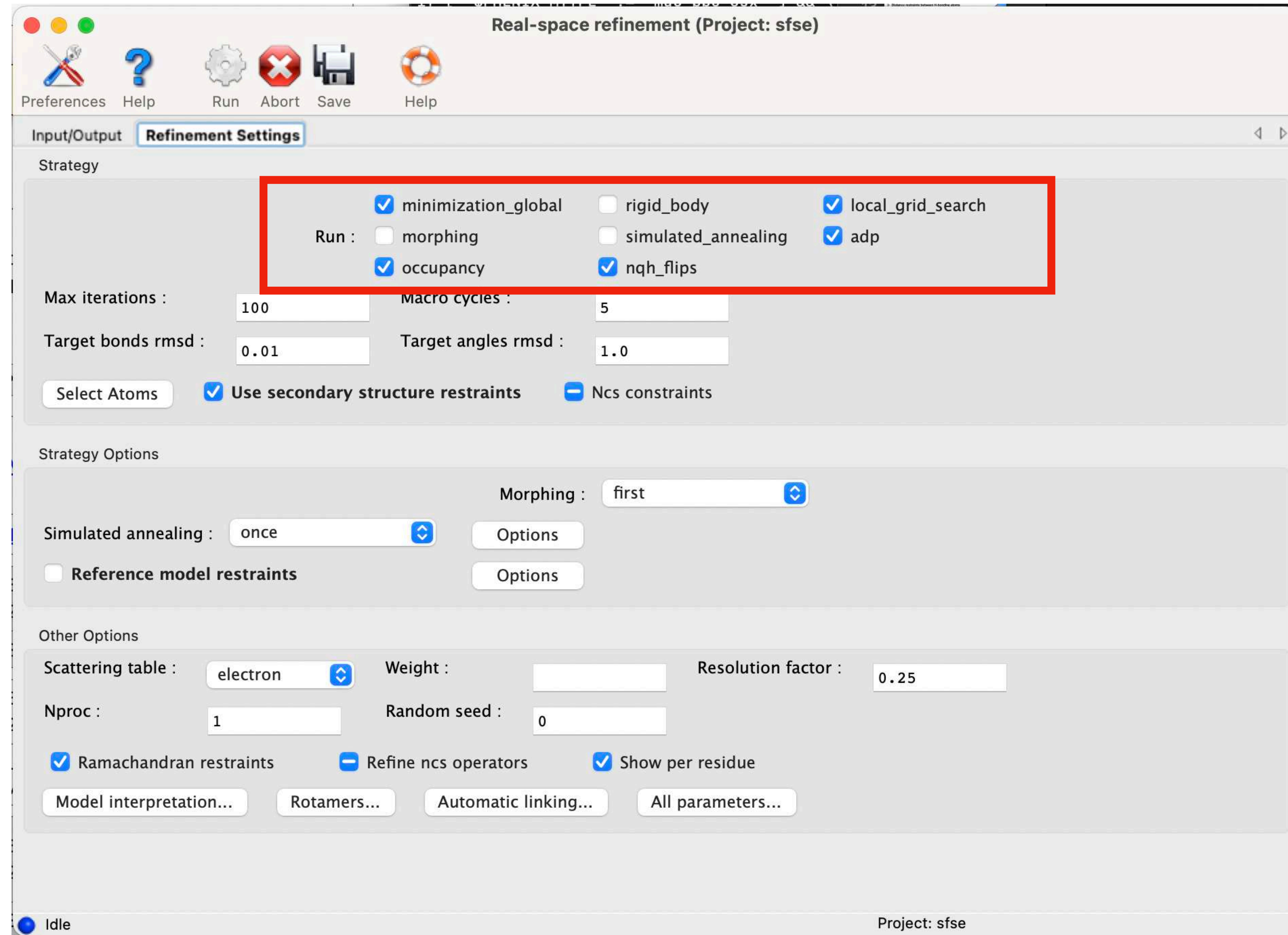


Refinement Target = (Model vs Data) + w_1 (Model vs Stereo) + w_2 (Model vs ForceField) + w_3 (Model vs NCS) + ...

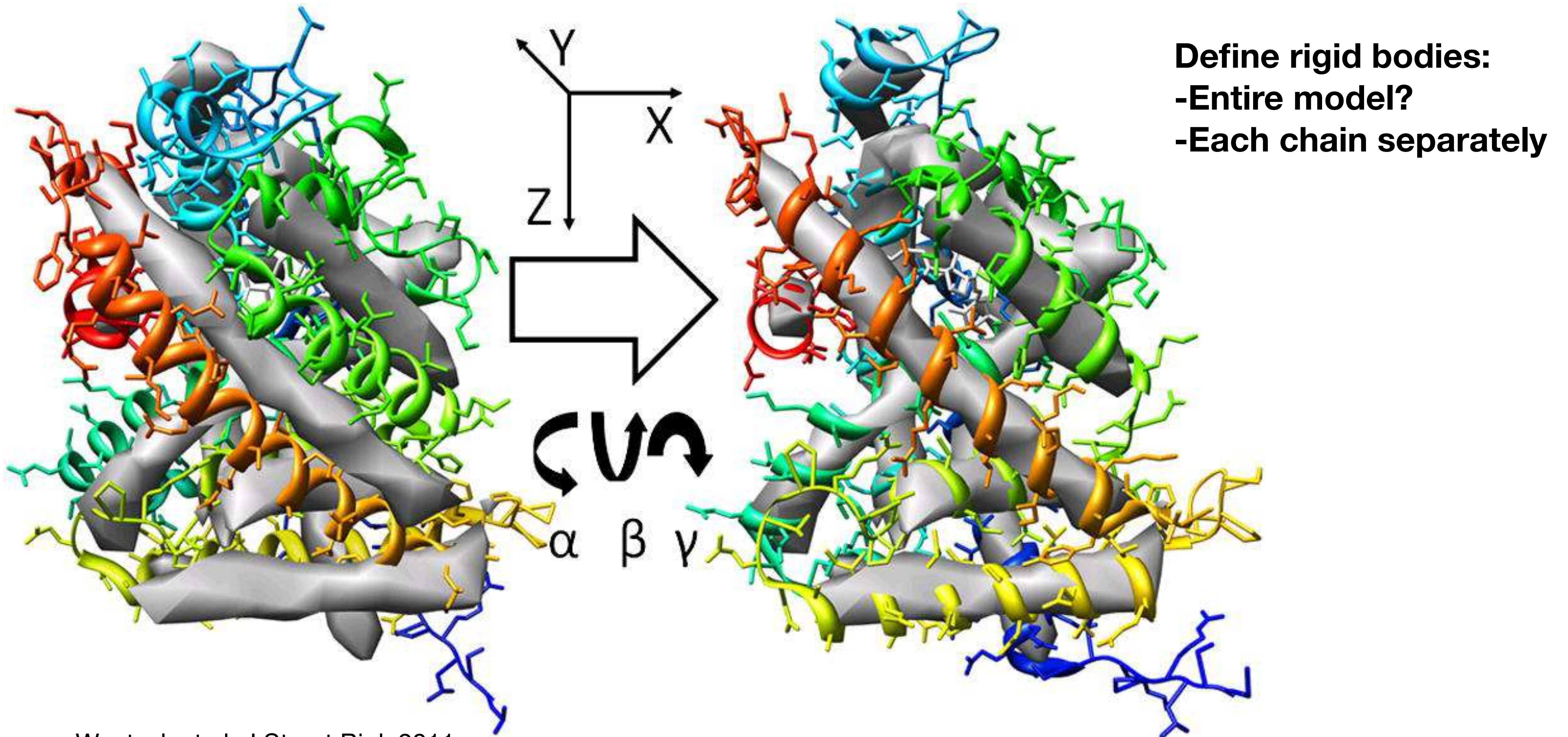
Refinement moves model towards local minimum



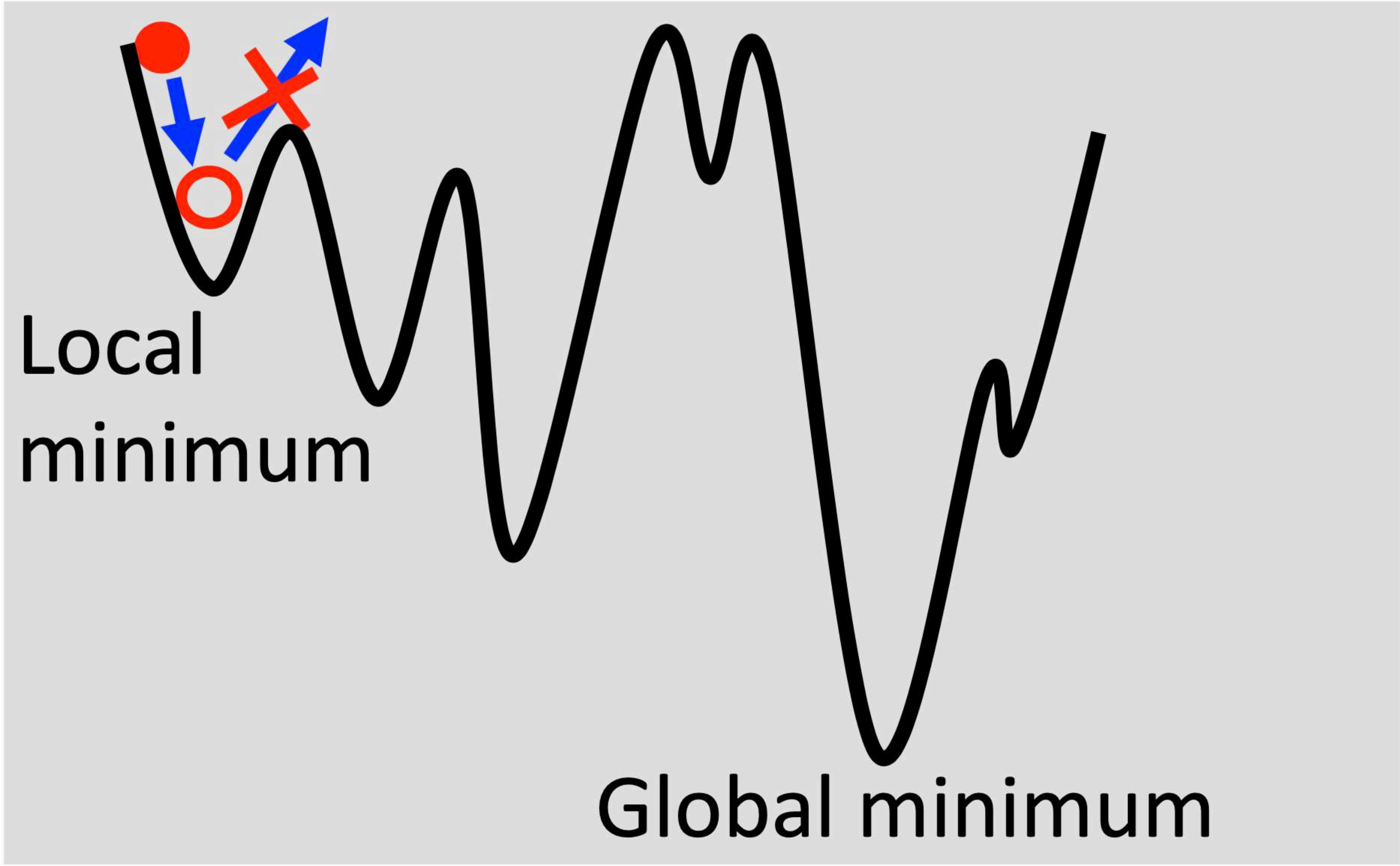
There are many refinement options to choose from!



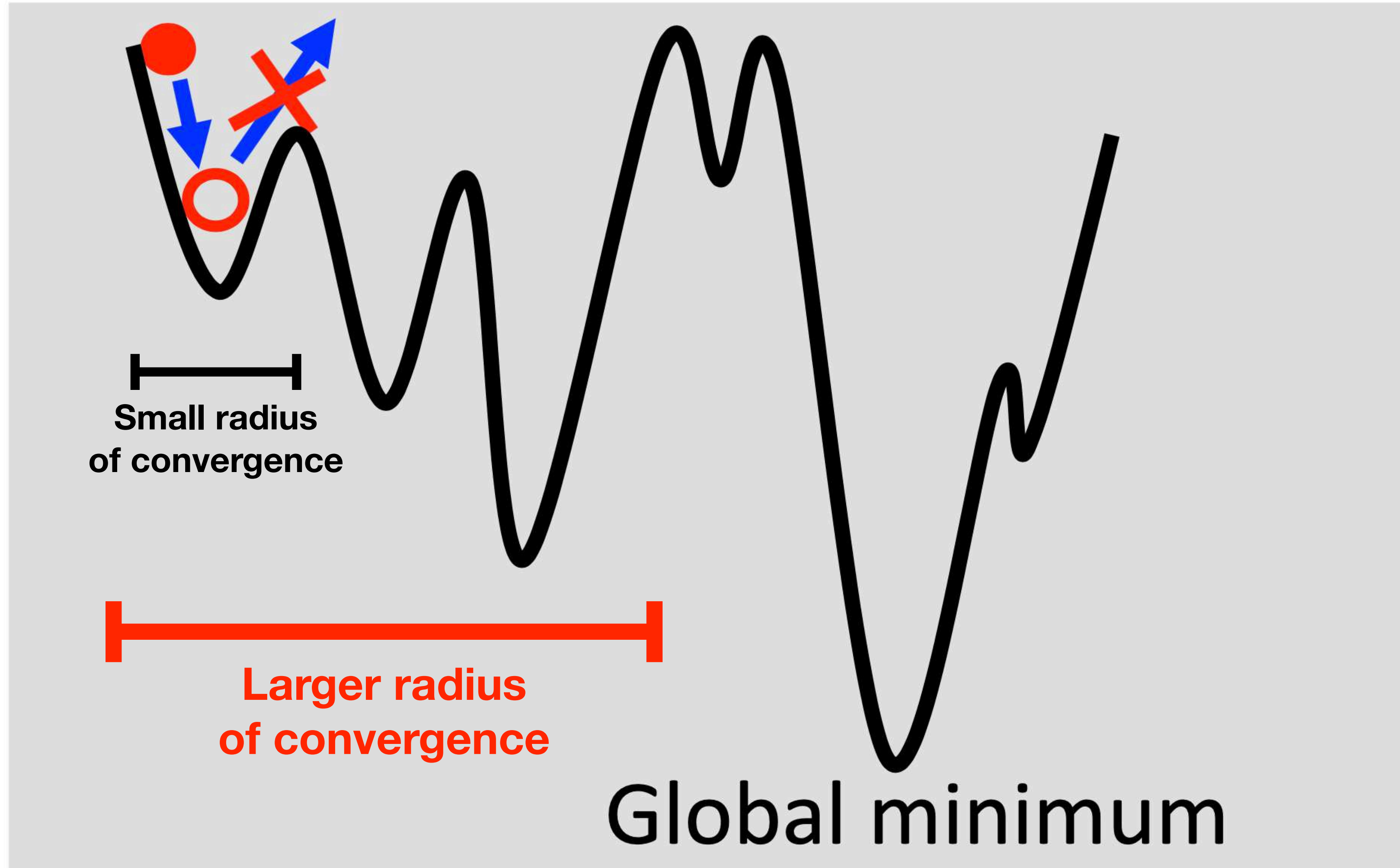
Optimization protocols: Rigid body refinement



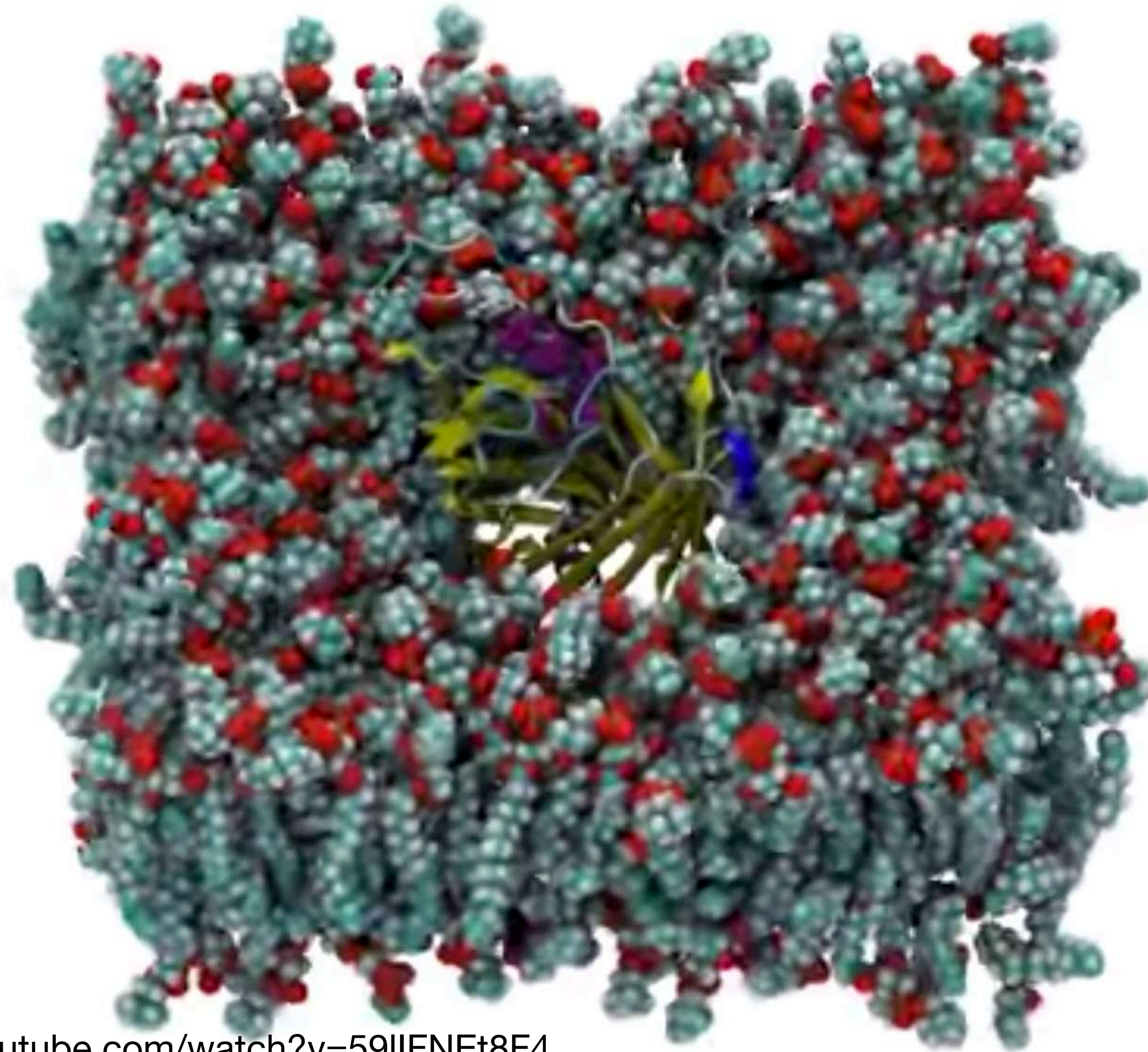
Optimization protocols: Gradient driven minimization



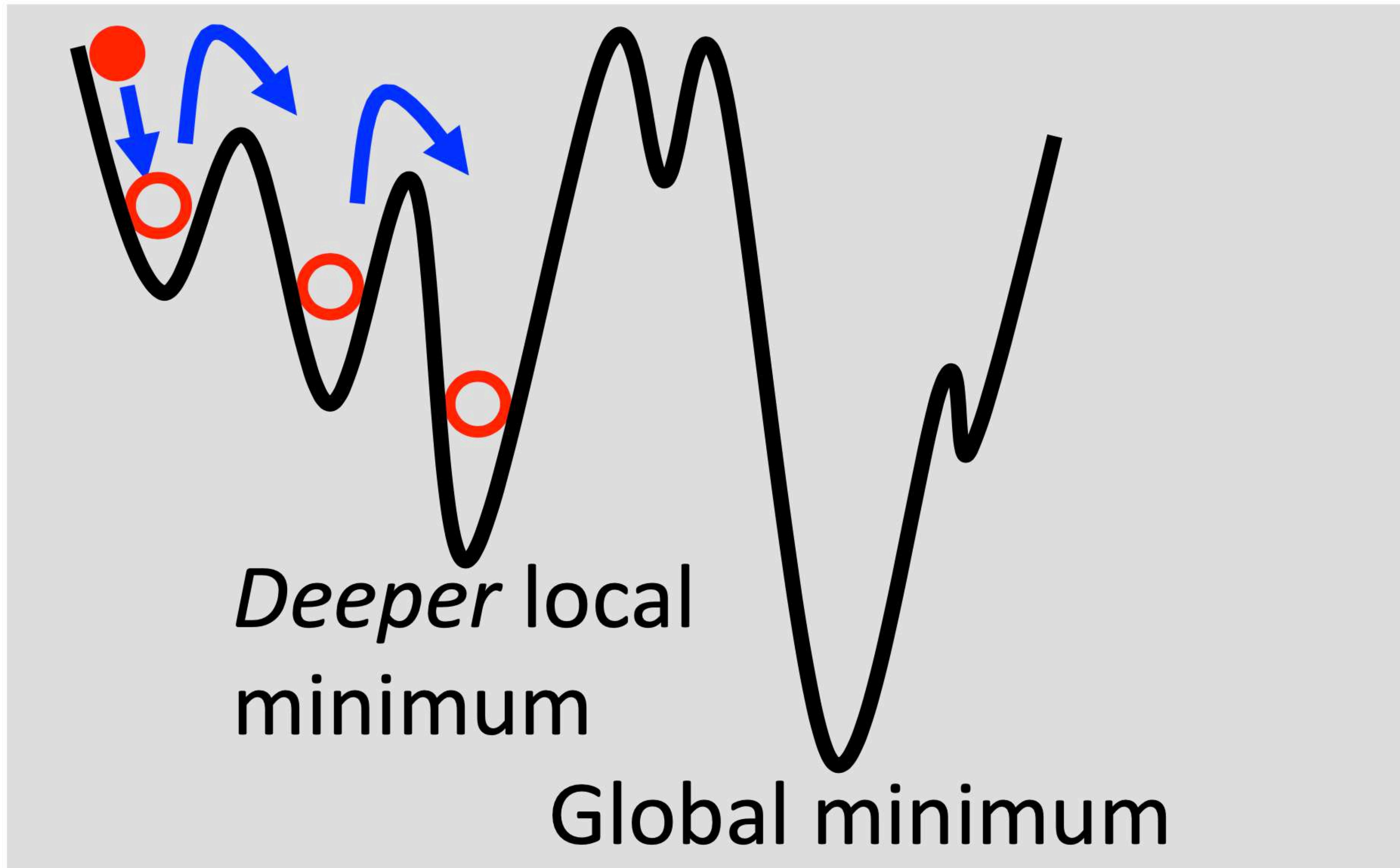
Refinement “radius of convergence”



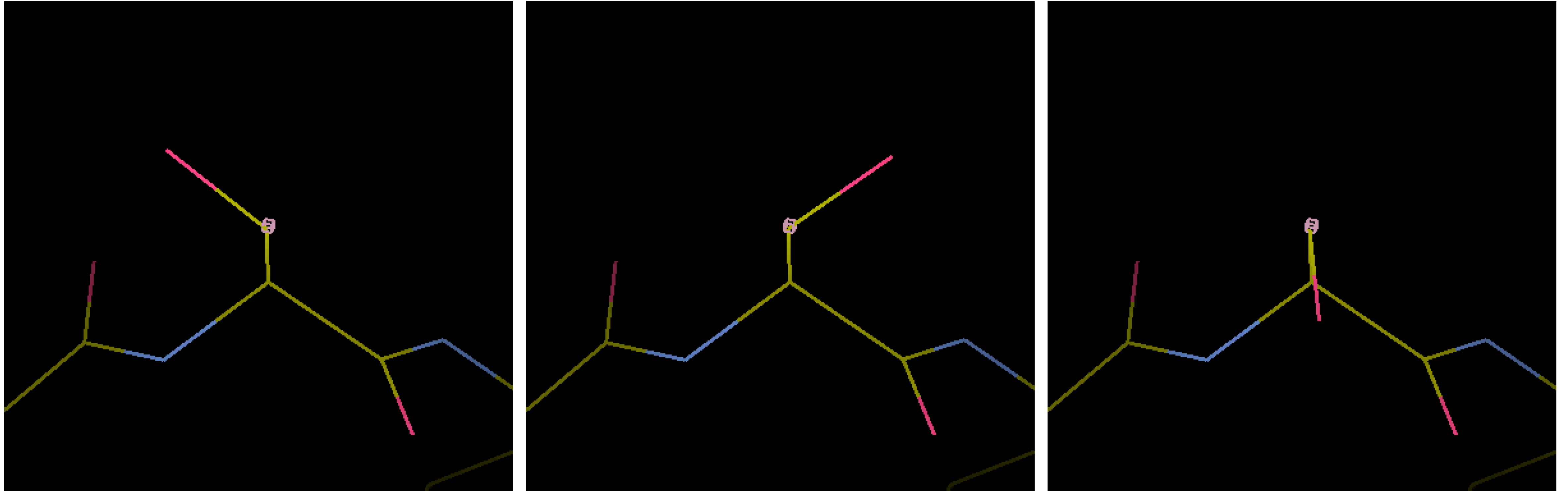
Optimization protocols: Simulated annealing



Optimization protocols: Simulated annealing

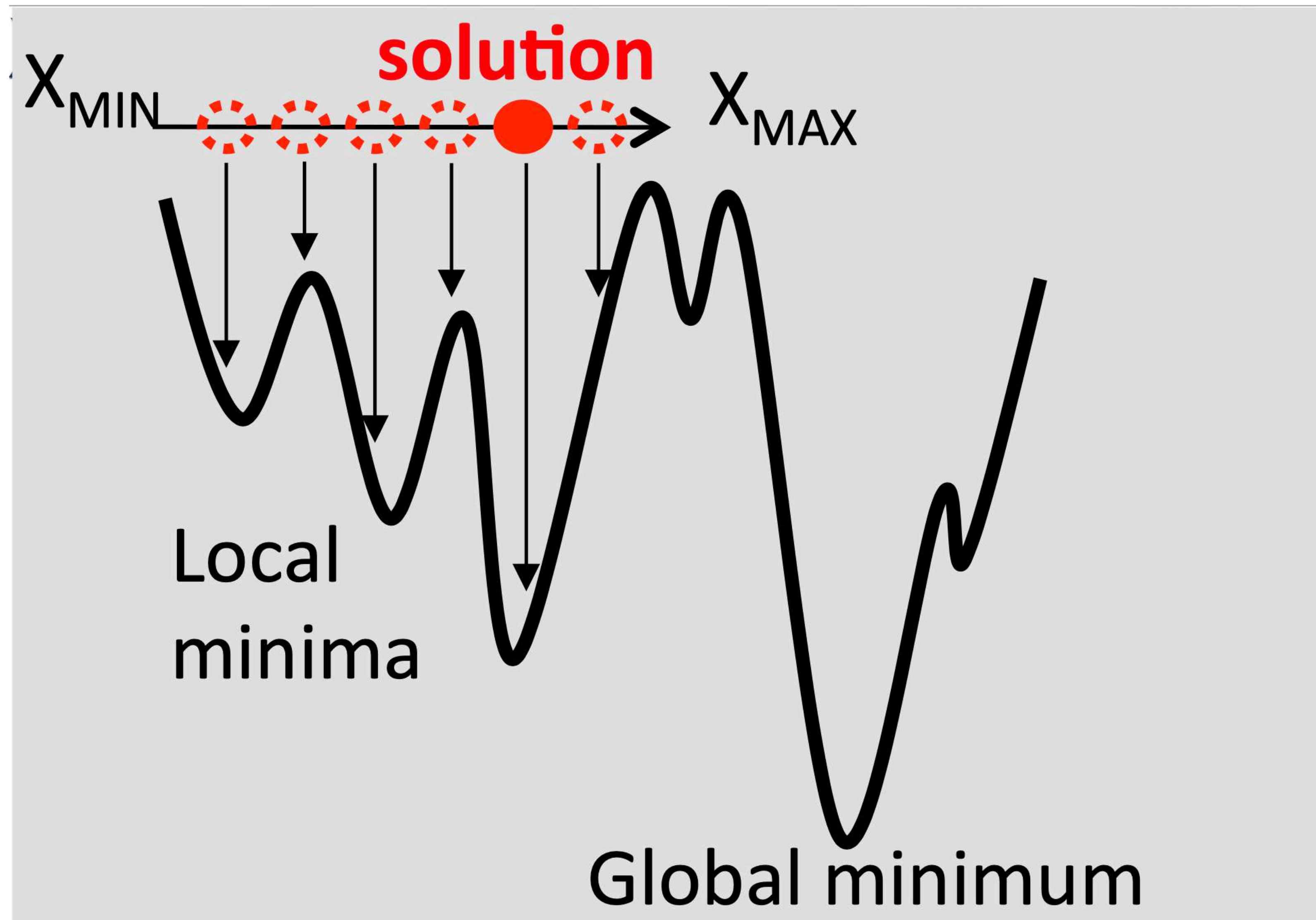


Optimization protocols: Torsion-angle Grid Search

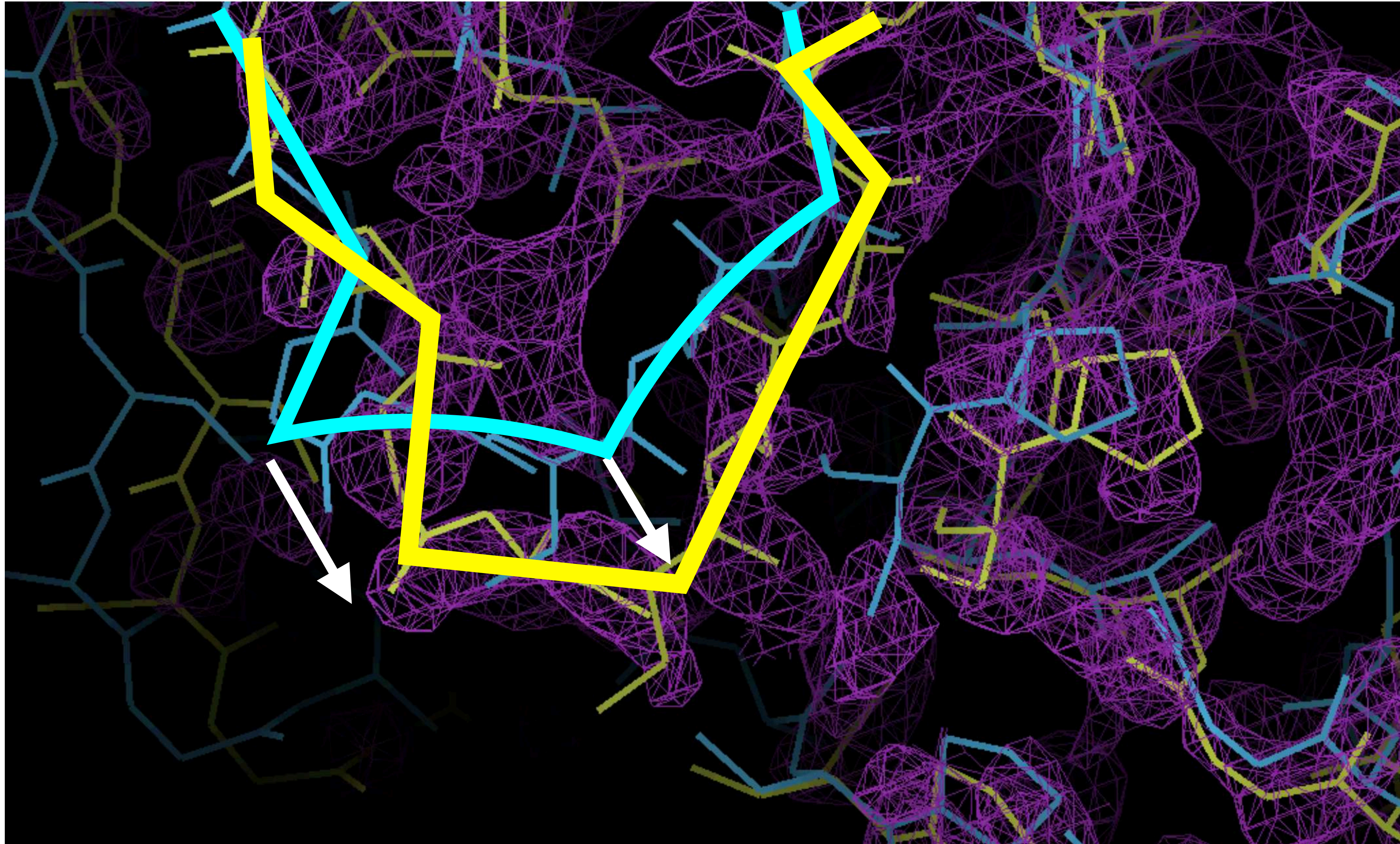


- **Can allow for larger shifts in model than simple minimization**

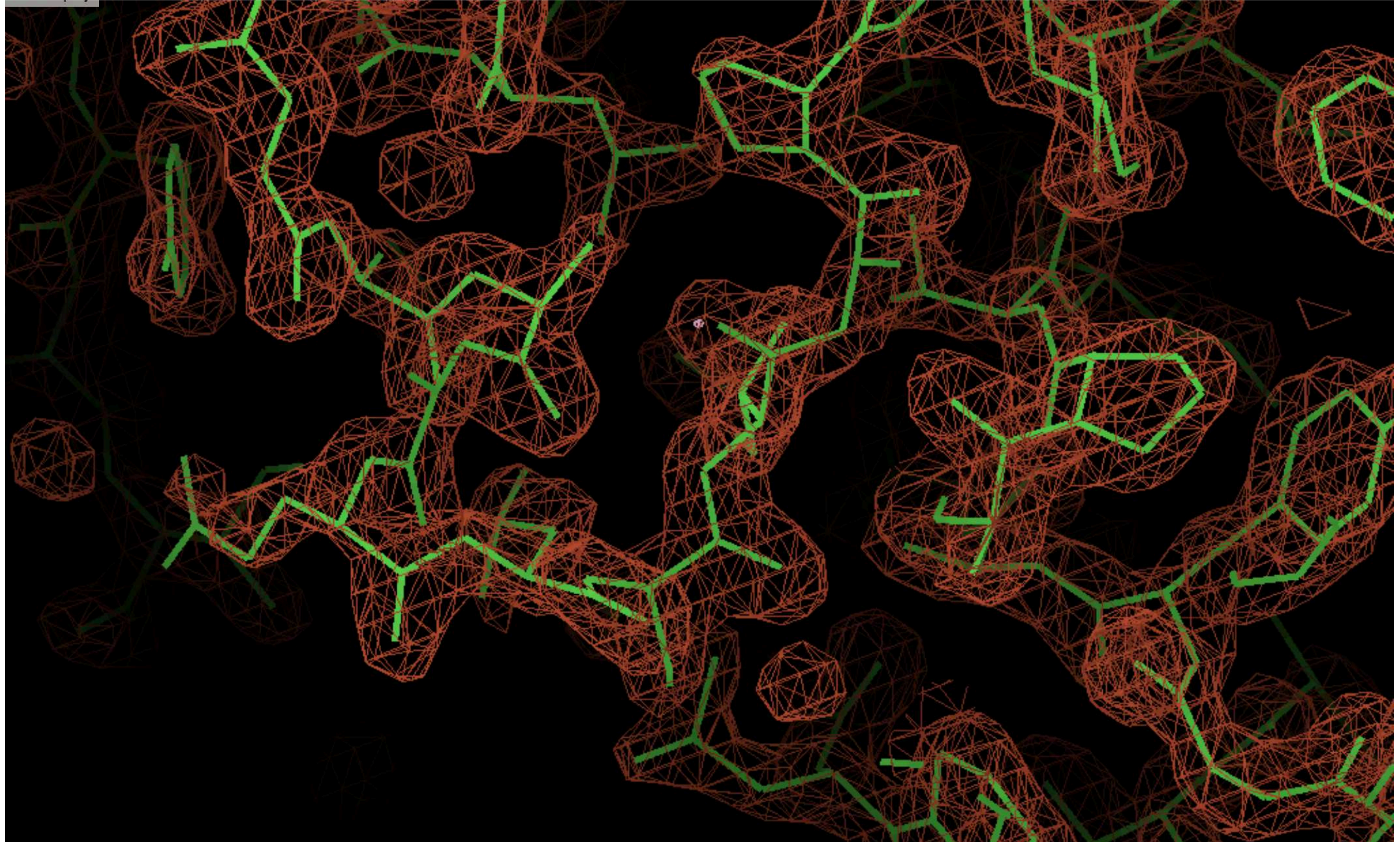
Optimization protocols: Torsion-angle Grid Search



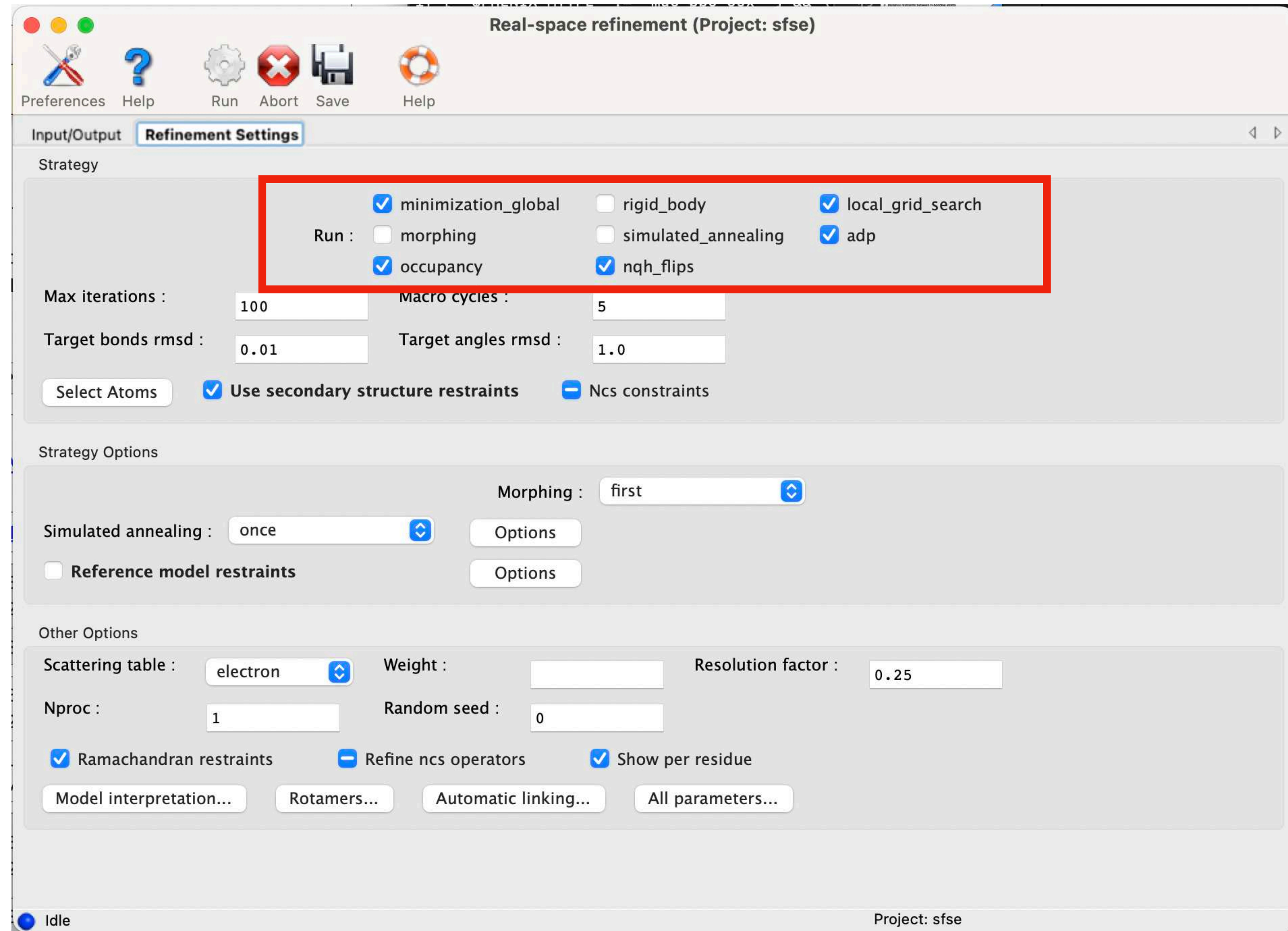
Optimization protocols: Morphing



Optimization protocols: Morphing



There are many refinement options to choose from!



What parameters should I use?

How aggressive do I want to be in refinement?

--How much do I trust my starting model vs my data?

--How different is my starting model from my data?

More aggressive = larger radius of convergence, potentially less manual rebuilding; but changes the model a lot (morphing, grid search, simulated annealing)

Less aggressive = smaller radius of convergence, will change model less (rigid body, minimization_global)

What parameters should I use?

Before doing anything: rigid body refinement (overall, individual domains)

Early stages:

-Target structure very similar to starting model: minimization_global, adp, grid_search(?)

-Target structure very different: minimization_global, adp, grid_search, morph, simulated annealing(?)

Late stages:

-Go easy: minimization_global, adp

Restraints

https://phenix-online.org/documentation/reference/real_space_refine.html

Real-space refinement (Project: tmp)

Preferences Help Run Abort Save Ask for help

Input/Output **Refinement Settings** RealSpaceRefine_4

Strategy

Run : minimization_global rigid_body local_grid_search
 morphing simulated_annealing adp

Max iterations : 100 Macro cycles : 5

Target bonds rmsd : 0.01 Target angles rmsd : 1.0

Use secondary structure restraints Use NCS

Strategy Options

Morphing : once

Simulated annealing : once Options

Reference model restraints Options

Other Options

Scattering table : electron Weight : Resolution factor : 0.25

Nproc : 1 Random seed : 0

Rotamer restraints Ramachandran restraints Ncs constraints

Refine ncs operators Show per residue

1 job(s) running Project: tmp

What restraints should I use?

Resolution dependent:

- High res, use less restraints and trust map more
- Low res, use more restraints and trust map less

High res: Often only basic stereochemical restraints are sufficient

Low res: Try different combinations of secondary structure restraints, ncs, reference model, rotamers (ramachandran(?))

- Sometimes using too many restraints can prevent efficient refinement because model can't move; experiment and see what results in best fit to map while maintaining good geometry

Goal of model validation:

- 1) To assess refinement strategies and progress**
- 2) To identify problem areas requiring manual intervention**
- 3) To assess overall and local model quality/reliability**

“Self-assessment”: We want to create the most accurate and reliable model we can, and validation stats clue us in to regions of the model that may have issues

Model validation metrics - By Problem type

- Overall Quality Indicators
 - Model/Map CC
 - RMS deviations
 - Unmodeled densities
 - Molprobity score and clash score
- Backbone issues
 - Ramachandran plot
 - Cis peptide bonds
- Side chain issues:
 - Rotamer outliers
 - Cbeta deviations
- B factor / ADP outliers

Model validation metrics - By source of problem

- Model building problem
 - Unmodeled densities
 - Cbeta deviations
 - Ramachandran plot
 - Cis peptide bonds
 - Model/Map CC
- Refinement problem
 - RMS deviations
 - B factor / ADP outliers
- Either/both?
 - Molprobit score and clash score
 - Rotamer outliers

Model/Map CC

Table 3

Summary of map correlation coefficients used in this work.

Metric	Region of the map used in calculation	Purpose
CC_{box}	Whole map	Similarity of maps
CC_{mask}	Jiang & Brünger (1994) mask with a fixed radius	Fit of the atomic centers
CC_{volume}	Mask of points with the highest values in the model map	Fit of the molecular envelope defined by the model map
CC_{peaks}	Mask of points with the highest values in the model and in the target maps	Fit of the strongest peaks in the model and target maps
$CC_{\text{vr_mask}}$	Same as CC_{mask} but atomic radii are variable and function of resolution, atom type and ADP	Fit of the atomic images in the given map

Root mean square (RMS) deviations

Covalent bond lengths and angles exhibit known, narrow distributions

Typical RMS Bonds for protein structure:
0.005 - 0.015 Å

Typical RMS Angles for protein structure:
0.5 - 1.5 Å

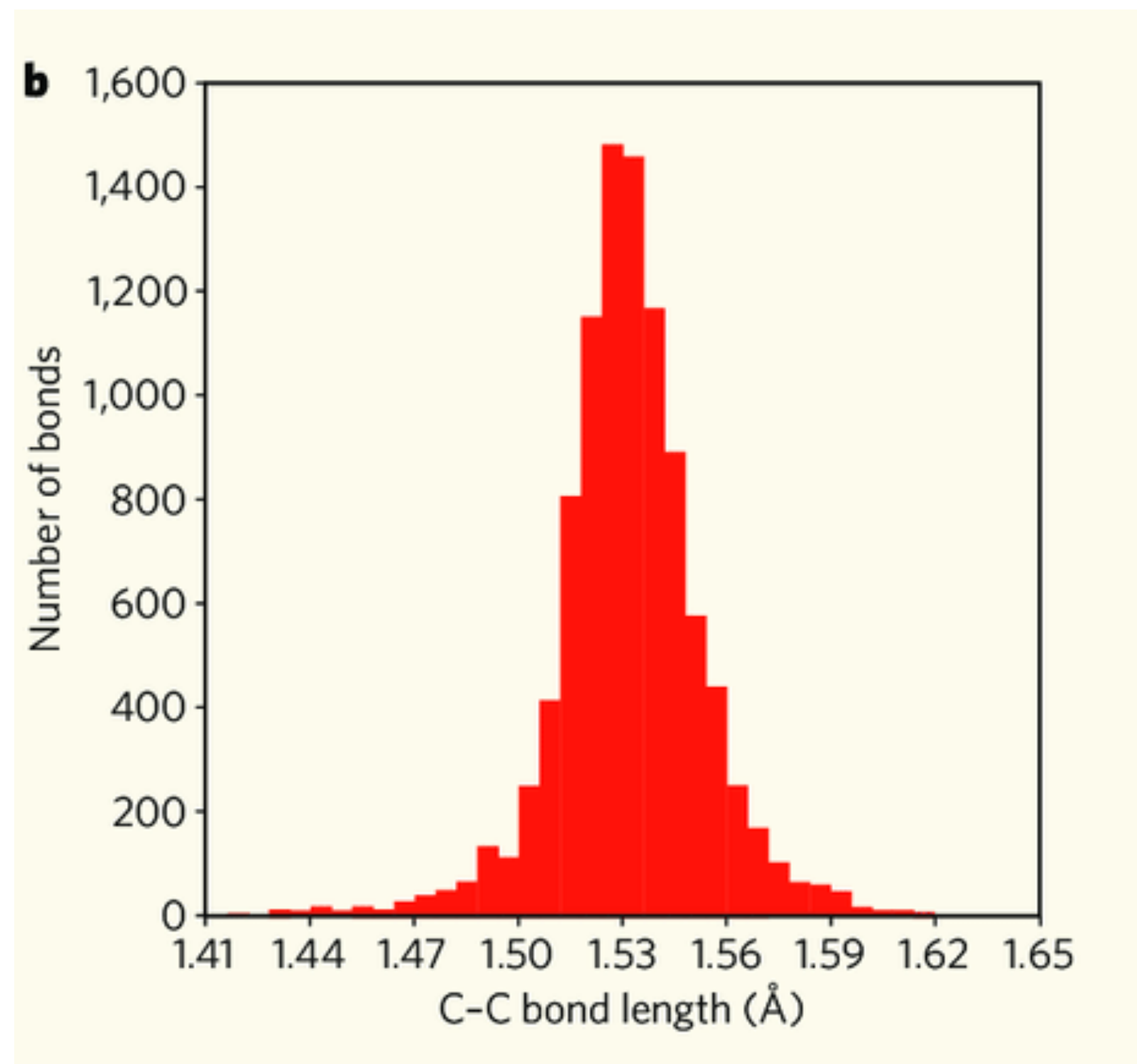
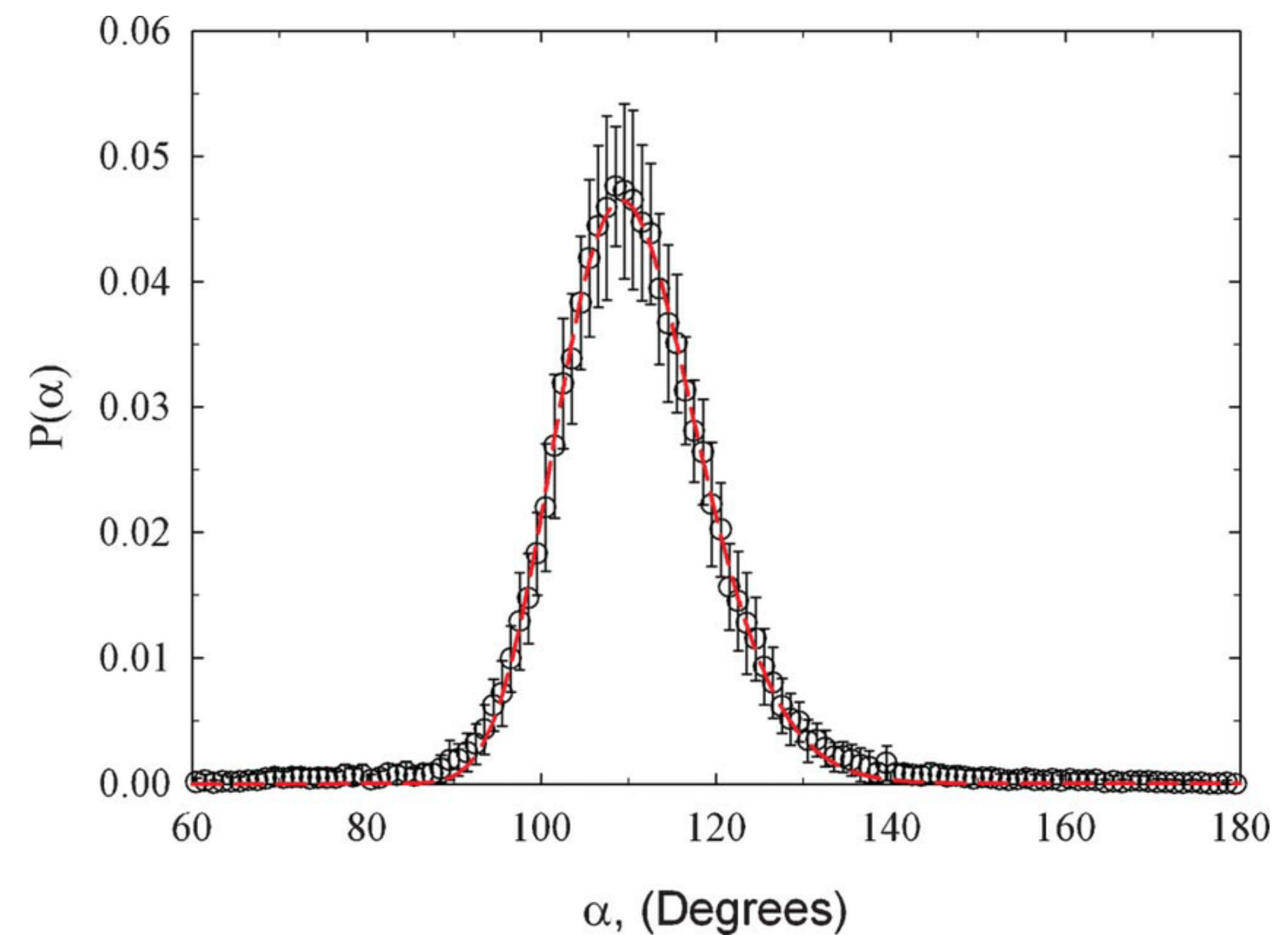


Image: Janez Stepisnik



C-C-C angle distribution of sp³ carbon
Kowalczyk, et al. RSC Advances, 2012.

Unmodeled densities

The image shows a screenshot of the Coot software interface. The main window displays a 3D molecular model with a purple mesh representing unmodeled density. A menu is open over the 'Validate' tab, with 'Unmodelled blobs...' selected. To the right, a panel titled 'Unmodelled blobs of density:' lists 21 individual blobs, numbered from 1 to 21. The text above the list reads: 'There are unexplained blobs of density (too big to be waters):'. A 'Dismiss' button is located at the bottom right of the panel.

Coot 0.8.9.2-pre EL (revision count 7766)

File Edit Calculate Draw Measures Validate HID About Ligand Extensions

Reset View Display Manager

- Ramachandran Plot
- Kleywegt Plot...
- Incorrect Chiral Volumes...
- Unmodelled blobs...
- Difference Map Peaks...
- Check/Delete Waters...
- Geometry analysis
- Peptide omega analysis
- Temp. fact. variance analysis
- Average Temp. fact. analysis
- GLN and ASN B-factor Outliers
- Rotamer analysis
- Density fit analysis
- Probe clashes
- NCS Differences
- Highly coordinated waters...
- Pukka Puckers...?
- Alignment vs PIR...

Unmodelled blobs of density:

There are unexplained blobs of density (too big to be waters):

Blob 1
Blob 2
Blob 3
Blob 4
Blob 5
Blob 6
Blob 7
Blob 8
Blob 9
Blob 10
Blob 11
Blob 12
Blob 13
Blob 14
Blob 15
Blob 16
Blob 17
Blob 18
Blob 19
Blob 20
Blob 21

Dismiss

(mol no: 3) CA (1/4/160 ASP occ: 1.00 bf: 83.75 ele: C pos: (166.29 151.54 102.31)

Unmodeled densities

The image shows a screenshot of the Coot software interface. The main window displays a 3D molecular model with a density map. A dialog box titled "Find Unmodelled Blobs of density" is open, allowing the user to select a map and protein model, and specify a threshold for finding blobs. The threshold is set to 5 r.m.s.d. The "Find Blobs" button is highlighted with a red box. To the right, a panel titled "Unmodelled blobs of density:" lists 21 unexplained blobs, from Blob 1 to Blob 21. A "Dismiss" button is located at the bottom right of this panel.

Coot 0.8.9.2-pre EL (revision count 7766)

File Edit Calculate Draw Measures Validate HID About Ligand Extensions

Reset View Display Manager Sphere Refine +

Find Unmodelled Blobs of density

Select Map:

- 4 /Users/dcekiert/Synch/PDBs/YebT_20180804/ring24_long/ring24_long_m24C6T30_cl2_locfilt.mrc

Select Protein Model:

- 0 /Users/dcekiert/Synch/PDBs/YebT_20180804/Full_close/FullClose_hexa-r10-coot-227_new2.pdb
- 1 /Users/dcekiert/Synch/PDBs/MlaD_hexameric_w-grafted-125loop_ABCDEF.pdb
- 2 /Users/dcekiert/xtal/bex2628_Pr1307/refinement/phenix1/bex2628_phenix1_001.pdb
- 3 /Users/dcekiert/Synch/PDBs/YebT_20180804/ring24_long/long_ring2to4hexa-r09-coot-161.pdb

Find Blobs above:

5 r.m.s.d.

Find Blobs Cancel

Unmodelled blobs of density:

There are unexplained blobs of density (too big to be waters):

- Blob 1
- Blob 2
- Blob 3
- Blob 4
- Blob 5
- Blob 6
- Blob 7
- Blob 8
- Blob 9
- Blob 10
- Blob 11
- Blob 12
- Blob 13
- Blob 14
- Blob 15
- Blob 16
- Blob 17
- Blob 18
- Blob 19
- Blob 20
- Blob 21

Dismiss

(mol no: 3) CA (1/160 ASP occ: 1.00 bf: 83.75 ele: C pos: (166.29 151.54 102.31)

Unmodeled densities

The image shows a screenshot of the Coot software interface. The main window displays a 3D molecular model with a purple wireframe density map overlaid. The model is colored in green, blue, and red. A label 'CA 160 ASP/A' is visible on the model. The interface includes a menu bar with options like File, Edit, Calculate, Draw, Measures, Validate, HID, About, Ligand, and Extensions. Below the menu bar are icons for 'Reset View', 'Display Manager', and 'Sphere Refine +'. A coordinate system (x, y, z) is shown in the top left corner of the main window.

On the right side, there is a panel titled 'Unmodelled blobs of density:'. It contains the text: 'There are unexplained blobs of density (too big to be waters):'. Below this text is a list of 21 blobs, labeled 'Blob 1' through 'Blob 21'. A mouse cursor is pointing at 'Blob 13'. At the bottom right of the panel is a 'Dismiss' button.

At the bottom of the main window, there is a status bar with the following text: '(mol. no: 2) CA 160 ASP/A; 1.00 bf: 83.75; 6 pos: (166.29, 151.54, 102.31)'

Molprobity score and clash score

Summary statistics

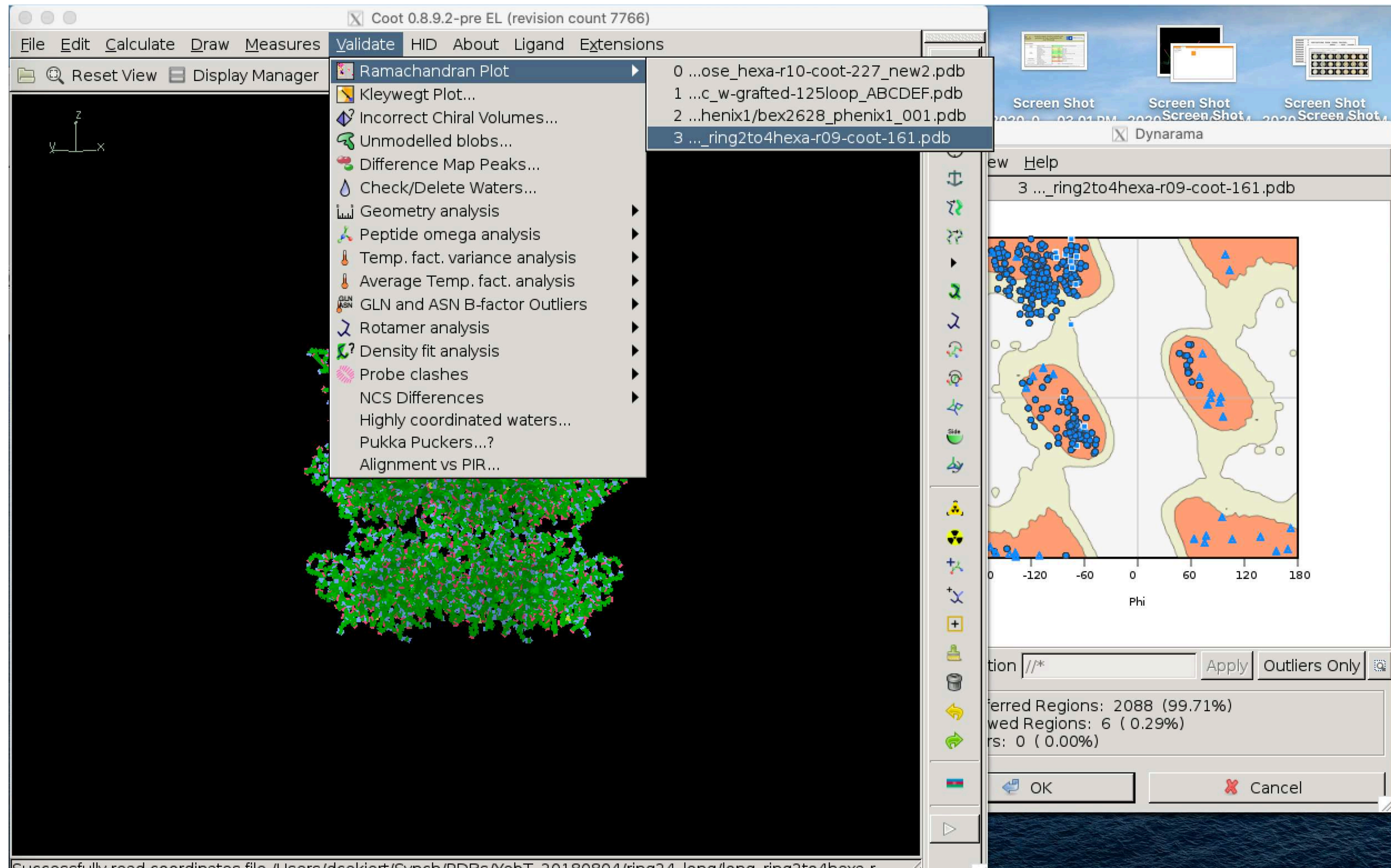
All-Atom Contacts	Clashscore, all atoms:	3.82		96 th percentile* (N=1784, all resolutions)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	12	0.68%	Goal: <0.3%
	Favored rotamers	1674	95.22%	Goal: >98%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	2070	98.85%	Goal: >98%
	MolProbity score [^]	1.17		99 th percentile* (N=27675, 0Å - 99Å)
	Cβ deviations >0.25Å	6	0.32%	Goal: 0
	Bad bonds:	0 / 16212	0.00%	Goal: 0%
	Bad angles:	24 / 21996	0.11%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 114	0.00%	Expected: ≤1 per chain, or ≤5%
Low-resolution Criteria	CaBLAM outliers	42	2.0%	Goal: <1.0%
	CA Geometry outliers	24	1.15%	Goal: <0.5%
Additional validations	Pseudochiral naming errors	6		
	Waters with clashes	0/0	0.00%	See UnDowser table for details

In the two column results, the left column gives the raw count, right column gives the percentage.

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.

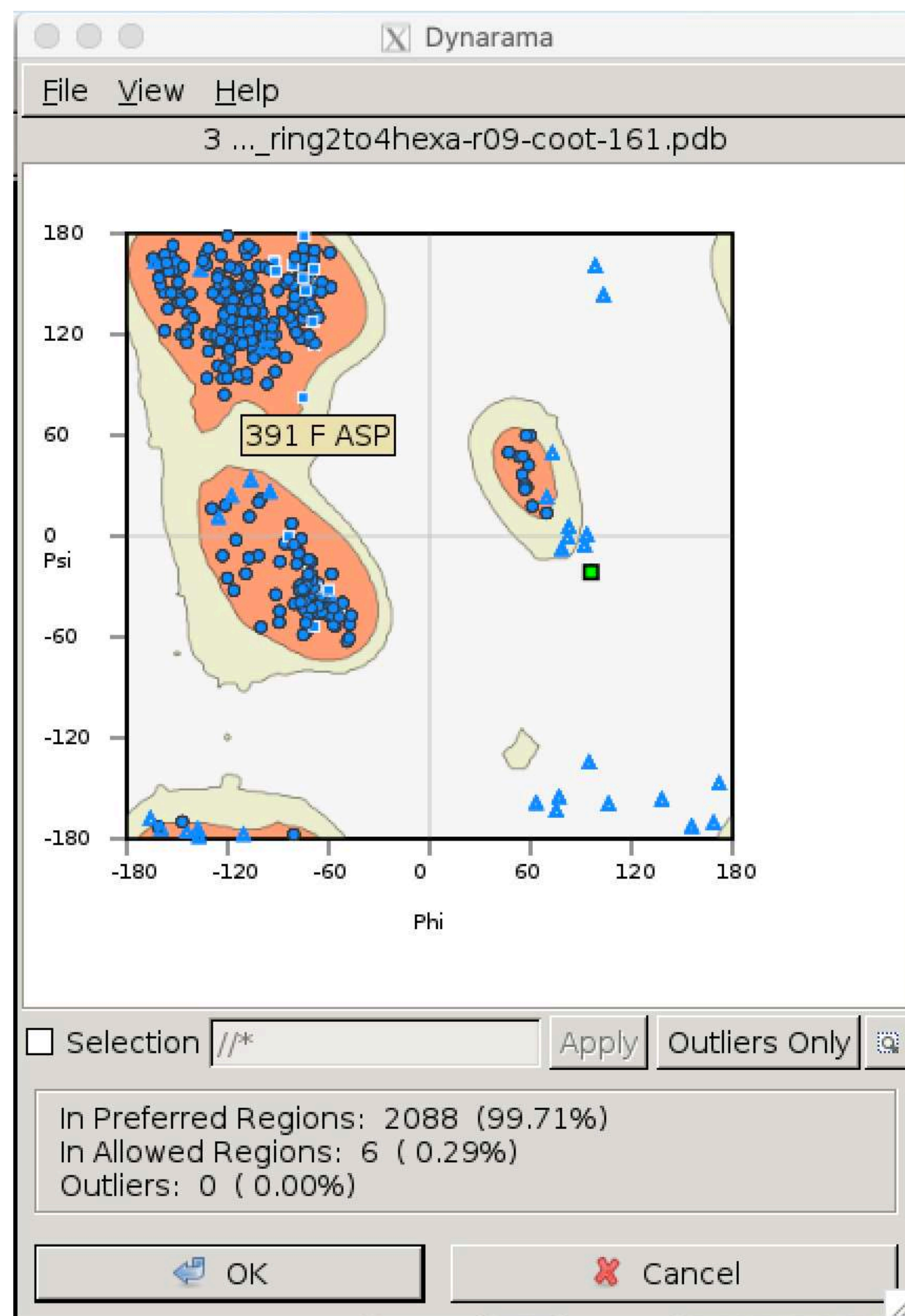
[^] MolProbity score combines the clashscore, rotamer, and Ramachandran evaluations into a single score, normalized to be on the same scale as X-ray resolution.

Ramachandran plot

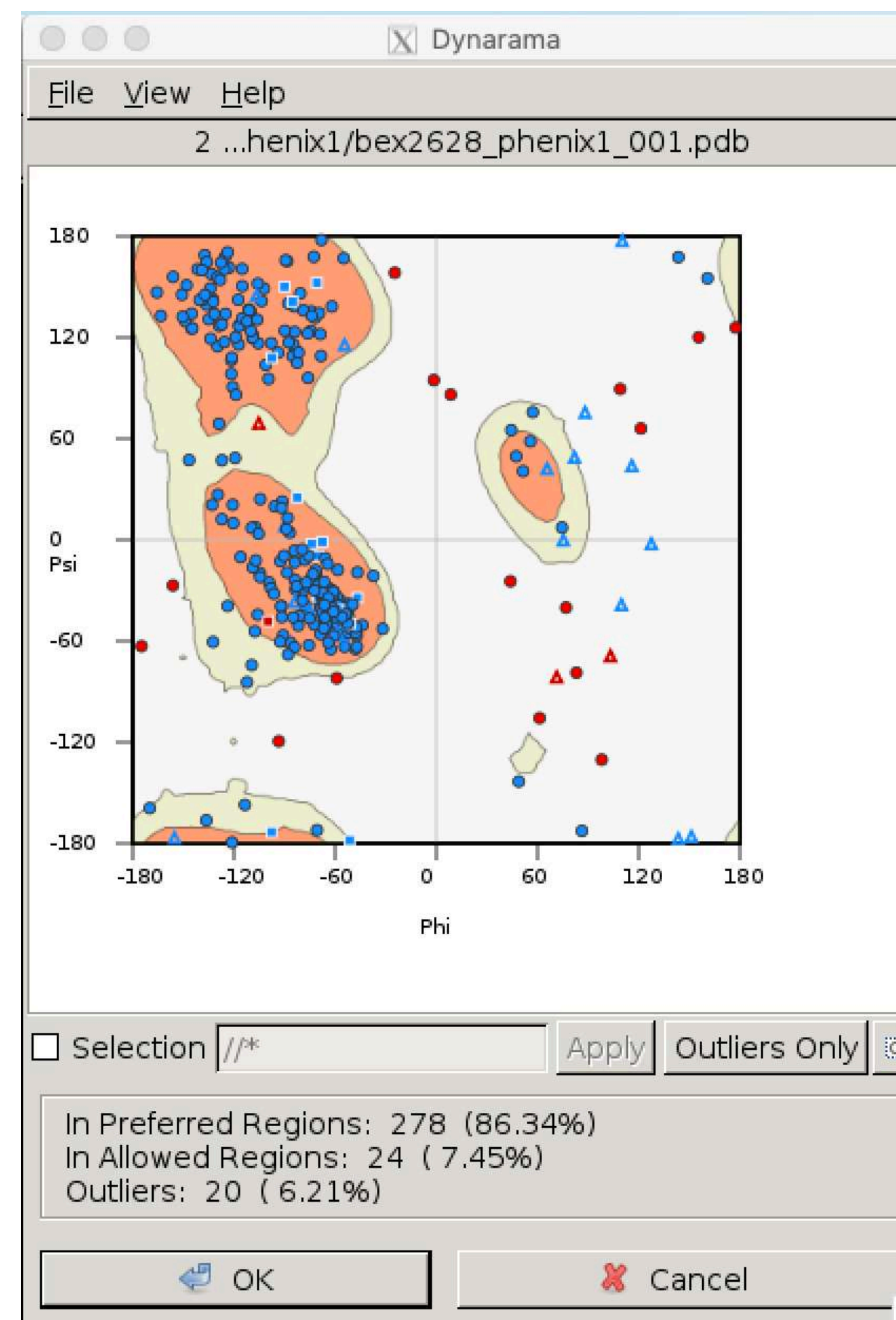


Ramachandran plot

Ideal plot

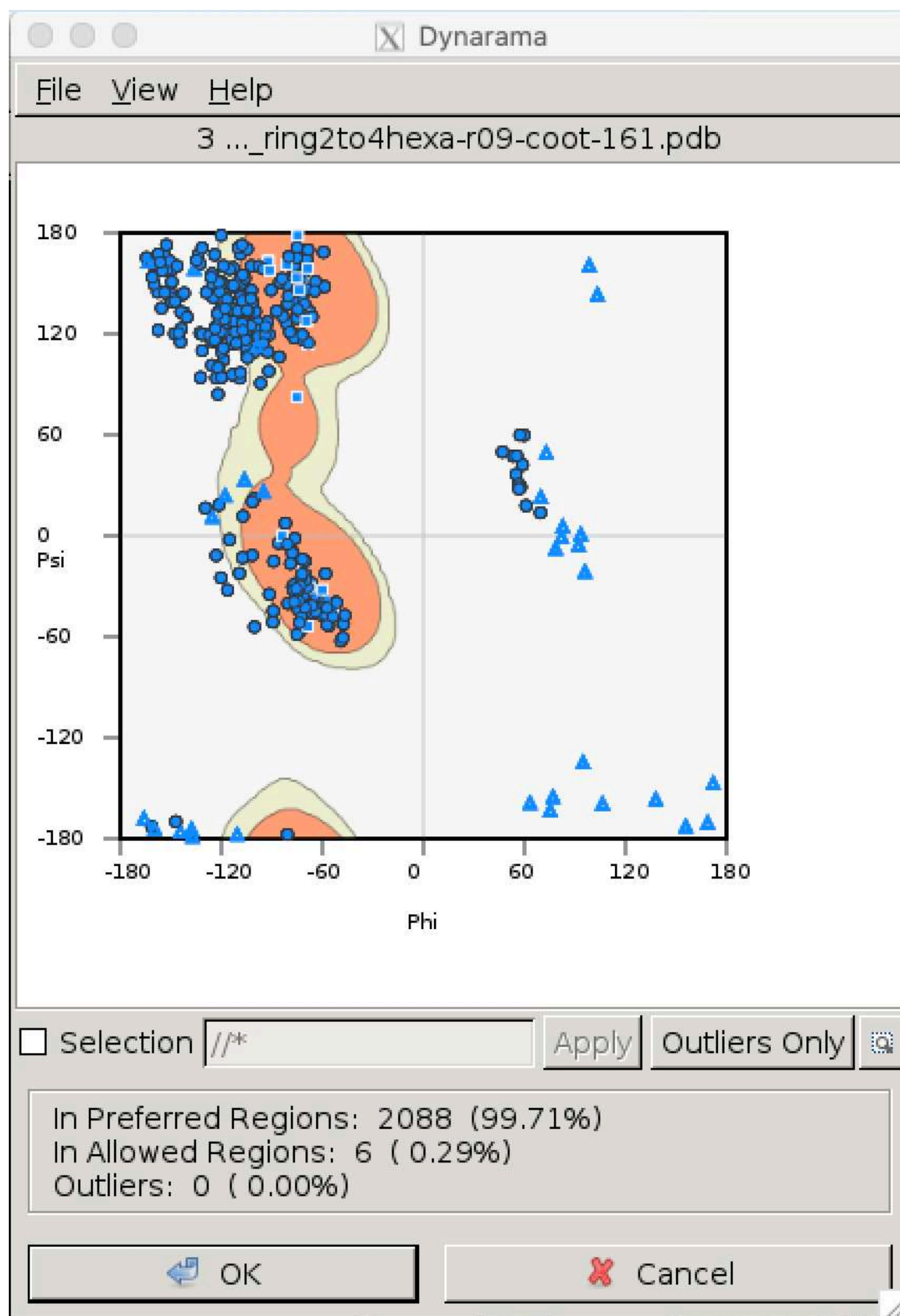


Problematic plot

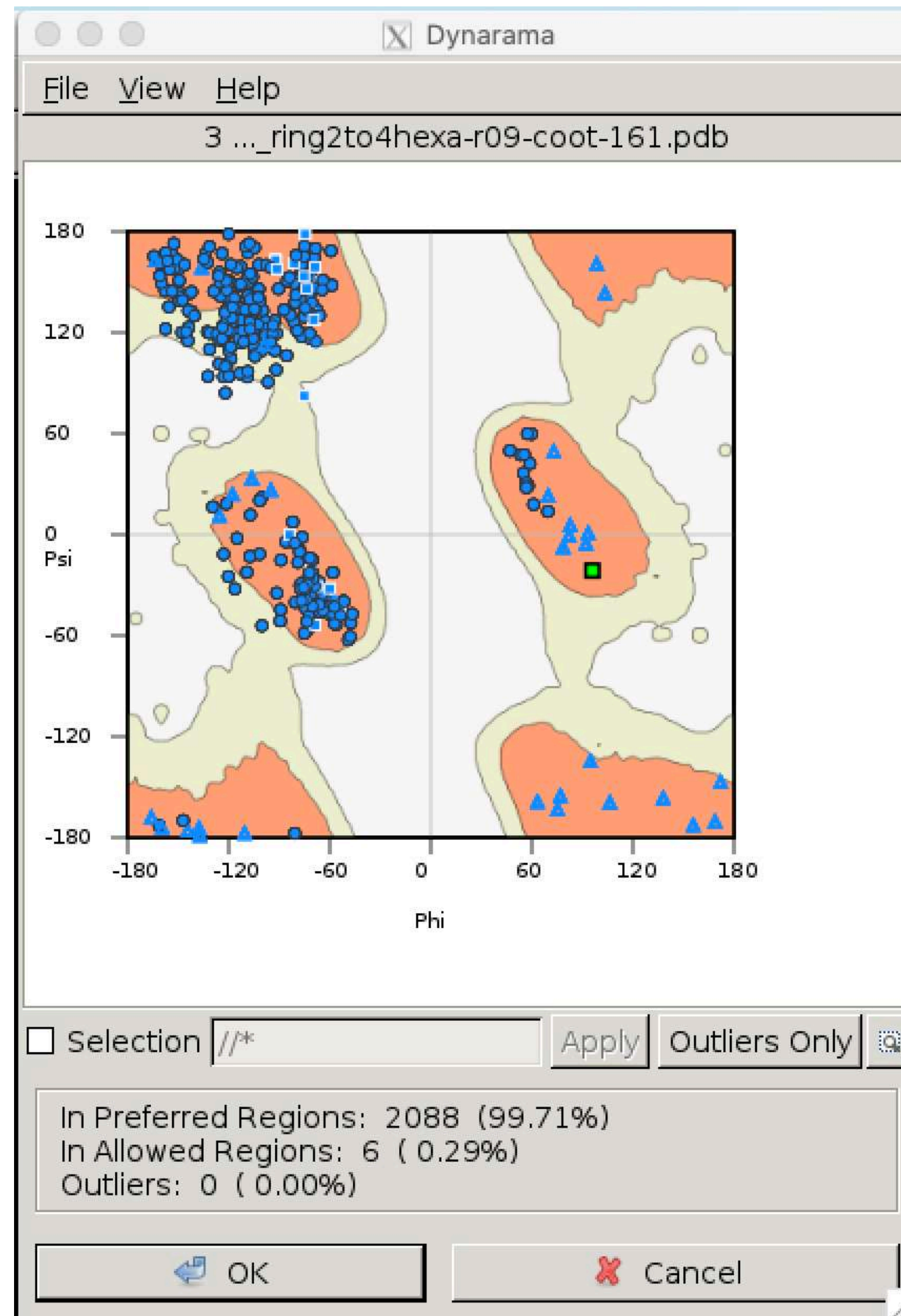


Ramachandran plot

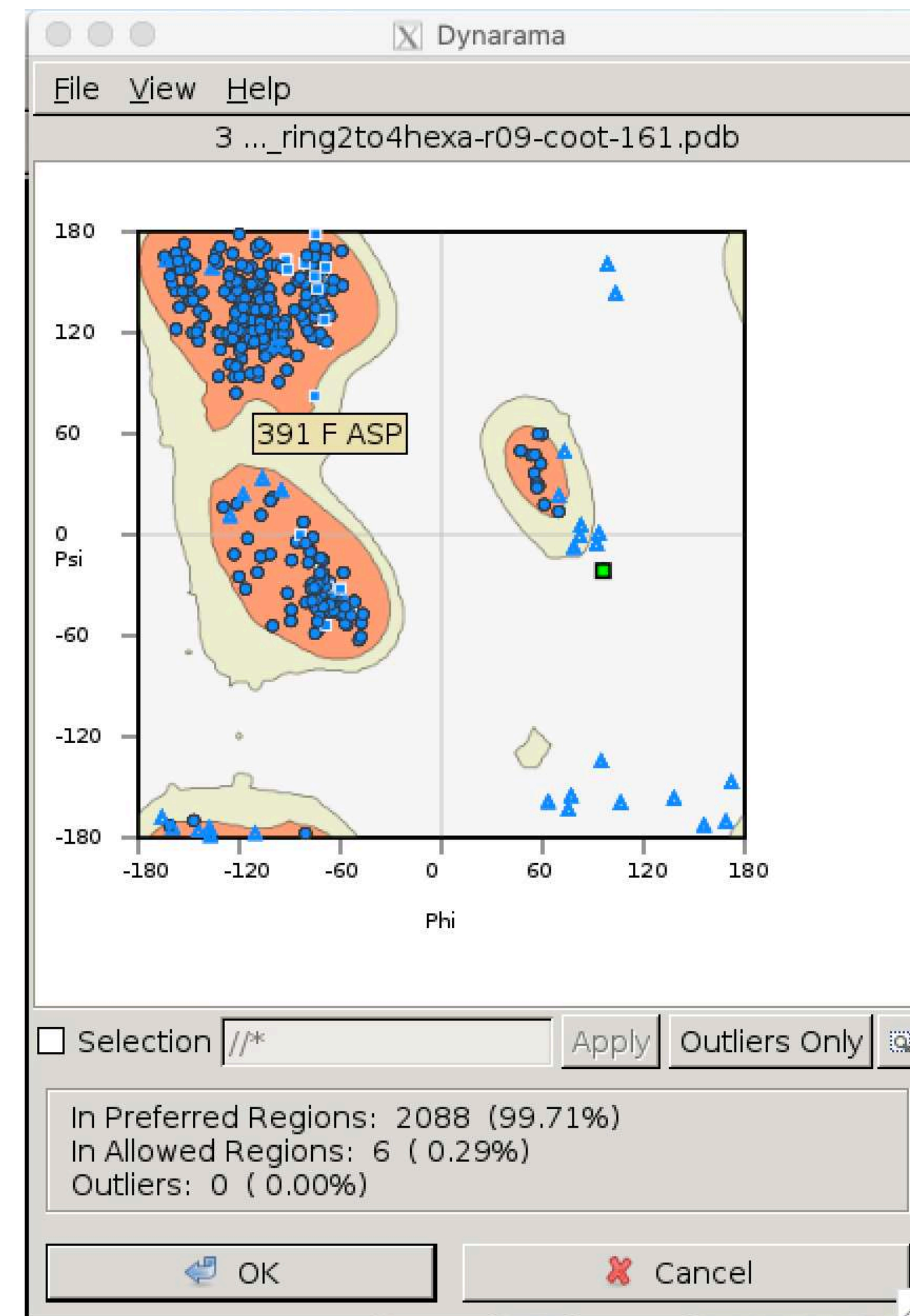
Proline Allowed Regions



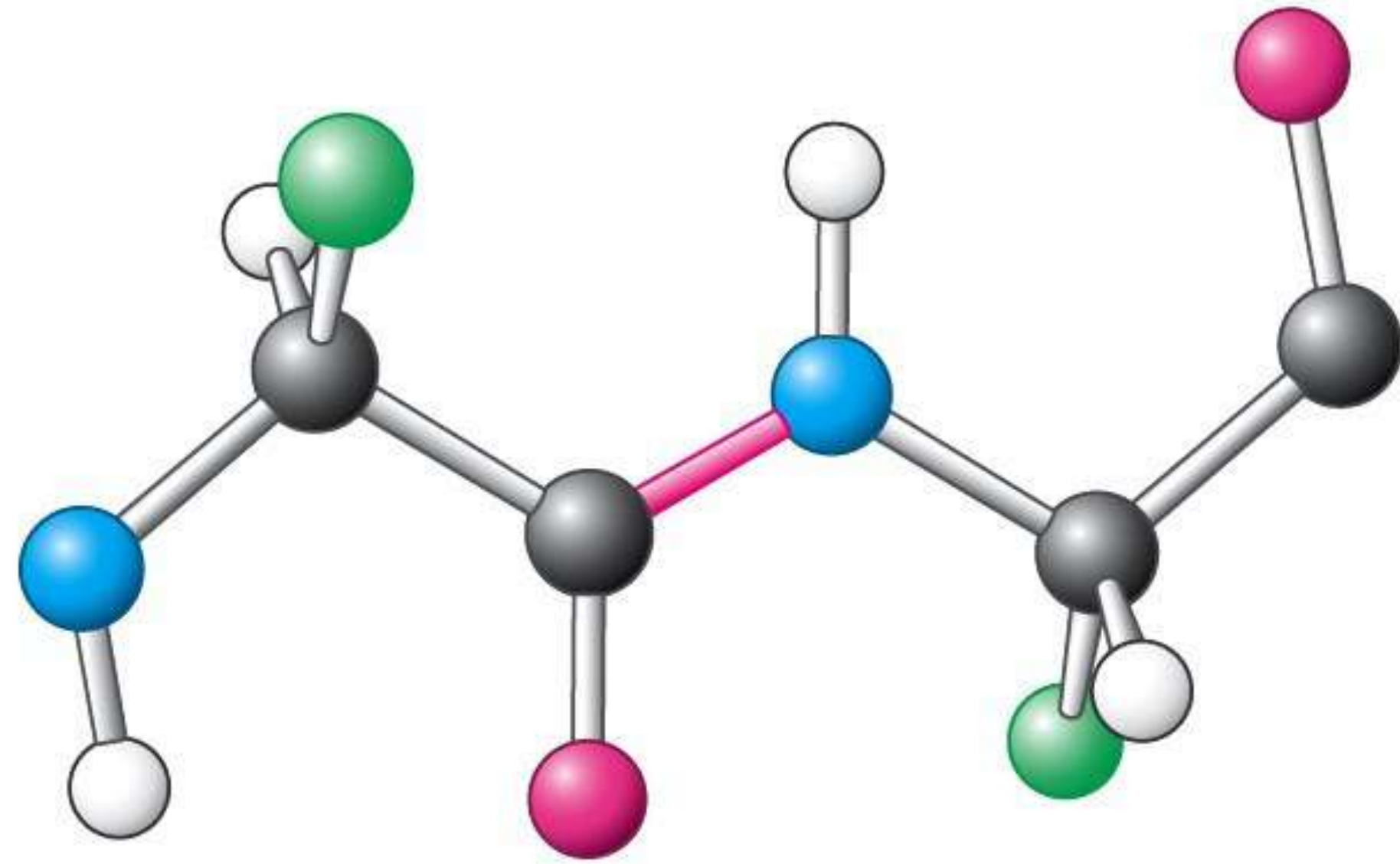
Glycine Allowed Regions



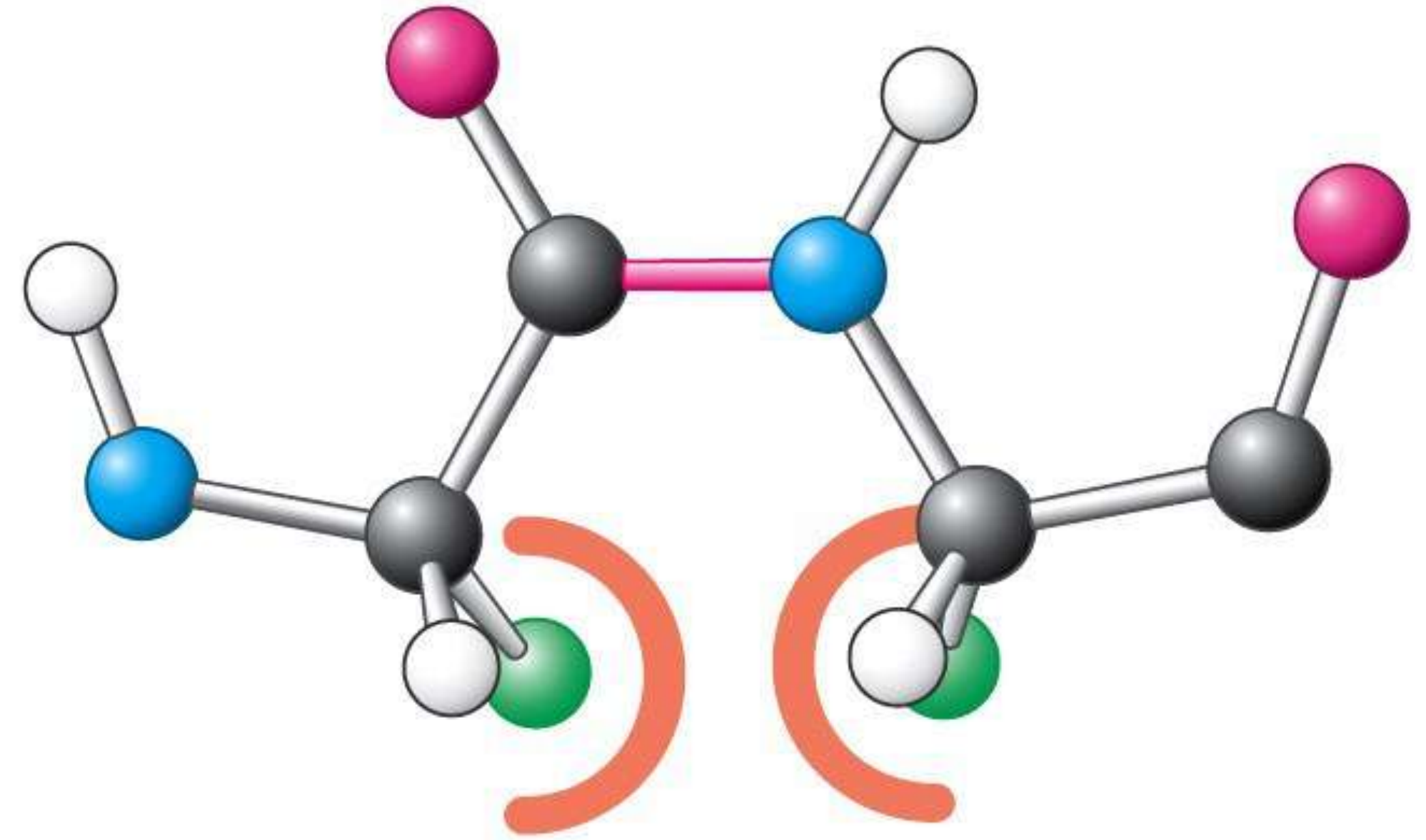
Allowed Regions (non-Pro/Gly)



Cis peptide bonds



Trans



Cis

Figure 2.20
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Cis peptide bonds

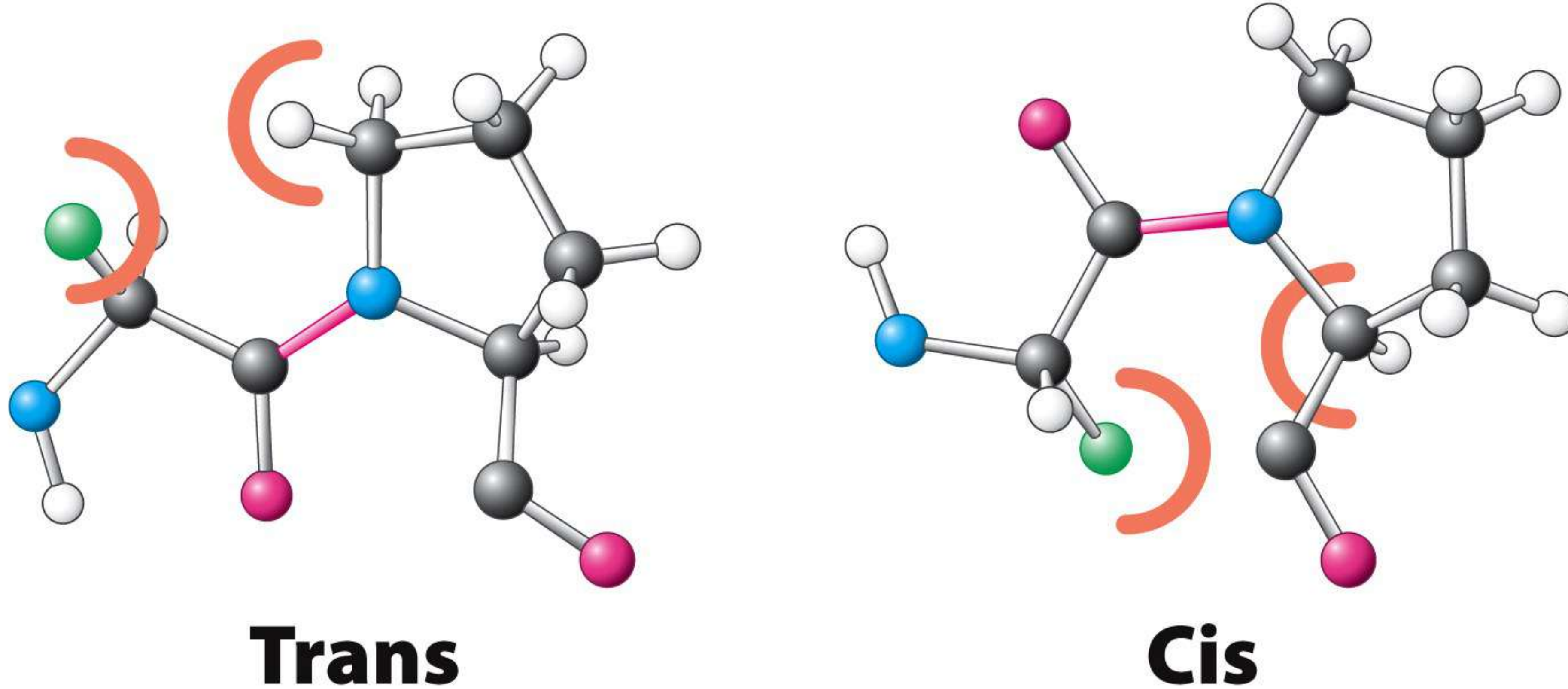
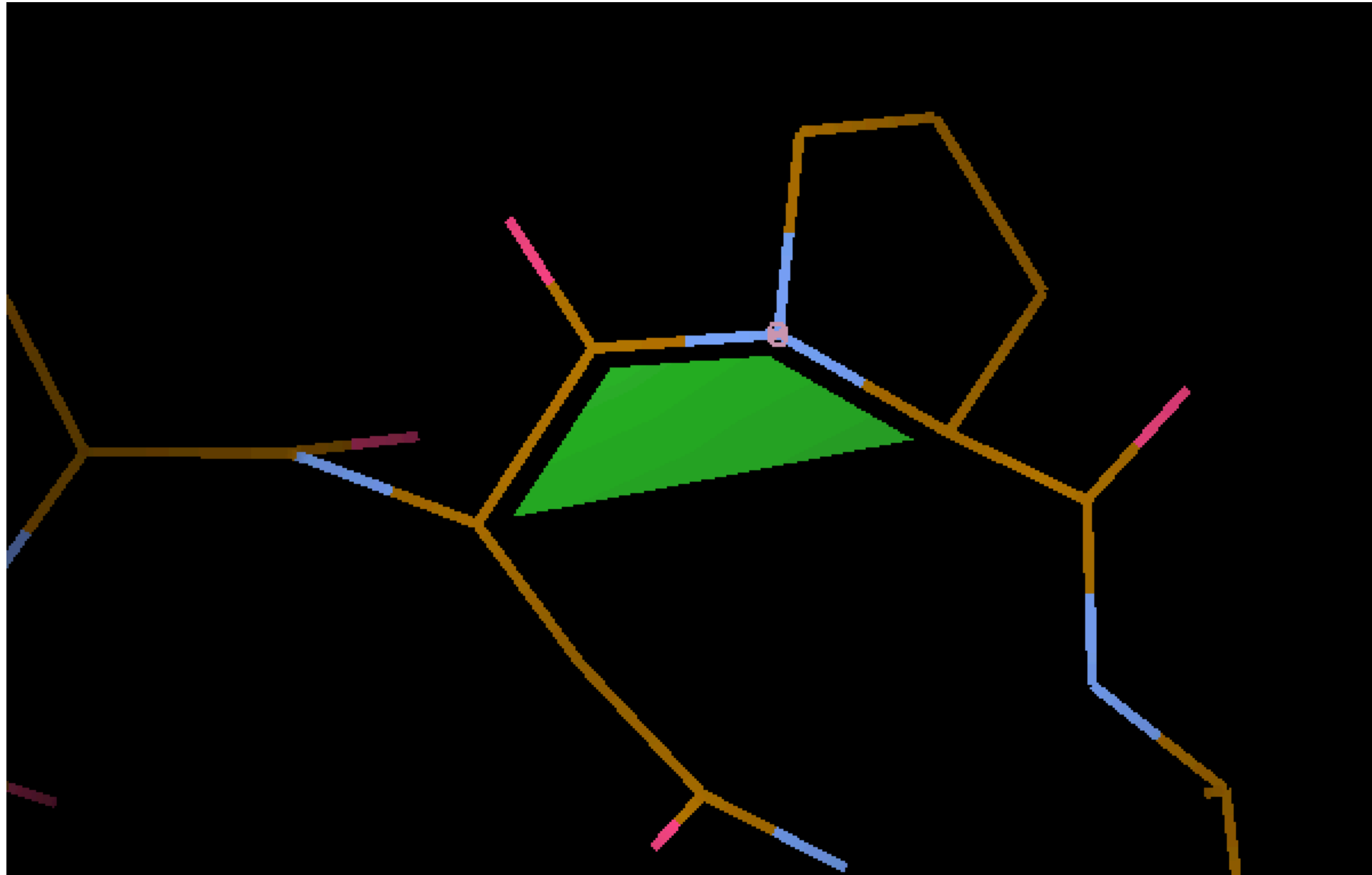


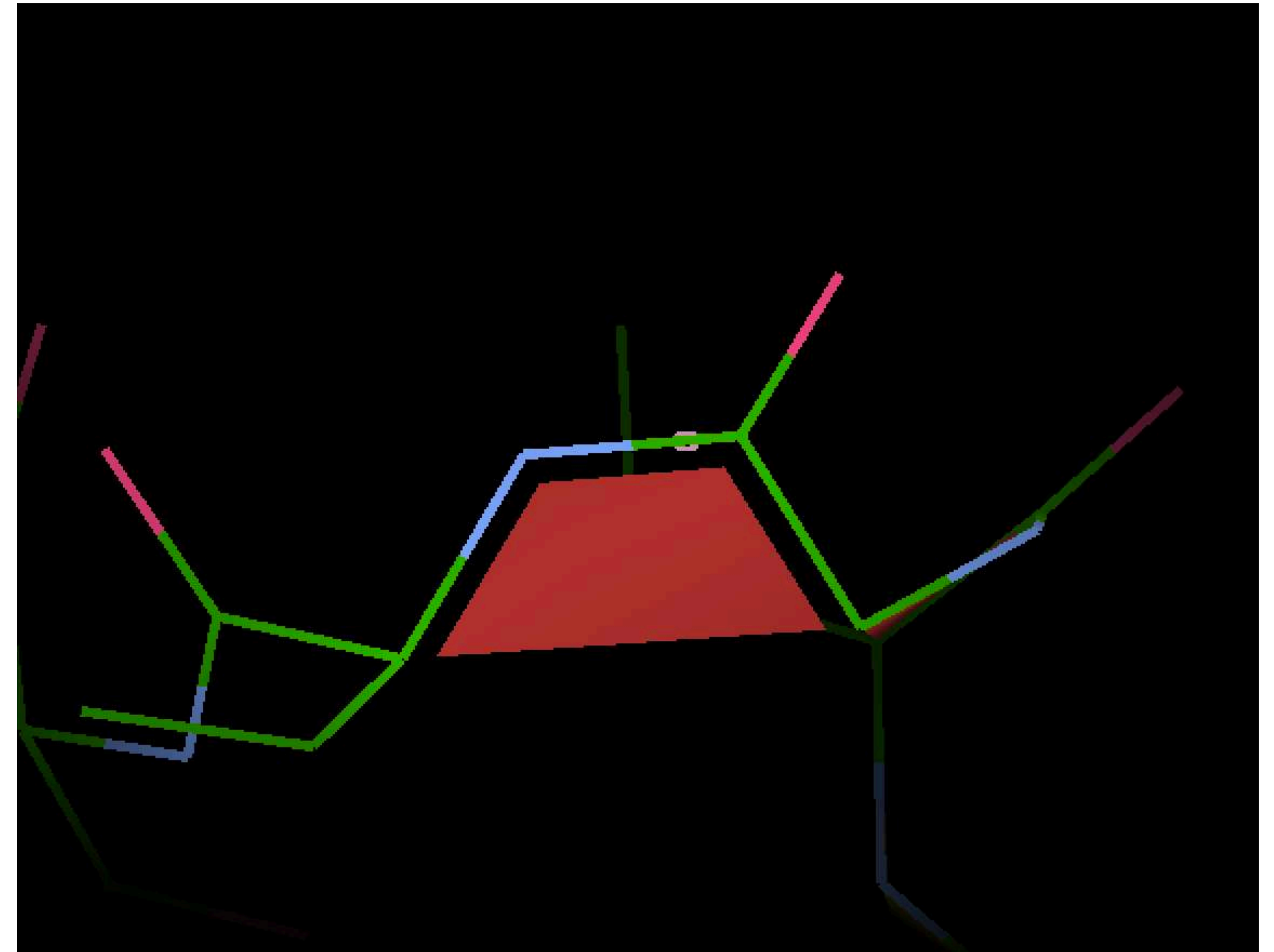
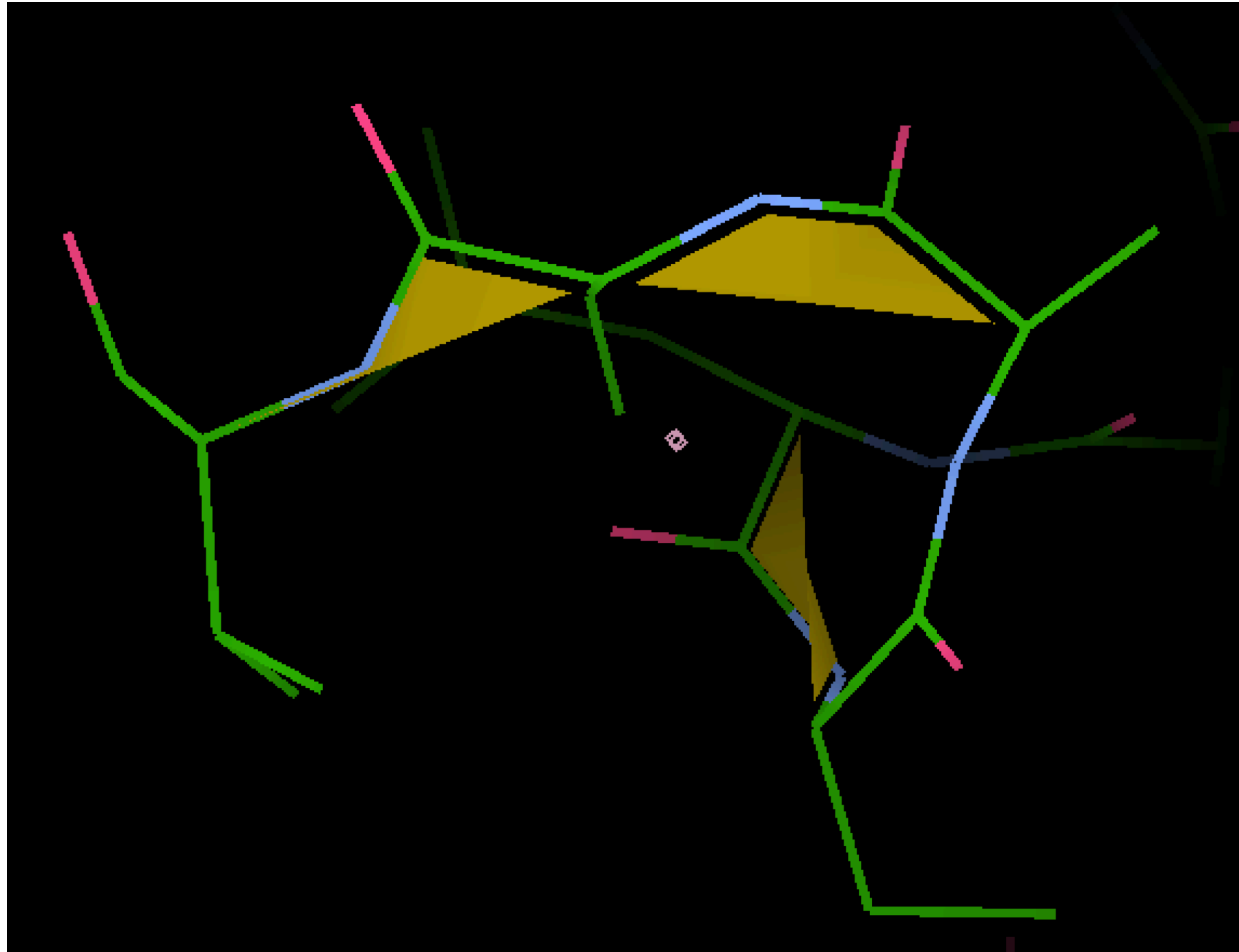
Figure 2.21
Biochemistry, Seventh Edition
© 2012 W. H. Freeman and Company

Cis peptide bonds



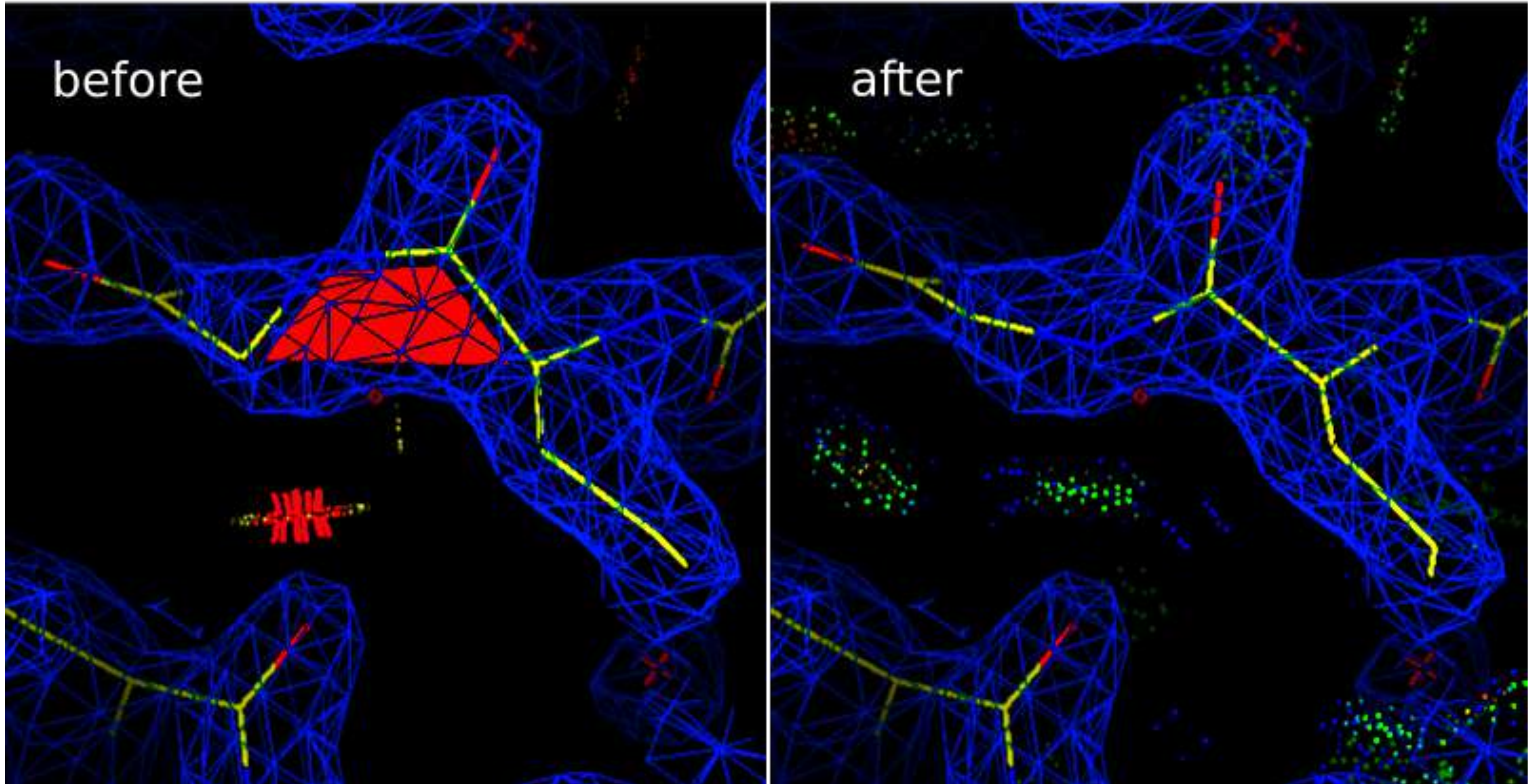
Coot highlights all cis and non-planar peptide bonds, and color codes them to make potential problems easy to ID
Green = cis-Proline (probably OK); Yellow = non-planar peptide bond (check!); Red = non-proline cis peptide bond (check!)

Cis peptide bonds



Coot highlights all cis and non-planar peptide bonds, and color codes them to make potential problems easy to ID
Green = cis-Proline (probably OK); Yellow = non-planar peptide bond (check!); Red = non-proline cis peptide bond (check!)

Cis peptide bonds



Rotamer outliers: Coot

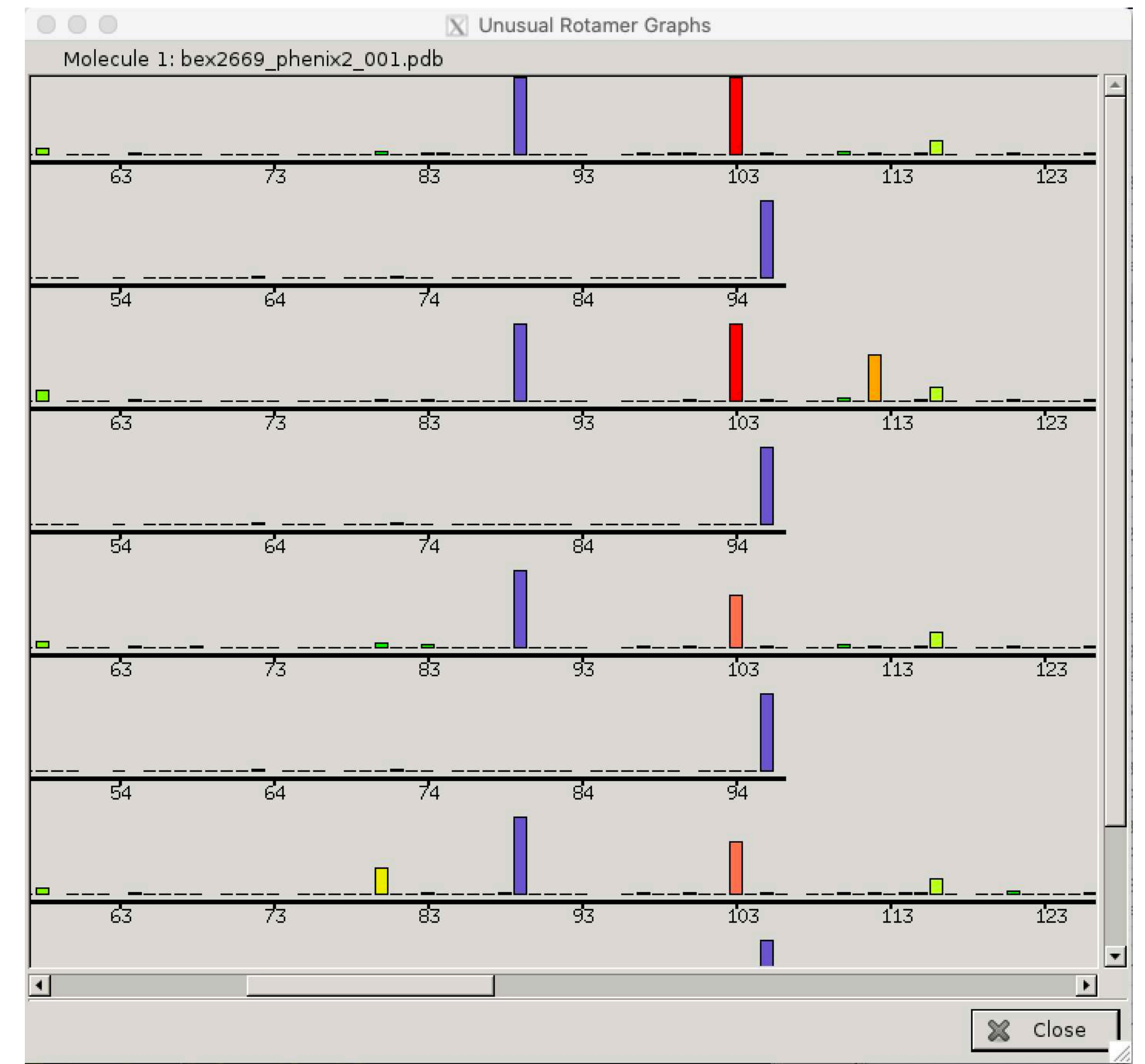
Coot 0.8.9.2-pre EL (revision count 7766)

File Edit Calculate Draw Measures Validate HID About Ligand Extensions

Reset View Display Manager

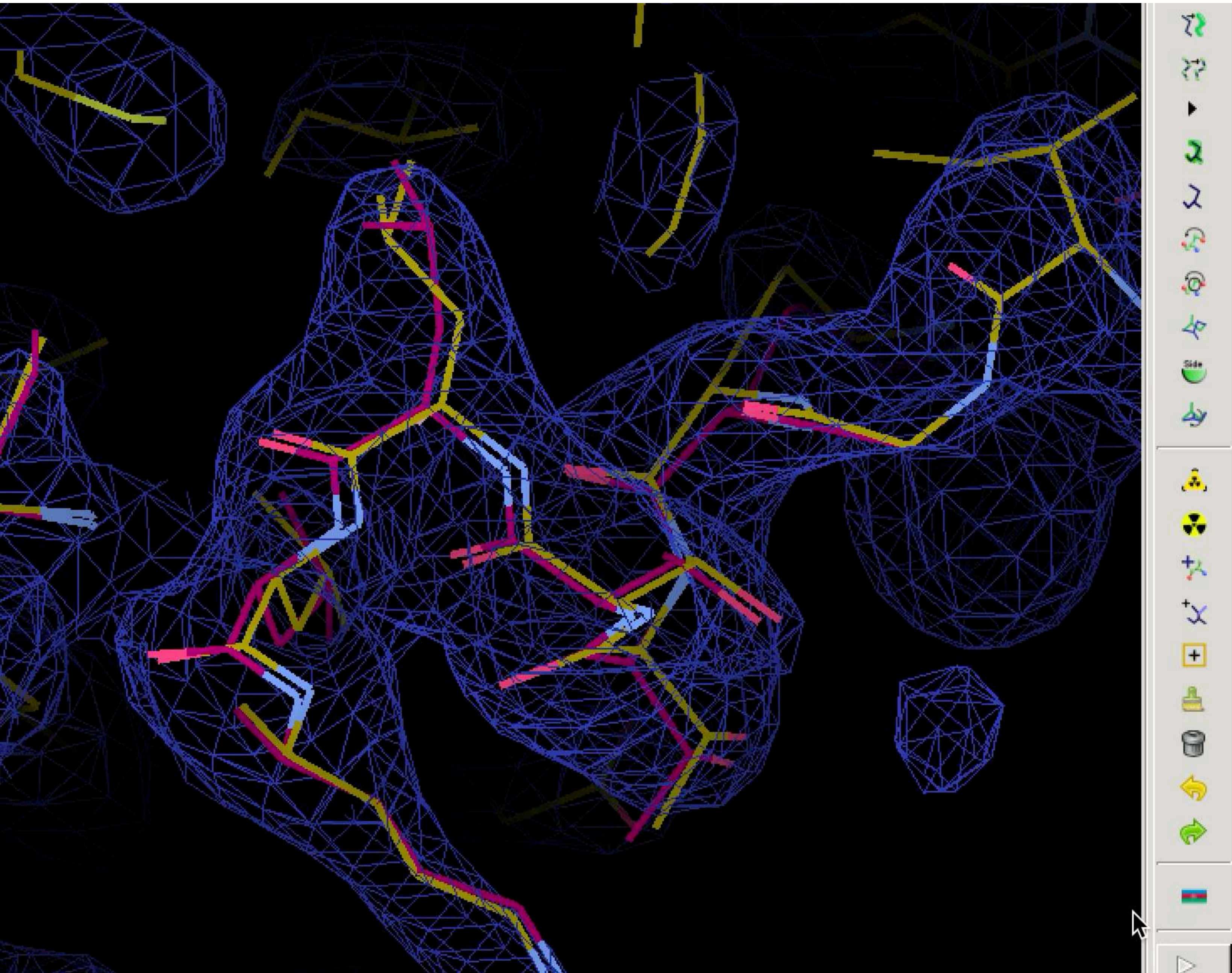
- Ramachandran Plot
- Kleywegt Plot...
- Incorrect Chiral Volumes...
- Unmodelled blobs...
- Difference Map Peaks...
- Check/Delete Waters...
- Geometry analysis
- Peptide omega analysis
- Temp. fact. variance analysis
- Average Temp. fact. analysis
- GLN and ASN B-factor Outliers
- Rotamer analysis**
 - 0 ..._ring2to4hexa-r09-coot-161.pdb
 - 1 ...henix2/bex2669_phenix2_001.pdb
- Density fit analysis
- Probe clashes
- NCS Differences
- Highly coordinated waters...
- Pukka Puckers...?
- Alignment vs PIR...

Successfully read coordinates file /Users/dcekiert/xtal/bex2669_Pr1307/refinement_20191126/phenix2/bex2669_phenix2_001....



Red/Tall - Rotamer outlier
Green/Short - Happy rotamer
Lilac(?): Missing side chain atoms

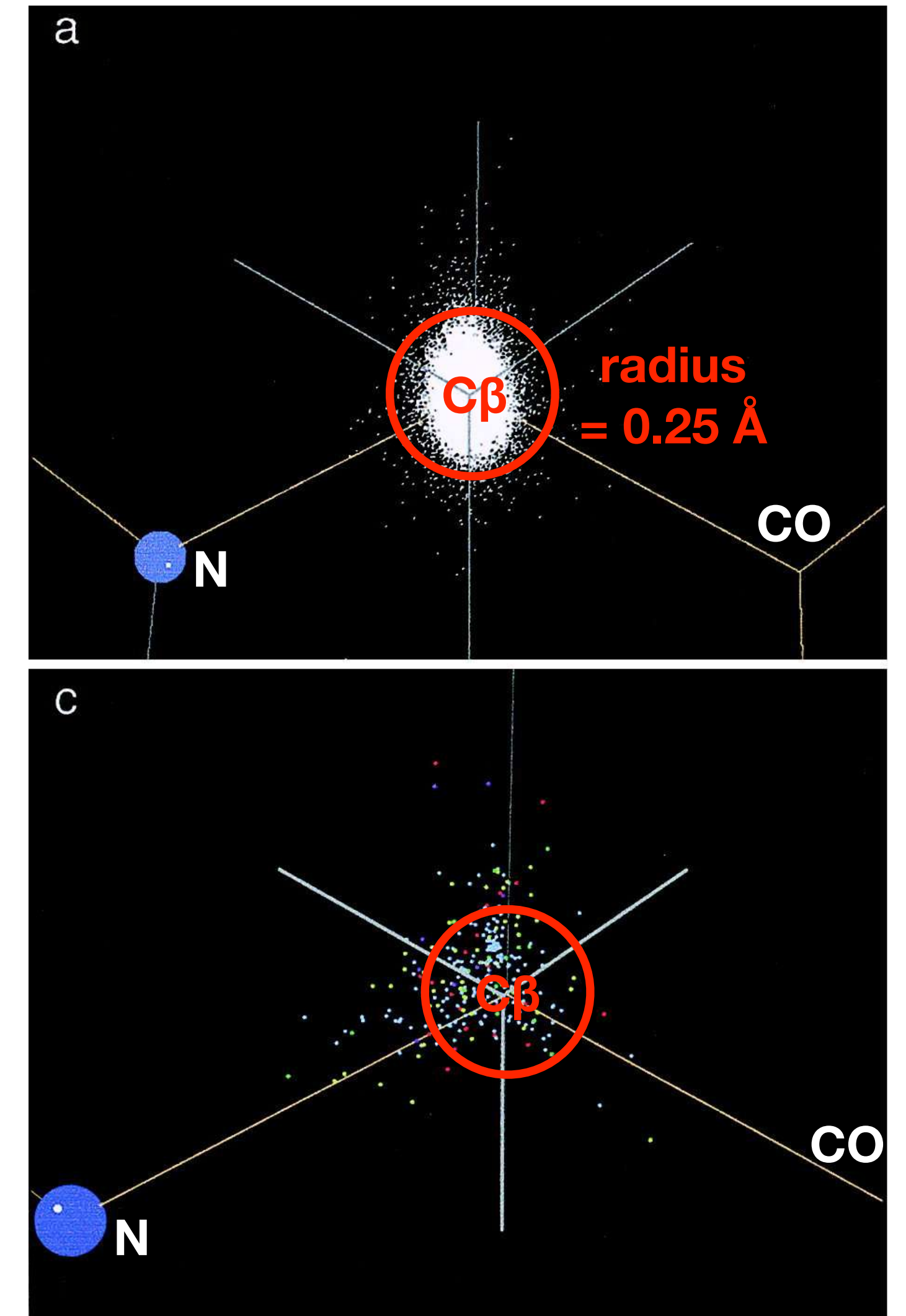
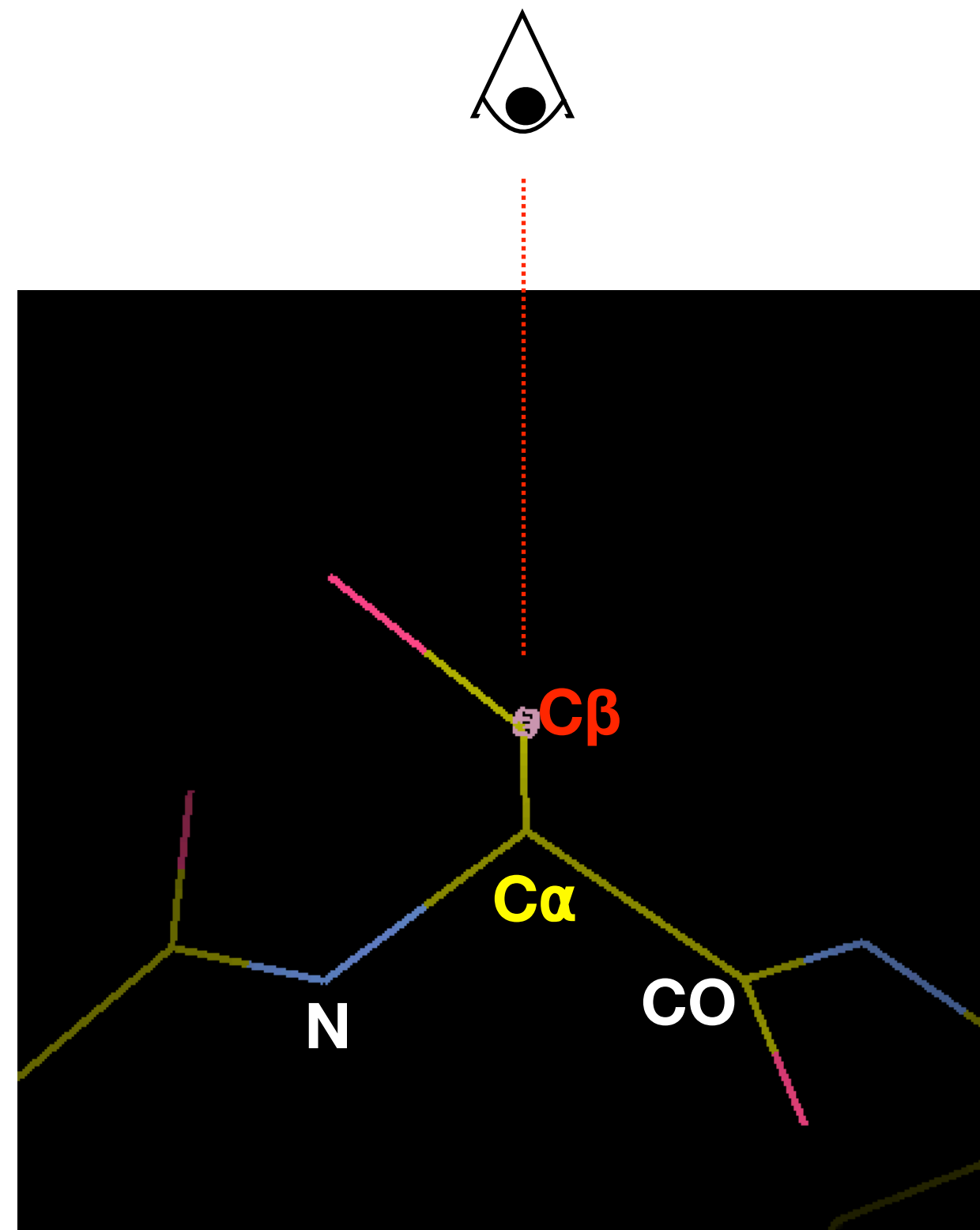
Cbeta deviations



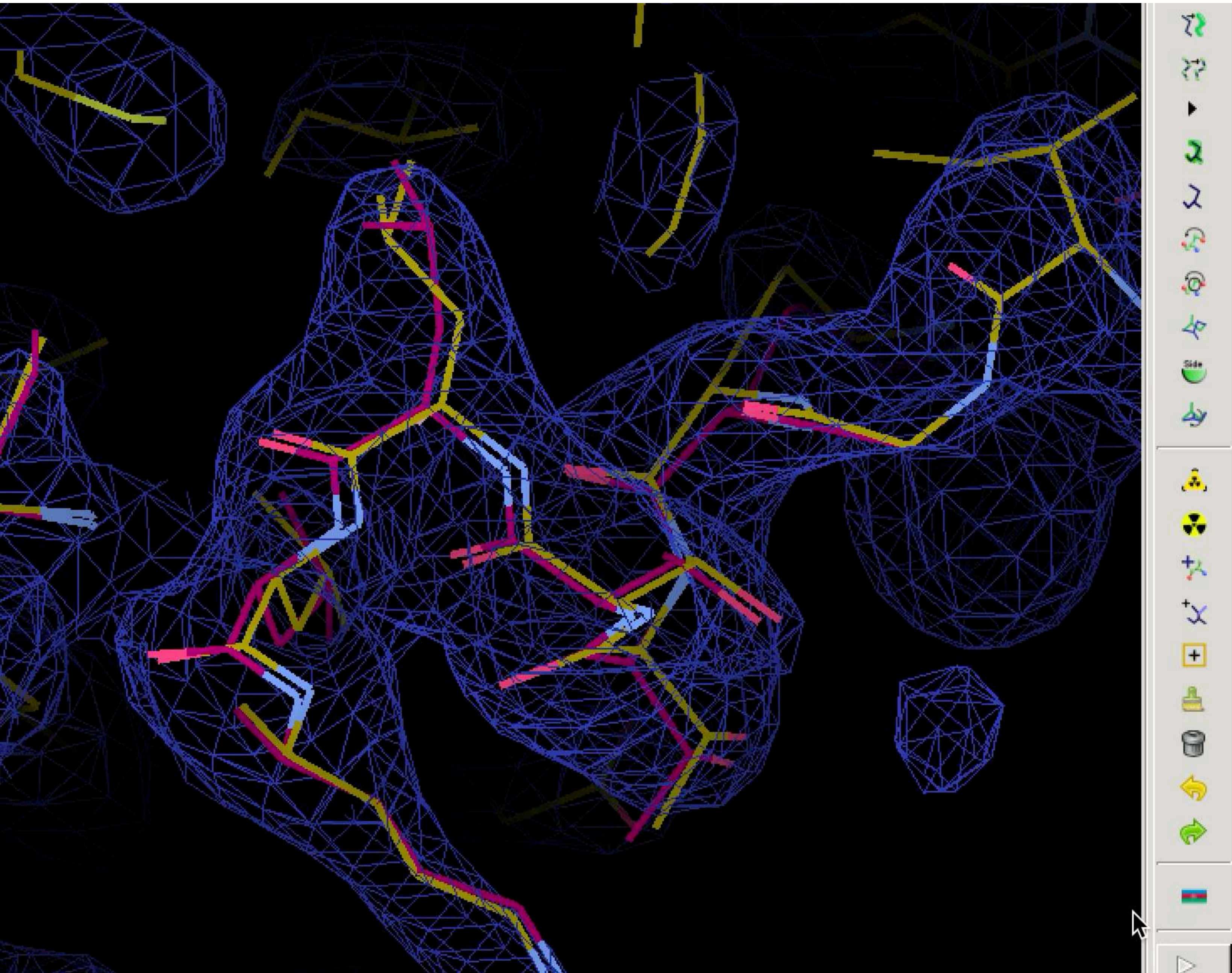
Which Leu rotamer is correct???

Cbeta deviations

Cbeta deviations report on a combination of backbone and side chain problems, frequently when an incorrect side chain rotamer is leading to a distortion of the backbone conformation as well.



Cbeta deviations



Which Leu rotamer is correct???

Correct answer:
YELLOW

ADP outliers

Summary		MolProbity		Model vs. Data		Data	
Composition (#)				Box			
Chains	6	Lengths (Å)	113.97, 103.49, 120.52				
Atoms	15942 (Hydrogens: 0)	Angles (°)	90.00, 90.00, 90.00				
Residues	Protein: 2106 Nucleotide: 0	Supplied Resolution (Å)	3.0				
Water	0	Resolution Estimates (Å)	Masked	Unmasked			
Ligands	0	d FSC (half maps; 0.143)	---	---			
Bonds (RMSD)				d 99 (full/half1/half2)	3.4/---/---		3.4/---/---
Length (Å) (# > 4σ)	0.002 (0)	d model	3.3			3.4	
Angles (°) (# > 4σ)	0.573 (0)	d FSC model (0/0.143/0.5)	3.0/3.1/3.4		3.1/3.2/3.4		
MolProbity score	2.26	Map min/max/mean	-0.29/0.53/0.00				
Clash score	7.67	Model vs. Data					
Ramachandran plot (%)				CC (mask)	0.78		
Outliers	0.00	CC (box)	0.69				
Allowed	4.30	CC (peaks)	0.64				
Favored	95.70	CC (volume)	0.78				
Rotamer outliers (%)	5.12	Mean CC for ligands	---				
Cβ outliers (%)	0.00						
Peptide plane (%)							
Cis proline/general	0.0/0.0						
Twisted proline/general	0.0/0.0						
CaBLAM outliers (%)	2.02						
ADP (B-factors)							
Iso/Aniso (#)	15942/0						
min/max/mean							
Protein	60.42/162.58/90.50						
Nucleotide	---						
Ligand	---						
Water	---						
Occupancy							
Mean	1.00						
occ = 1 (%)	100.00						
0 < occ < 1 (%)	0.00						
occ > 1 (%)	0.00						

Expected mean B-factor at 3-4 Å resolution:
Roughly 100-200 ?
Currently, ability to adjust B-factor parameterization is limited (e.g. group B was individual atoms; TLS)

- If your B factors seem very high or low:
- Check for regions with very high B's; out of density? Weak density? Delete?
 - Try resetting all B's to same low value (e.g., 30) and try refining again.

JCSG QC Server

Date: 03/17/22 16:22:07

The Quality Control Check was developed at the Joint Center for Structural Genomics. It is Maintained by Bridge Structural Biology Center and CARC at USC.
If you use this tool in preparing a structure, please reference this URL: <https://qc-check.usc.edu>

[Refinement Stats](#)

[PDB Check](#)

[Nomenclature Check](#)

[ADIT Results](#)

[Molprobit Results](#)

[NCS Check](#)

[Sequence Check](#)

[Real Space CC](#)

More Resources

- <http://molprobity.biochem.duke.edu/>
- <https://www.phenix-online.org/documentation/index.html>
- <https://www.ccpem.ac.uk/>
- <https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/>
- Afonine, et al. “New tools for the analysis and validation of cryo-EM maps and atomic models” *Acta Cryst.* 2018
- Wang, et al. “Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta” *eLife* 2016
- Nicholls, et al. “Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM” *Acta Cryst.* 2018