# Model Refinement and Validation

Damian Ekiert and Gira Bhabha
NYU School of Medicine
March 18, 2022

# 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

$\Delta G_{unfolding}$

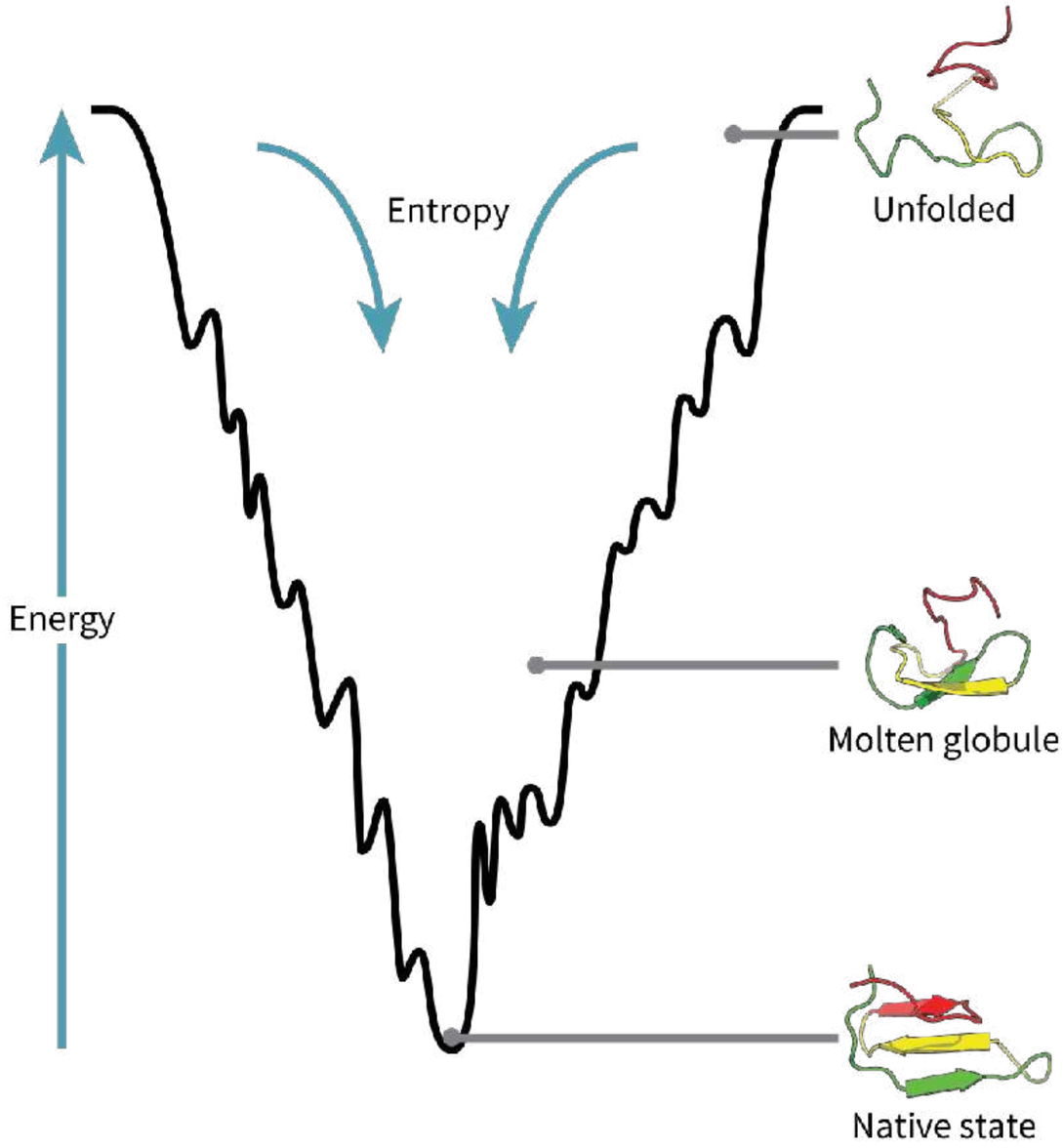



Image: Thomas Splettstoesser

# Refinement target describes differences between model and data

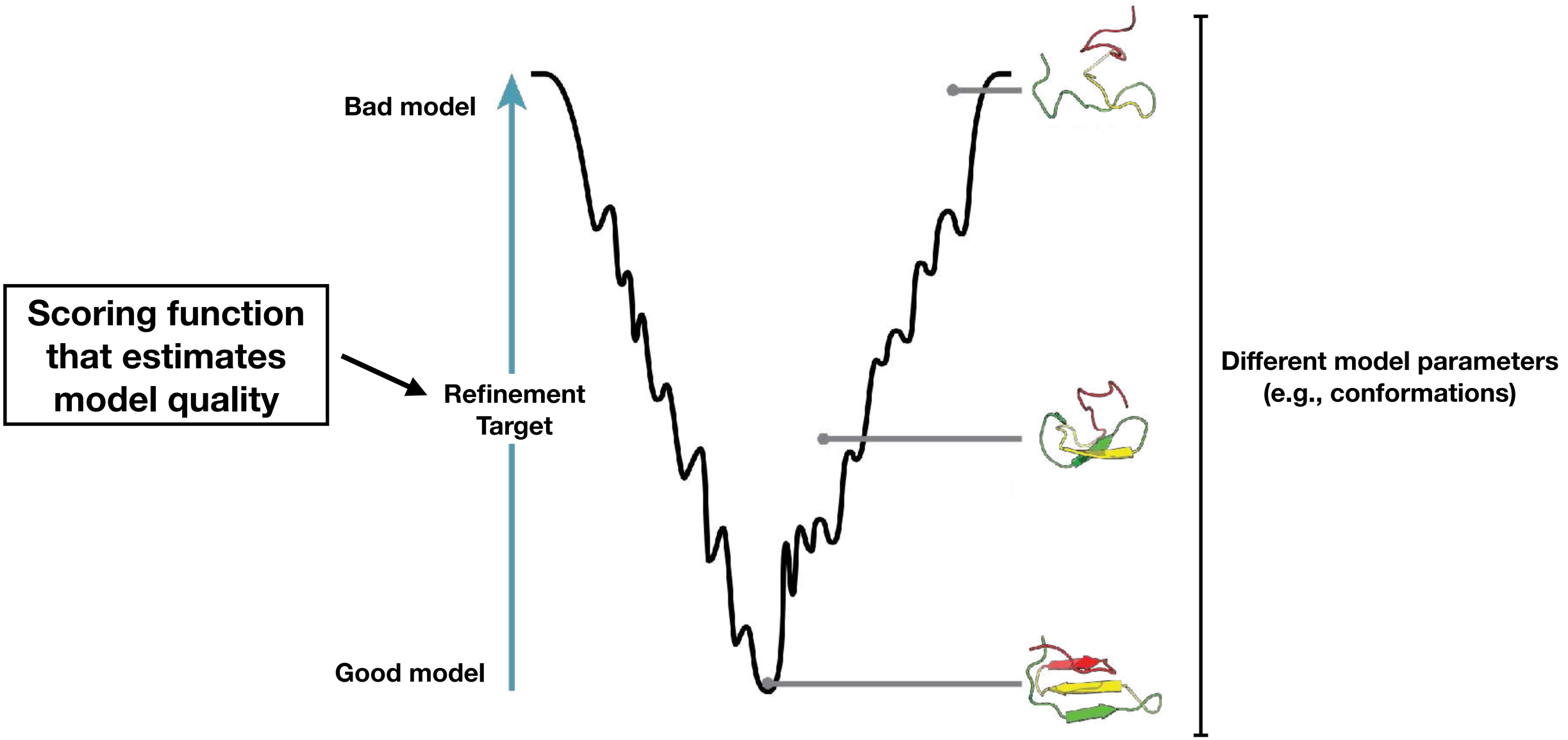
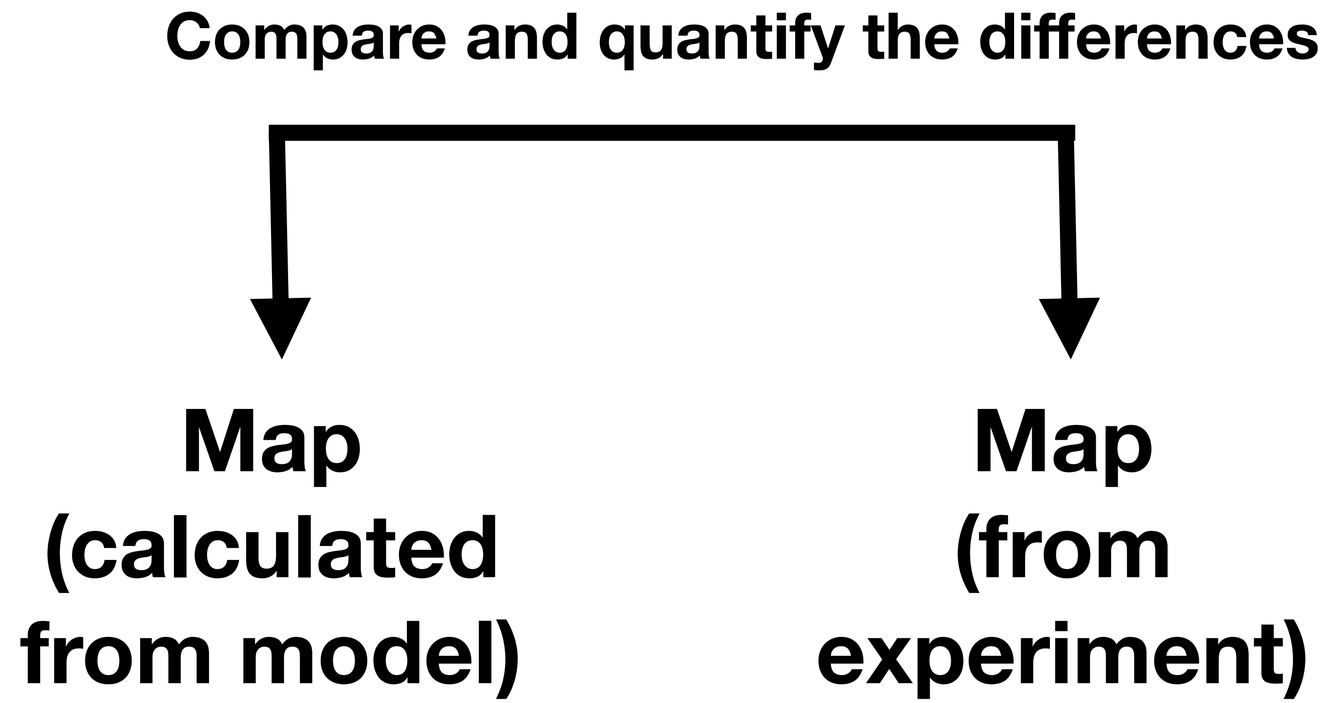
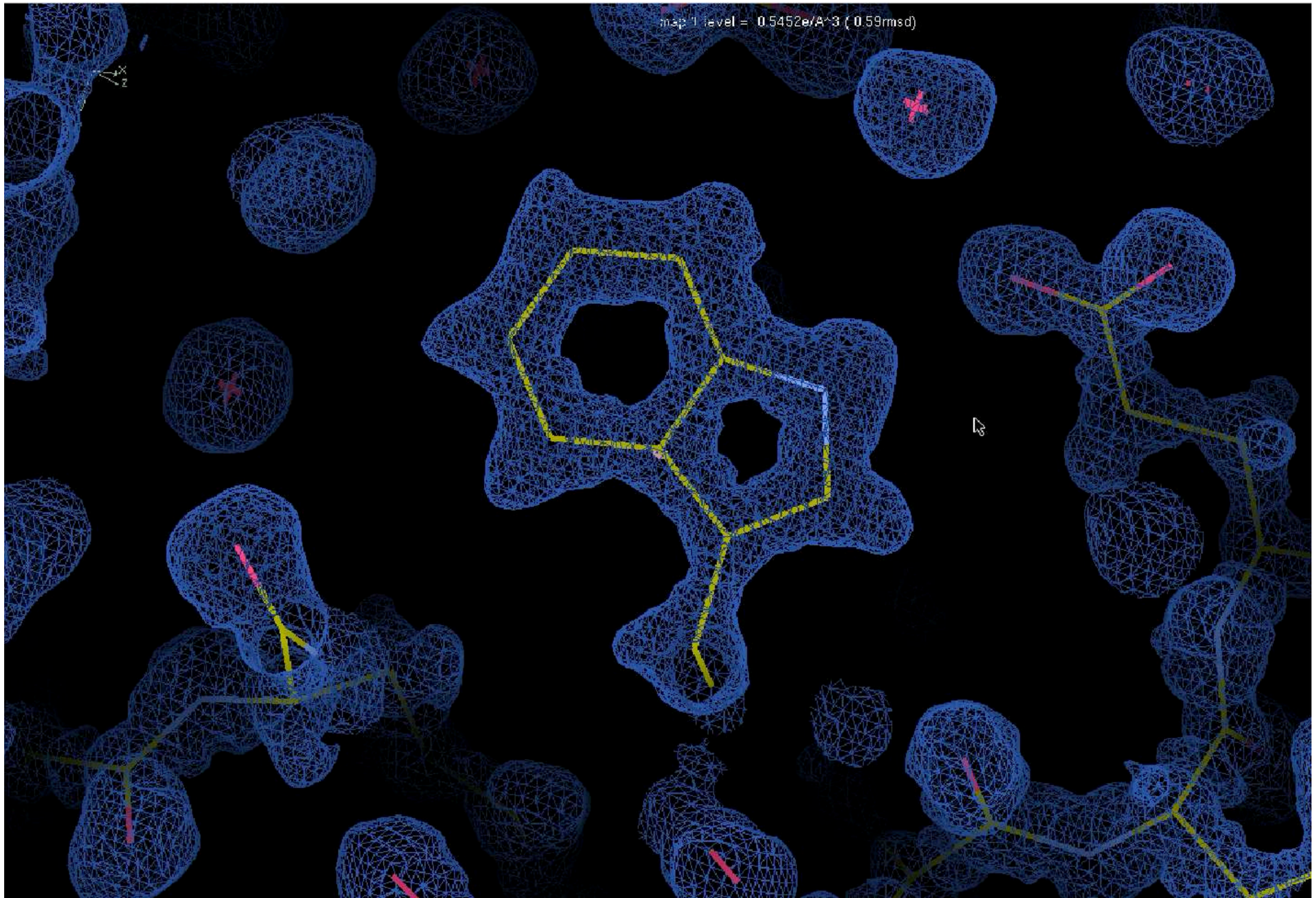



Image (adapted): Thomas Splettstoesser

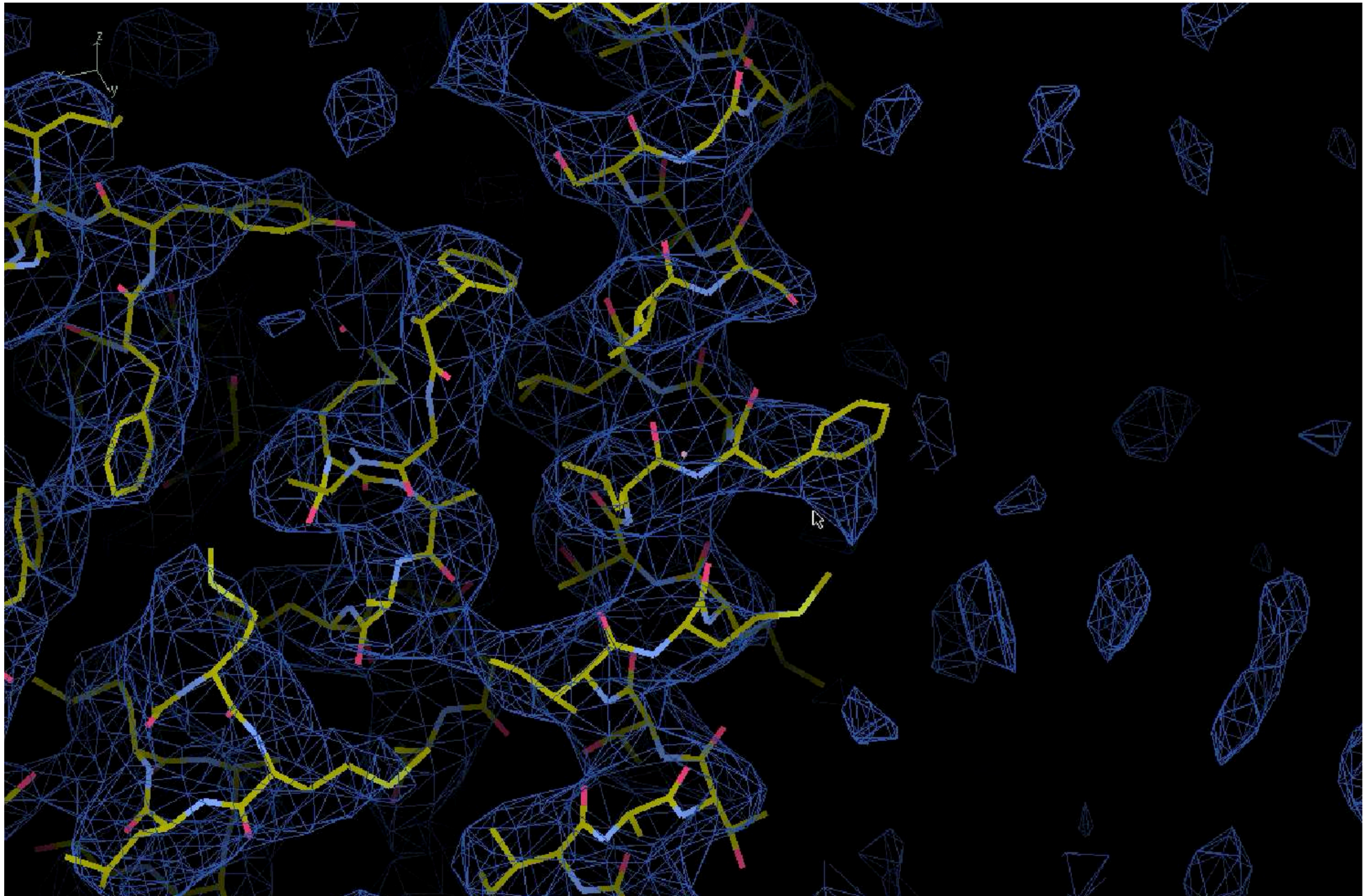
# Simplest: Refinement Target = (Model vs Data)



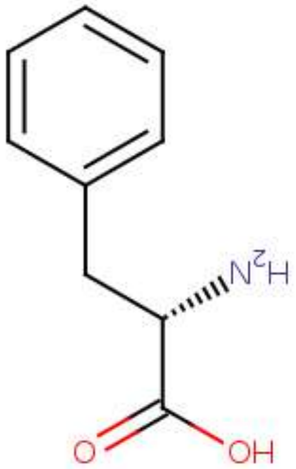
# At atomic resolution, position of individual atoms is well-defined

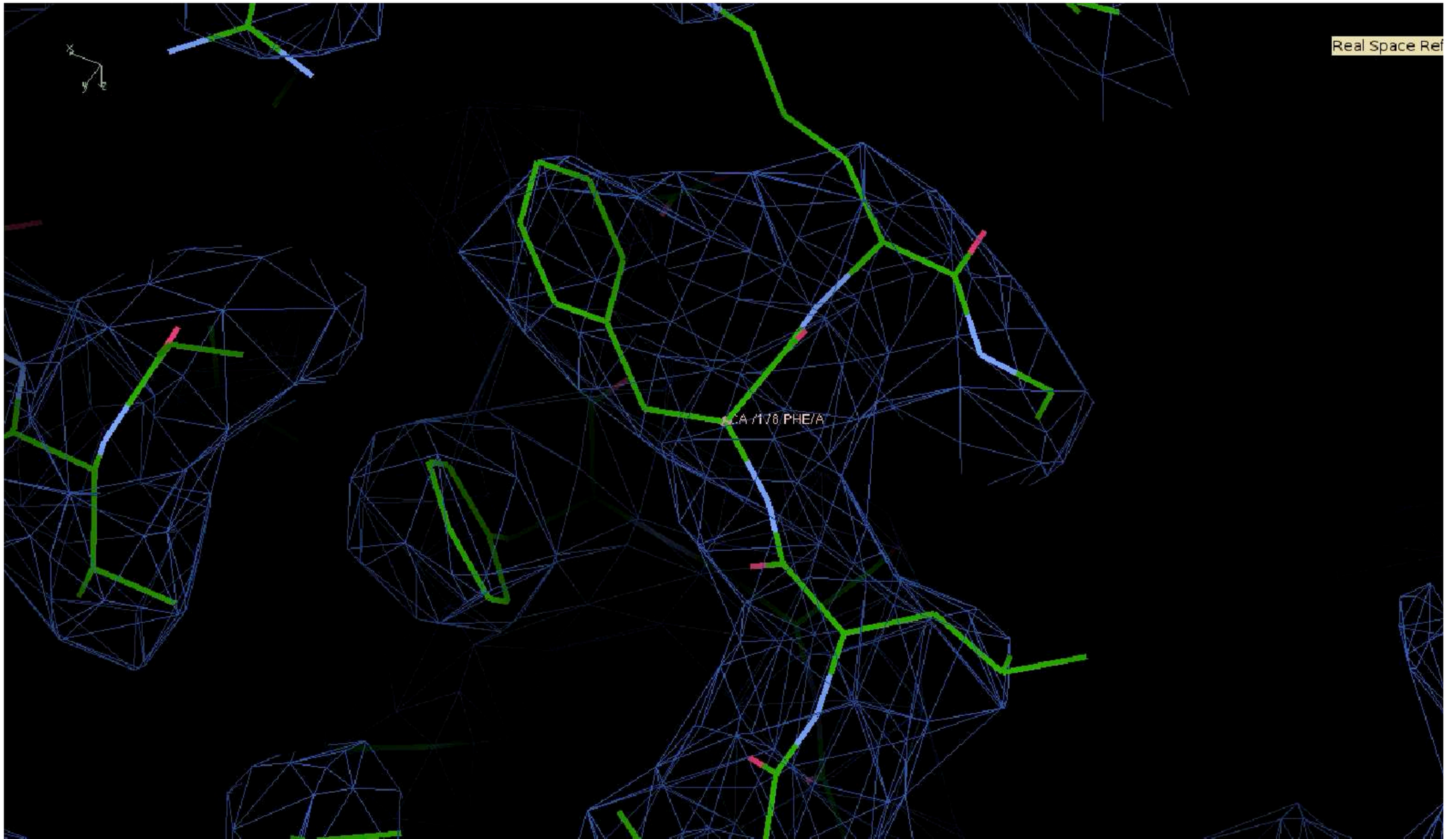


# But at “near-atomic” resolution, the position of residues and side chains is not always clear

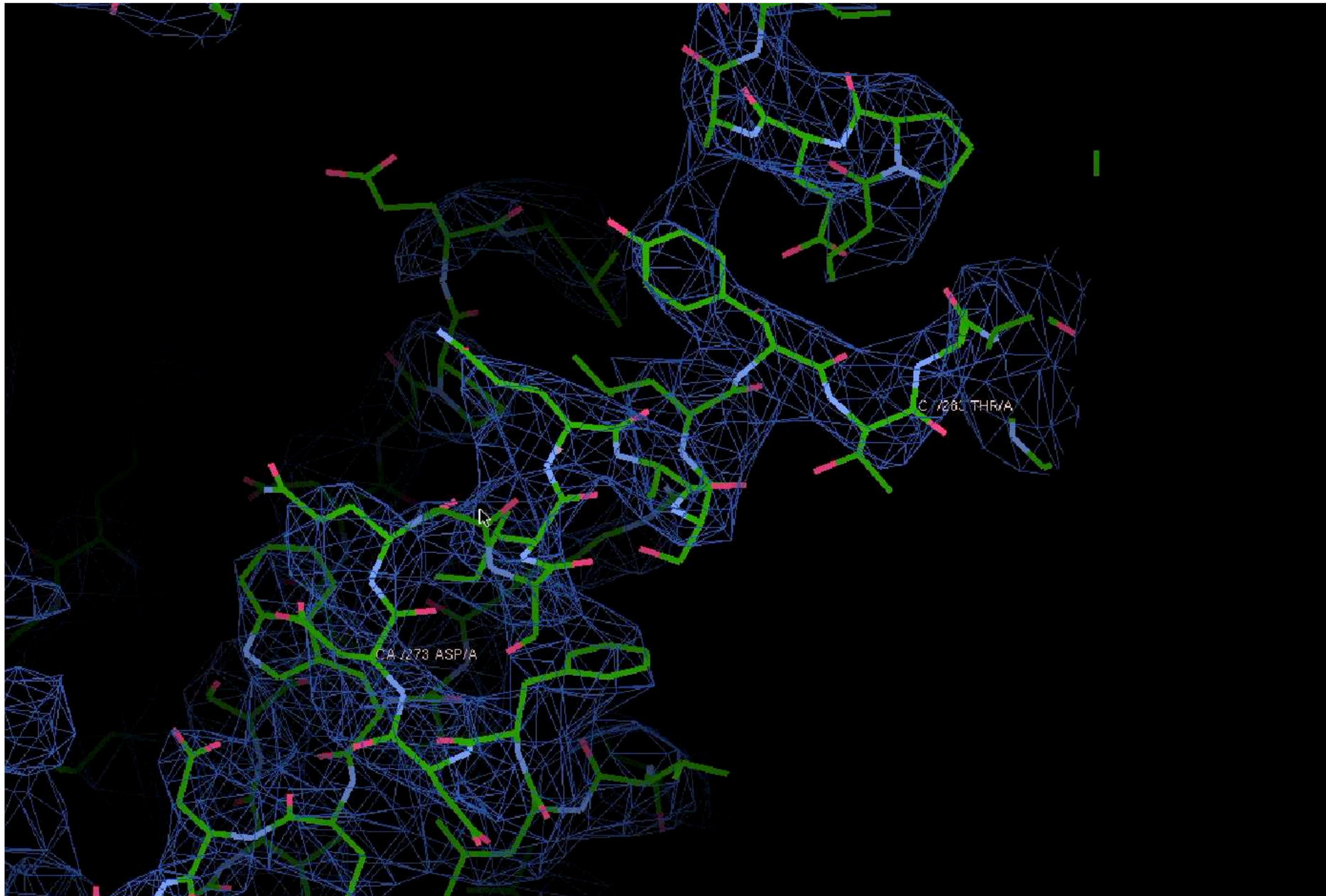


# Refinement using <u>only data</u>

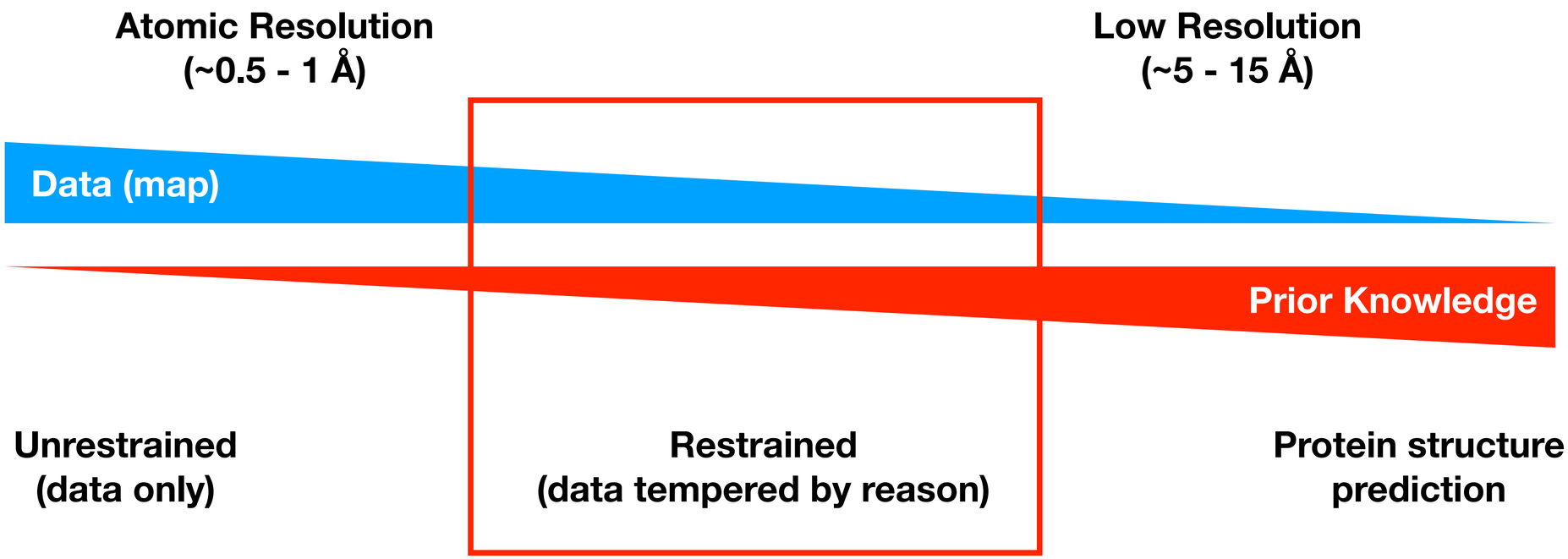


# Refinement using only data

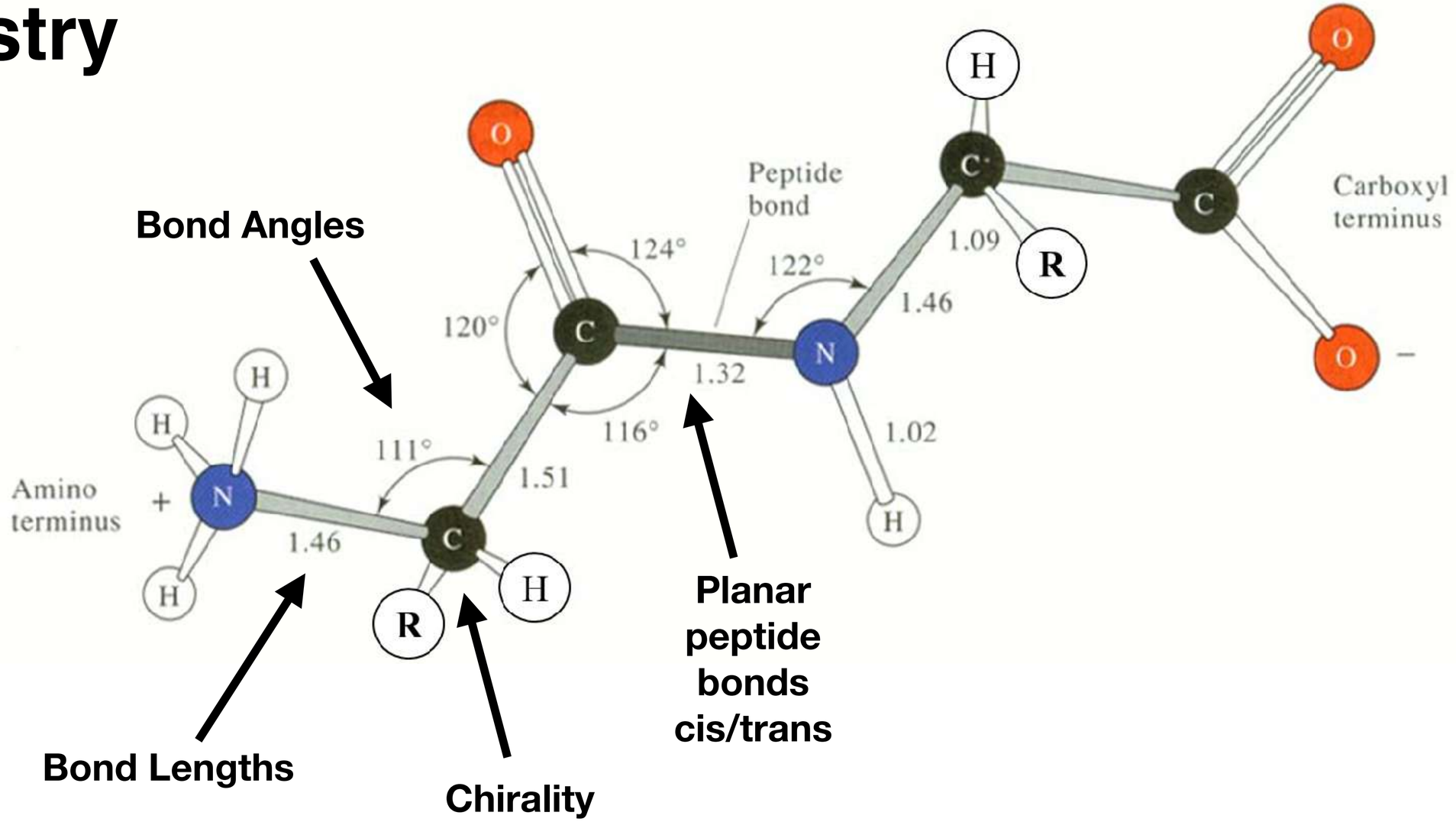


# Harnessing prior knowledge of protein structure to bridge the gap

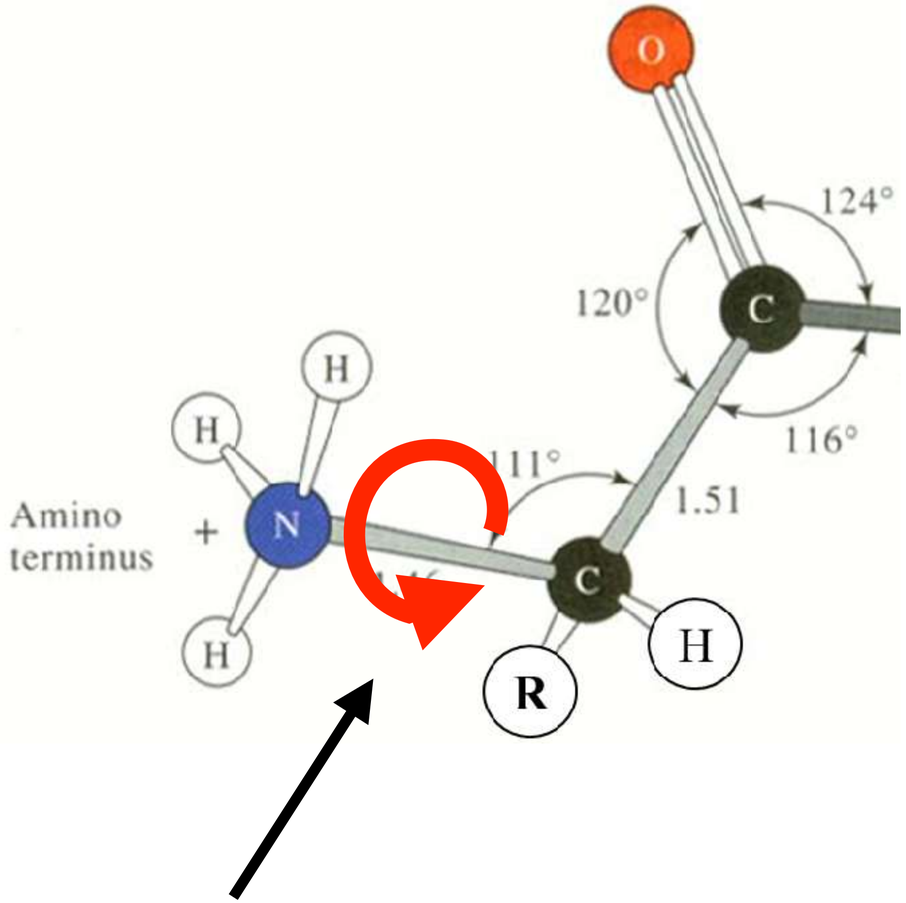



$$\text{Refinement Target} = (\text{Model vs Data}) + w_1(\text{Model vs PriorKnowledge})$$

# Stereochemistry



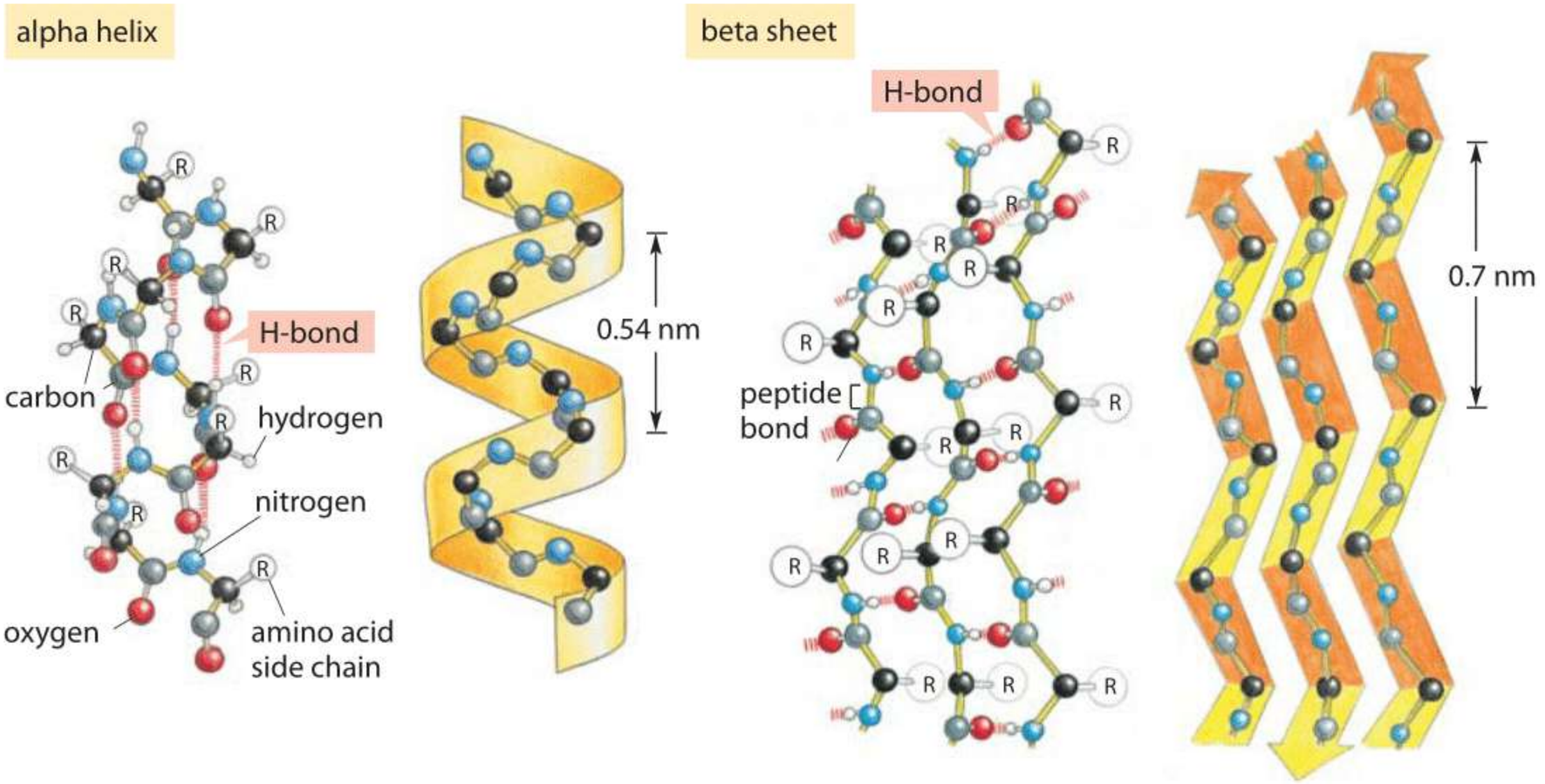
# Stereochemistry




**Torsion/dihedral Angles**

**Constraints backbone conformations as well as side chain rotameric states**

# Secondary Structure and Hydrogen Bonds

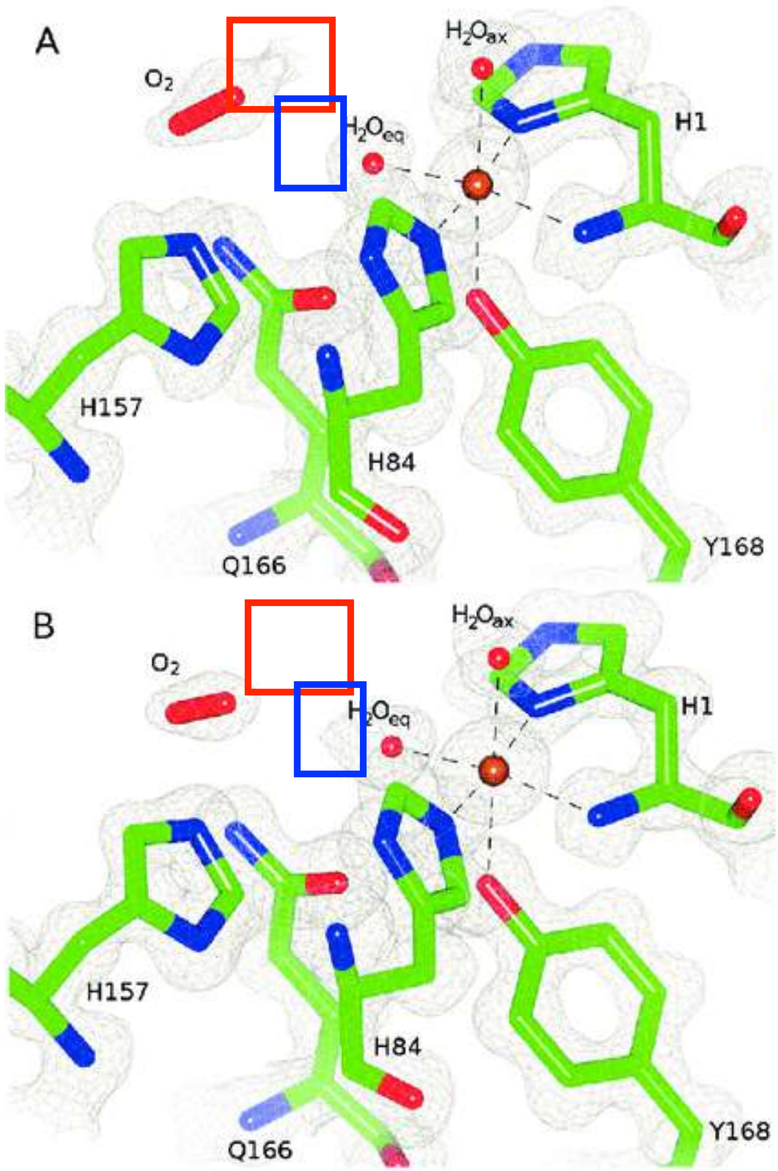



- **Torsion angle restraints to maintain appropriate backbone conformation**
- **Distance restraints between H-bonding atoms**

**Image: http://book.bionumbers.org/what-is-the-energy-of-a-hydrogen-bond/**

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

Active Site #1




Active Site #2

**<u>Restrain</u> to be similar**
**<u>Constrain</u> 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)**

Image: O’Dell, et al. Angewandte Chemie (2016).

# B factor / ADP restraints

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

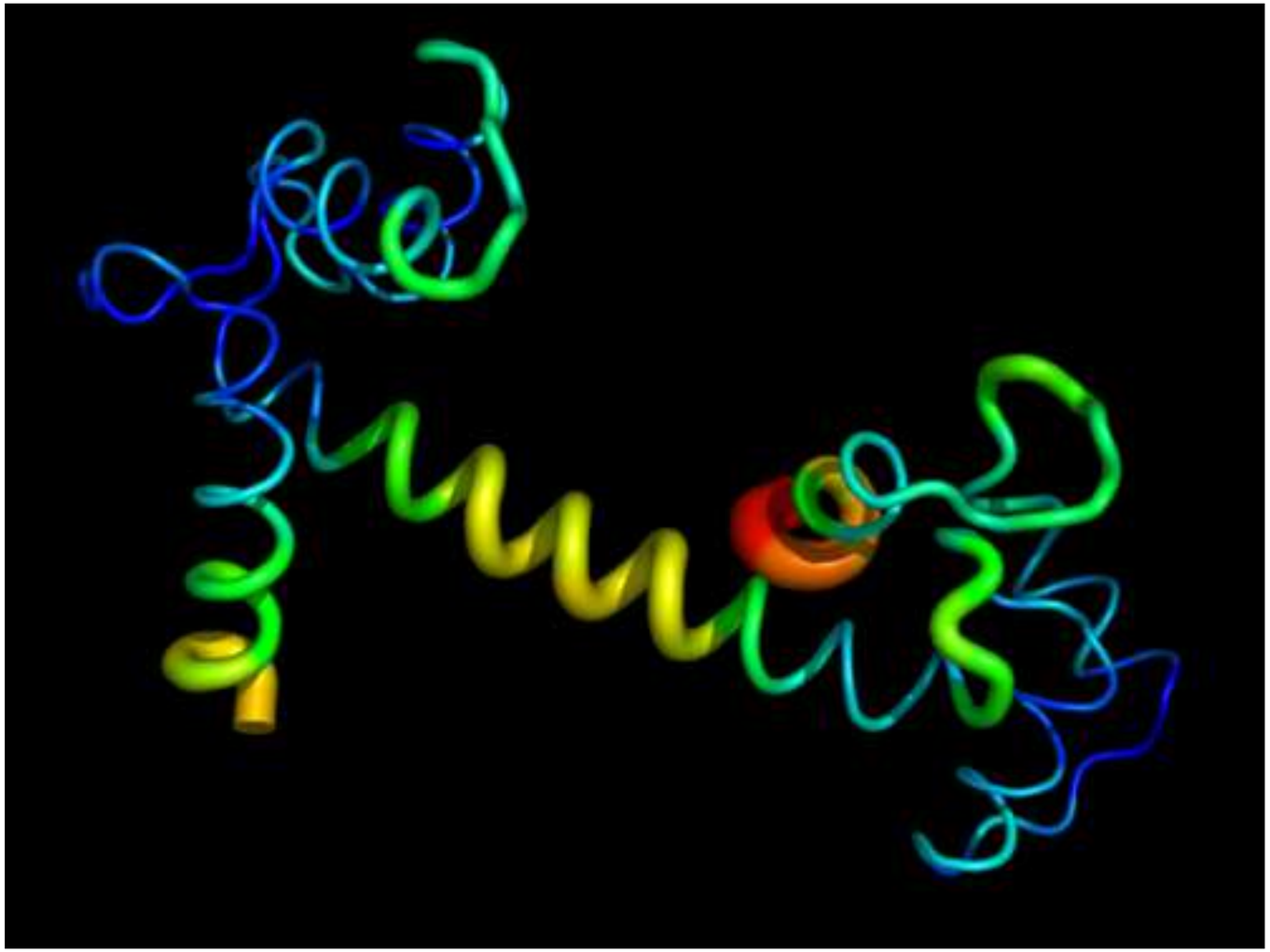


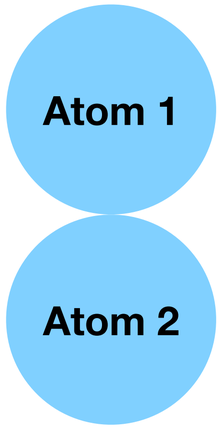
Image: Harry Jubb, https://github.com/arose/ngl/issues/291

**B factors are not randomly distributed**

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

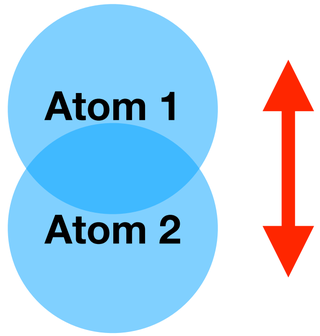
**Therefore, we can retrain B factors such that connected atoms/residues must have similar B factors**

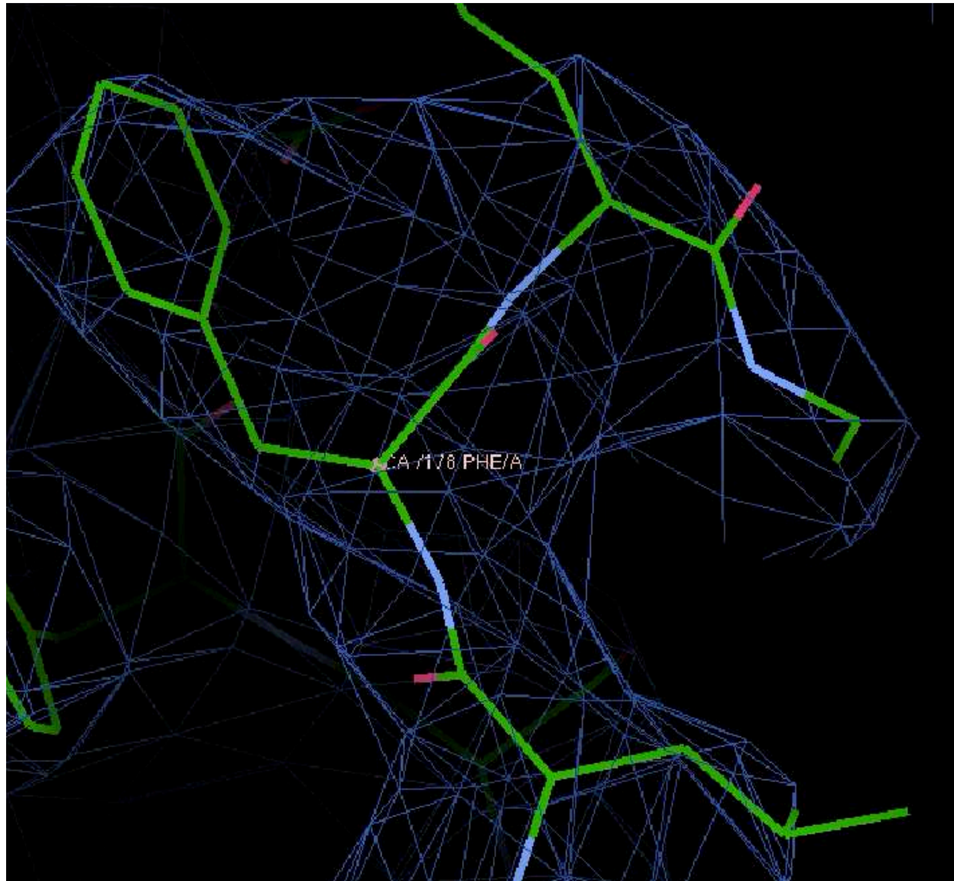
# Steric repulsion



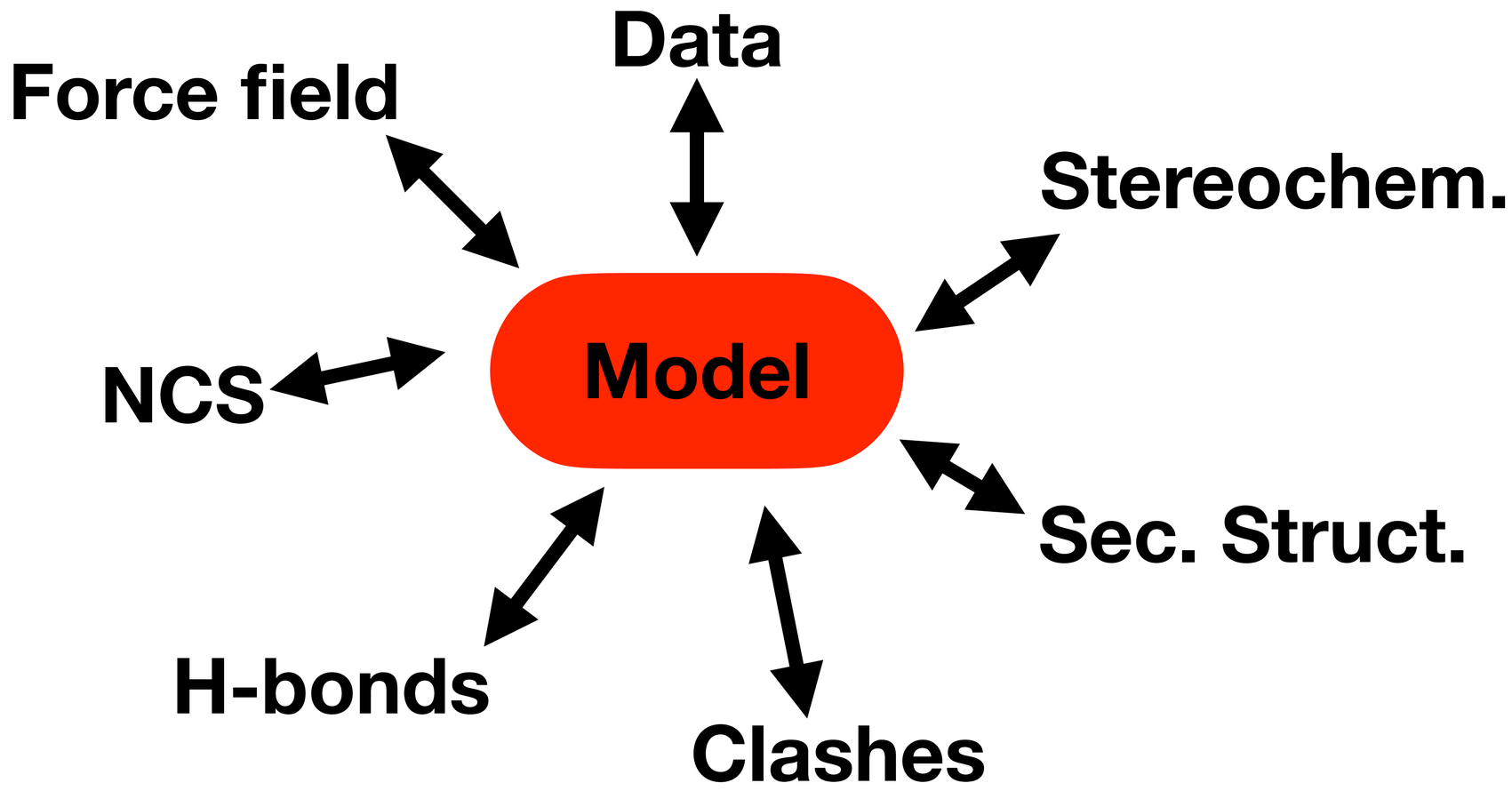

Optimal center-to-center distance ~ sum of VDW radii

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

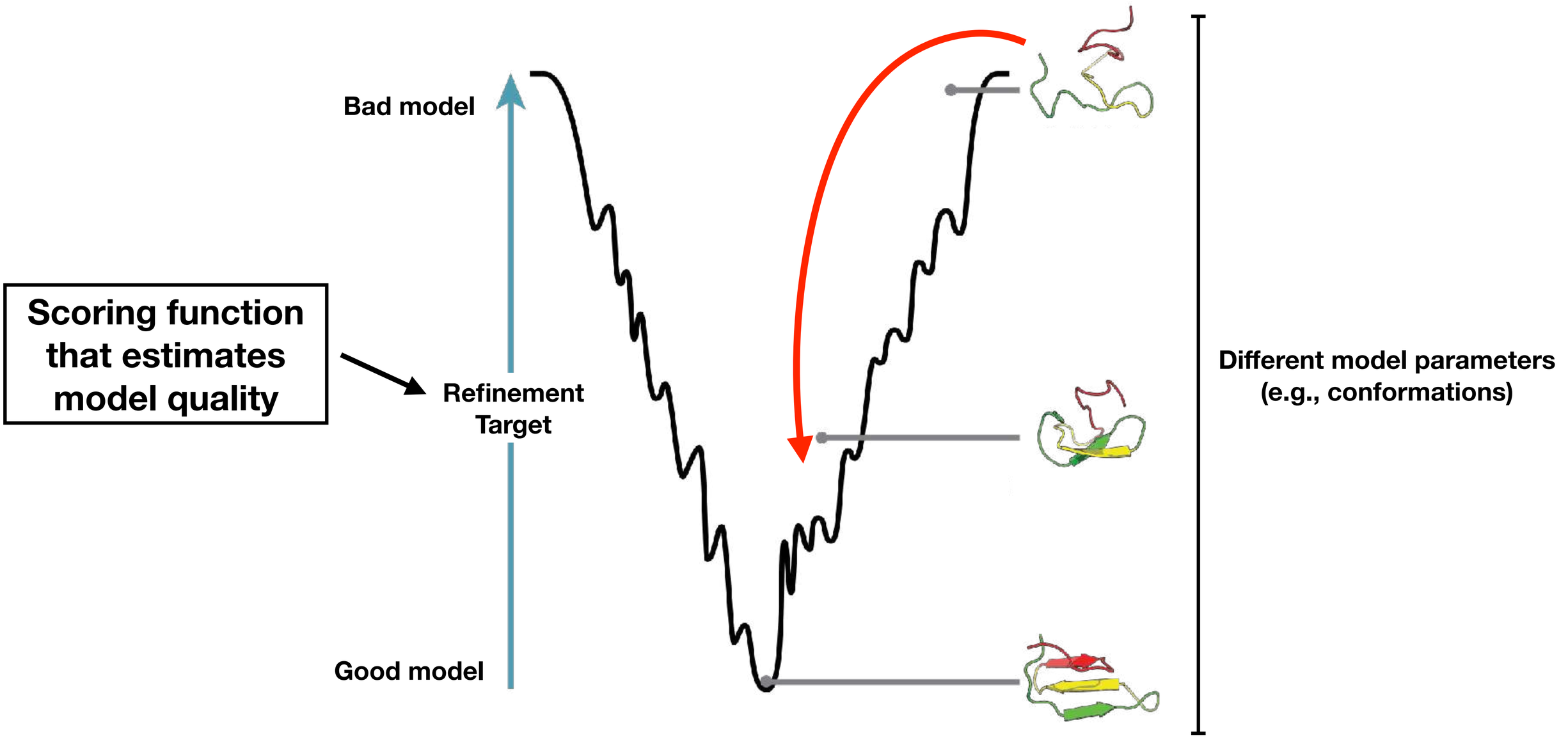
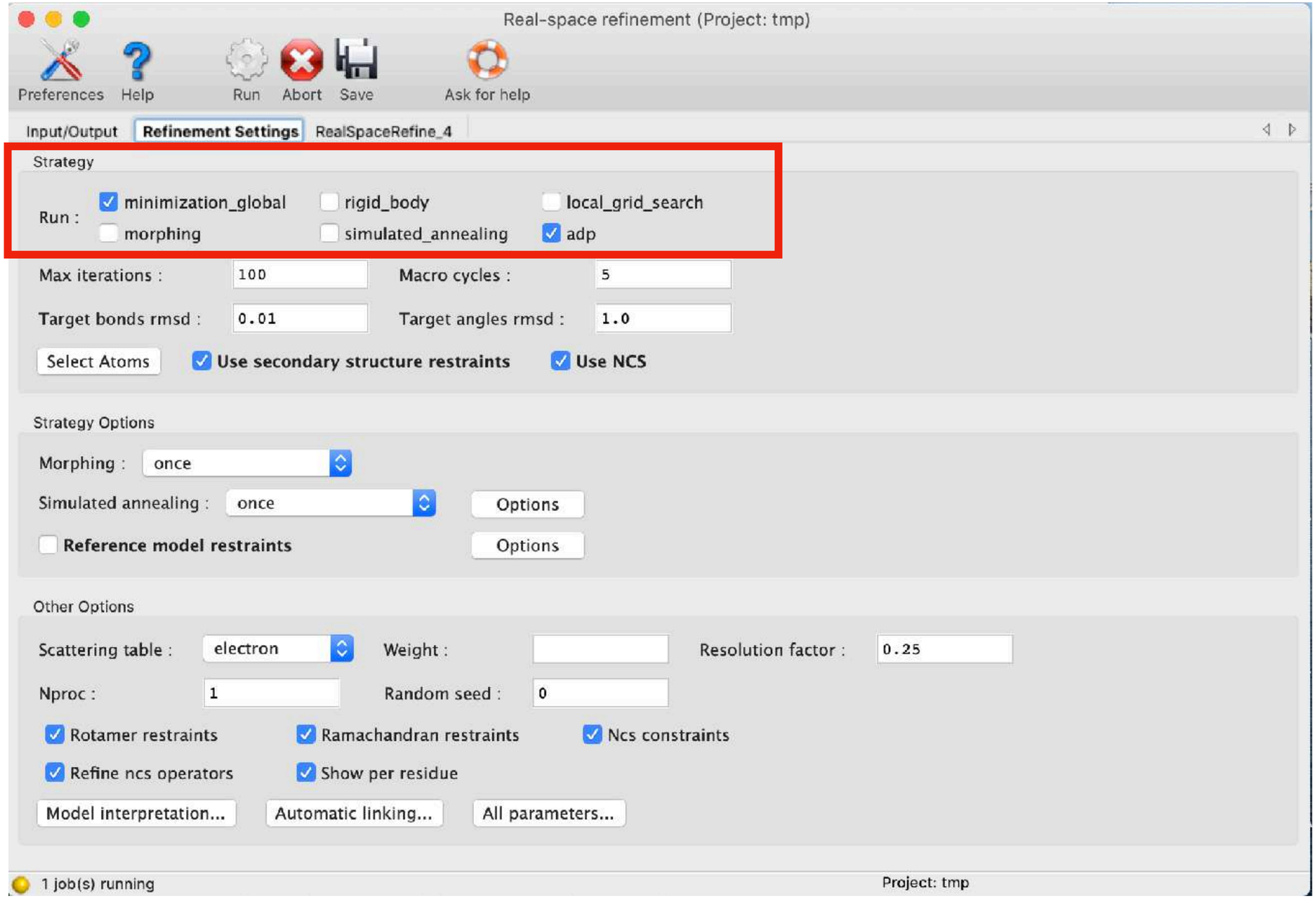


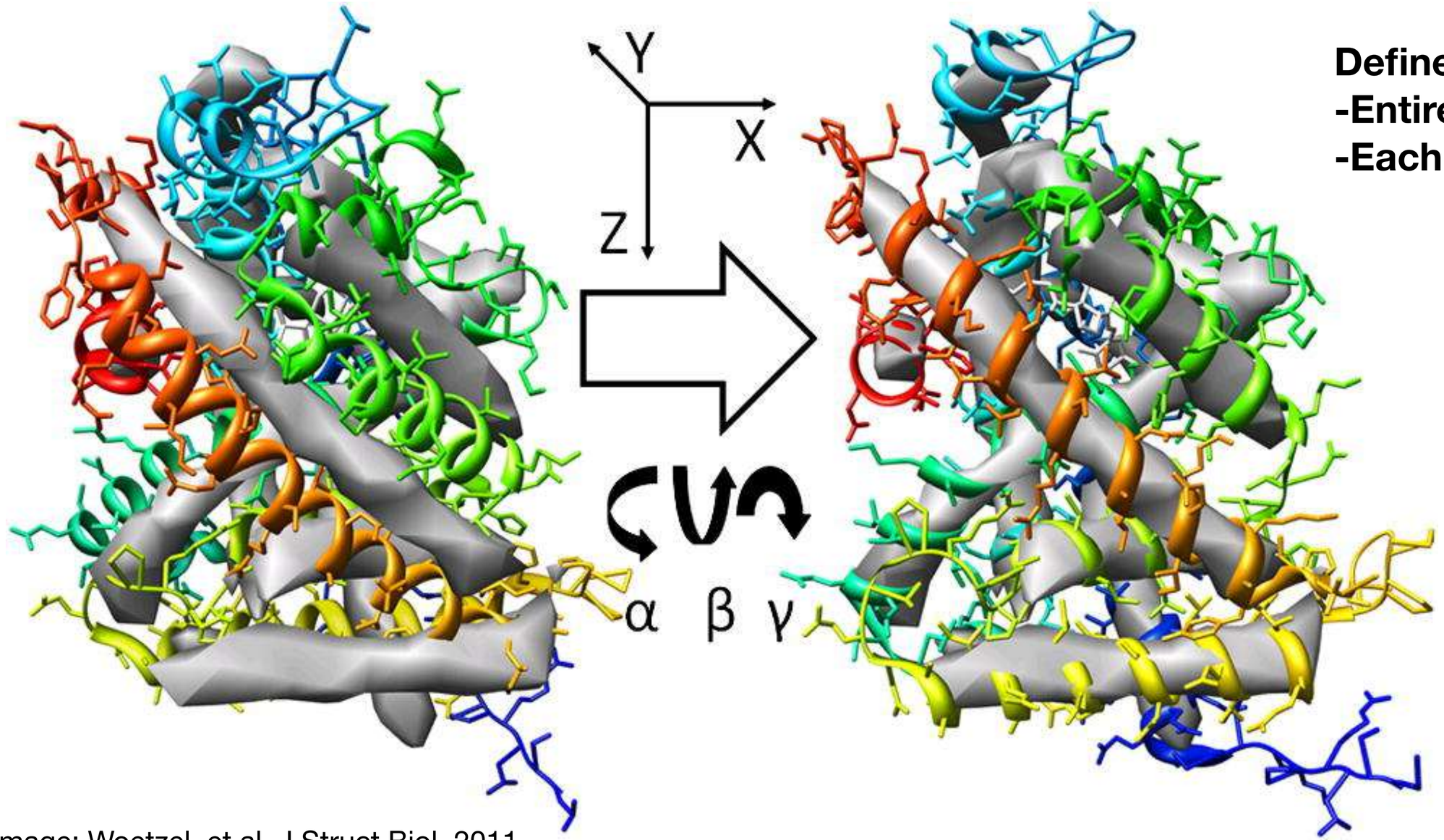

Image (adapted): Thomas Splettstoesser

# There are many refinement options to choose from!




**phenix.real_space_refine**

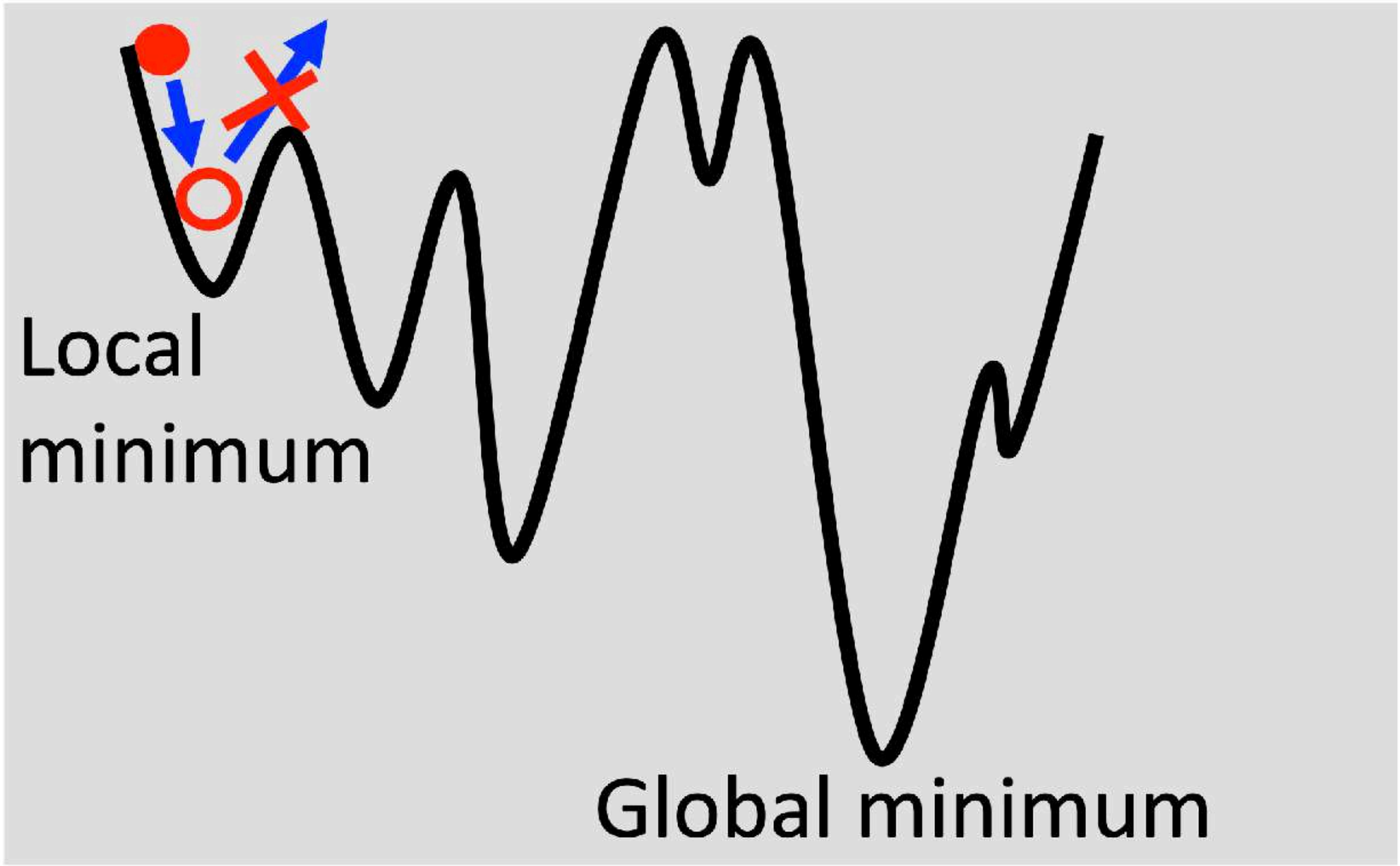
# Optimization protocols: Rigid body refinement




**Define rigid bodies:**
**-Entire model?**
**-Each chain separately**

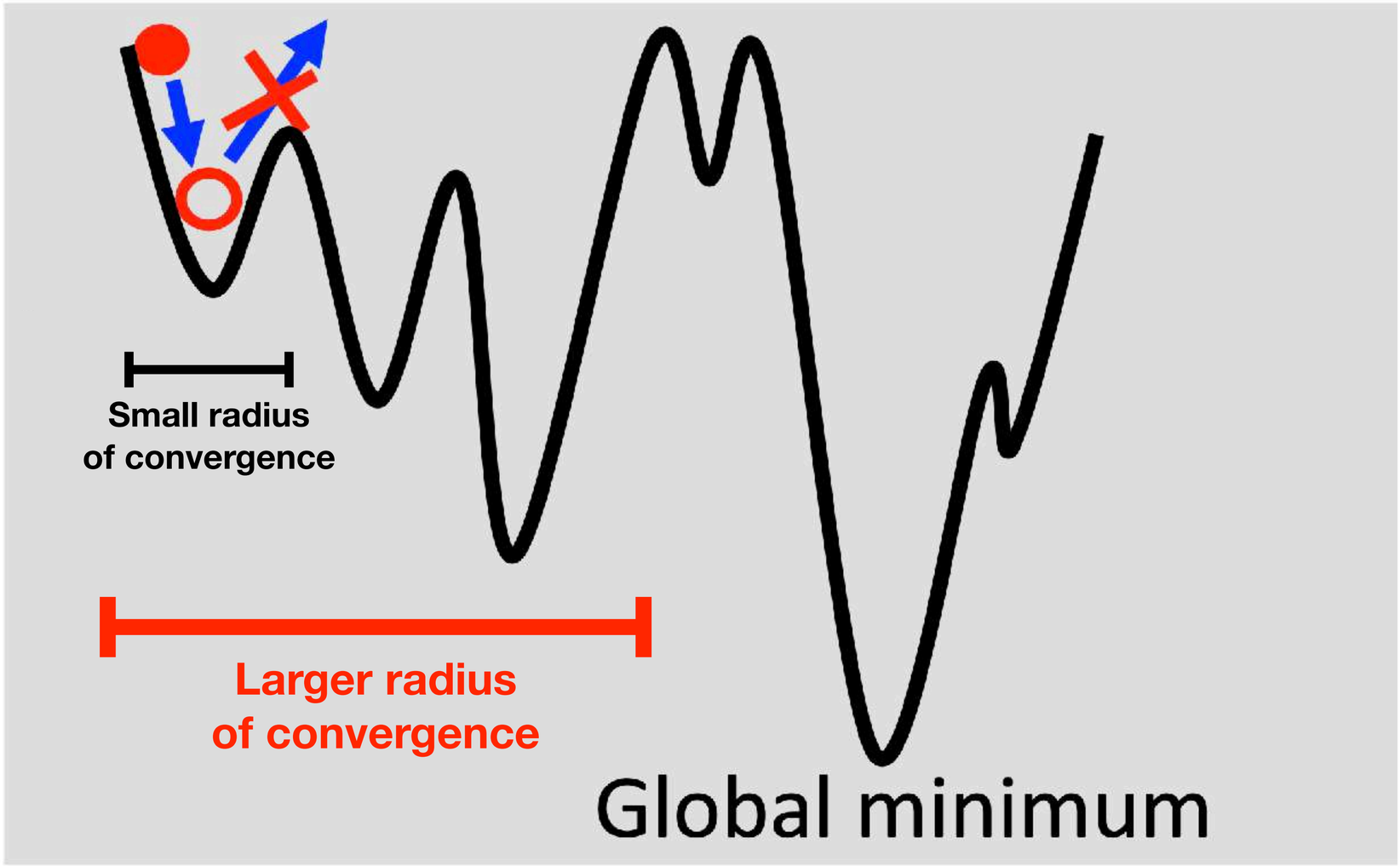
Image: Woetzel, et al. J Struct Biol. 2011.

# Optimization protocols: Gradient driven minimization

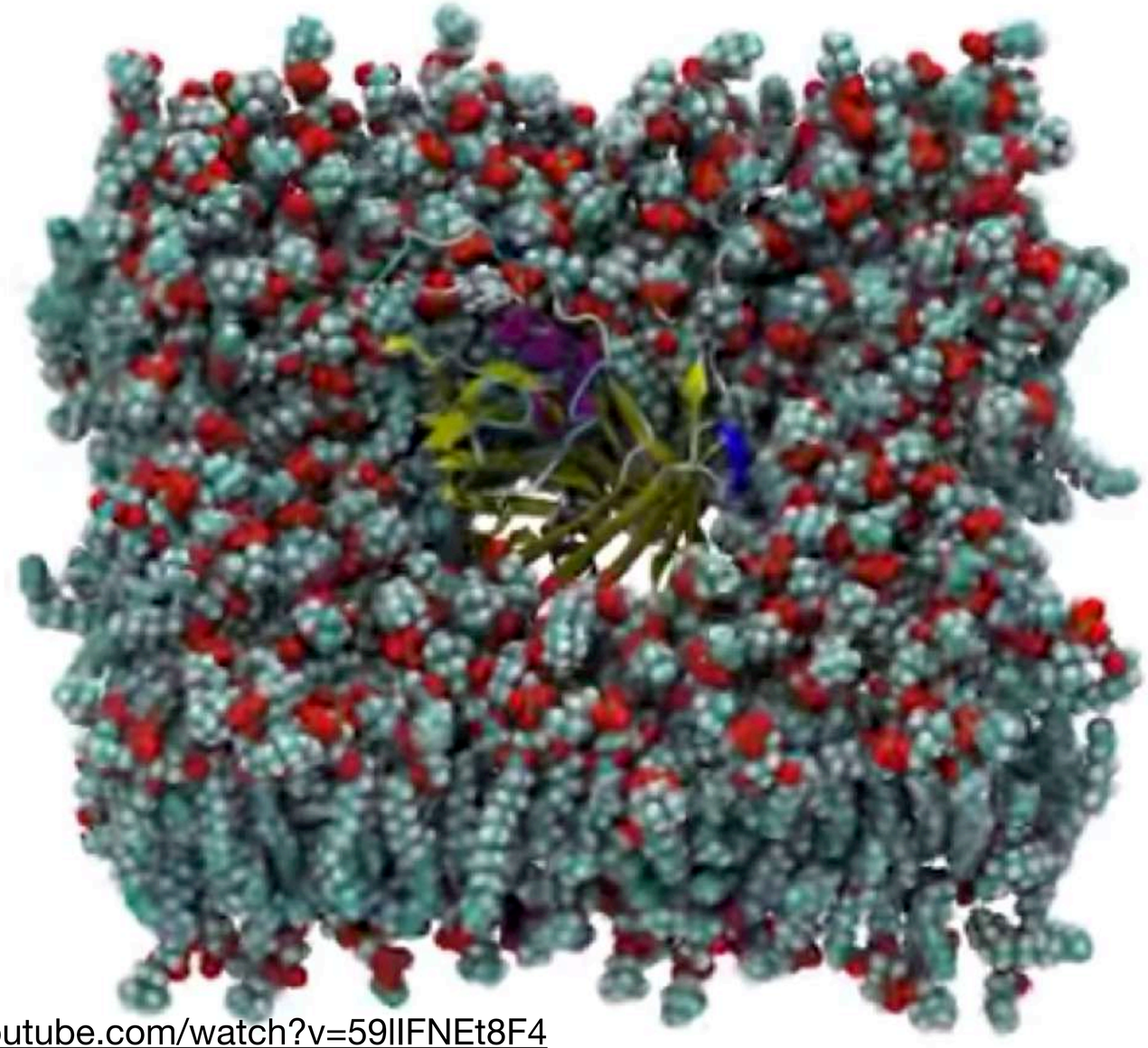



Slide adapted from: Pavel Afonine, LBNL (Phenix)

# Refinement “radius of convergence”

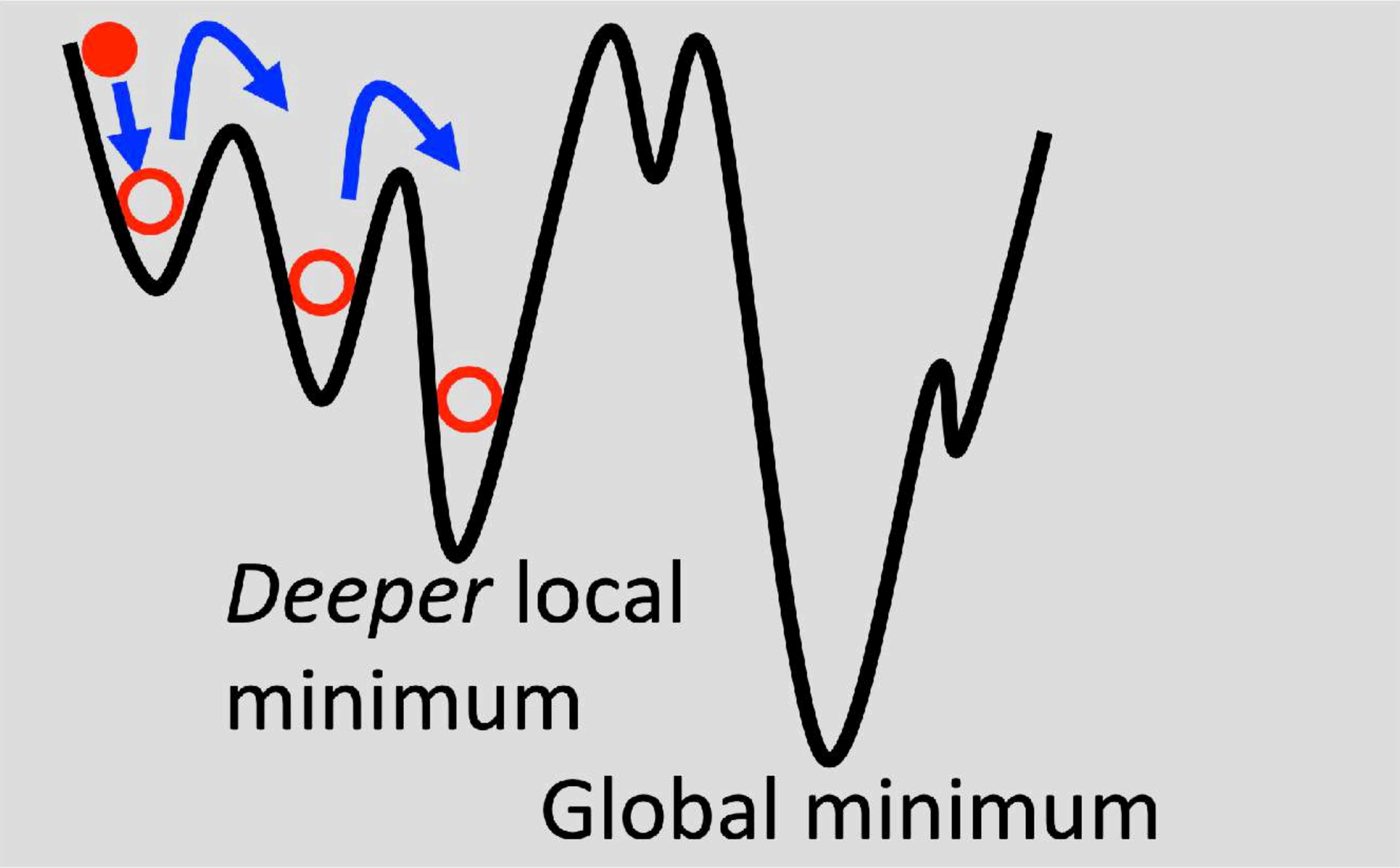


# Optimization protocols: Simulated annealing



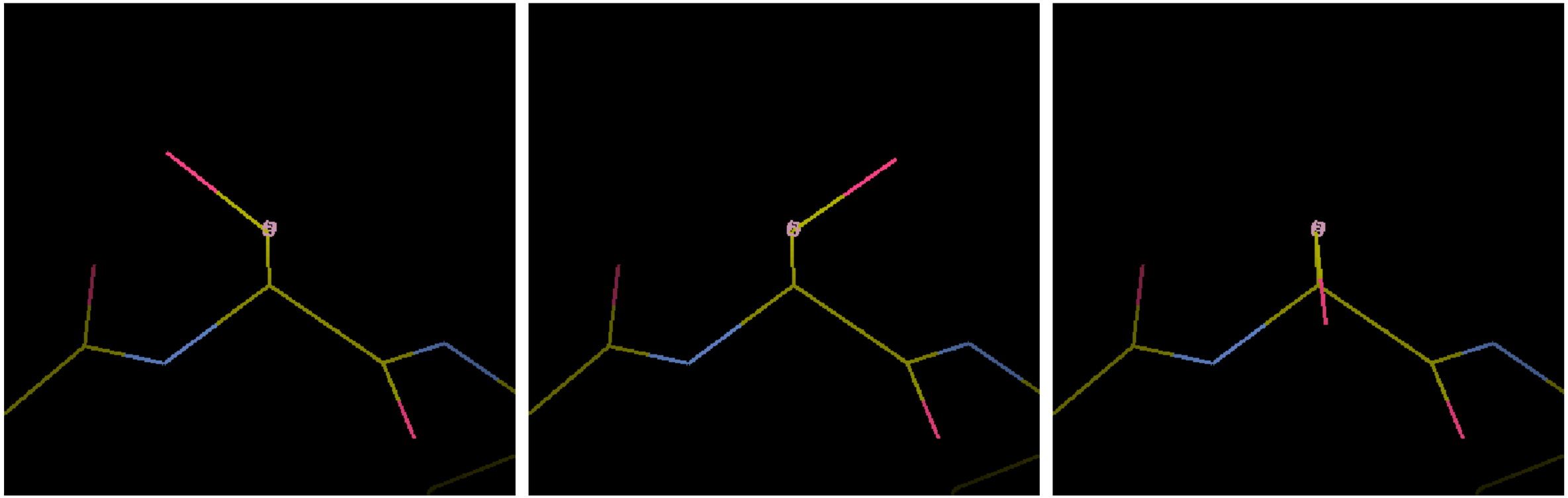
Video: Darrell Hurt, https://www.youtube.com/watch?v=59lIFNEt8F4

# Optimization protocols: Simulated annealing



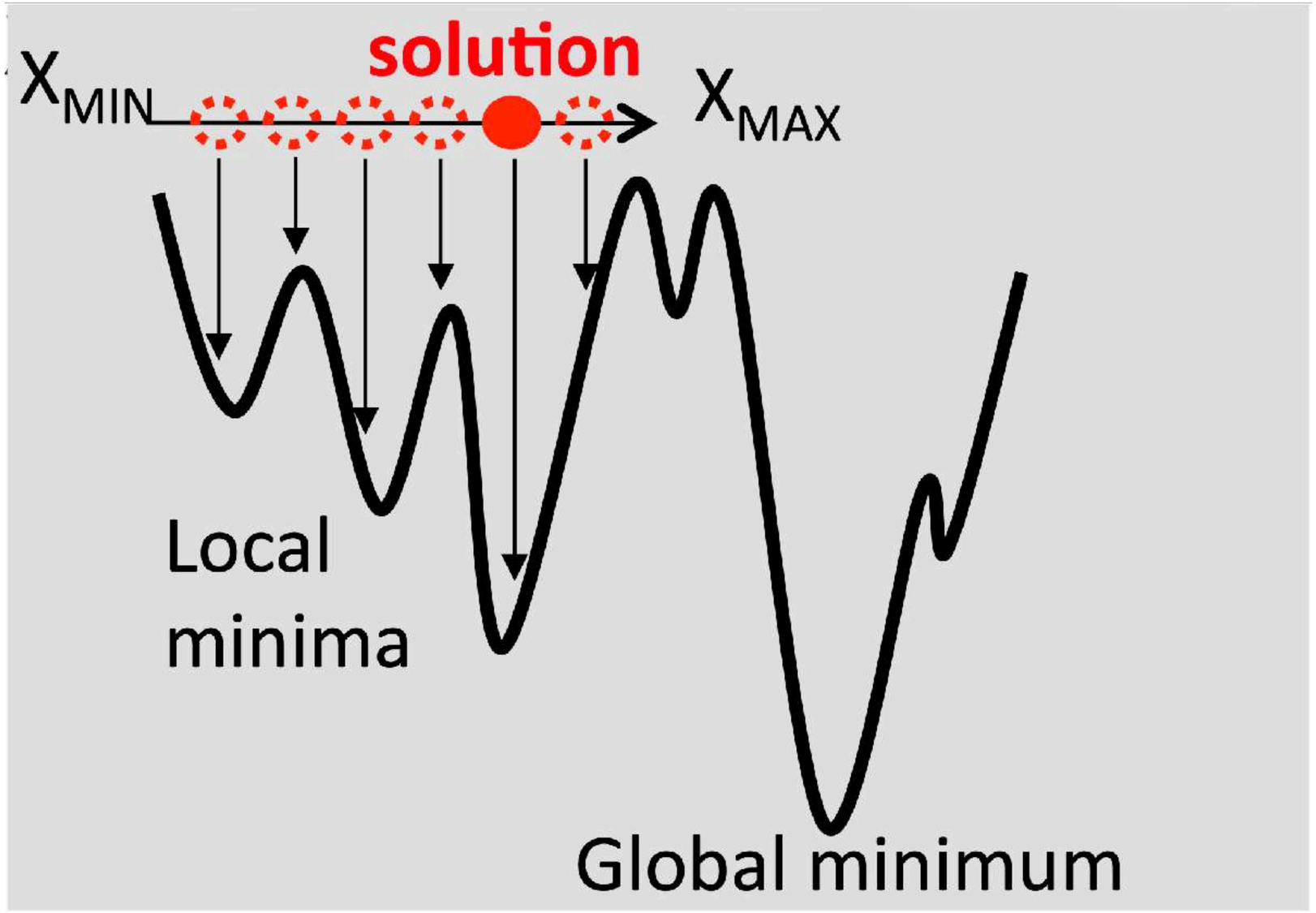

Slide adapted from: Pavel Afonine, LBNL (Phenix)

# Optimization protocols: Torsion-angle Grid Search



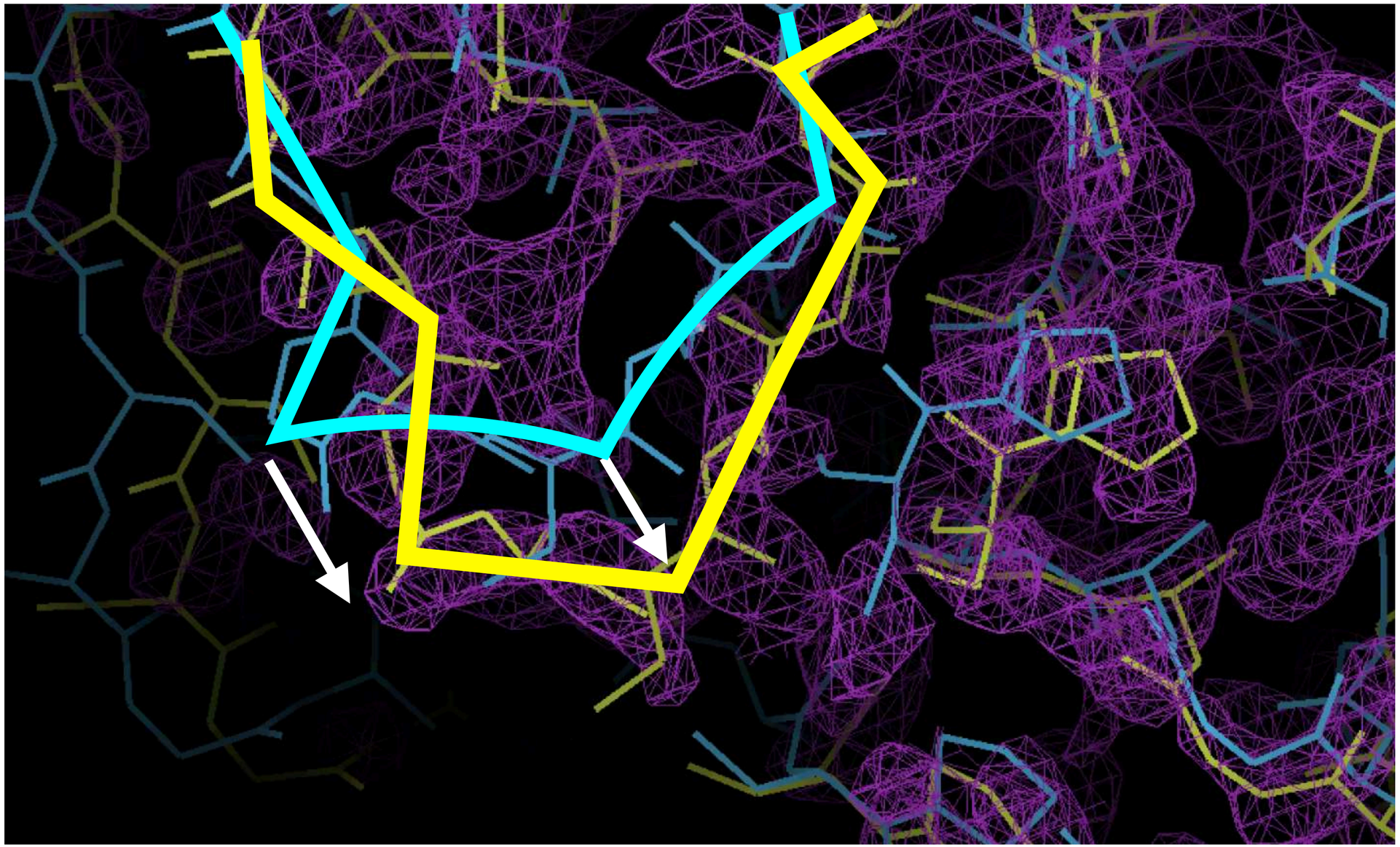
- **Can allow for larger shifts in model than simple minimization**

# Optimization protocols: Torsion-angle Grid Search



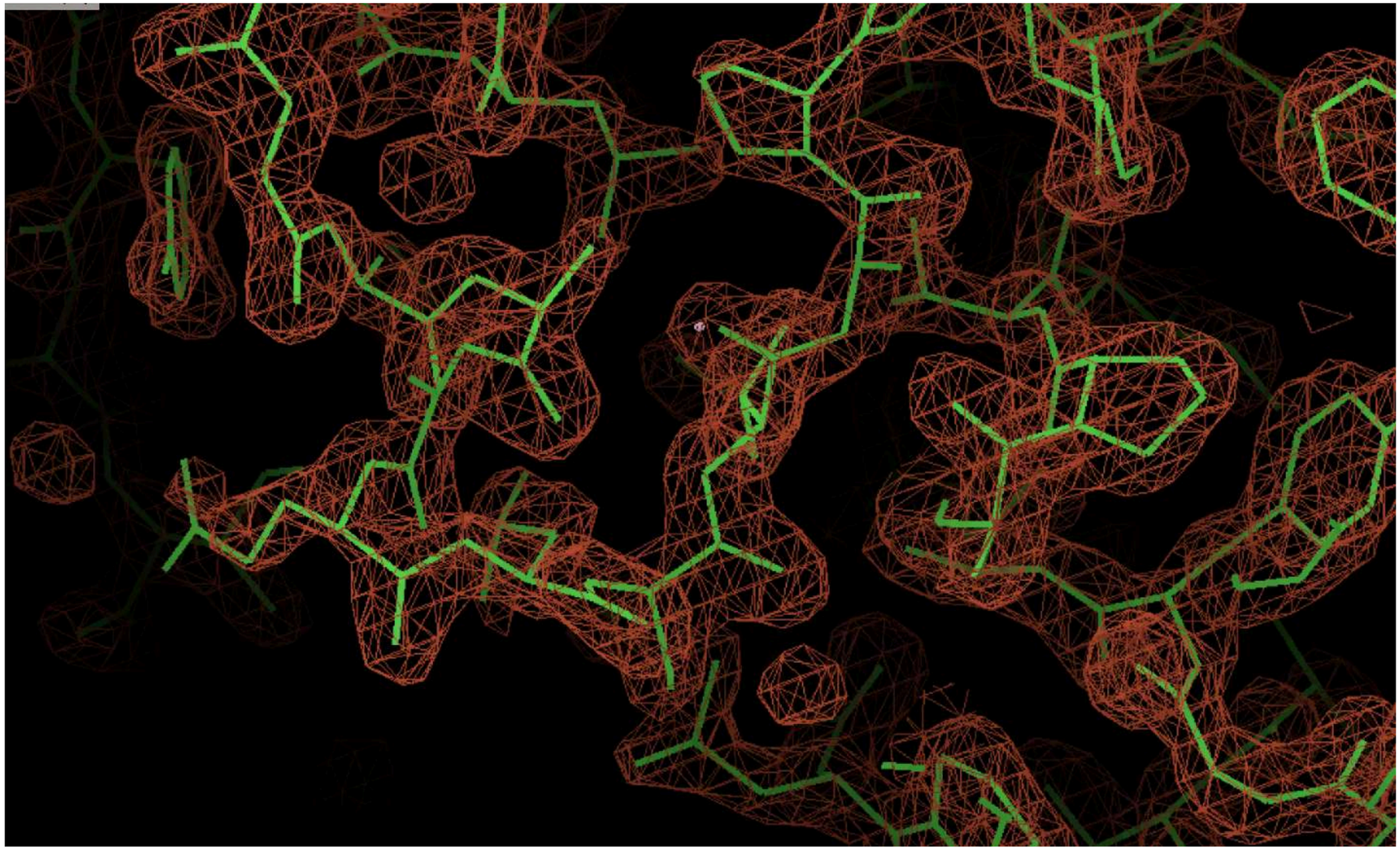

Slide adapted from: Pavel Afonine, LBNL (Phenix)

# Optimization protocols: Morphing



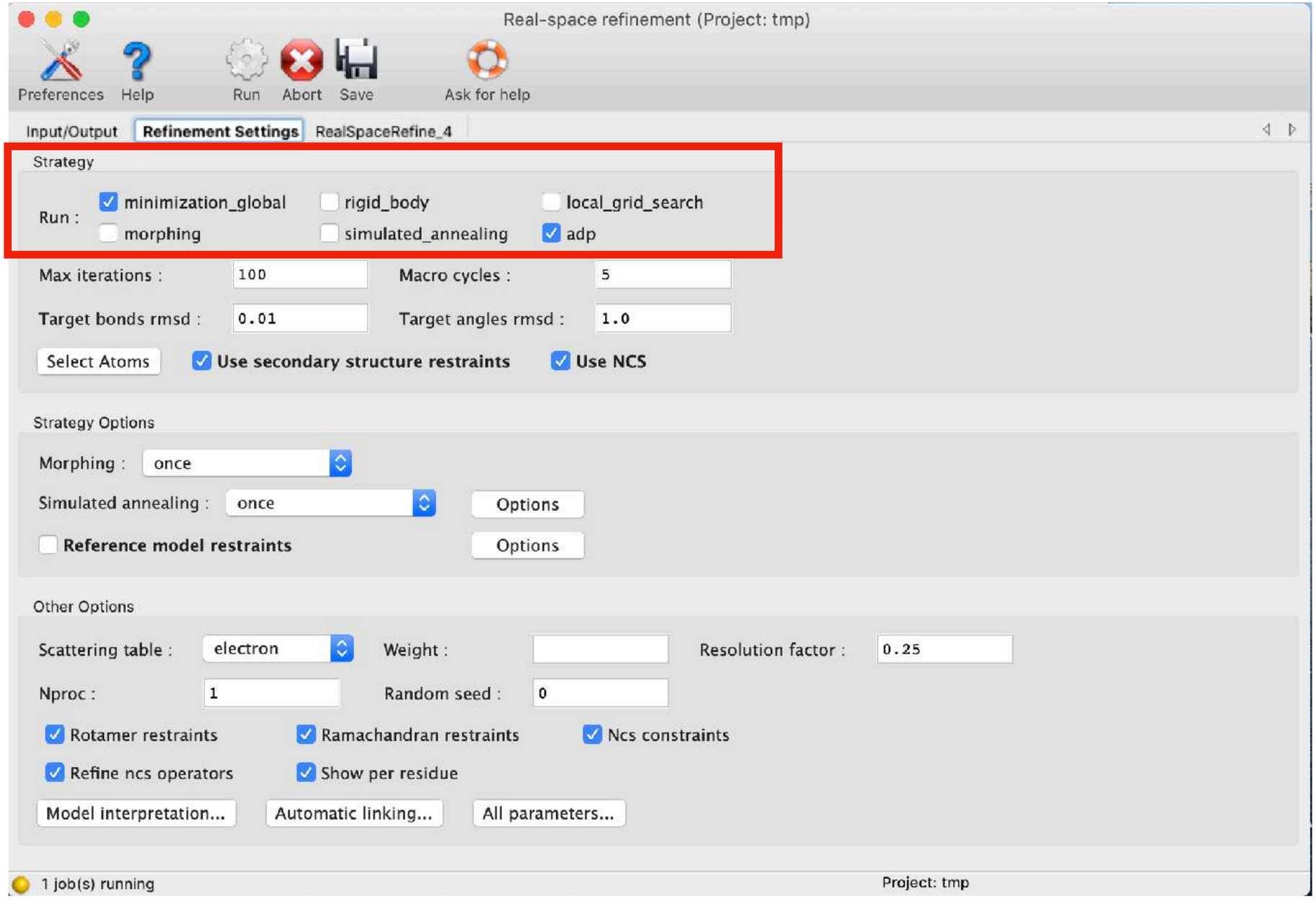
Slide adapted from: Tom Terwilliger, Los Alamos (Phenix)

# Optimization protocols: Morphing



Slide adapted from: Tom Terwilliger, Los Alamos (Phenix)

# There are many refinement options to choose from!




**phenix.real_space_refine**

# 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(?)**

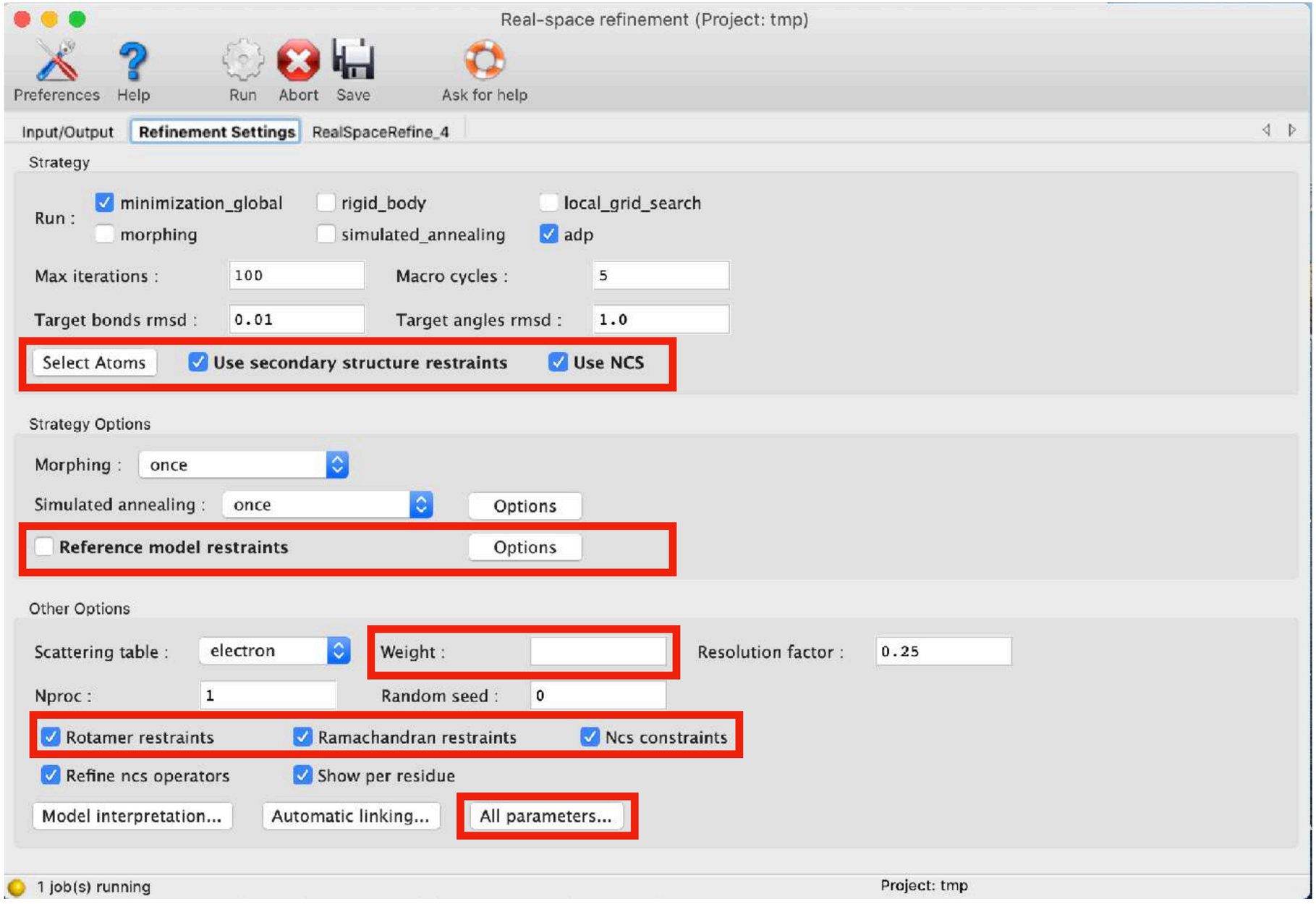
**-Target structure very different: minimization_global, adp, grid_search, morph, simulated annealing(?)**

**Late stages:**

**-Go easy: minimization_global, adp**

# Restraints

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



# 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 sterochemical 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?
    - Molprobity 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 |
| $CC_{vr\_mask}$ | 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 |

**Afonine, et al. “New tools for the analysis and validation of cryo-EM maps and atomic models” Acta Cryst. 2018**

# 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**

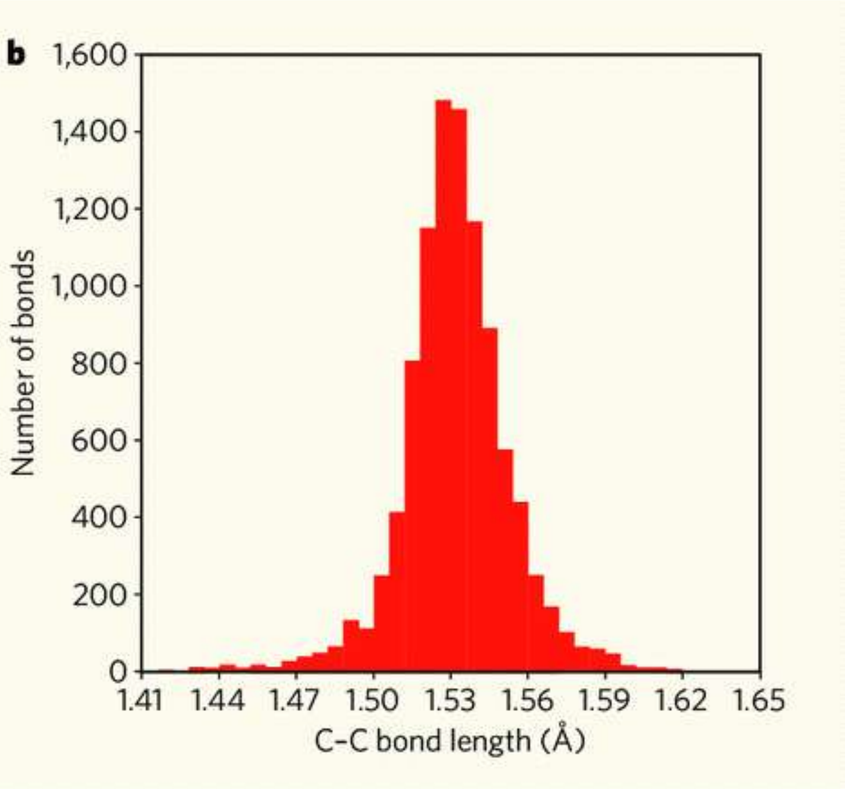
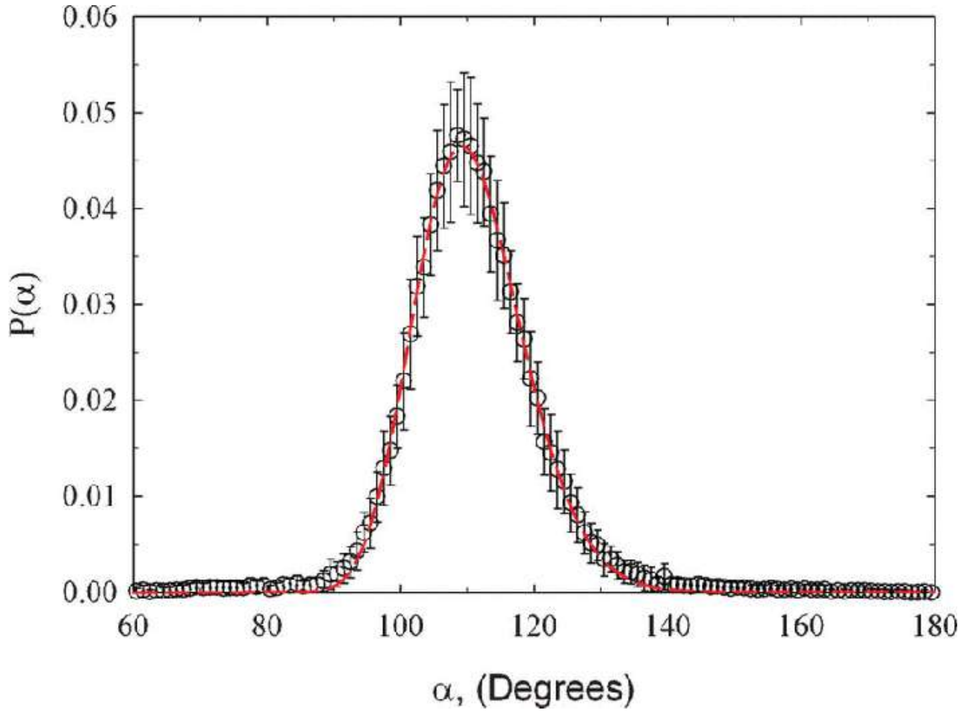


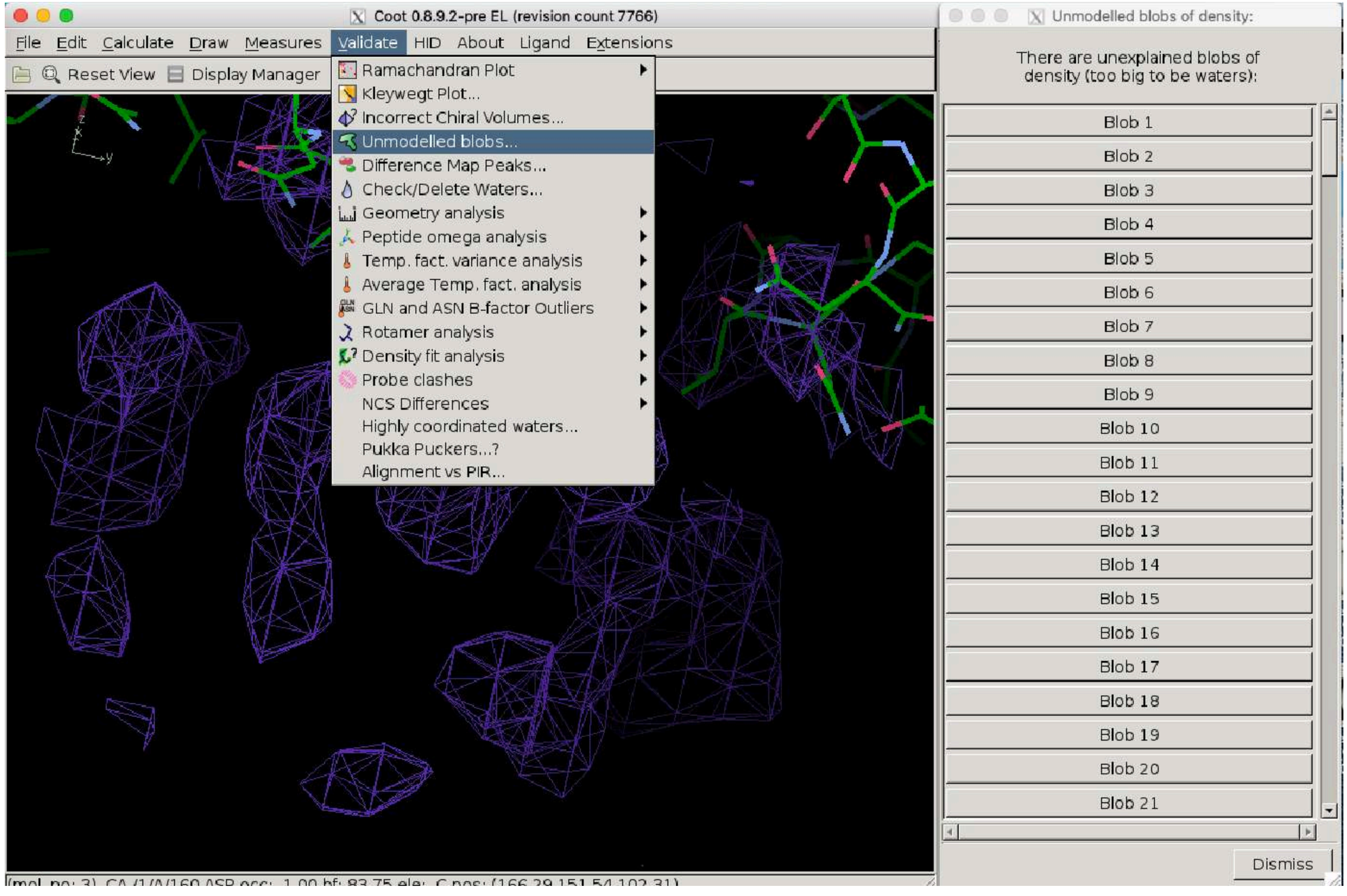

**Image: Janez Stepisnik**

**Typical RMS Angles for protein structure:**
**0.5 - 1.5 Å**

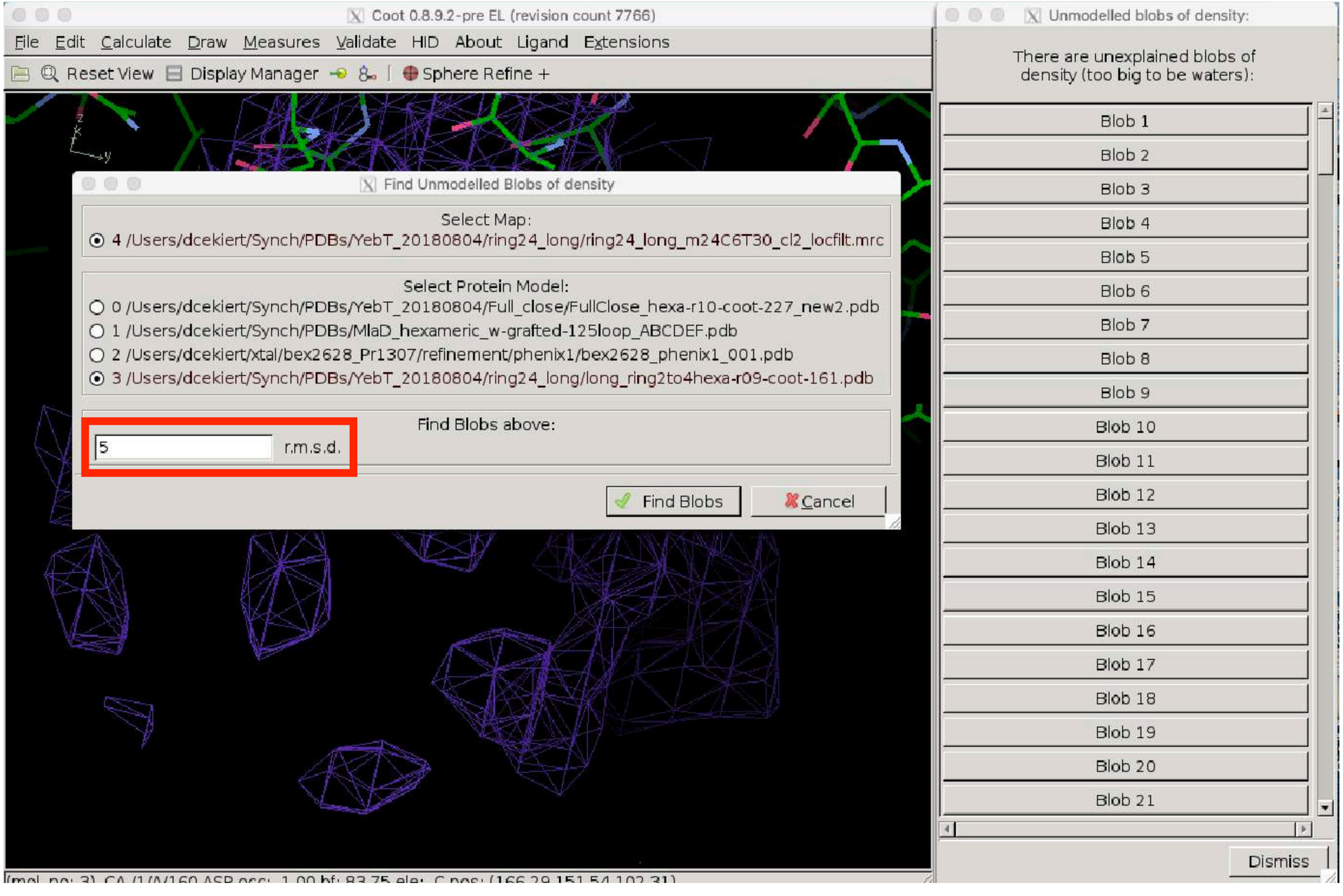



**C-C-C angle distribution of sp3 carbon**
**Kowalczyk, et al. RSC Advances, 2012.**

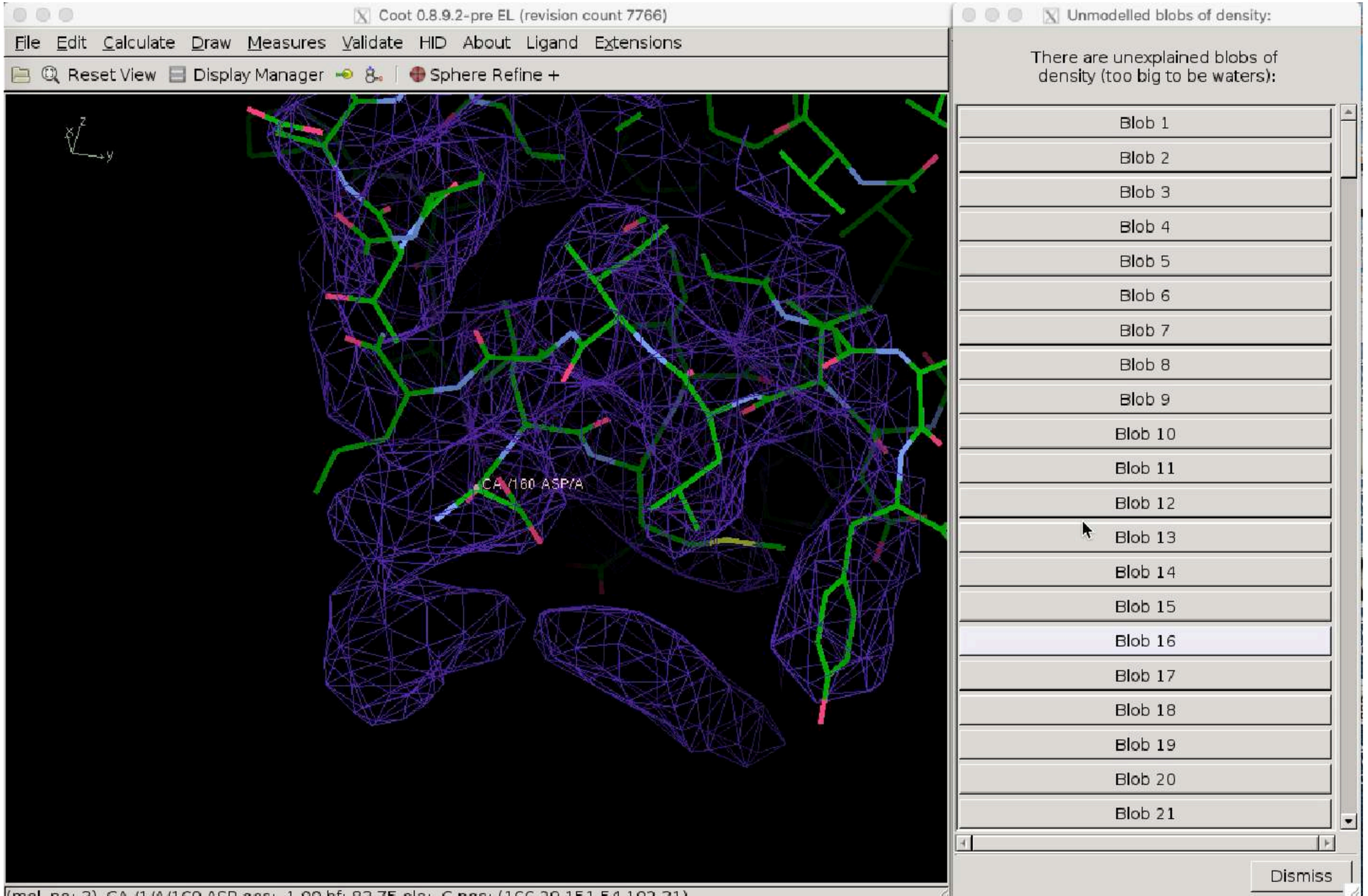
# Unmodeled densities



# Unmodeled densities



# Unmodeled densities



# Molprobity score and clash score

## Summary statistics

| | | | | |
|---|---|---|---|---|
| All-Atom Contacts | Clashscore, all atoms: | 3.82 | | 96th 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 | | 99th 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.

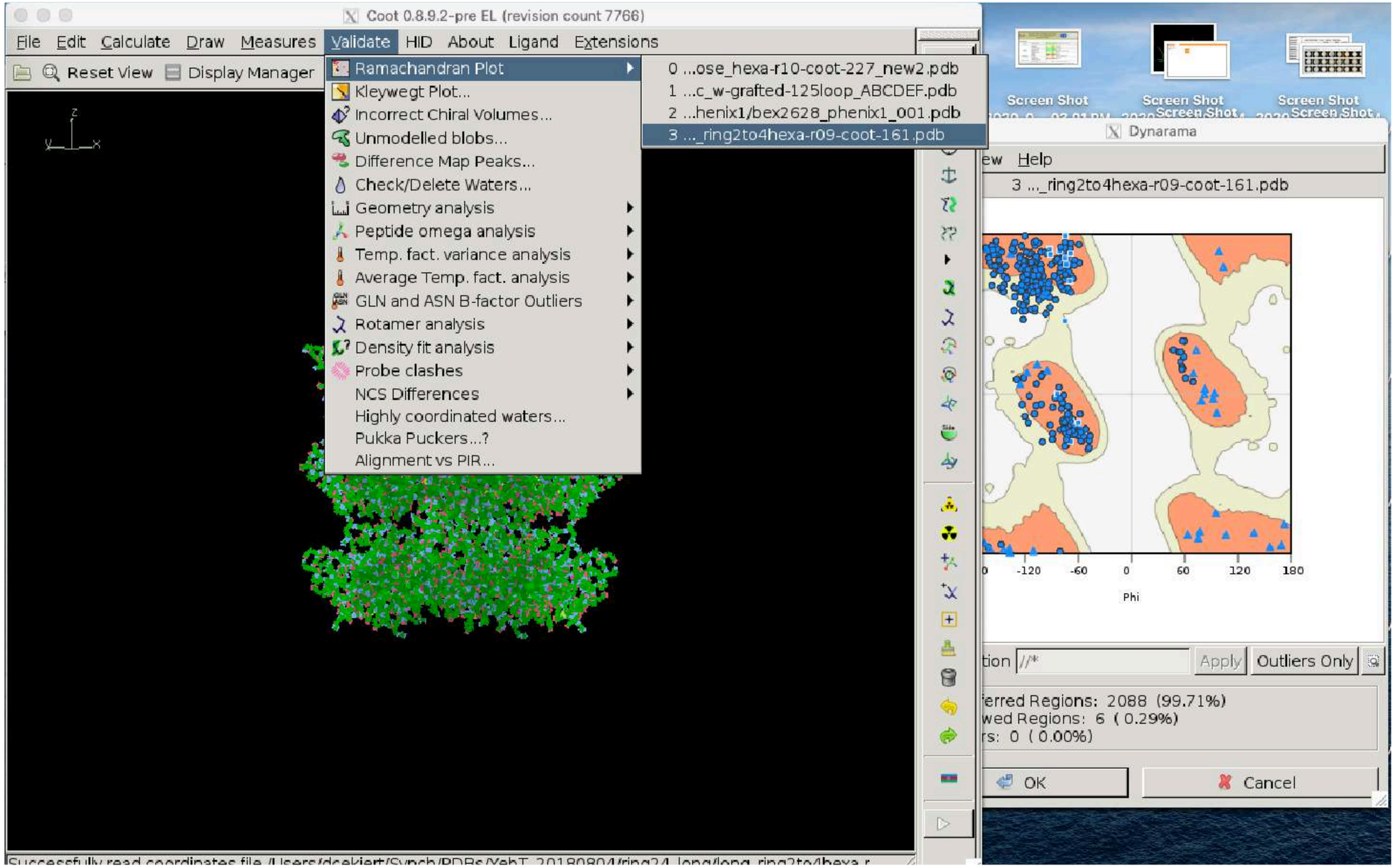
* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.

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

**http://molprobity.biochem.duke.edu/**

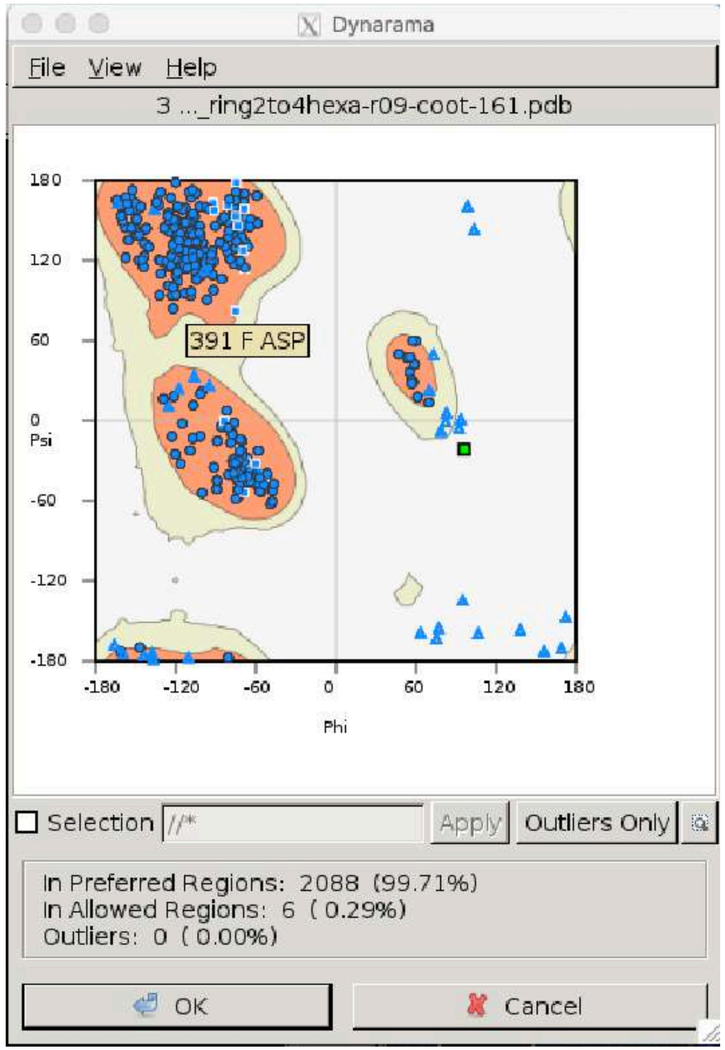
**Williams et al. Protein Science (2018).**

# Ramachandran plot

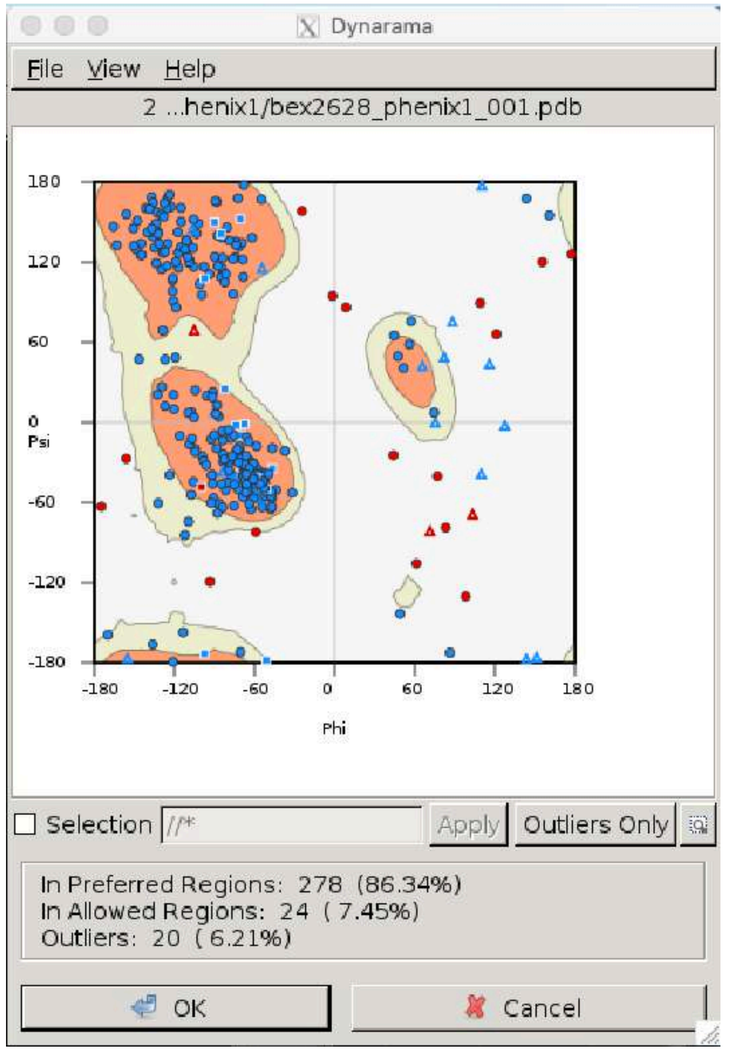


# Ramachandran plot

## Ideal plot

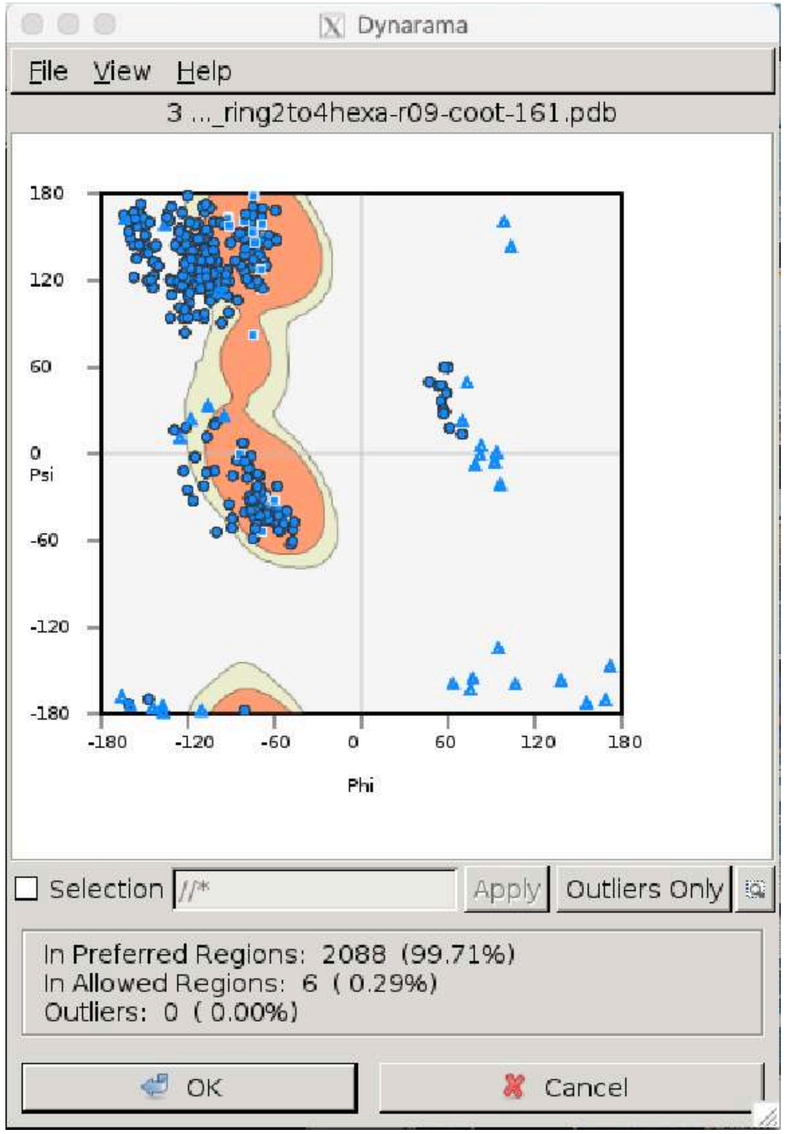



## Problematic plot

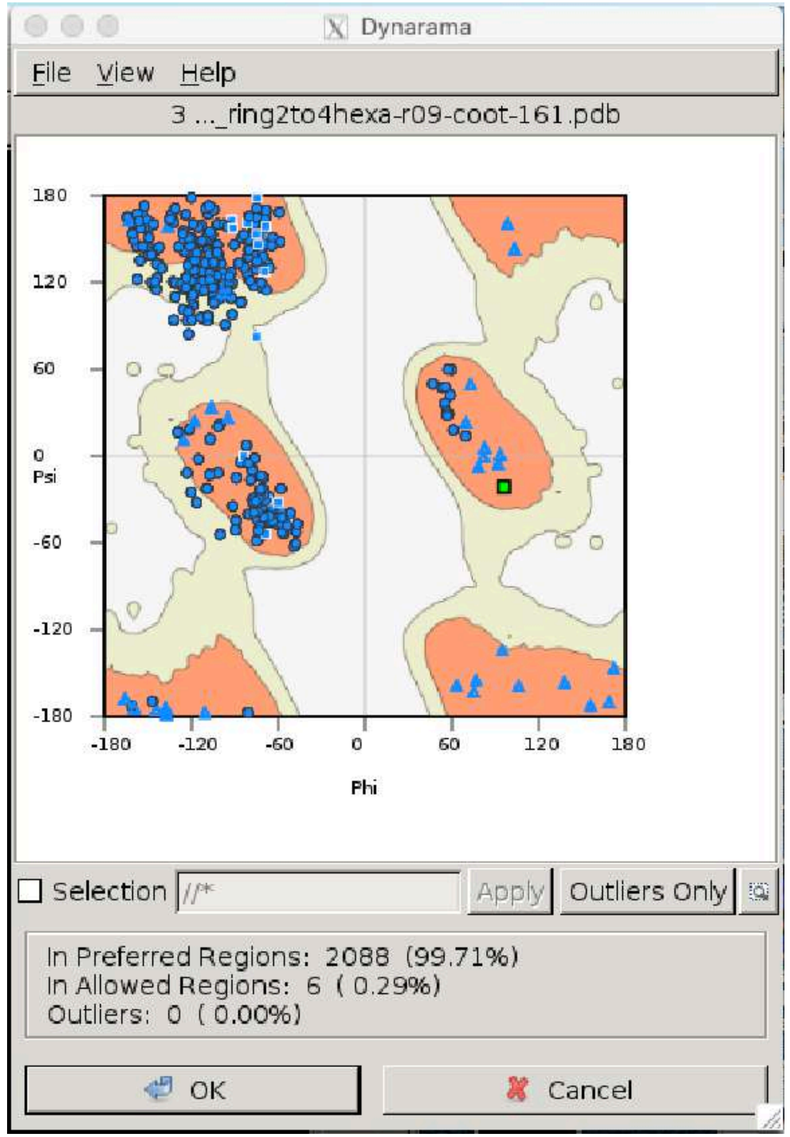


# Ramachandran plot

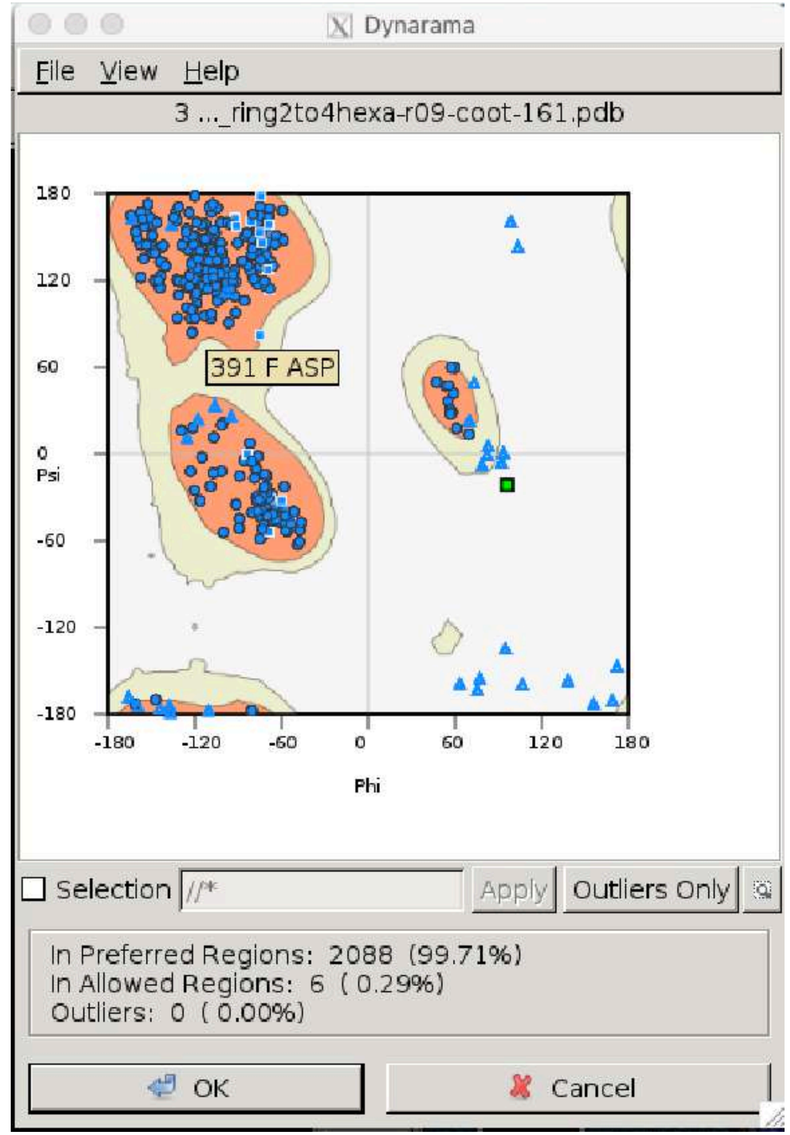
## Proline Allowed Regions




## Glycine Allowed Regions




## Allowed Regions (non-Pro/Gly)



# Cis peptide bonds

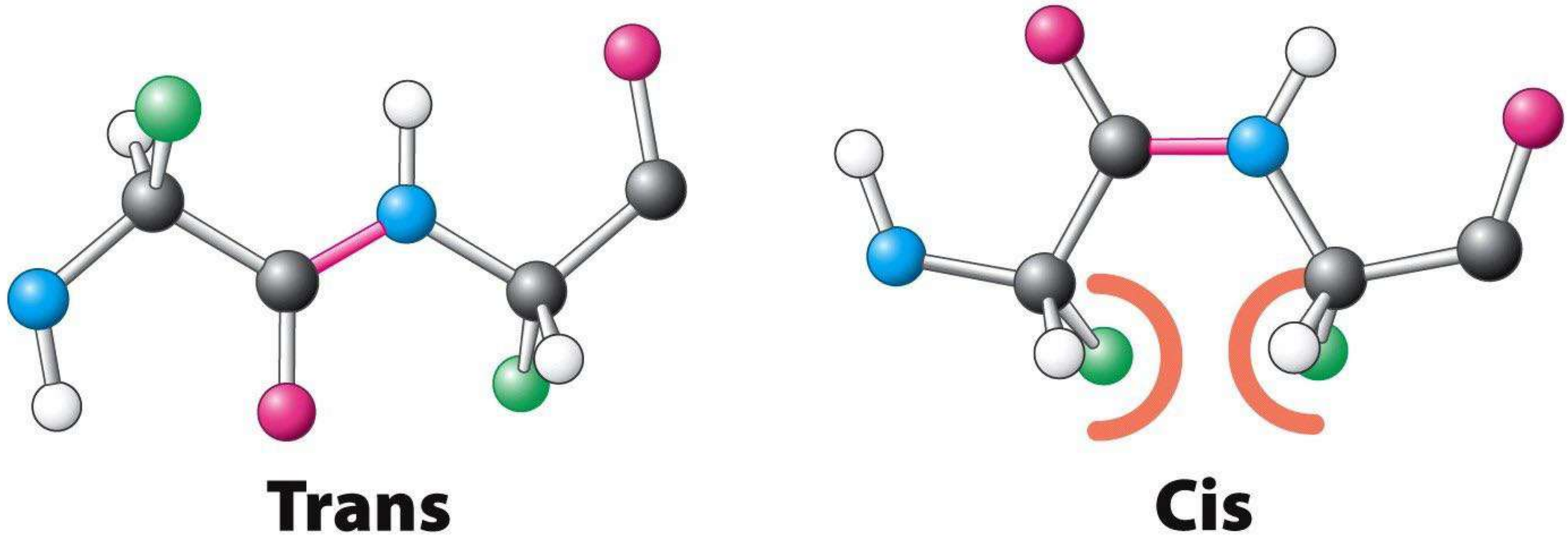



**Figure 2.20**
***Biochemistry*, Seventh Edition**

# Cis peptide bonds

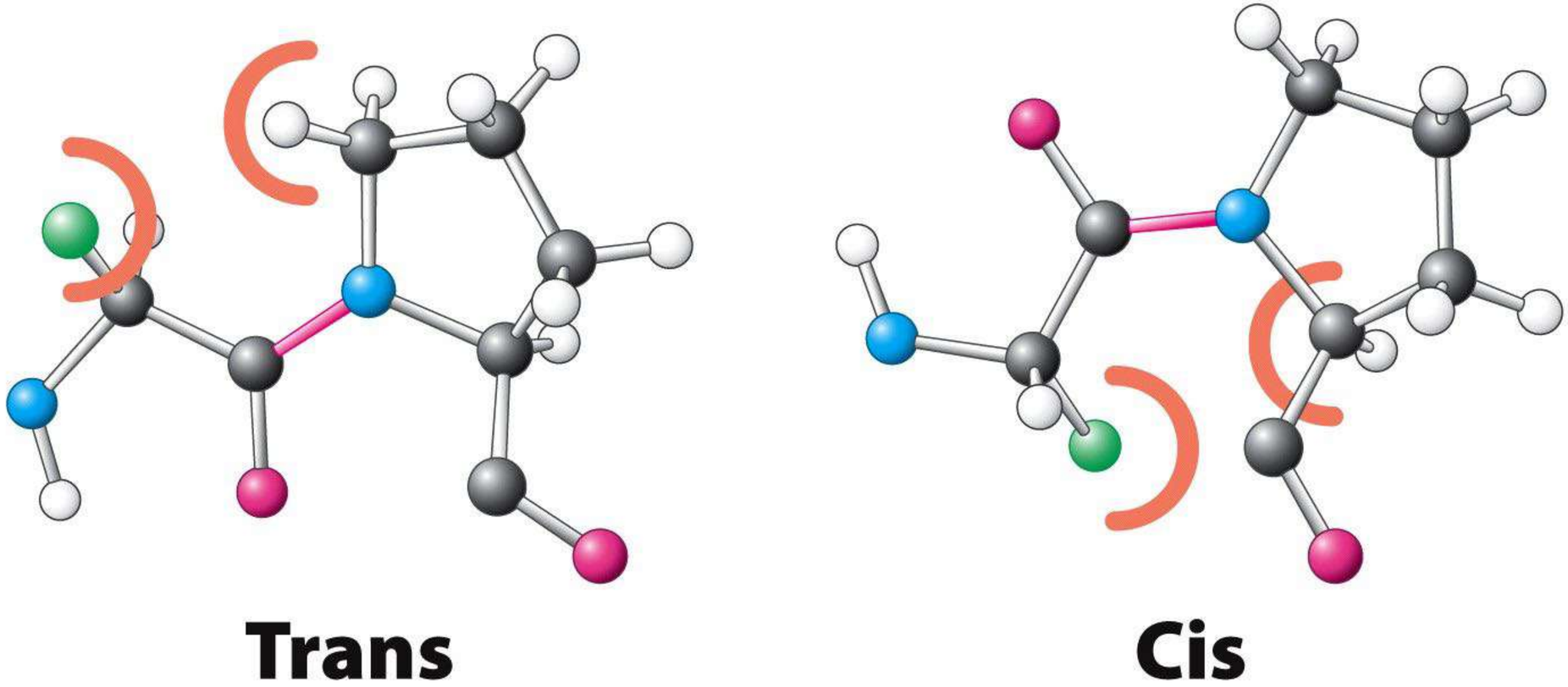
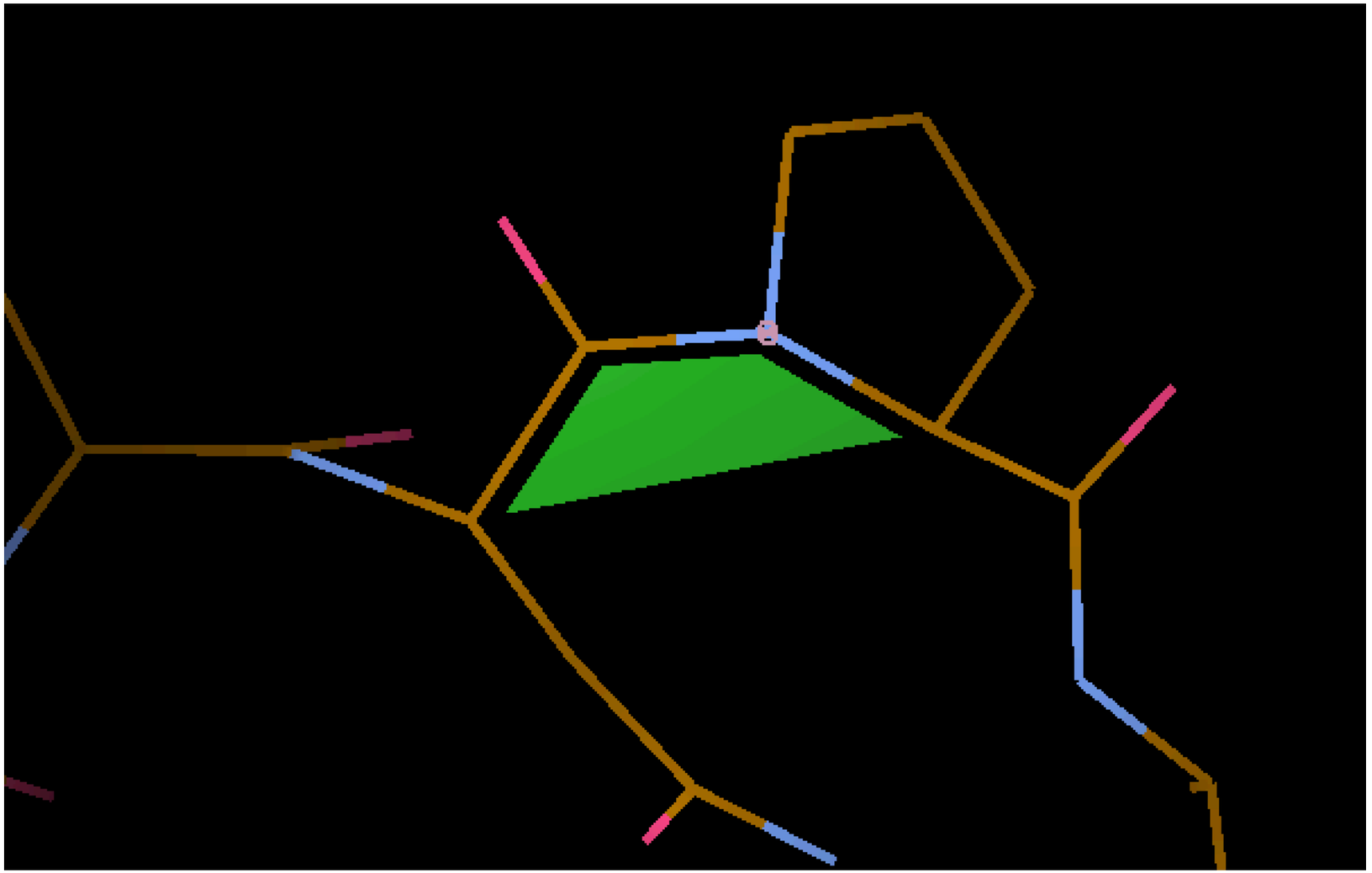



**Figure 2.21**
***Biochemistry*, Seventh Edition**

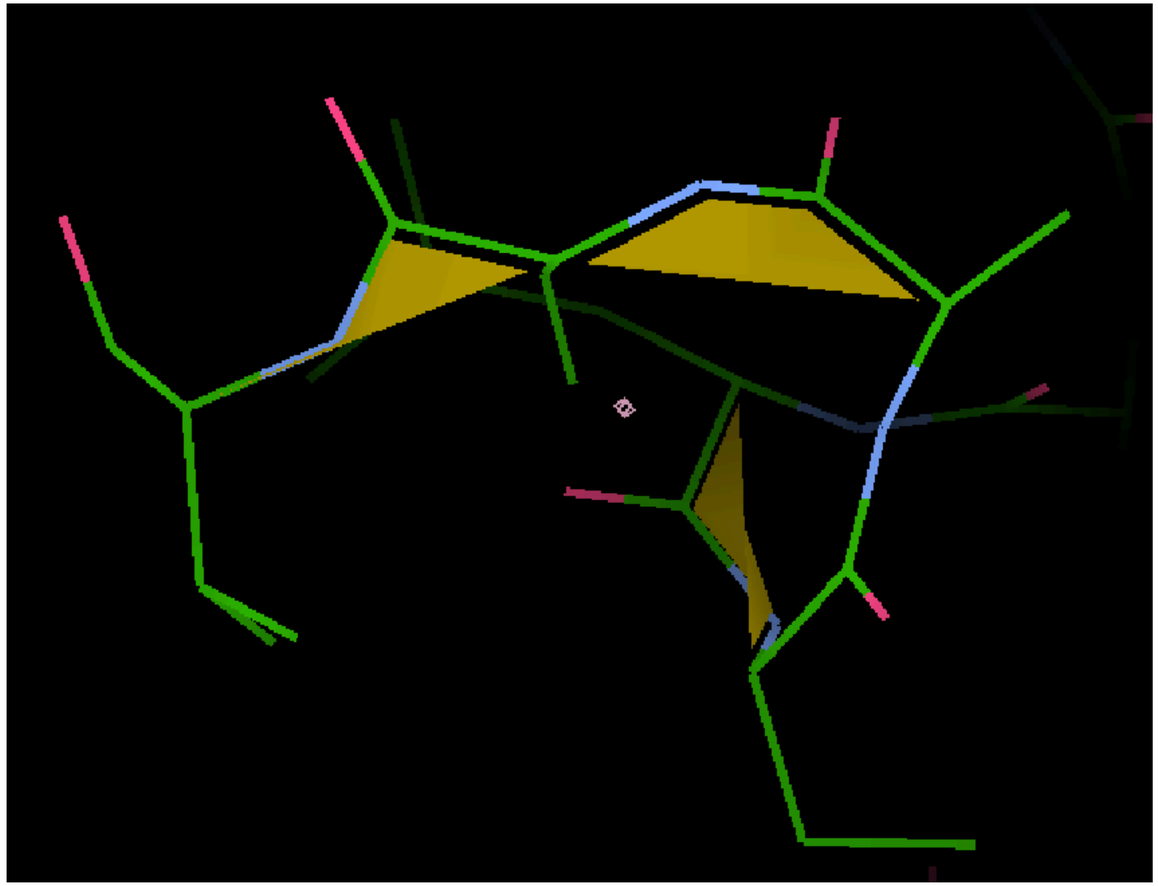
# Cis peptide bonds

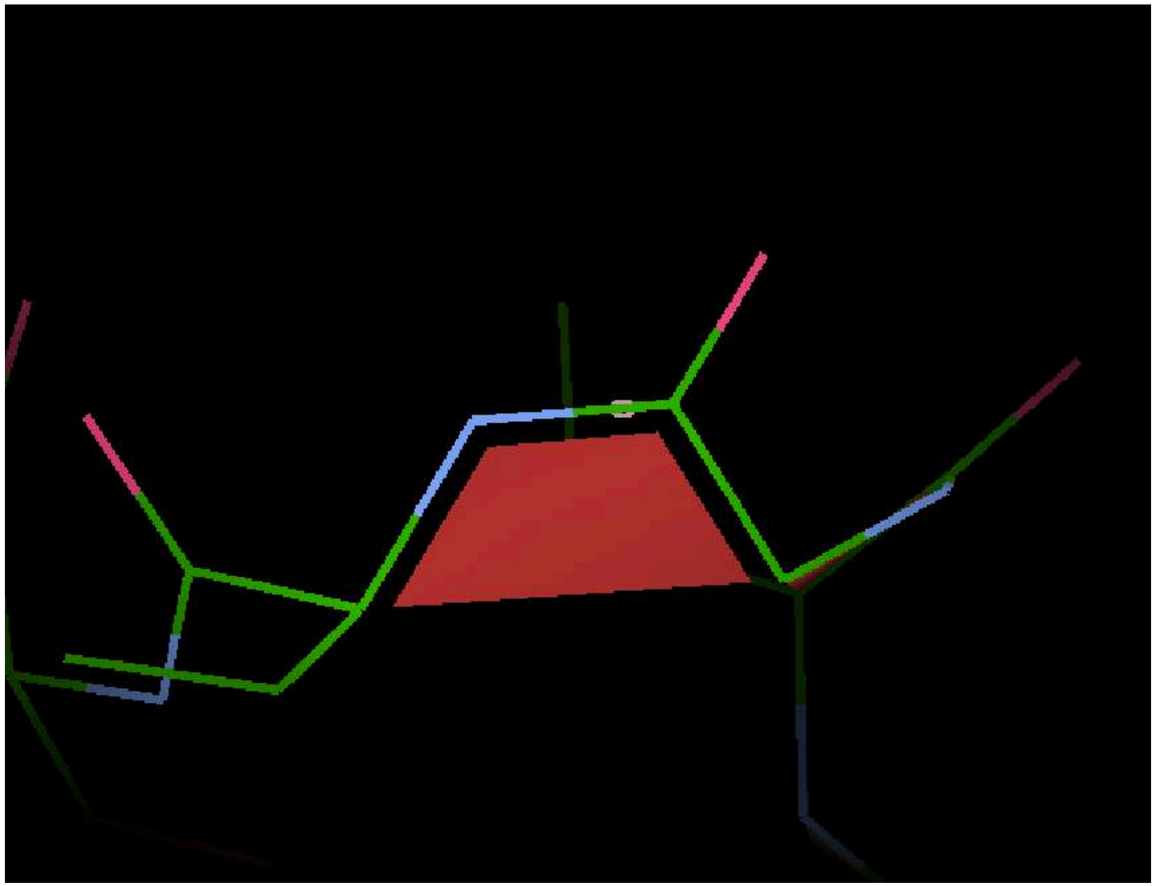


**Coot highlights all cis and non-planar peptide bonds, and color codes them to make potential problems easy to ID**

**Green = cis-Proline (probably OK);  Yellow = non-planar peptide bond (check!);  Red = non-proline cis peptide bond (check!)**

# Cis peptide bonds

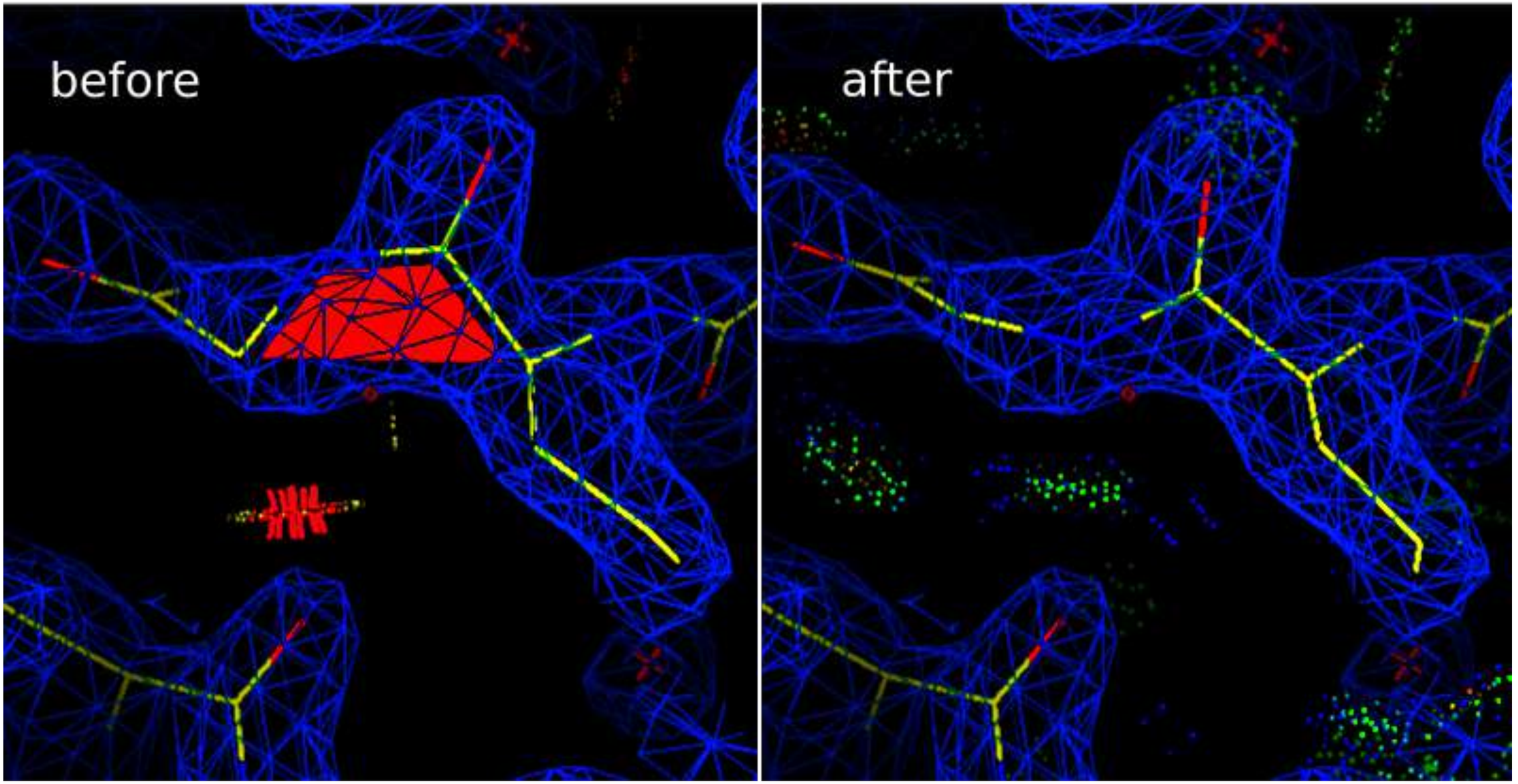




**Coot highlights all cis and non-planar peptide bonds, and color codes them to make potential problems easy to ID**

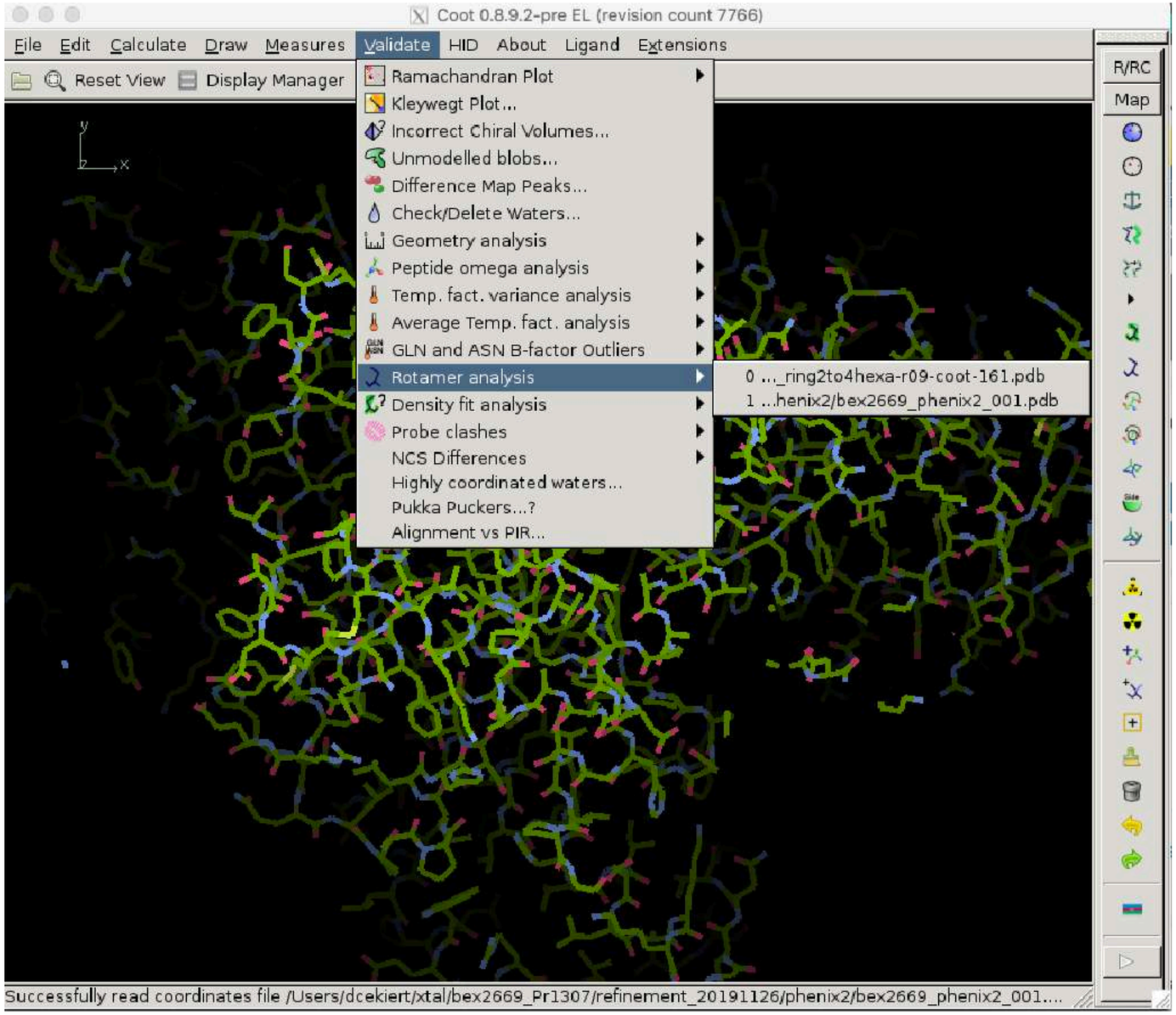
**Green = cis-Proline (probably OK);  Yellow = non-planar peptide bond (check!);  Red = non-proline cis peptide bond (check!)**

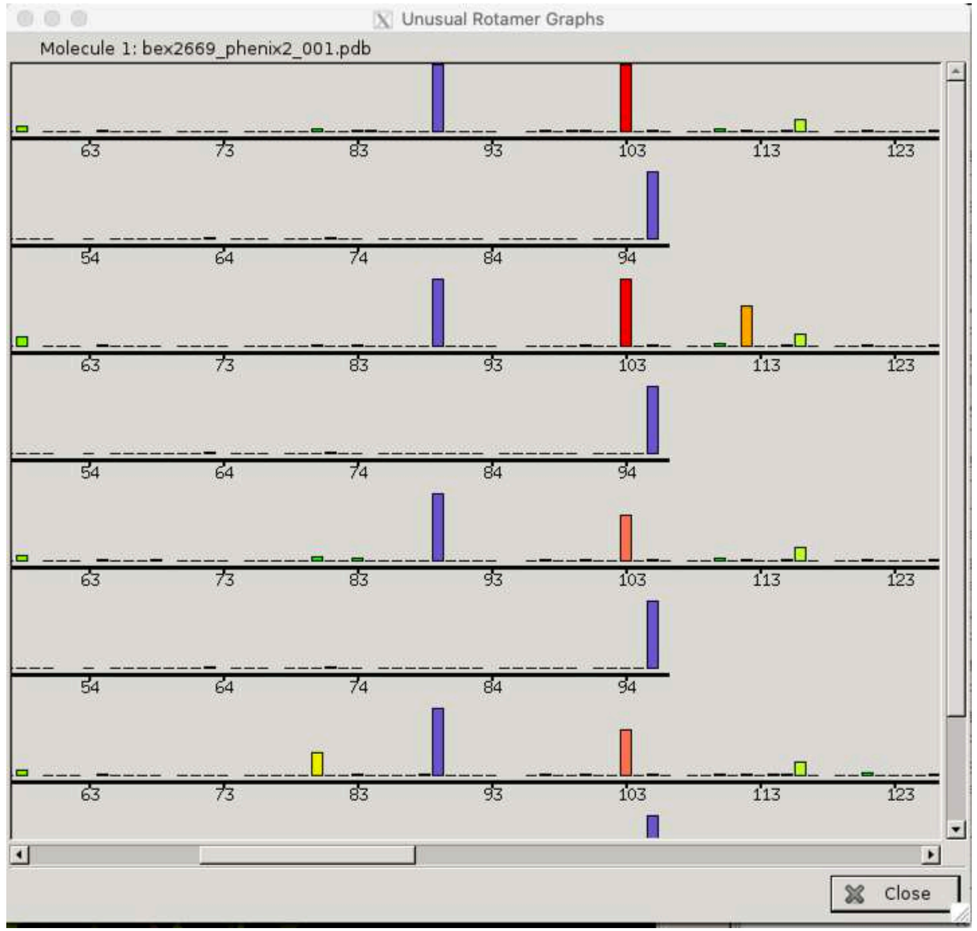
# Cis peptide bonds




https://www.phenix-online.org/documentation/tutorials/molprobity.html

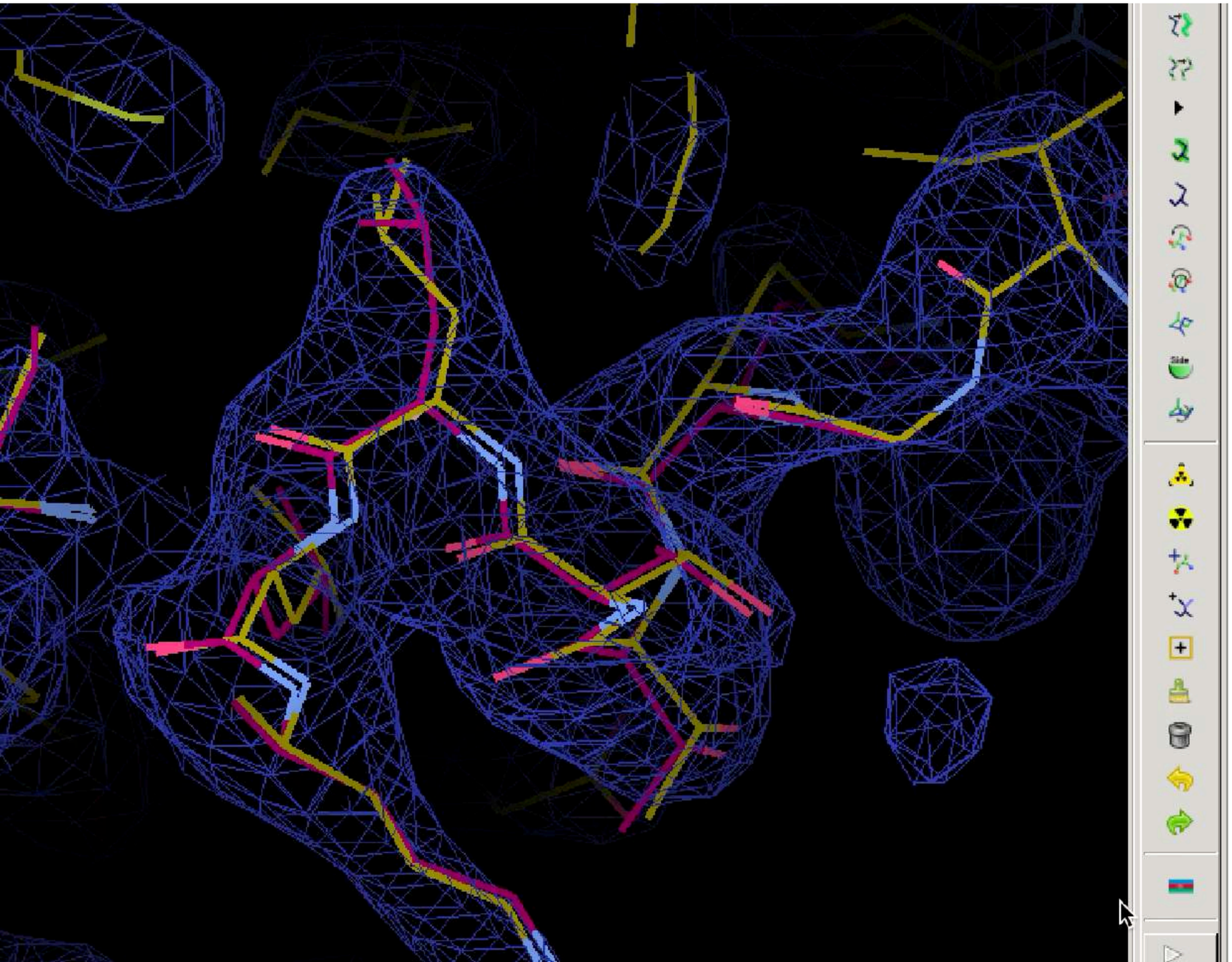
# Rotamer outliers: Coot




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

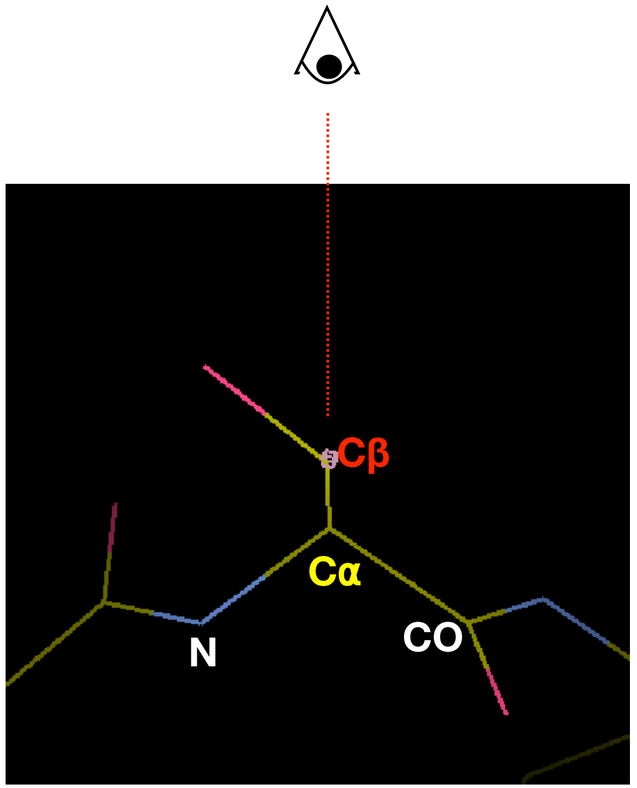
# Cbeta deviations

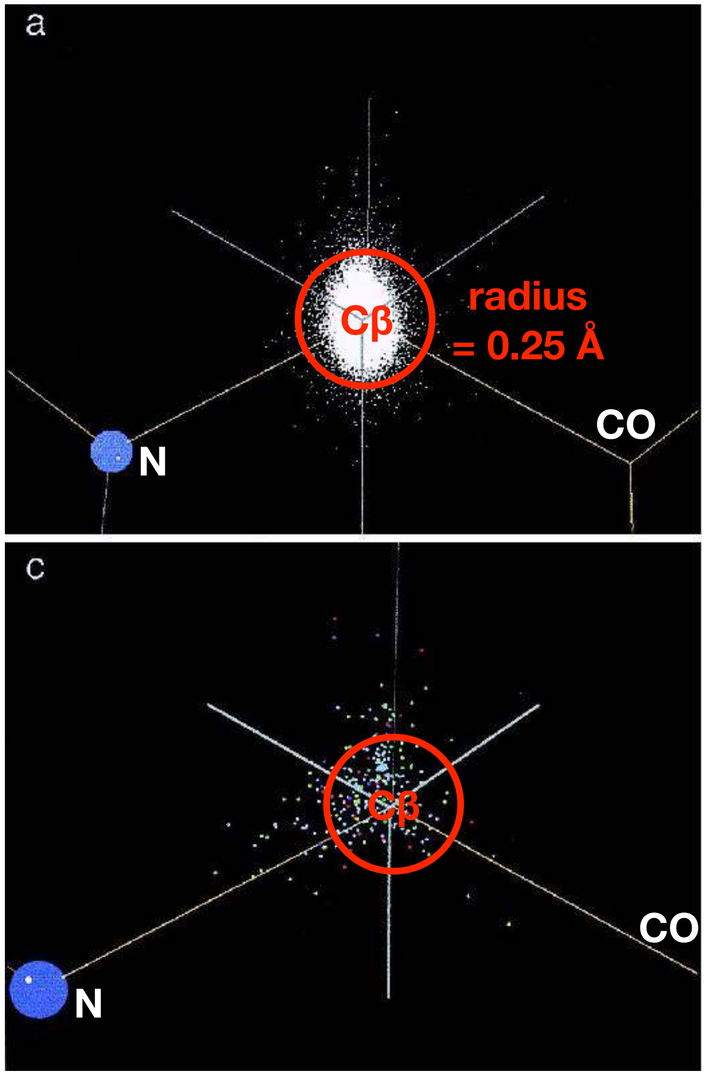


**Which Leu rotamer is correct???**

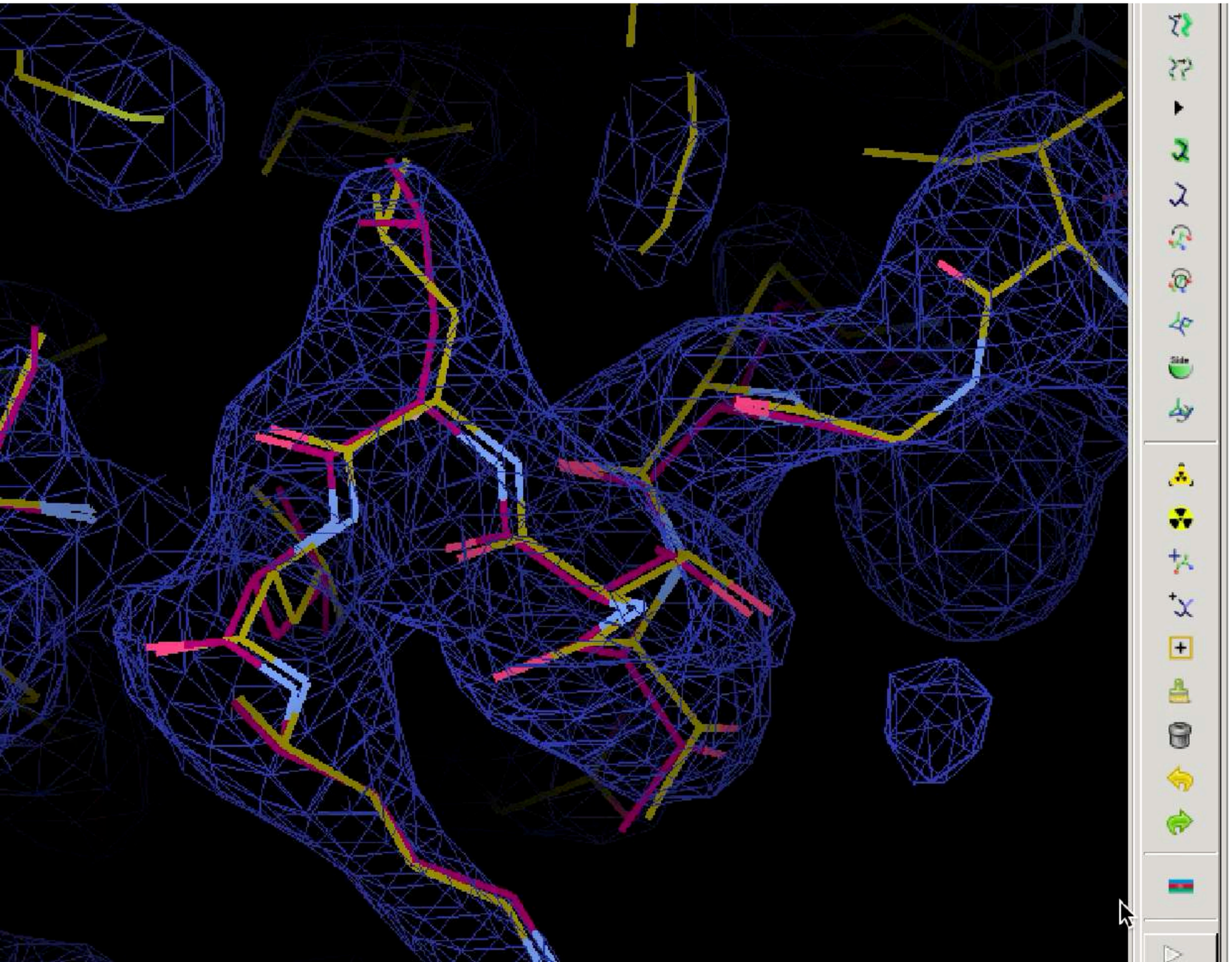
# Cbeta deviations

**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.**



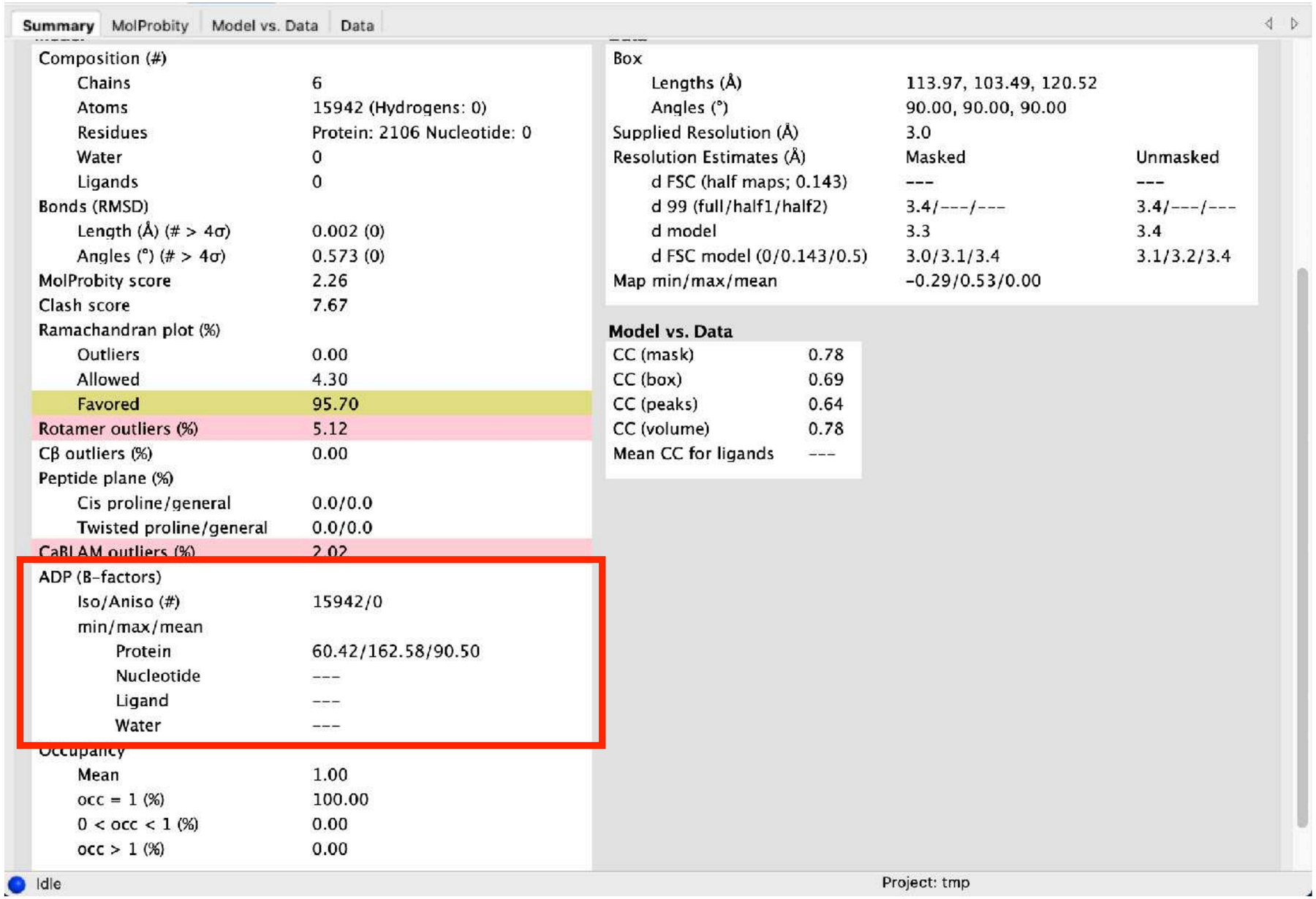
# Cbeta deviations



**Which Leu rotamer is correct???**

Correct answer:
YELLOW

# ADP outliers



Summary | MolProbity | Model vs. Data | Data

| | |
|---|---|
| Composition (#) | |
| Chains | 6 |
| Atoms | 15942 (Hydrogens: 0) |
| Residues | Protein: 2106 Nucleotide: 0 |
| Water | 0 |
| Ligands | 0 |
| Bonds (RMSD) | |
| Length (Å) (# > 4σ) | 0.002 (0) |
| Angles (°) (# > 4σ) | 0.573 (0) |
| MolProbity score | 2.26 |
| Clash score | 7.67 |
| Ramachandran plot (%) | |
| Outliers | 0.00 |
| Allowed | 4.30 |
| Favored | 95.70 |
| Rotamer outliers (%) | 5.12 |
| 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 |

| | | |
|---|---|---|
| Box | | |
| Lengths (Å) | 113.97, 103.49, 120.52 | |
| Angles (°) | 90.00, 90.00, 90.00 | |
| Supplied Resolution (Å) | 3.0 | |
| Resolution Estimates (Å) | Masked | Unmasked |
| d FSC (half maps; 0.143) | --- | --- |
| d 99 (full/half1/half2) | 3.4/---/--- | 3.4/---/--- |
| d model | 3.3 | 3.4 |
| d FSC model (0/0.143/0.5) | 3.0/3.1/3.4 | 3.1/3.2/3.4 |
| Map min/max/mean | -0.29/0.53/0.00 | |

**Model vs. Data**

| | |
|---|---|
| CC (mask) | 0.78 |
| CC (box) | 0.69 |
| CC (peaks) | 0.64 |
| CC (volume) | 0.78 |
| Mean CC for ligands | --- |

Idle

Project: tmp

**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 reseting 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
Molprobity Results
NCS Check
Sequence Check
Real Space CC

**https://qc-check.usc.edu/QC/qc_check.pl**

# 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