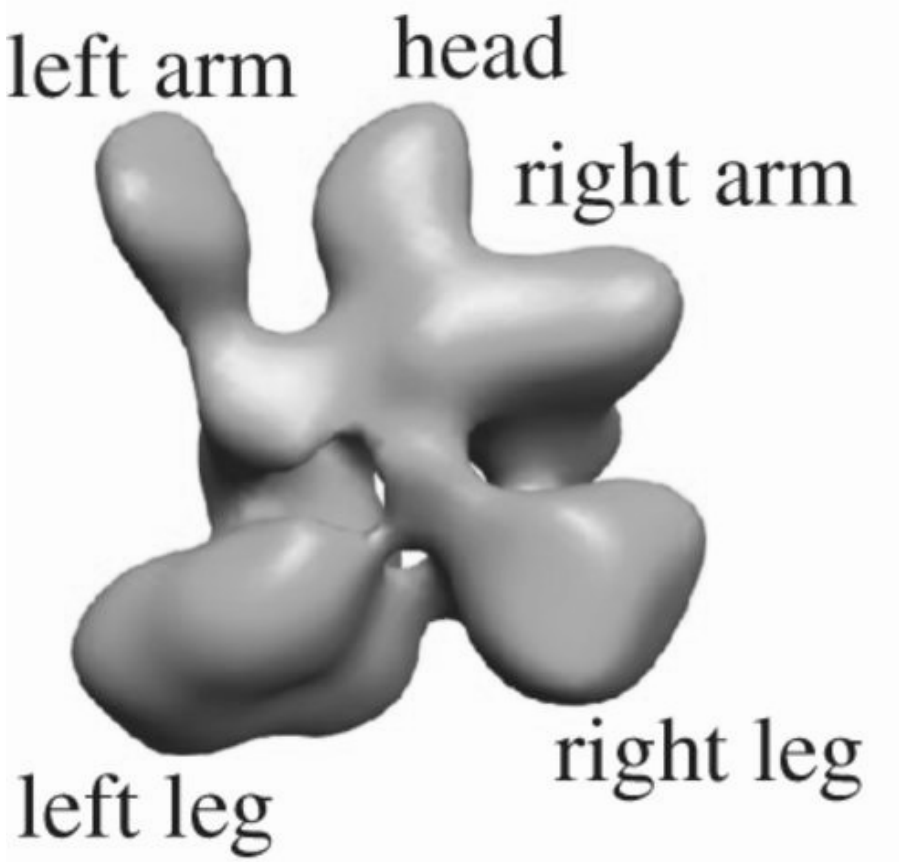
# Model building, refinement, and model validation

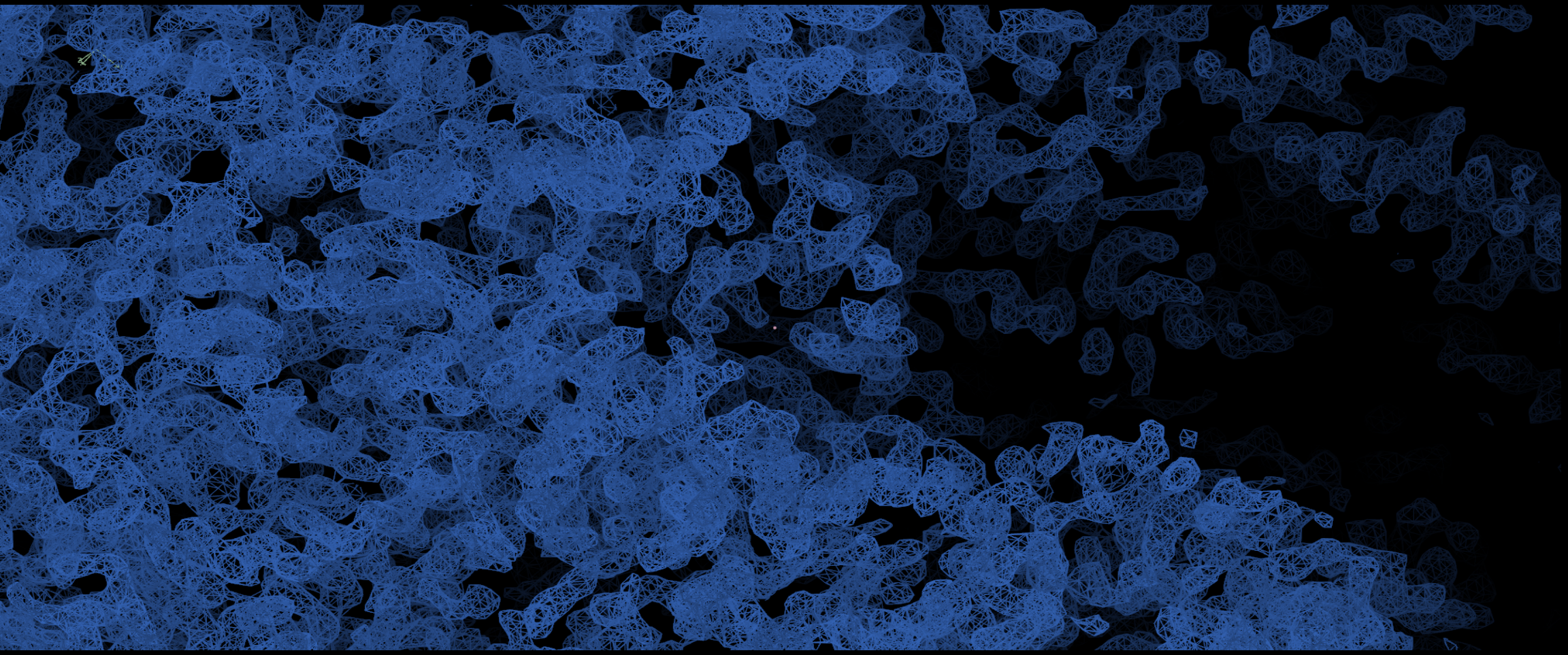
Alan Brown

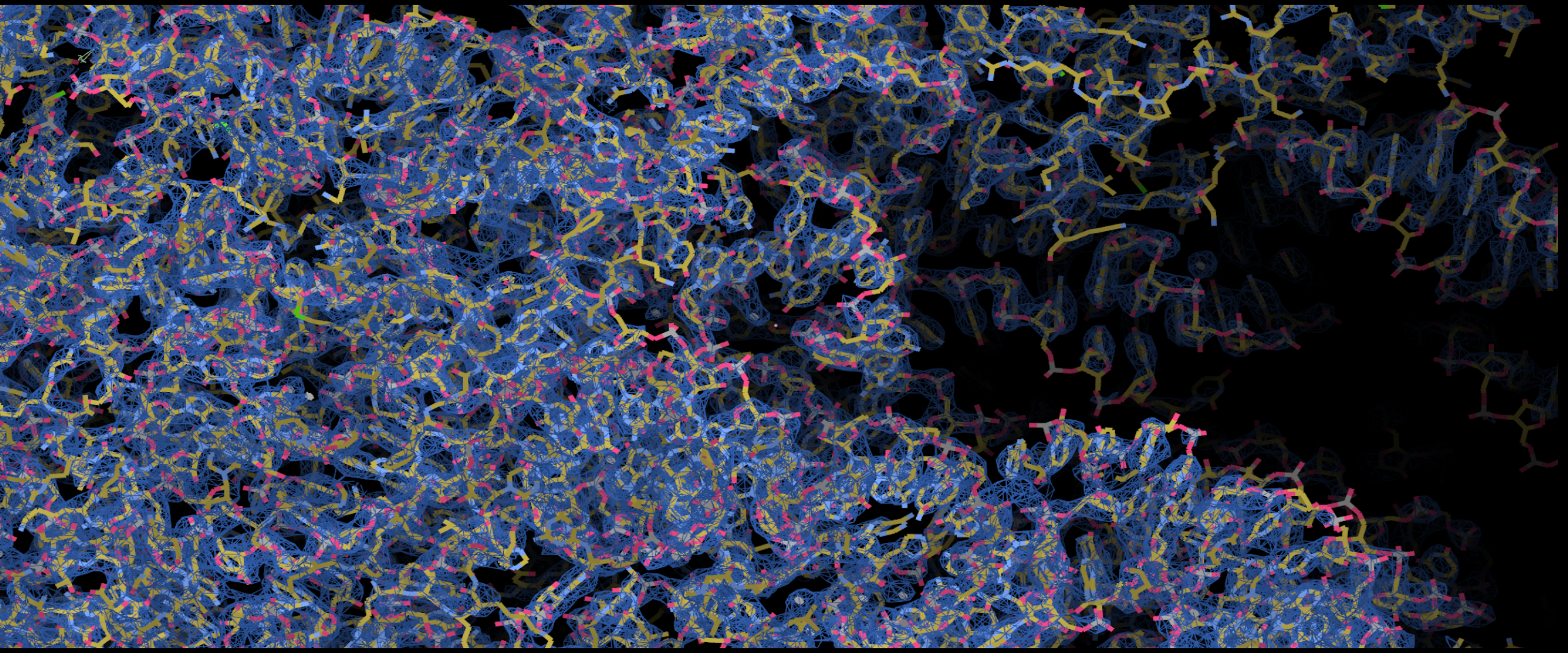
# Then



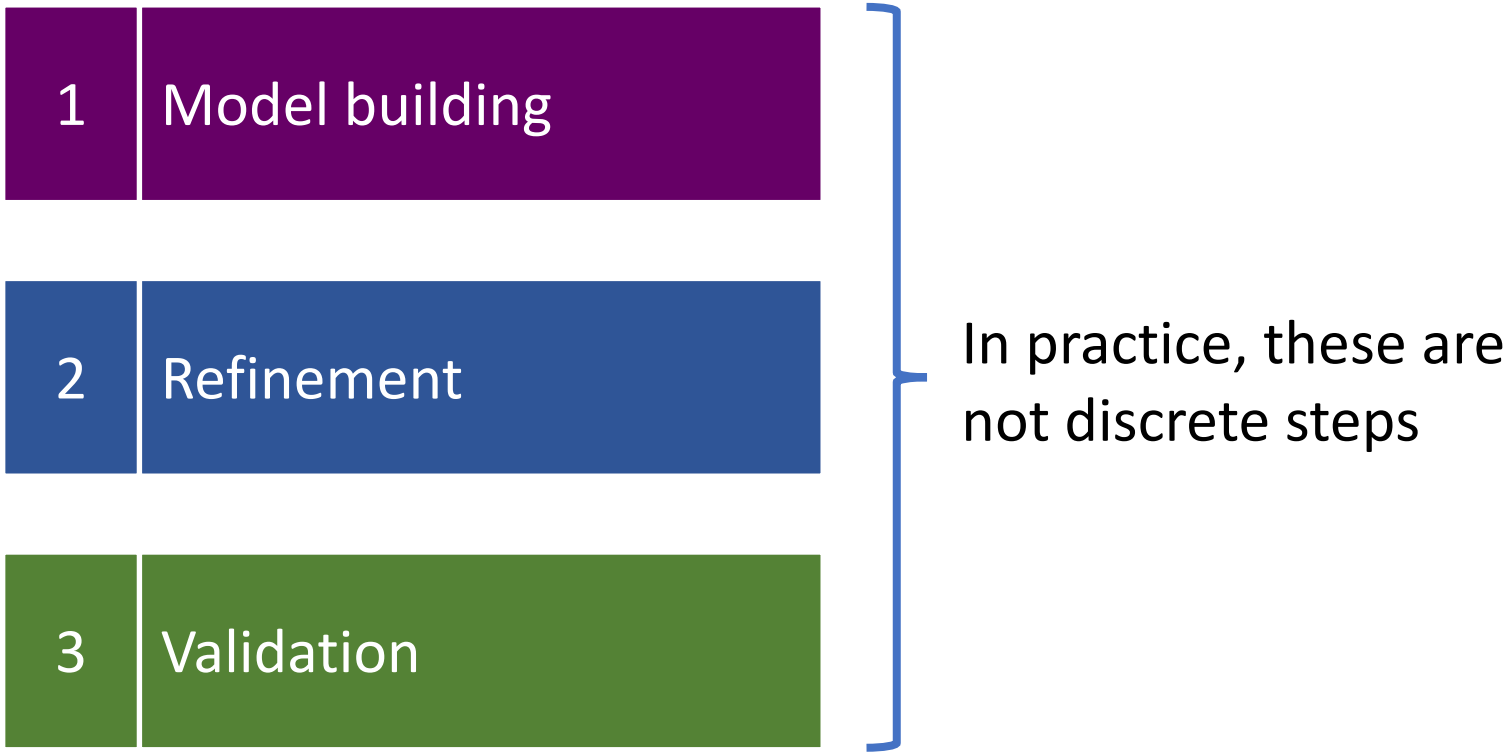

Sun et al (2011) PNAS 108, 20473-20478

# Now





# Organization



# Useful packages for cryo-EM


http://www.ccpem.ac.uk/


http://www.phenix-online.org/


https://www.rosettacommons.org/software/

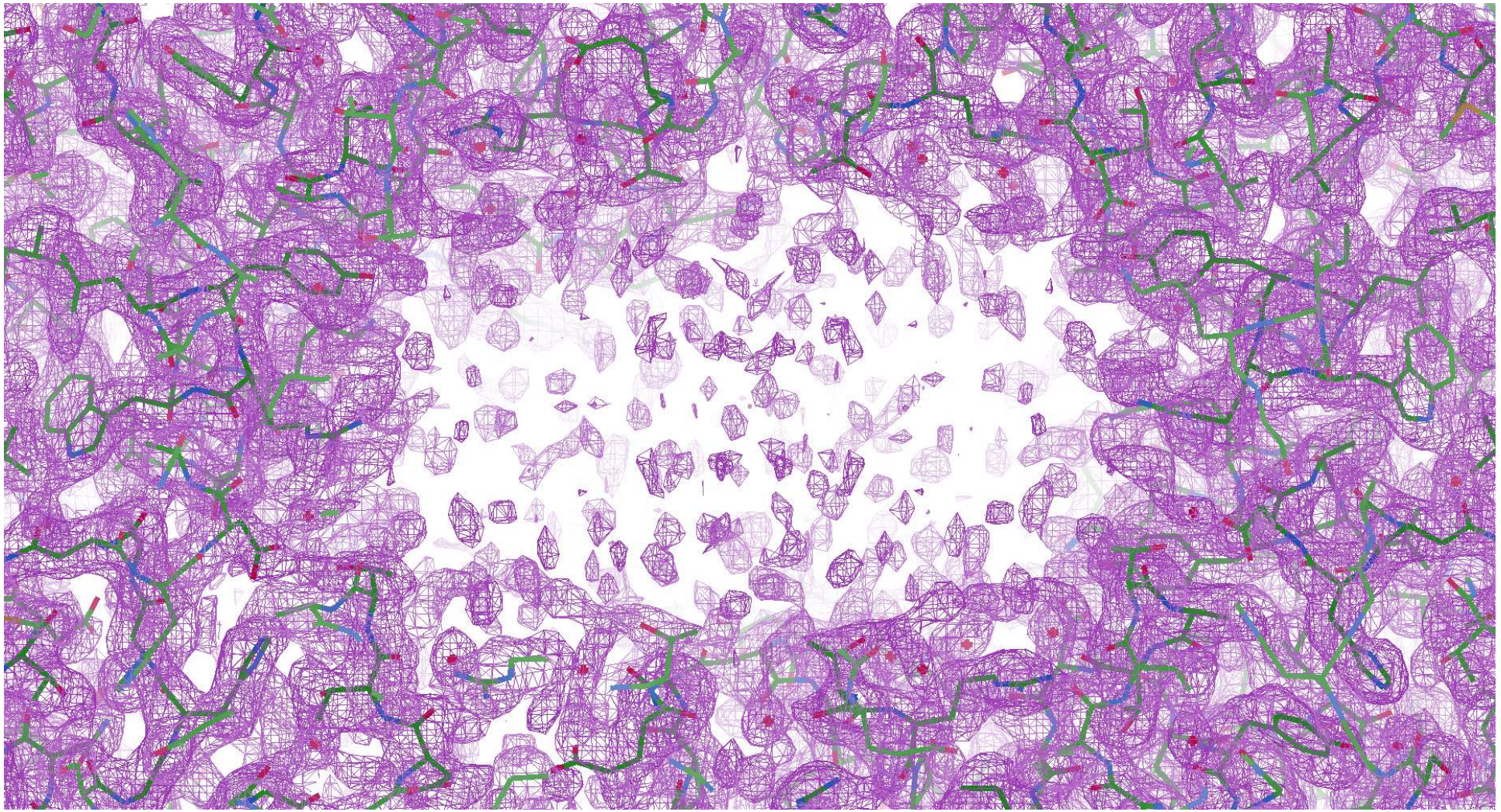
# Maps for model building

# EM maps for model building

- A single B-factor applied to the whole map may not be best for model building /refinement
- Local map sharpening can be extremely useful for model building

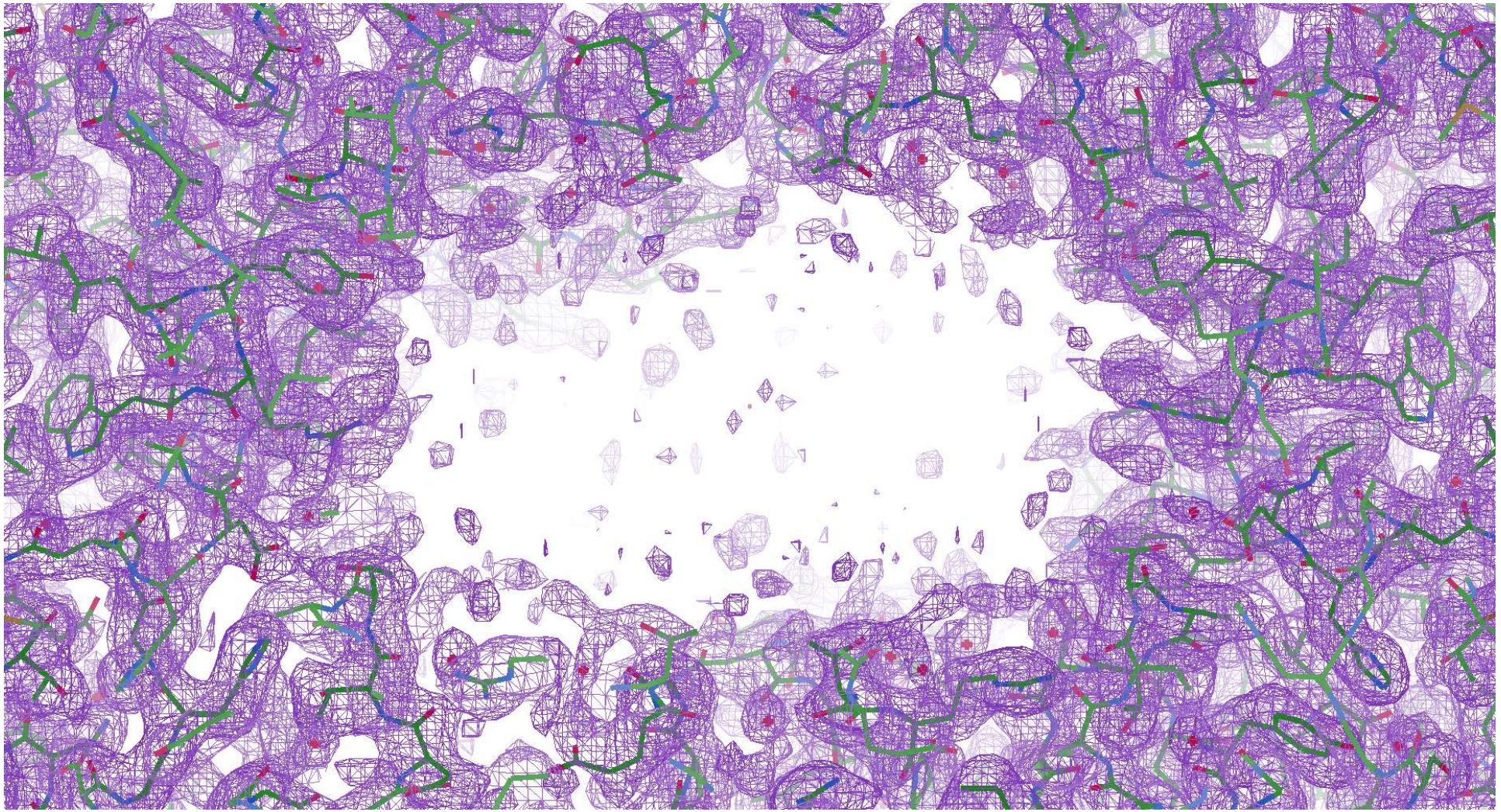
# Map sharpening/blurring

Deposited map (EMD-2984)



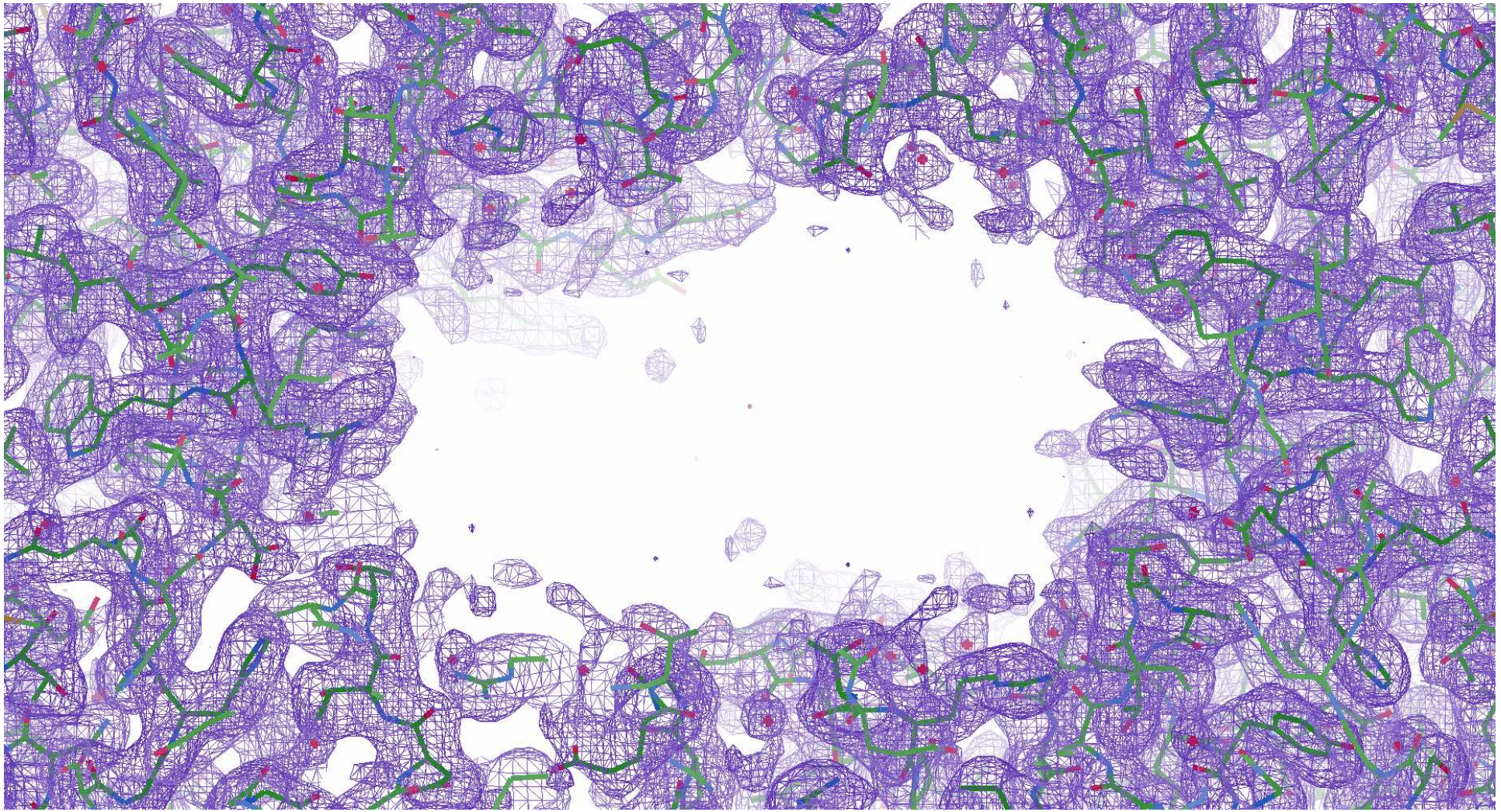
# Map sharpening/blurring

Blur 20 Å$^2$



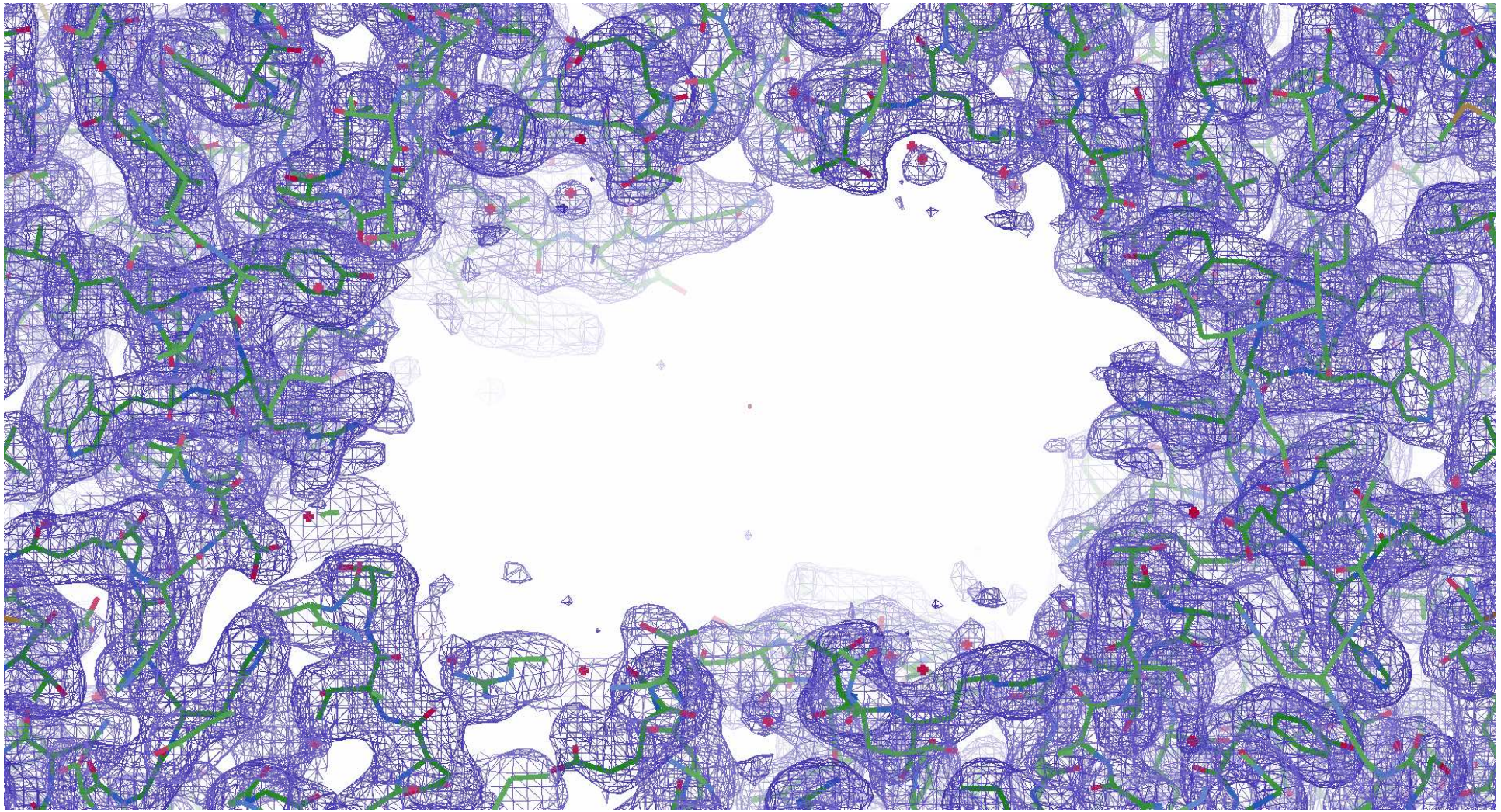
# Map sharpening/blurring

Blur 40 Å$^2$



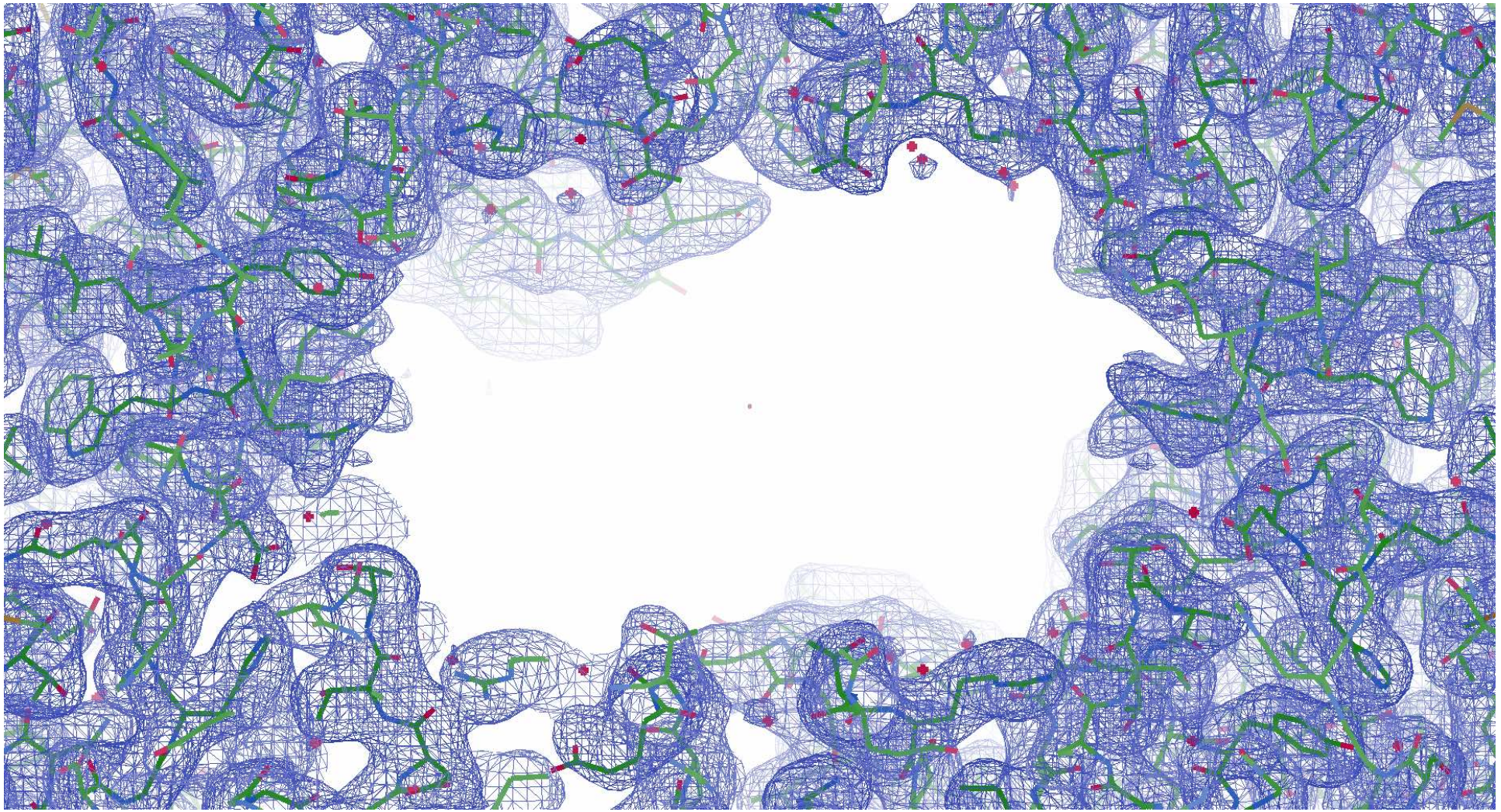
# Map sharpening/blurring

Blur 60 Å$^2$



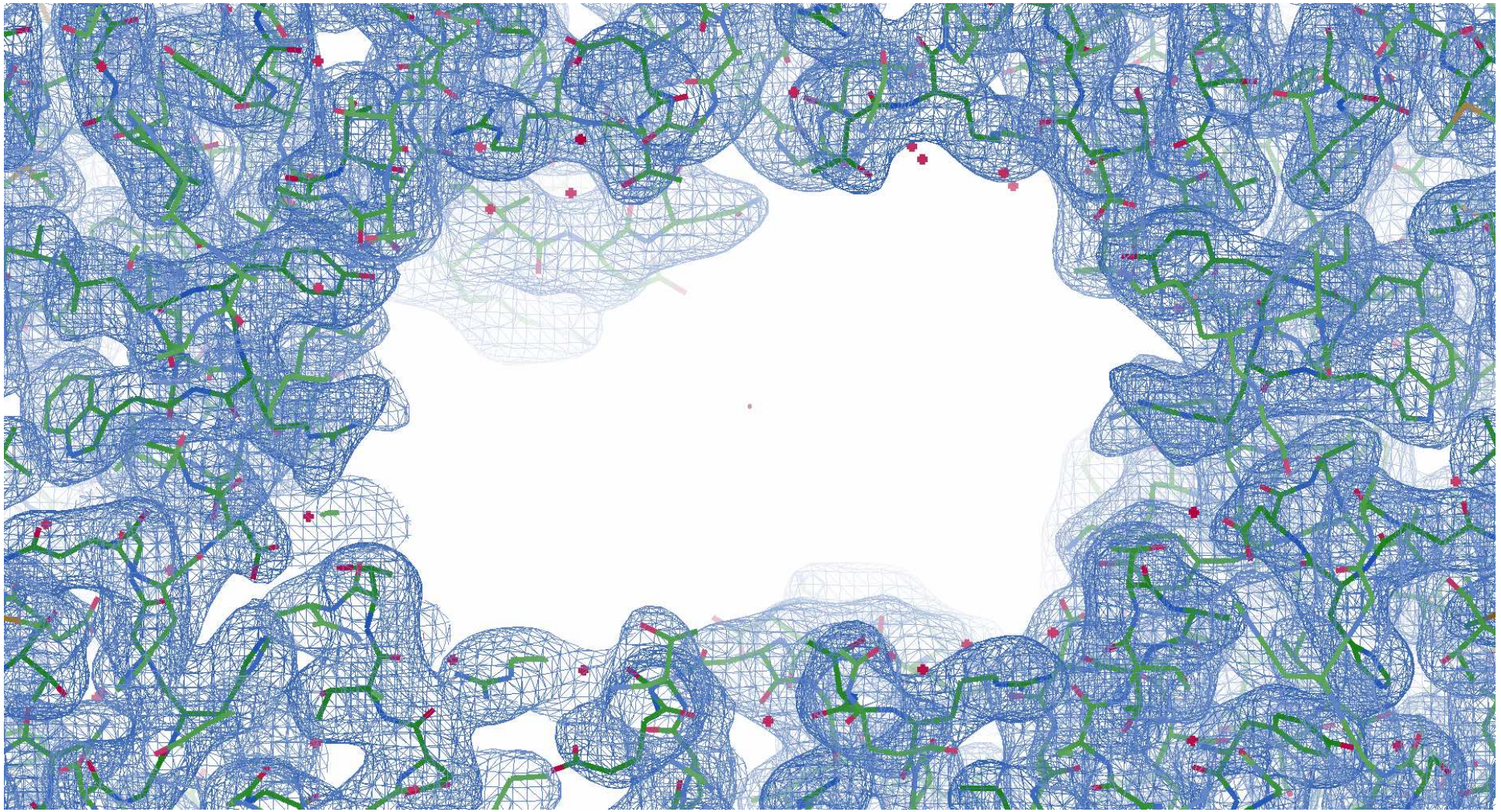
# Map sharpening/blurring

Blur 80 Å$^2$



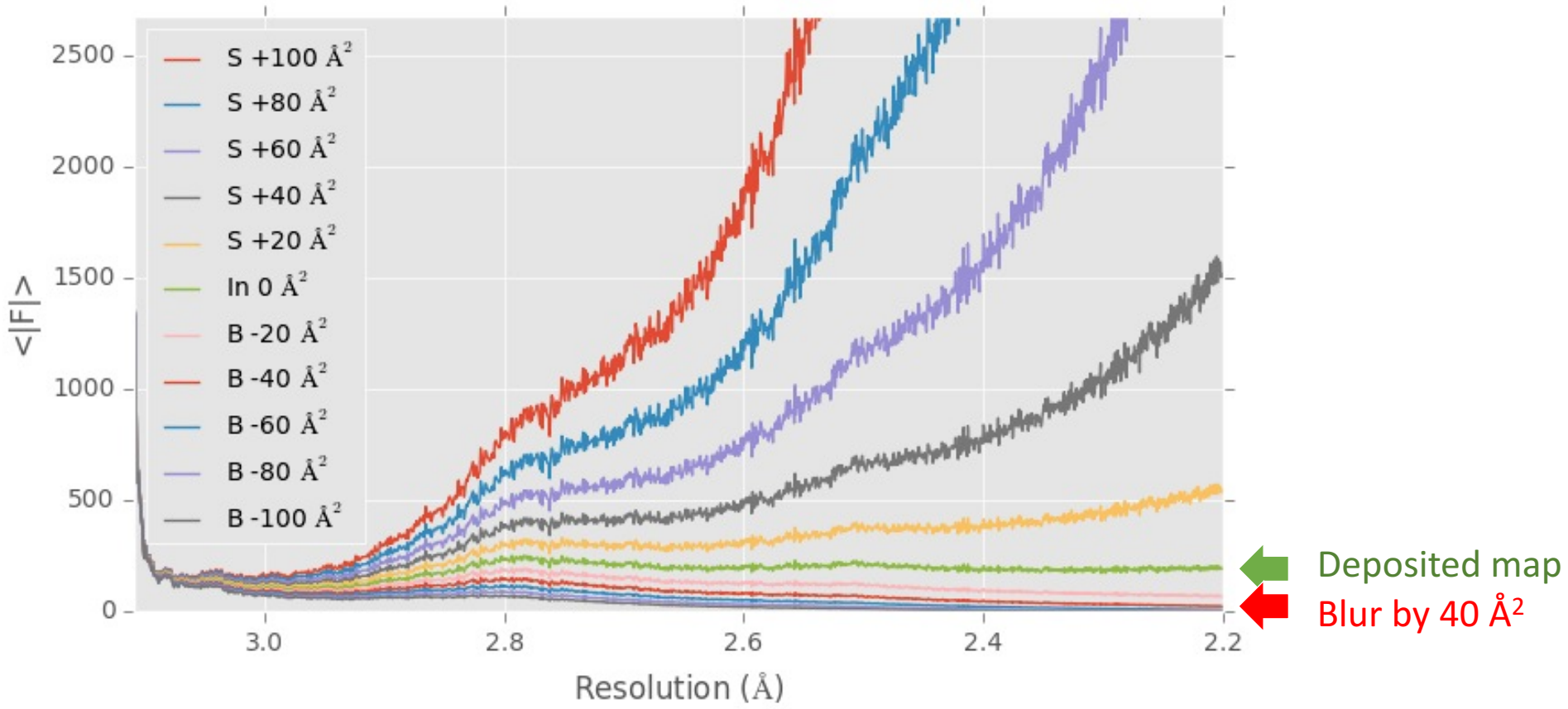
# Map sharpening/blurring

Blur 100 Å$^2$



# Map sharpening/blurring in CCP-EM

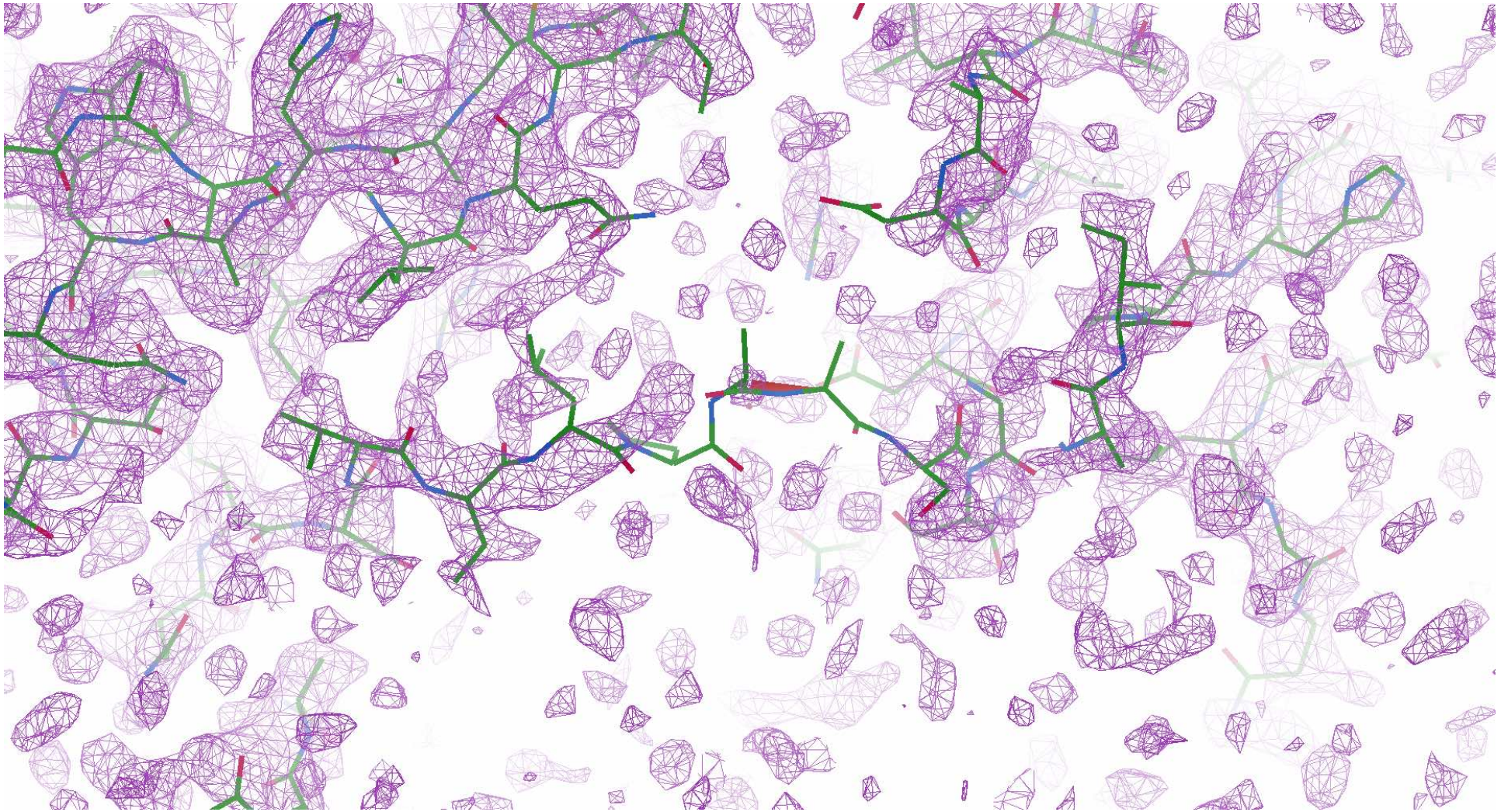
CCP-EM > MRC to MTZ




- The plot shows the mean structure factor amplitude (<|F|>) vs resolution (1 / Å) plot
- It should reach 0 at high resolution

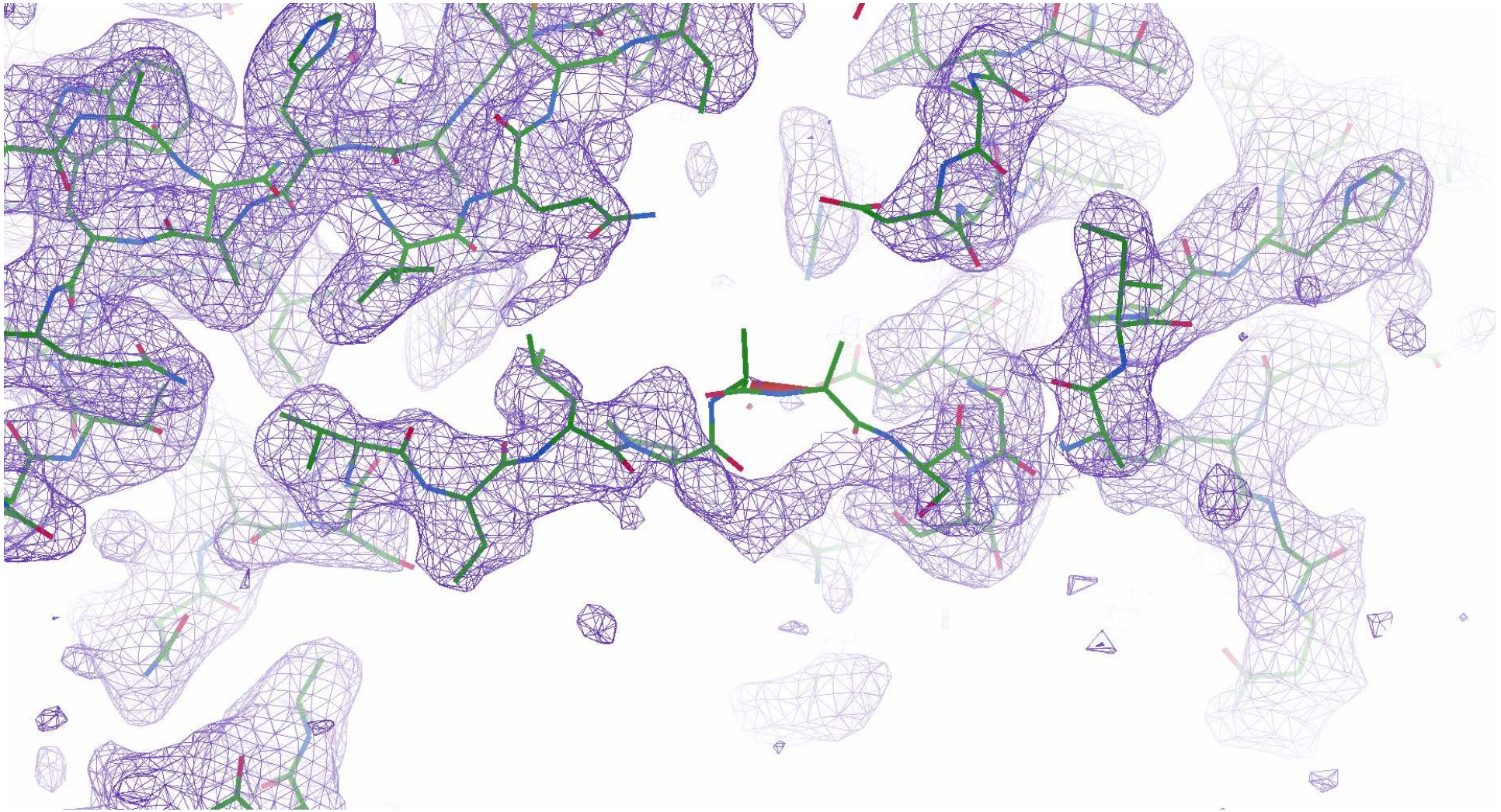
# Map sharpening/blurring aids model building

Deposited map (EMD-2984)



# Map sharpening/blurring aids model building

Blur 40 Å$^2$

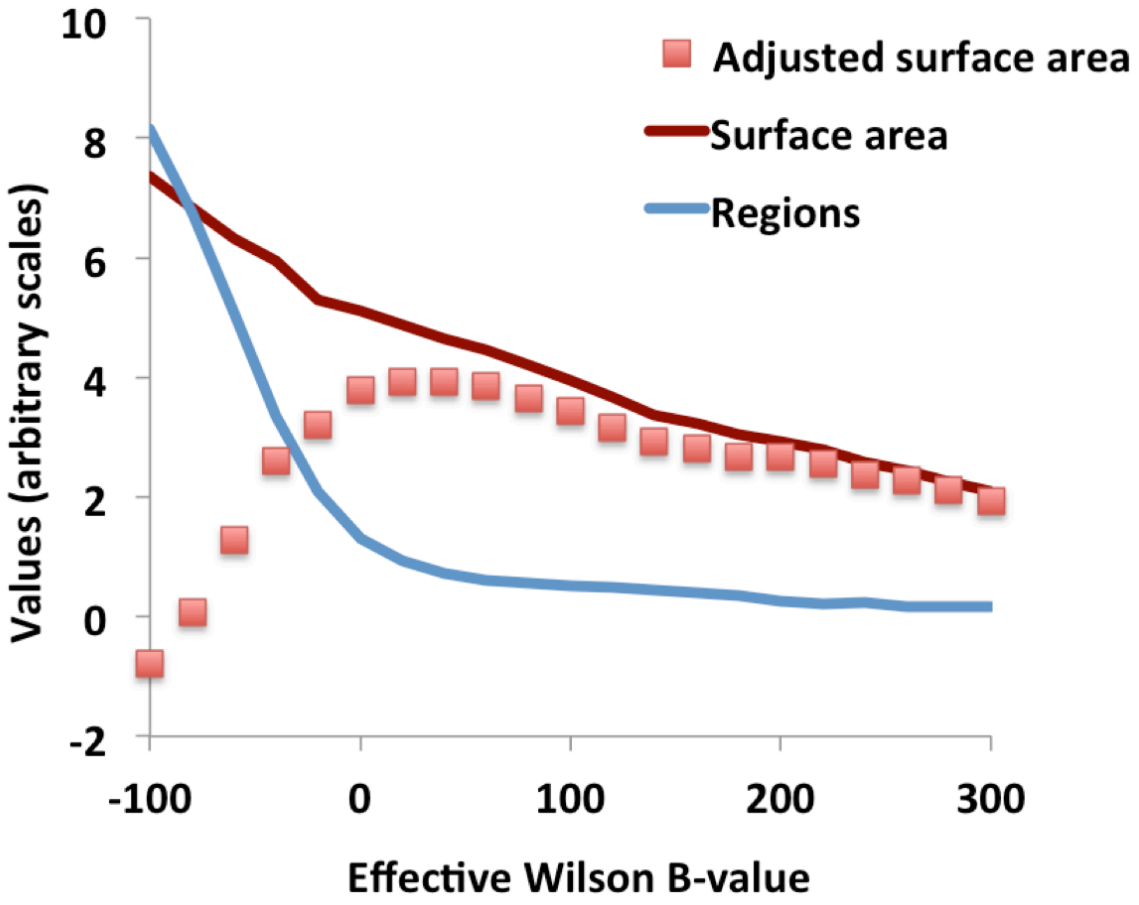


# Automated map sharpening

Developer: Tom Terwilliger

**Aim**:

- Maximize surface are
- As fewer as possible continuous regions




Adjusted surface area = surface area - weight*no. of regions

# Other map manipulations

- strongly recommended to avoid standard crystallographic procedures for map modification with cryo-EM maps
- For example, 2Fo-Fc maps are necessarily model biased
- Any new method of “map improvement” must be rigorously tested
- The safest approach is always to use the observed maps; these maps are the last link between the data and the atomic models.

- For an in-depth discussion of the potential misuse of crystallographic maps and density modification in cryo-EM see:
  - Murshudov (2016) Methods in Enzymology, 579:277-305

# Template generation and fold recognition

# Generating starting models

**Structure exists in PDB**

If the model is from X-ray data, use PDB_redo to re-refine the model using the latest software

**Similar structures exist in PDB**

Generate homology models using template-based modeling

**There's nowt like it**

Generate secondary structure predictions (e.g. JPRED)

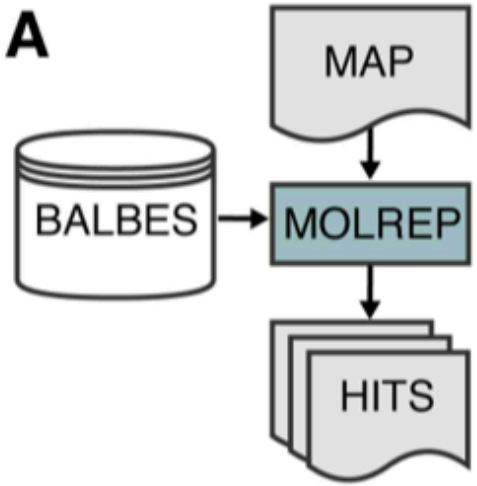
*Ab initio* structure predictions

Some useful programs:

iTASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/)
Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/)
Evfold (http://evfold.org/evfold-web/evfold.do)

# Fold recognition : the BALBES-MOLREP pipeline

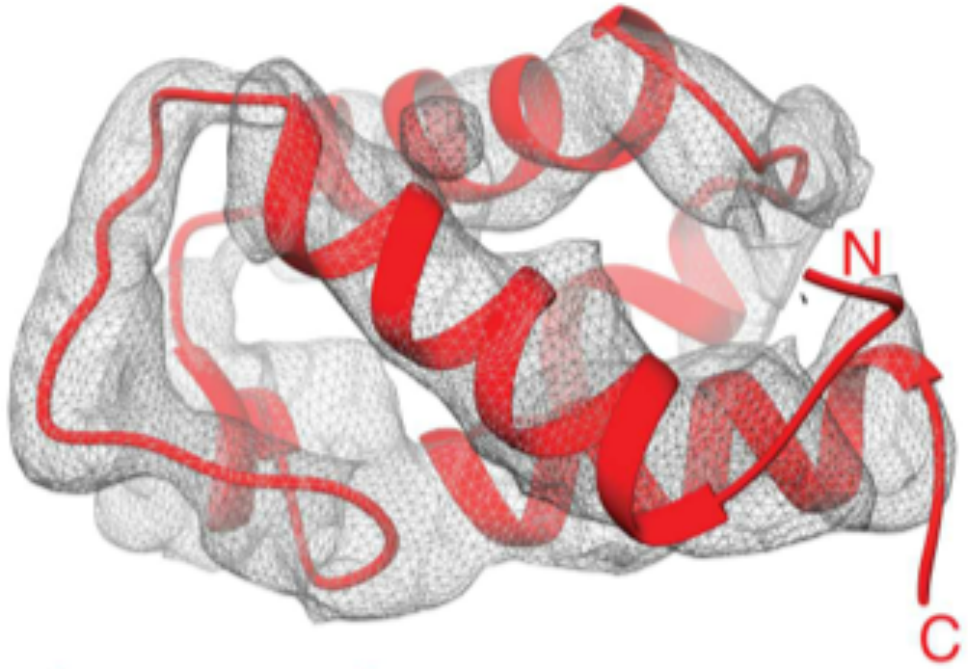
A




B

| Rank | Score | PDB:chain:domain |
|---|---|---|
| 1 | 3.95 | 3EJB : A : 1 |
| 2 | 3.50 | 1K4T : A : 5 |
| 3 | 3.45 | 1JQ5 : A : 1 |
| 4 | 3.28 | 4A2P : A : 2 |
| 5 | 3.20 | 1VH4 : A : 1 |
| 6 | 3.17 | 3DDR : A : 1 |
| 7 | 2.99 | 3C96 : A : 1 |
| 8 | 2.93 | 2HRA : A : 1 |
| 9 | 2.92 | 2NYF : A : 2 |
| 10 | 2.90 | 2DT9 : A : 2 |

*Escherichia coli* Acyl Carrier Protein (ACP)

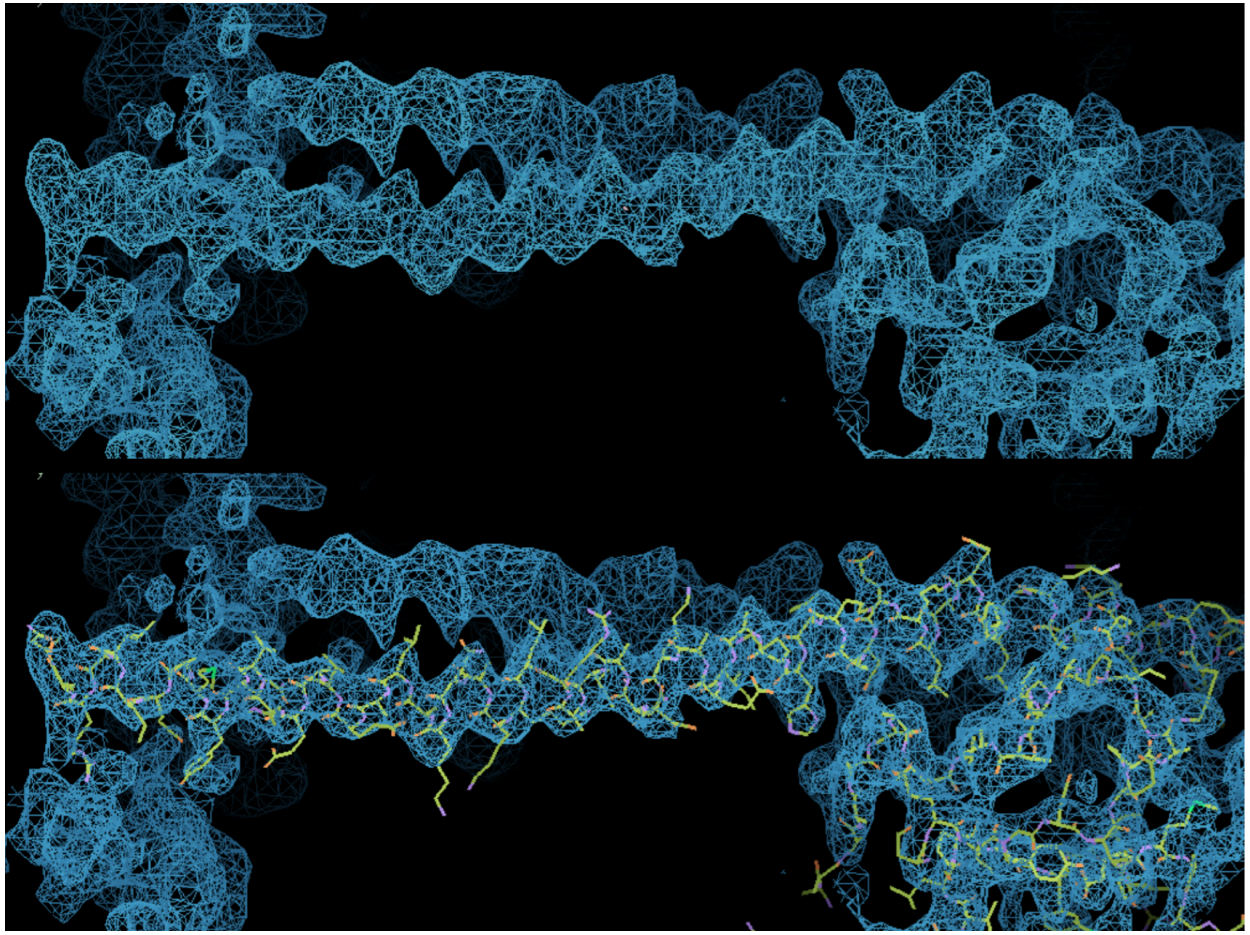



ACP fitted to map

Brown et al (2017) [hopefully…]

# Automated model building

# Automated model building



automated

# Automated model building

- Less time consuming than building models manually
- Removing human judgment can minimize errors

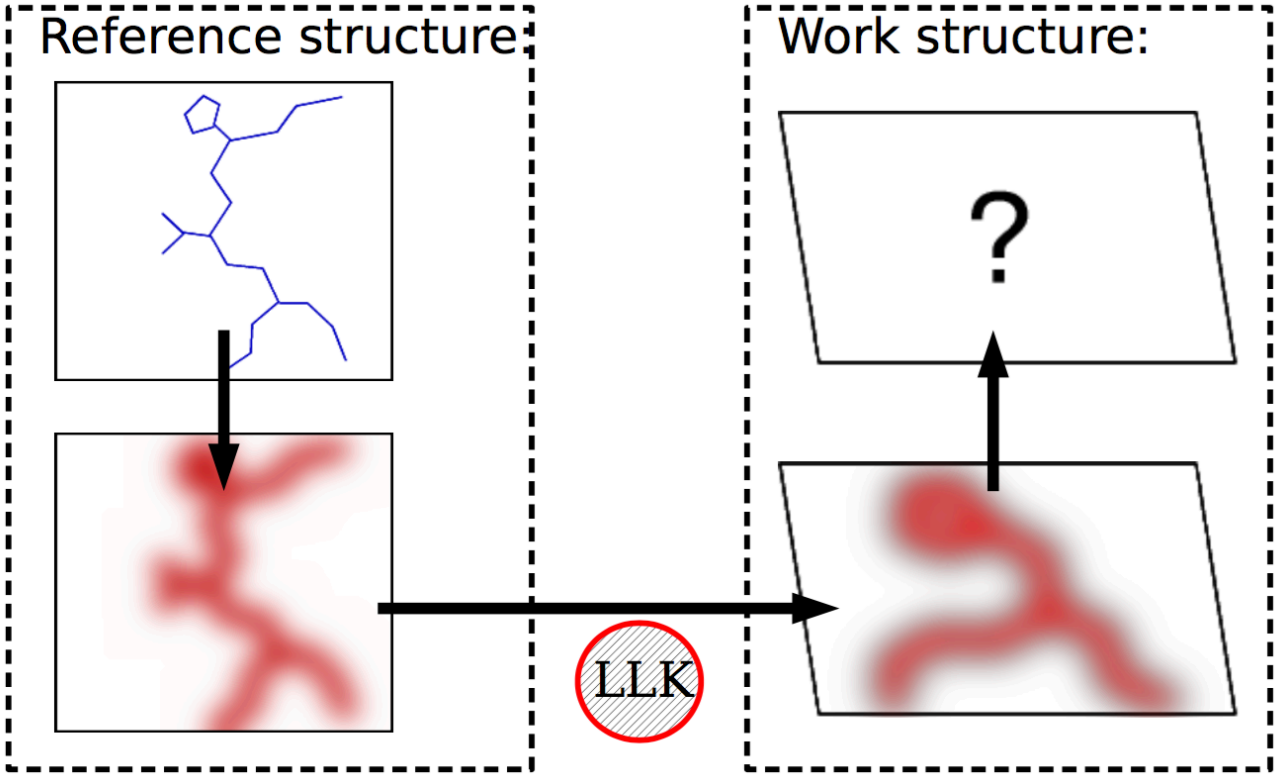
- Current methods:
  - Buccaneer
  - Phenix.map_to_model
  - Rosetta

# Buccaneer

- **Key developer**: Kevin Cowtan (University of York)
- **Basic premise**: trace protein structures in density maps by identifying connected alpha-carbon positions using a likelihood-based density target
- **Availability**: through CCPEM
- **References**: http://www.ccp4.ac.uk/newsletters/newsletter44/articles/buccaneer.html

# Buccaneer: how it works

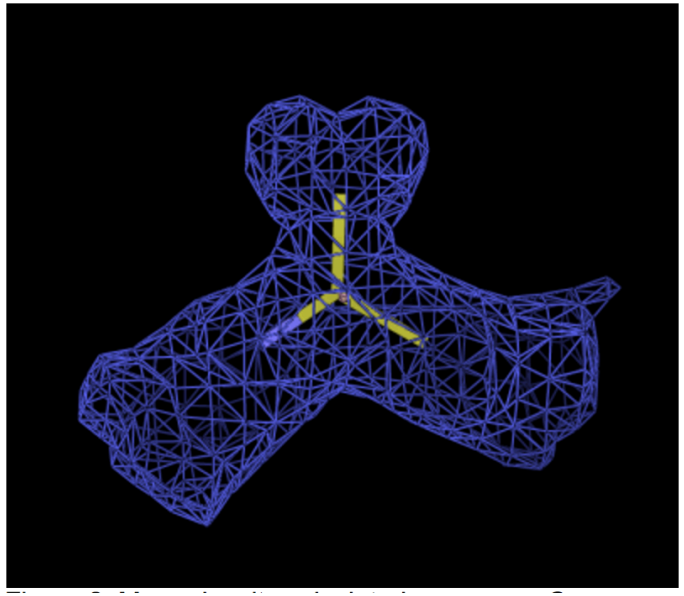
- Uses a simulated map from a known (reference) model to obtain likelihood target, and then search for this target in the unknown map



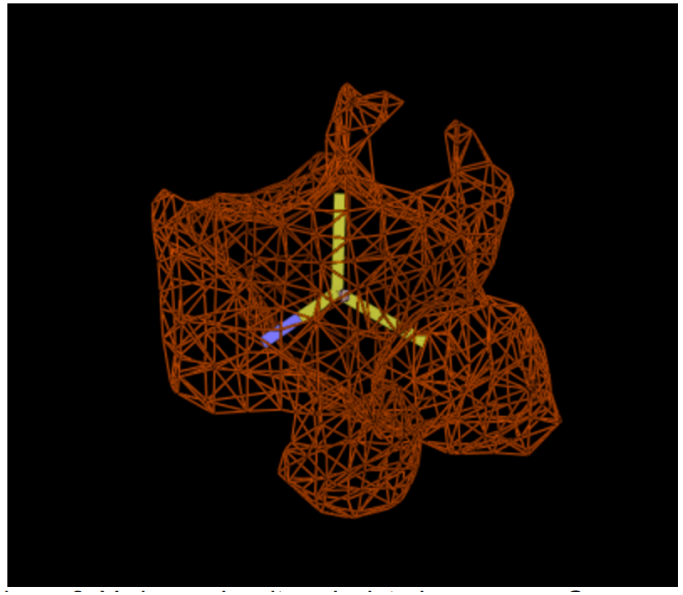

Developed by Kevin Cowtan at the University of York

# Buccaneer: 10 stages

1. Find candidate Cα positions
   - superimposed Cα positions from a known reference structure
   - uses a 4 Å sphere around Cα position
   - the likelihood function can therefore be described in terms of an expected density and a weighting



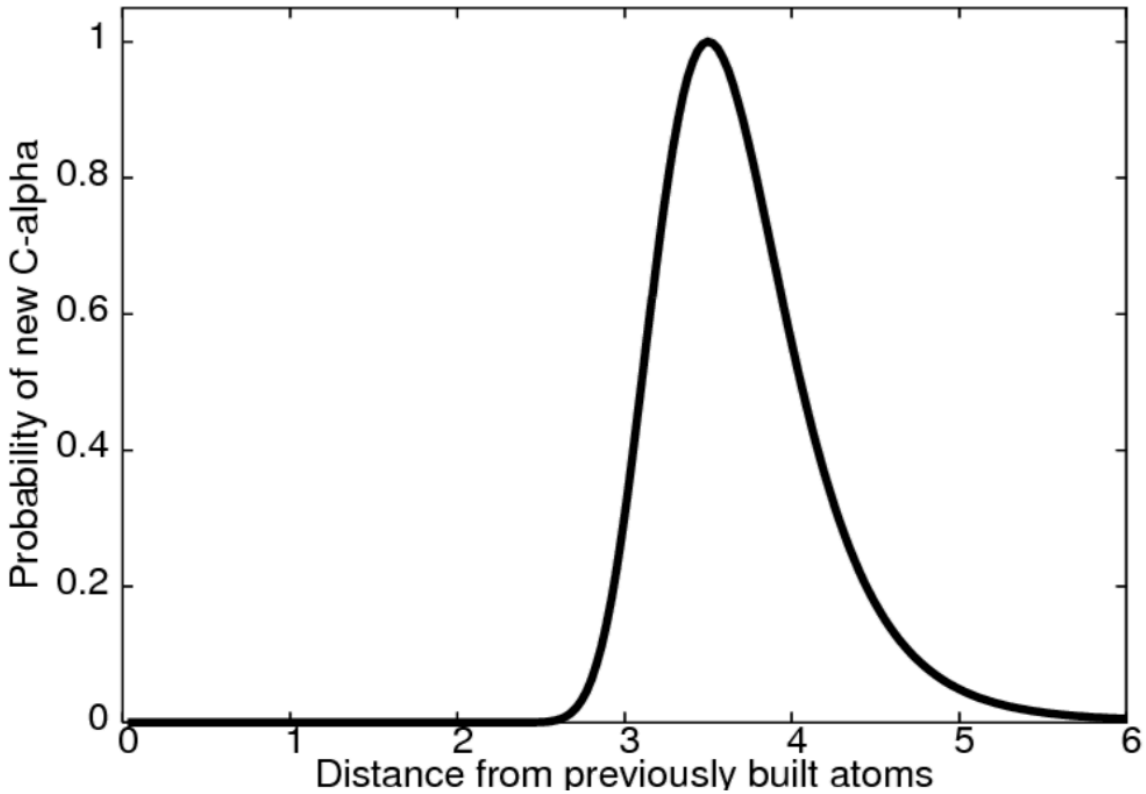
Mean density calculated over many Cα groups



Variance density calculated over many Cα groups

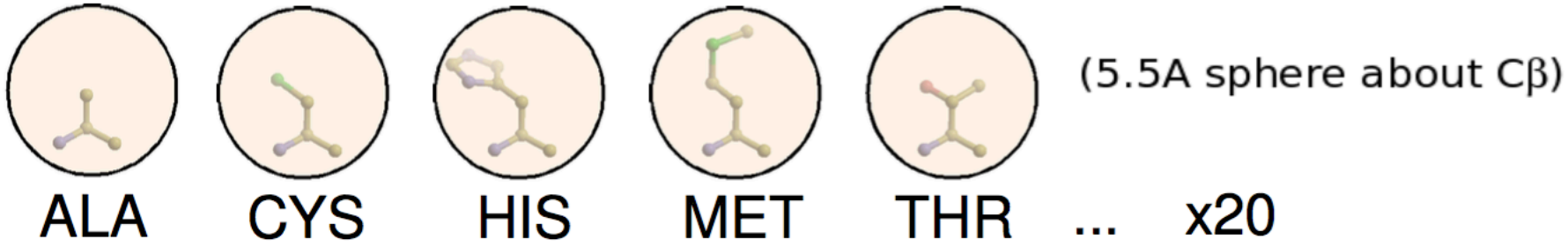
# Buccaneer: 10 stages

2. Grow Cαs into chain fragments
   - Model grows sideways from existing chain fragments by looking for new Cαs at an appropriate distance from the existing chain
3. Join and merge the fragments, resolving branches
4. Link nearby N and C terminii (if possible)



# Buccaneer: 10 stages

5. Sequence the chains (i.e. dock sequence)
   - Looks for Cβ environment
   - Likelihood comparison between the density of each residue in the work structure and the residues of the reference structure allows sequence to be assigned to longer fragments



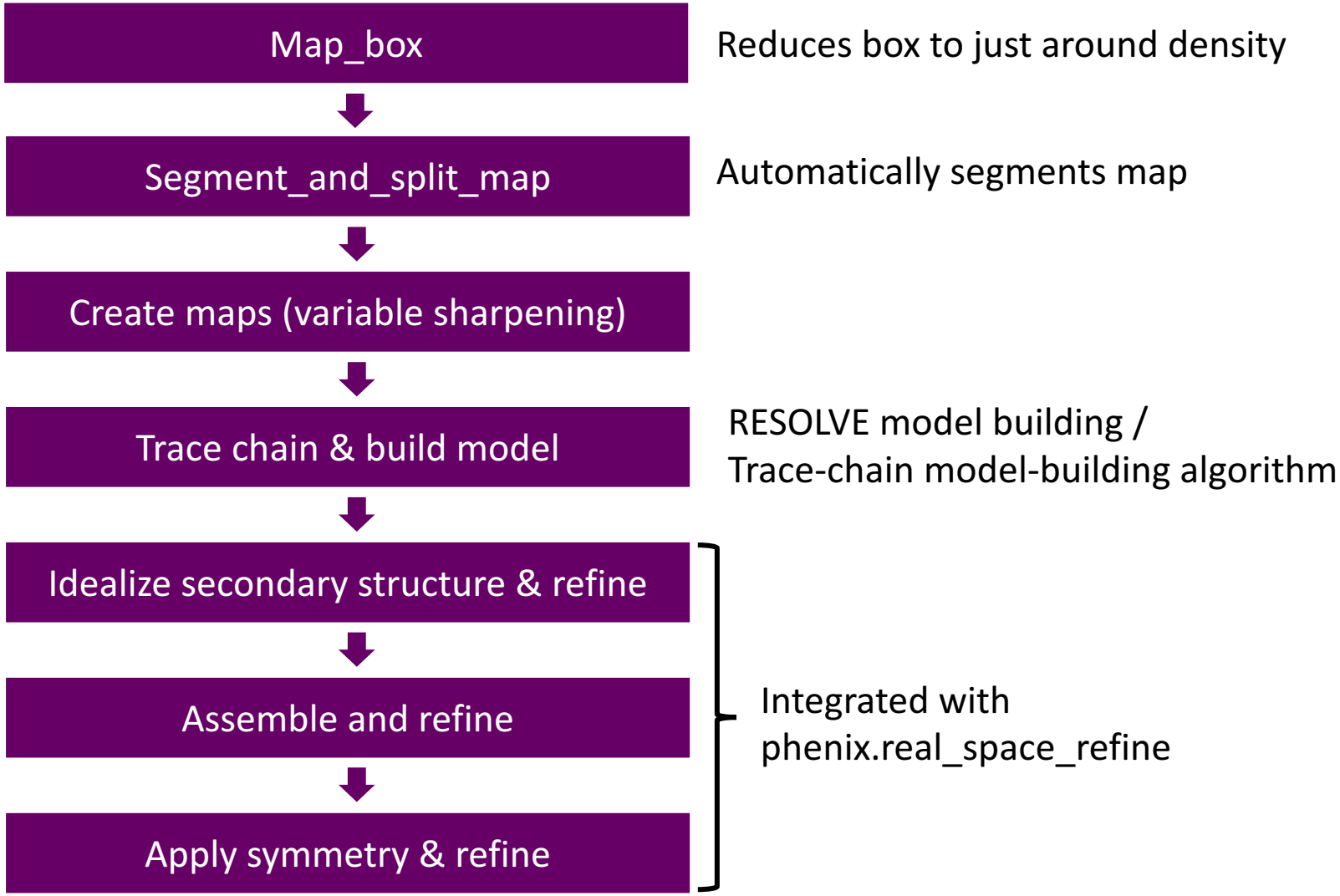

6. Correct insertions/deletions
7. Filter based on poor density
8. NCS Rebuild to complete NCS copies of chains [optional]
9. Prune any remaining clashing chains
10. Rebuild side chains

# Phenix.map_to_model

- **Key developer**: Tom Terwilliger (Los Alamos National Laboratory)
- **Basic premise**: builds protein/RNA/DNA into EM maps.
- **Availability**: through Phenix. Currently command line only.
- **References**: https://www.phenix-online.org/version_docs/dev-2428/reference/map_to_model.html

# Phenix.map_to_model

Can build both proteins and nucleic acids (type of chain to be built will be based on the supplied sequence file)



# Phenix.map_to_model : locating fragments

- Method based on RESOLVE
- Pattern matching algorithm
- Uses fragments larger than individual atoms – secondary structure
- Starts by FFT-based identification of helices and strands
  - Helical template: 6 amino acids (average density from ~200 6-residue helical segments)
  - β-strand template: 4 amino acid, average density
- Superimpose on this template each fragment in a library (helix, sheet)
  - Helix fragment library: 53 helices 6-24 amino acid long
  - Beta-strand fragment library: 24 strands 4-9 amino acid long
- Identify longest segment in good density

# Phenix.map_to_model : growing fragments and assigning sequences

- Extends chains using a tri-peptide fragment library
  - N-terminal extension (3 full amino acids), 9232 members
  - C-terminal extension (CA C O + 2 full amino acids), 4869 members
- find fragment that can itself be optimally extended (look-ahead scoring)
- For each fragment:
  - superimpose CA C O on same atoms of last residue in chain (extending by 2 residues)
  - pick the 10 highest scoring fragments
  - For each of these extend again by 2 residues and pick best 1

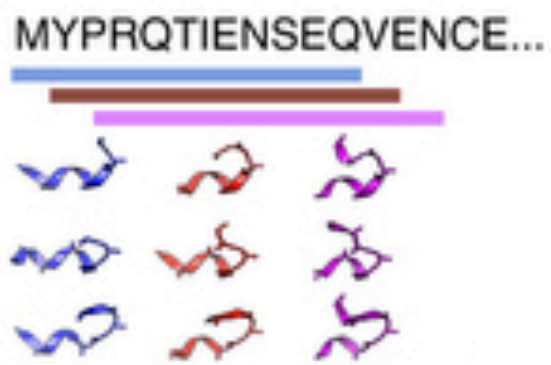
- Test all overlapping fragments as possible extensions
- Choose one that maximizes score when put together with current fragment
- When current fragment cannot be extended: remove all overlapping fragments, choose best remaining one, and repeat
- The sequence is assigned to the mainchain by determining the relative probability of every amino acid at each position (based on density and sequence composition)
- Rotamers are chosen based on correlation coefficient

# Rosetta (model building for cryo-EM)

- **Key developer**: Frank DiMaio (University of Washington)
- **Basic premise**: de novo structure determination from cryo-EM maps by combining conformational sampling with all-atom energy functions
- **Availability**: through the Rosetta software package
- **References**: Wang et al (2015). Nat. Methods. 12(4):335-8;  Frenz et al. (2007) Nat. Methods, doi:10.1038/nmeth.4340
- **Tutorial:** https://faculty.washington.edu/dimaio/wordpress/software/

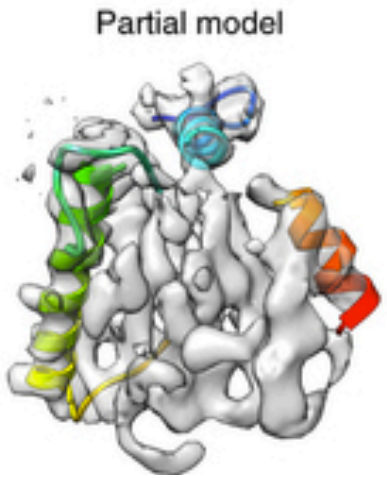
# Rosetta: step 1

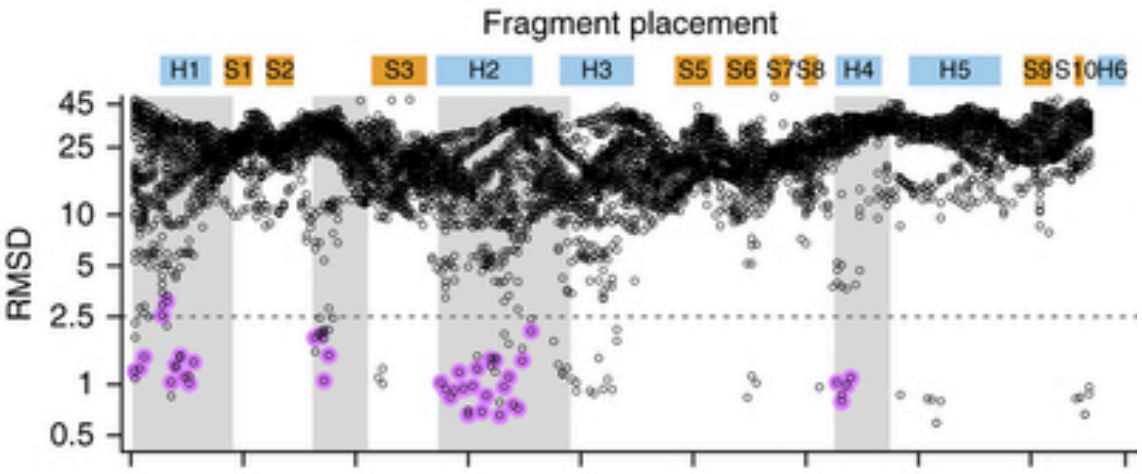
For overlapping 9-residue windows of sequence:




- identify “fragments” in the PDB with similar local sequences and predicted secondary structures
- Perform a 6-dimensional search (rotations and translations) to dock these into the map



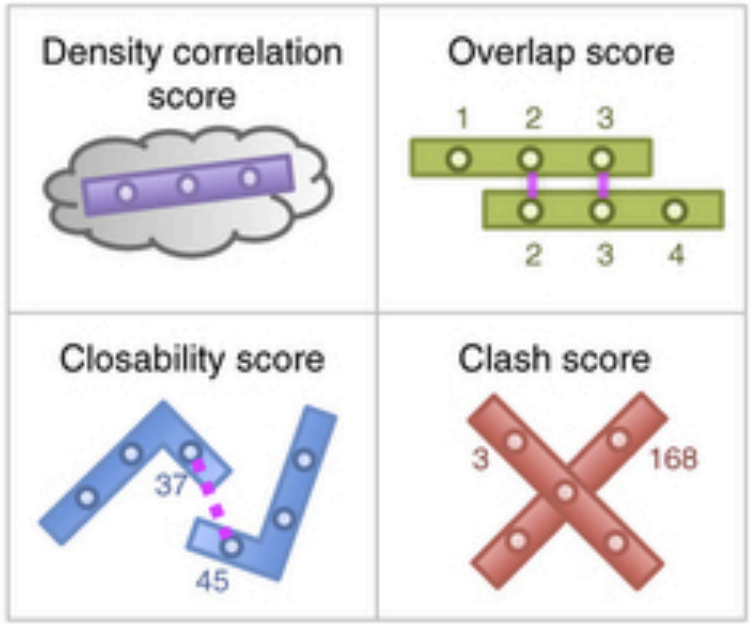



4.8 Å

# Rosetta : step 2

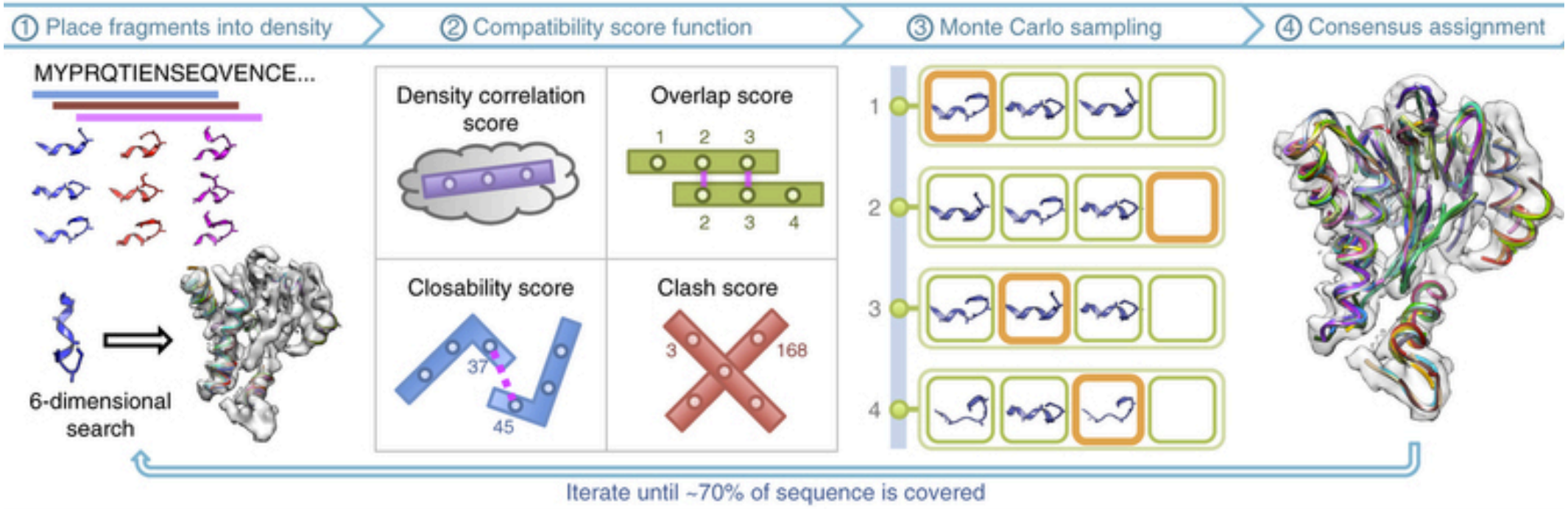
The resulting fragment placements are evaluated using a score function consisting of 4 terms:




1. A density correlation term assessing the agreement of fragment and map
2. An overlap term favoring fragment pairs assigning the same residue to the same location
3. A “closability” term favoring fragment pairs close in sequence that are close in space
4. A clash term preventing two residues from occupying the same place

# Rosetta : step 2

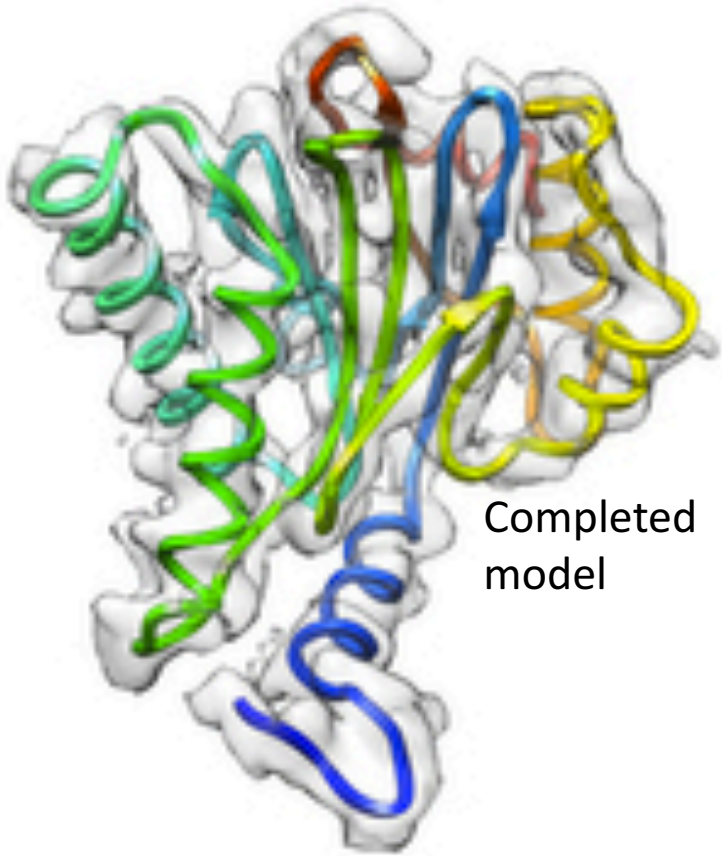
- Monte Carlo sampling guided by this score function finds the maximally consistent subset of fragment placements from this larger set. Sometimes no fragment is selected.
- These fragments are assembled into a partial model




- During iterations, density that has been assigned is masked out and only fragments that are unassigned are used.

# Rosetta :  step 3

- The partial model (70% complete) is completed using RosettaCM (comparative modeling) guided by the map
- For each partial model, 1,000 full-length models are generated.
- These are filtered by Rosetta energy
- And finally by best fit to the density



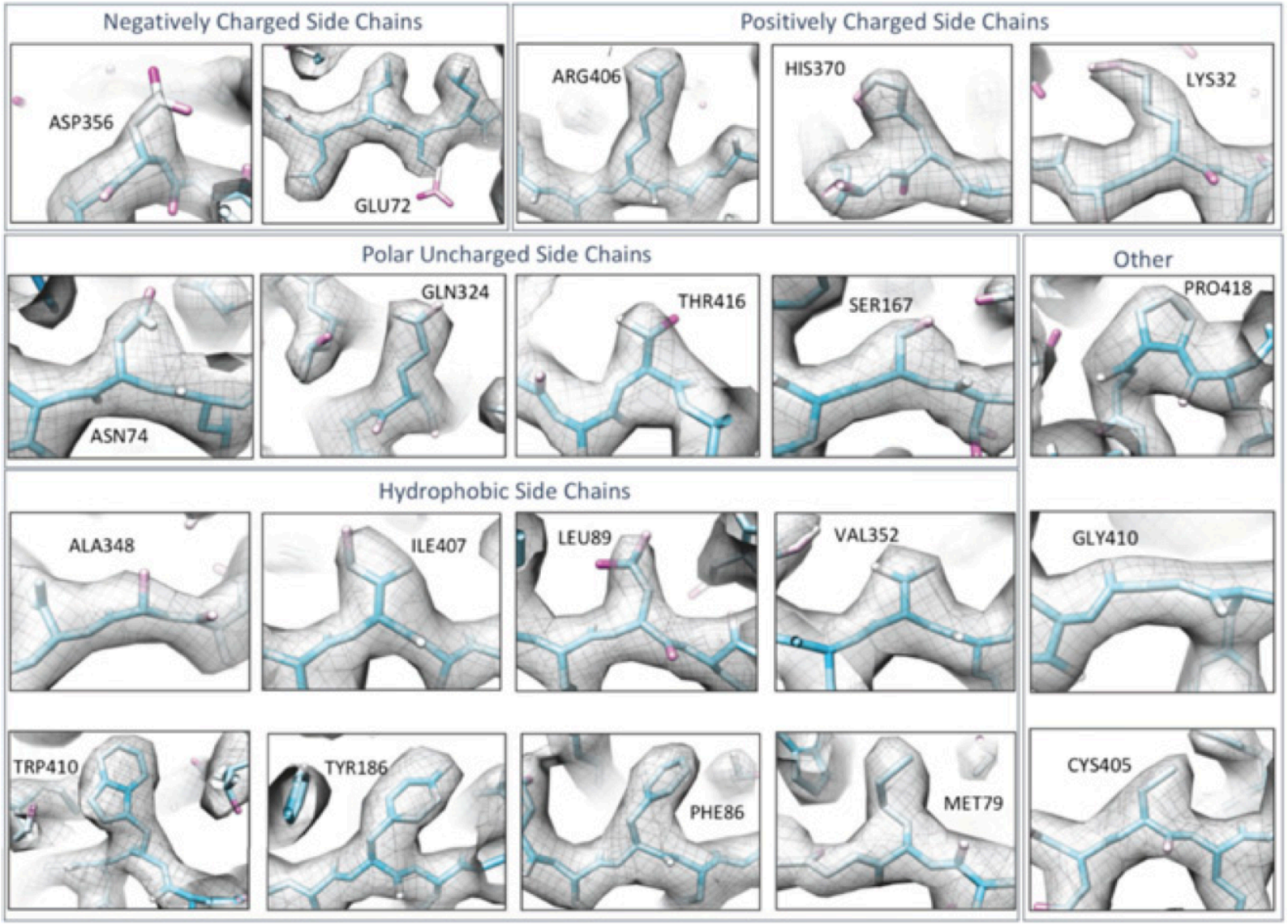
Completed model

# Rosetta :  step 3 (Rosetta enumerative sampling)

- A recent update (Frenz et al. (2017) doi:10.1038/nmeth.4340)
- Starts with an incomplete model
- Grows one residue at a time, starting with the terminal residue adjacent to the missing segment
- The conformation up to the previous 9 residues is sampled
- Each generated solution is evaluated against the experimental data and added to the 'beam'— the pool of partial models
- Following each sampling step, the model pool is culled to contain a set number of solutions
  - (usually 64 or 128)
- This process is repeated until all missing residues have been assigned.

# Manual model building

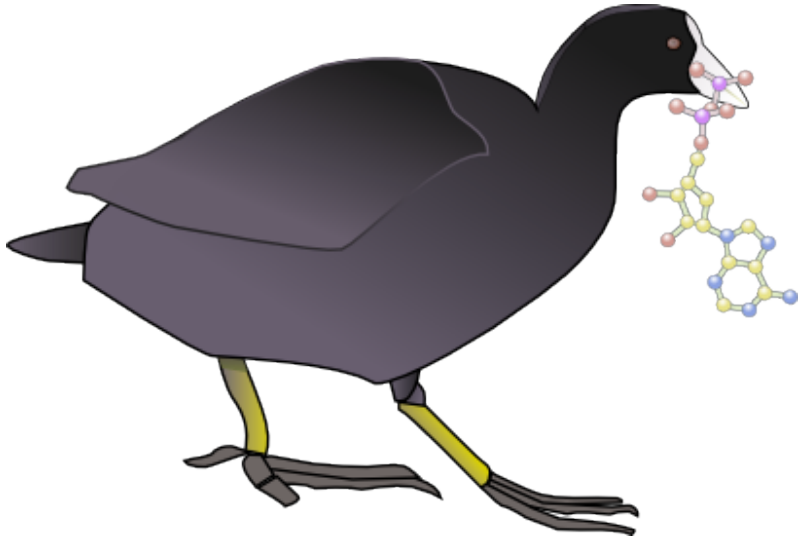
# Know your density




Hryc et al (2017) PNAS

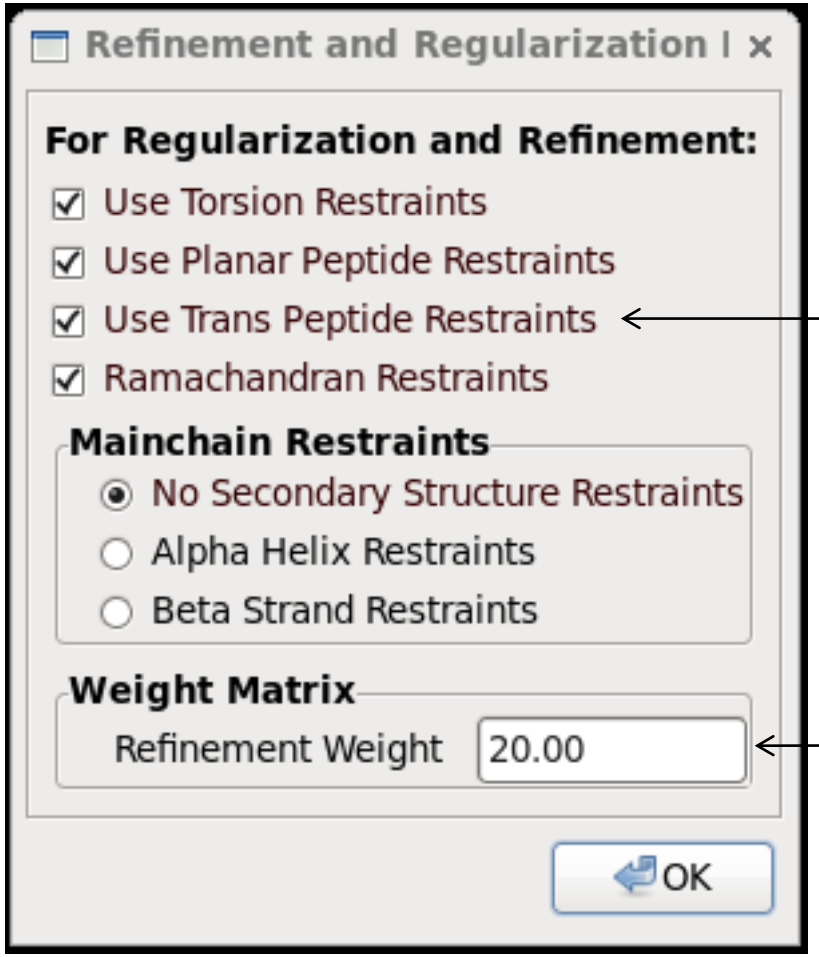
# Coot

- **Key developer**: Paul Emsley (MRC-LMB)
- **Basic premise**: macromolecular model building, model completion and validation
- **Availability**: CCP4/Phenix
- **References**: Emsley et al (2010) Acta Cryst D, 66;  Brown et al (2015) Acta Cryst D, 71: 136-153
- **Tutorial:** https://www2.mrc-lmb.cam.ac.uk/personal/pemsley/coot/files/EM-Tutorial-Coot-PE.pdf



# Coot : tools for building proteins

- Turn on restraints to ensure that help manual model building

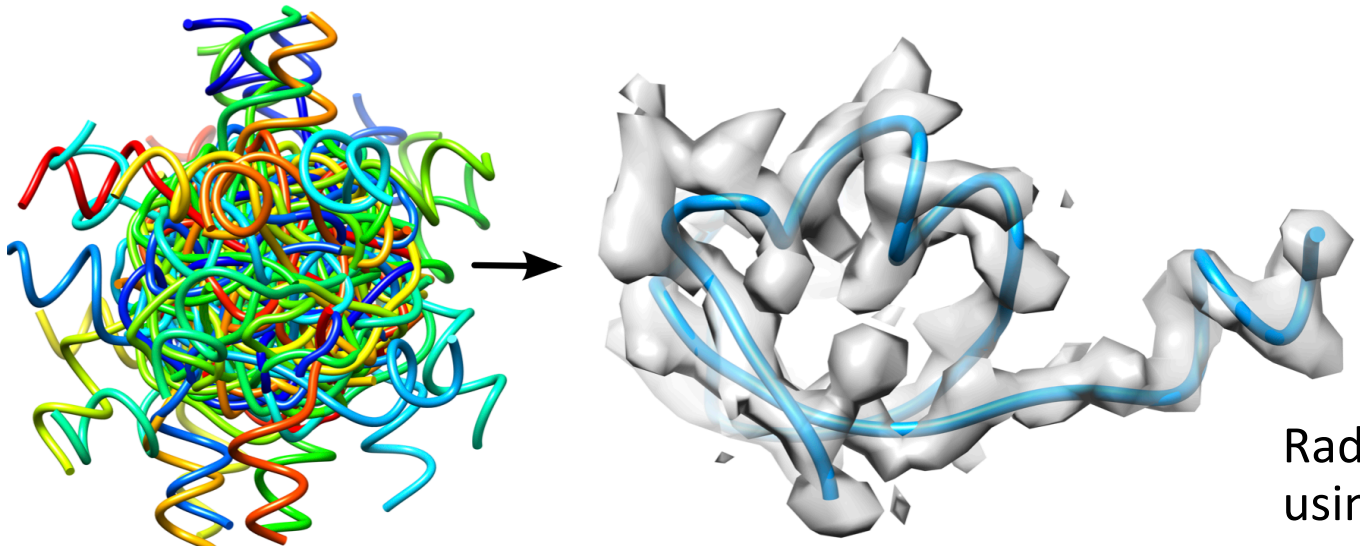



Cis-nonPro peptides are very rare (~0.03%), but you can turn off this restraint to model real cis-peptides

Smaller means better geometry

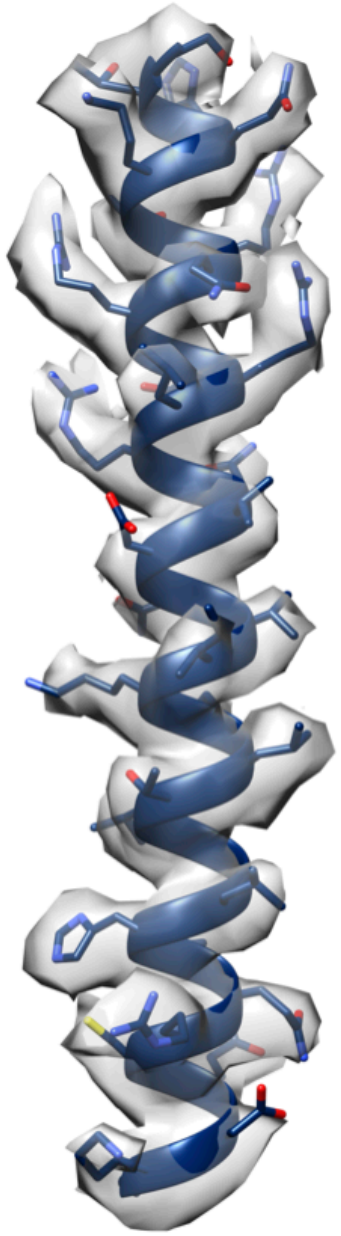
# Coot : Jiggle Fit

- Loop *n* (say 1000) times:
  - Generate random angles and translations
  - Transform atom selection by these rotations and translation
  - Score and store the fit to density
- Rank density fit scores
  - Pick top 20 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

# Coot : secondary structure elements



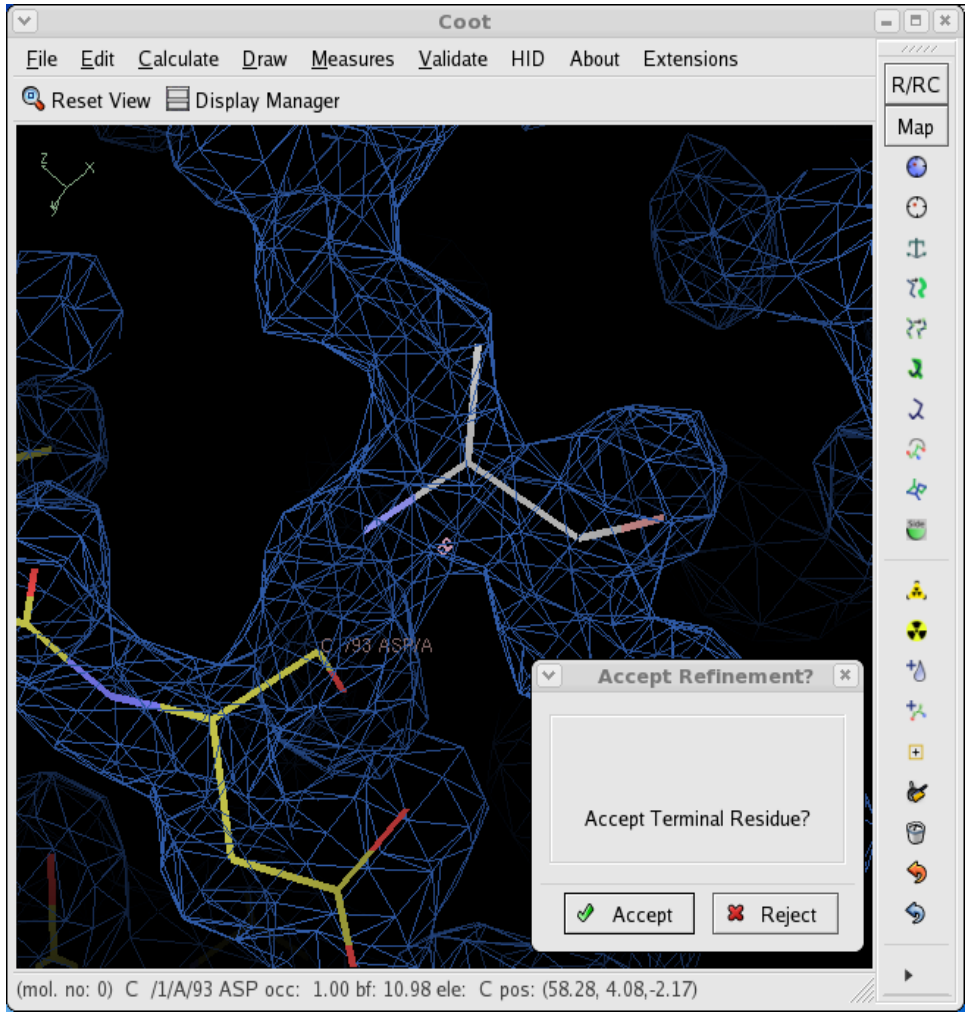
- Good starting point for *de novo* model building
- Coot has an option to automatically identify all secondary structure element in the map (**SSE identification**)
- Or add a helix/strand in a specific section of density (**Add helix/strand here**)

- In conjunction with jiggle fit finds correct orientation of α-helix every time at maps with resolution better than 4 Å

# Coot : tools for building proteins

## Add terminal residue

- Build one residue at a time starting at a previously positioned amino acid
- Remains the most popular way of model building
- either add as a alanine residue and mutate to correct residue afterwards
- Or assign a sequence to the model so that the identity of the next residue is known



# Coot : tools for building proteins

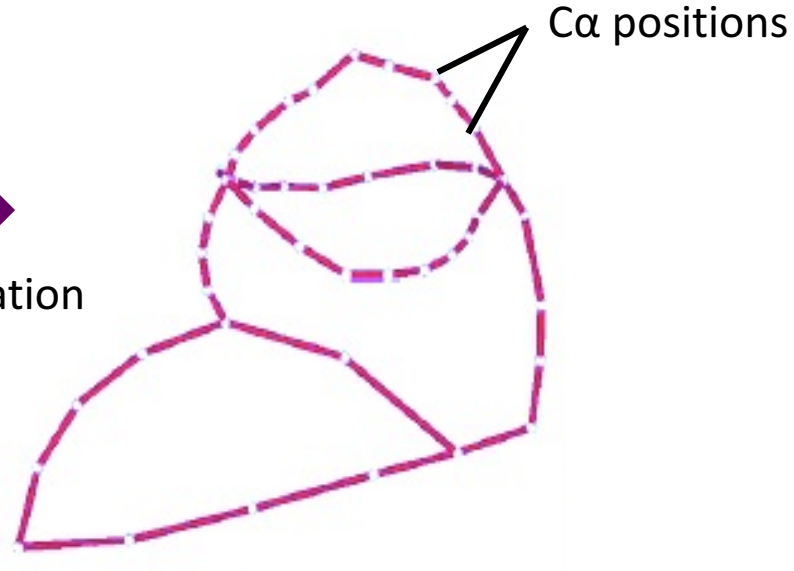
**Cα baton mode**

trace the main chain of a protein by placing correctly spaced α-carbon atoms

**Cα Zone -> main chain**

(this is not a real EM map)

skeletonisation

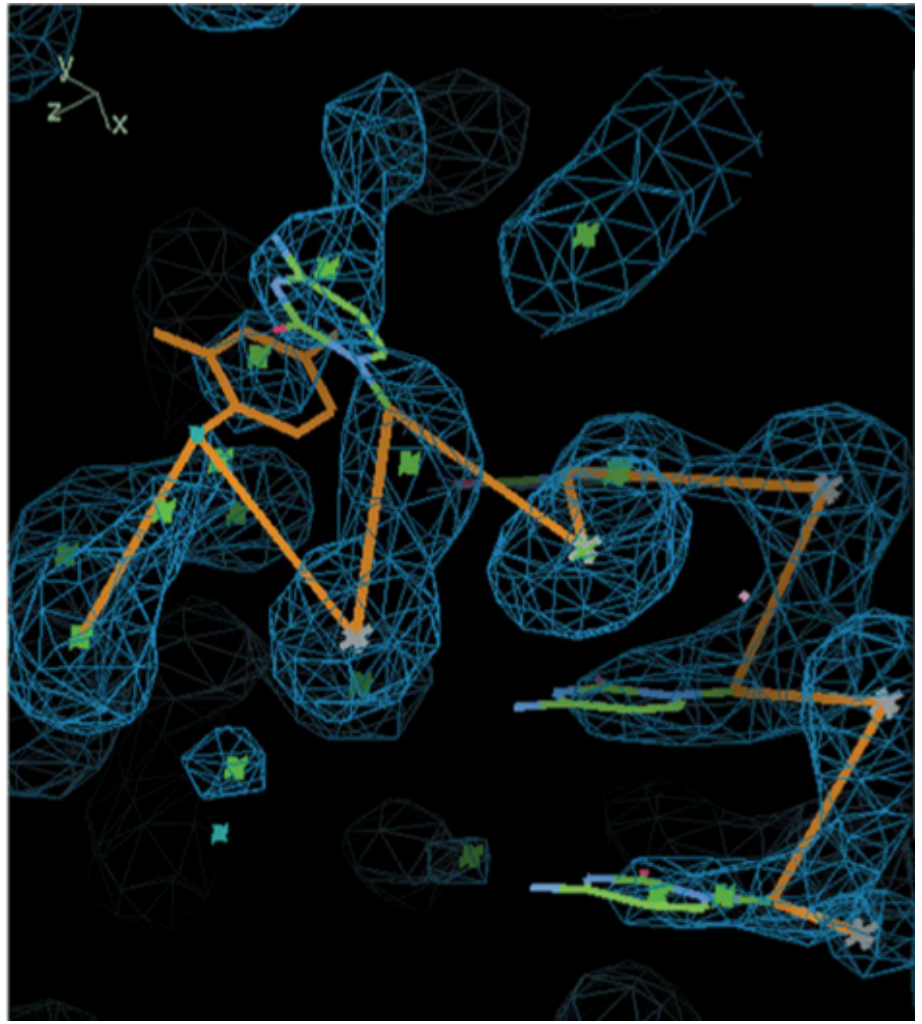


# Coot : tools for building nucleic acids

**Ideal DNA/RNA** - build an ideal DNA or RNA fragment

**Add terminal residue** - extend a nucleic acid

**Rcrane** (http://pylelab.org/software) – allows for semi-automated building of RNA models within Coot through the identification of phosphate positions within the density map



# Coot : tools for moving atoms around

**Real space refinement** - optimize the fit of the model to the density, while preserving stereochemistry

- **Sphere refine** – real-space refinement for an environment

**Regularize** - optimize stereochemistry

- **Sphere Regularize** - optimize stereochemistry for an environment

**Rigid body fit (local)** - optimize the fit of a rigid body to the density

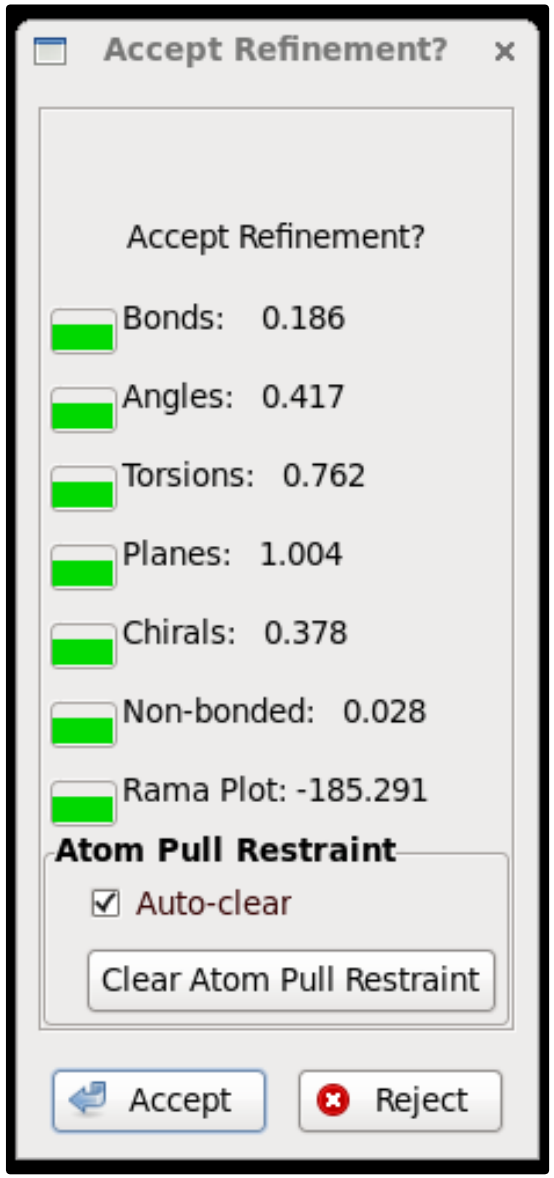
**Rotate/translate zone** - manually position a rigid body

**Rotamer tools** (auto fit rotamer, manual rotamer, mutate and autofit, simple mutate)

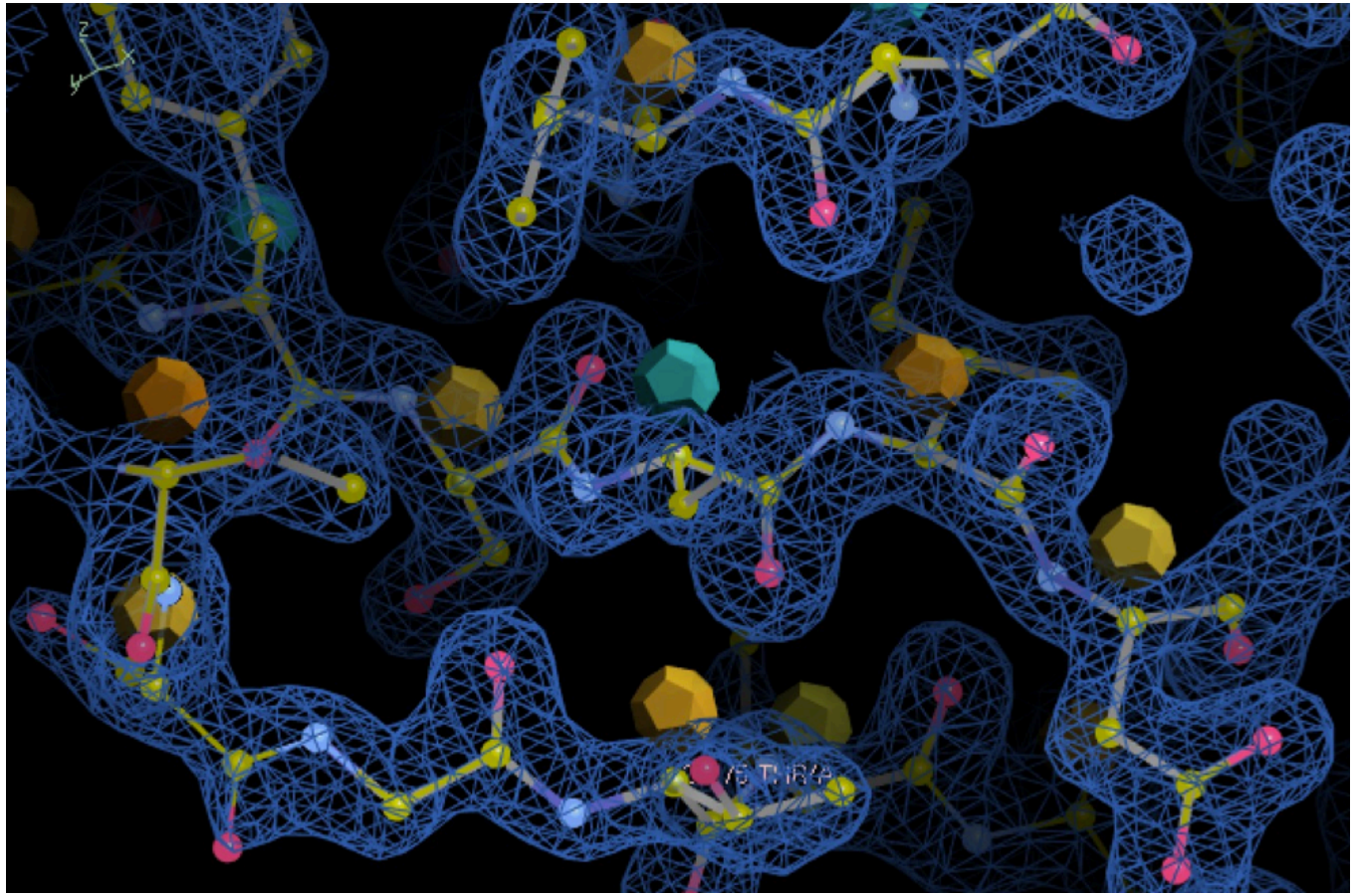
**Torsion editing** (edit chi angles, edit main chain torsions, general torsions)

**Other** (flip peptide, flip sidechain, cis <-> trans)

# Coot : live validation




Refinement in Coot gives immediate feedback



# iSOLDE : interactive molecular dynamics

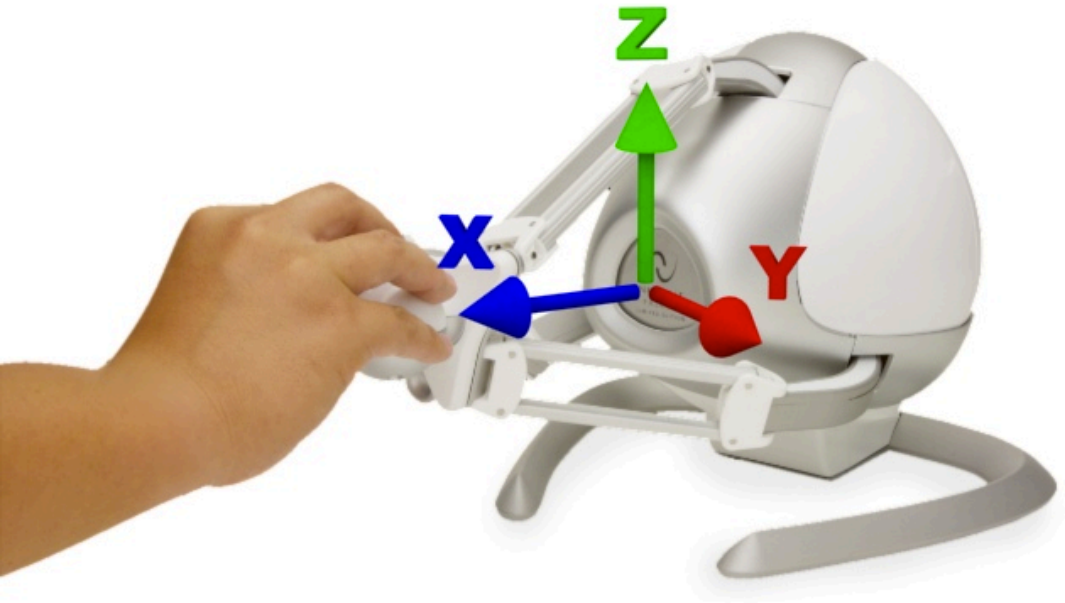
- **Key developer**: Tristan Croll (CIMR, University of Cambridge)
- **Basic premise**: Allows users to interact in real time with molecular dynamics simulations
- **Availability**: through VMD. In the future will be available through Chimera X.
- **References**: Croll et al (2016) Structure, 24:469-76
- **Tutorial**: https://www.youtube.com/watch?v=kqJpYIH0IdY

# iSOLDE : interactive molecular dynamics

- Select a region of interest to run molecular dynamics on
- This might be a small problem region (~10–20 contiguous residues and their immediate spatial neighbors)
- A further 8 Å shell of surrounding atoms is included in the simulation to maintain the physical context of the mobile atoms, but remains fixed in space
- Mask maps to within user-specified distances from the mobile atoms
- Map is converted to potential energy maps to which the simulation is then coupled
- Standard stereochemical restraints are included (bond, angle, torsions)
- Simulation also takes into consideration long-range interactions (electrostatics and van der Waals)

# iSOLDE : interactive molecular dynamics

- User can interact in real time with the molecular dynamics simulation
- coupled to a haptic interface – this allows the user to “pull” on any atom within a running simulation, while it (and its surroundings) responds in a manner akin to a real molecule
- This makes model building feel more like working in a physical environment



# Conundrums

- What to do about atoms without density?
  - Do I include loops with indistinct density?
  - Should I truncate sidechains?

- Is one model enough?
  - Cryo-EM is an averaging technique. Your final map **will** be the result of averaging an ensemble of states that the macromolecule exists in.
  - Is a single model therefore appropriate?

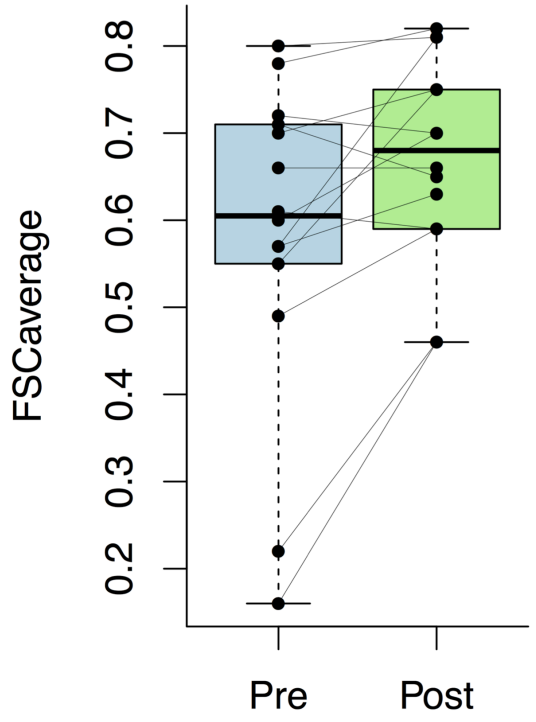
These same questions exist in X-ray crystallography and have similarly not been resolved.

# Refinement

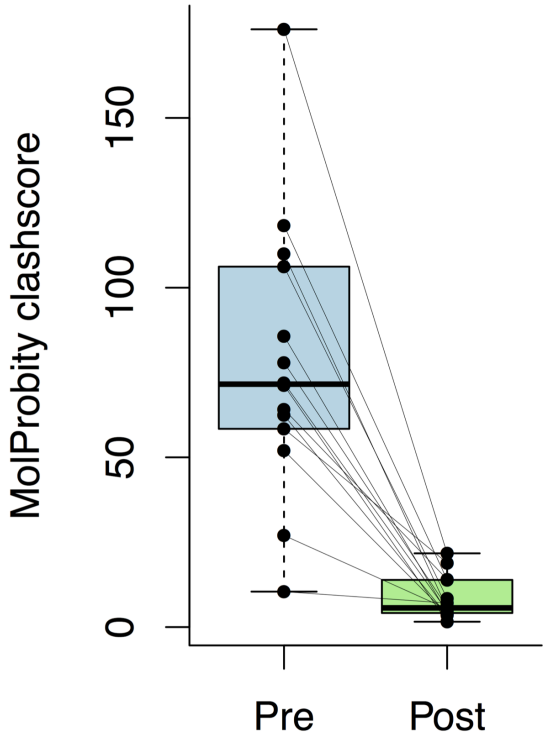
# The purpose of refinement

1. Improve the fit of your model to density
2. Ensures your molecule agrees with prior knowledge

Refinement improves fit

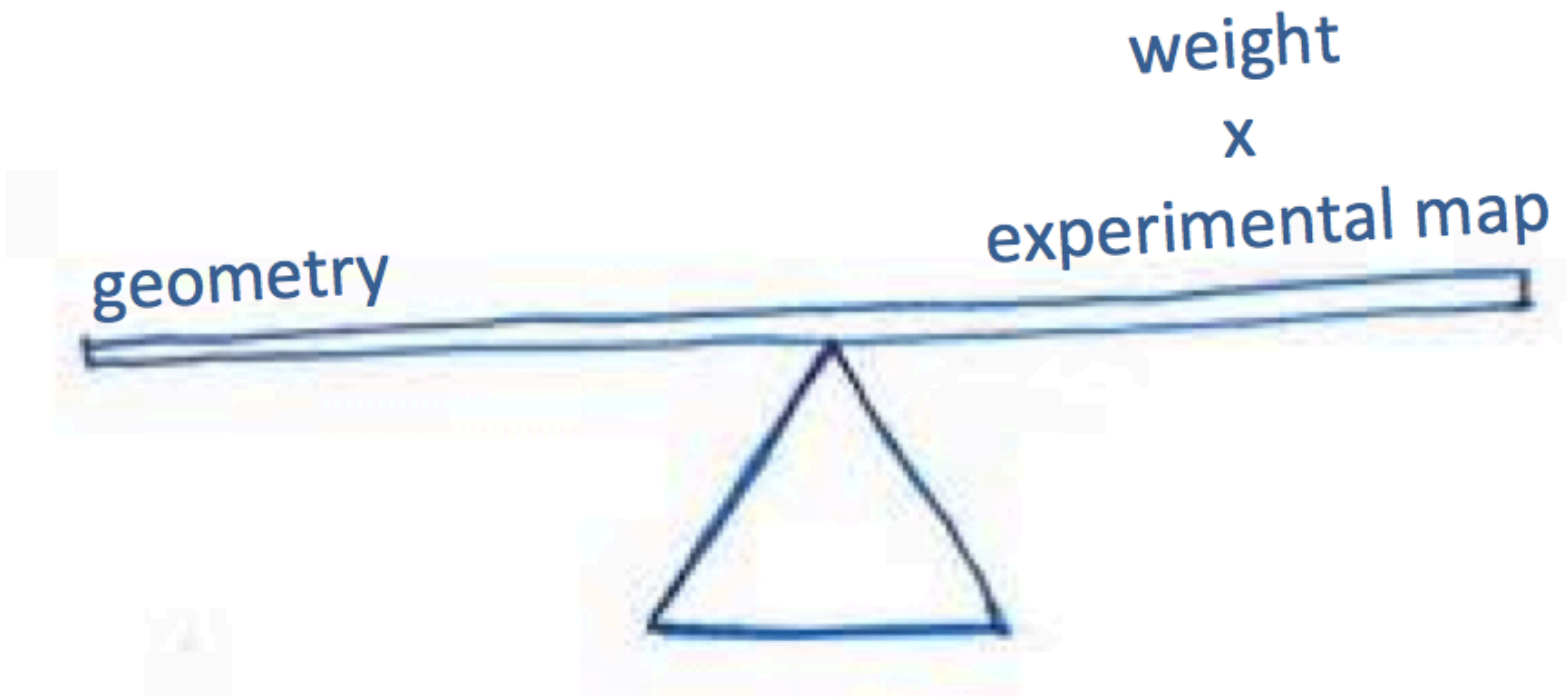



Refinement reduces clashes



# The purpose of refinement

Fitting to density is easy, fitting to density while maintaining good geometry is a delicate balance



# What changes during refinement…

1. Model coordinates (x, y, and z)
2. B-factors (these can be for each individual atom or group of atoms)
3. Occupancies (for multiple conformations)

# …and what doesn't

1. The map (currently)

# Restraints

Restraints are a way of specifying prior knowledge

Standard restraints (used by default) include:

1. Bond lengths
2. Angles
3. Chirals
4. Planes
5. Some torsion angles

} Atoms are bonded to each other in specific ways

6. B-values

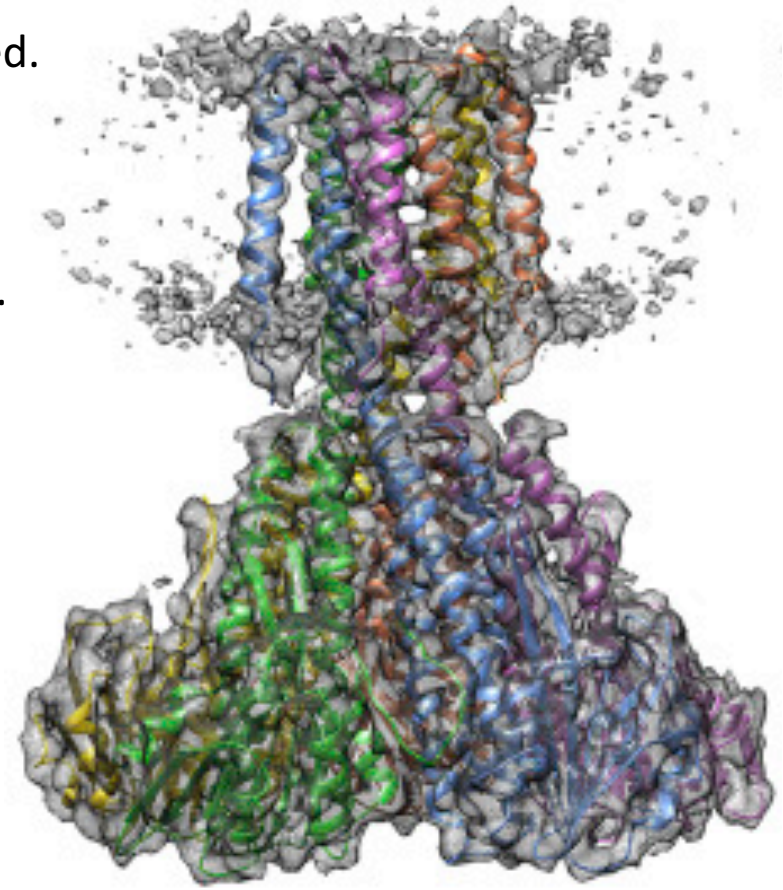
} Atoms close to one another cannot be dramatically different

| Bond | Bond Length (Å) |
|---|---|
| C—C | 1.54 |
| C=C | 1.34 |
| C≡C | 1.20 |
| C—N | 1.43 |
| C=N | 1.38 |
| C≡N | 1.16 |
| C—O | 1.43 |
| C=O | 1.23 |
| C≡O | 1.13 |

Restraints stabilize refinement, reduce the chance of overfitting, and ensures that the final model is consistent with prior knowledge

# NCS constraints / restraints

1. NCS constraints } If symmetry was applied during map reconstruction, the molecules will be exactly the same.
2. Global NCS restraints } Molecules are similar. Differences are minimized.
3. Local NCS restraints } Domains may be similar, but the orientation of the domains relative to one another may differ.



# External restraint generation

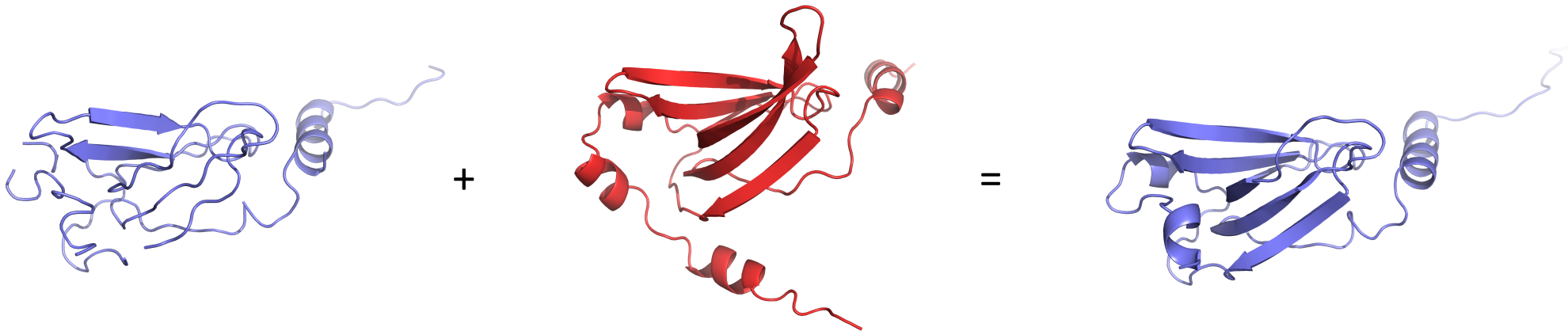
## For proteins and nucleic acids:

**ProSMART :** Rob Nicholls (CCP-EM)
**LIBG :** Fei Long (CCP-EM)
**Phenix.secondary_structure_restraints :** Oleg Sobolev / Pavel Afonine (Phenix)

## For ligands:

**ACEDRG :** Fei Long (CCP4 / CCP-EM)
**Phenix.elbow :** Nigel Moriarty (Phenix)

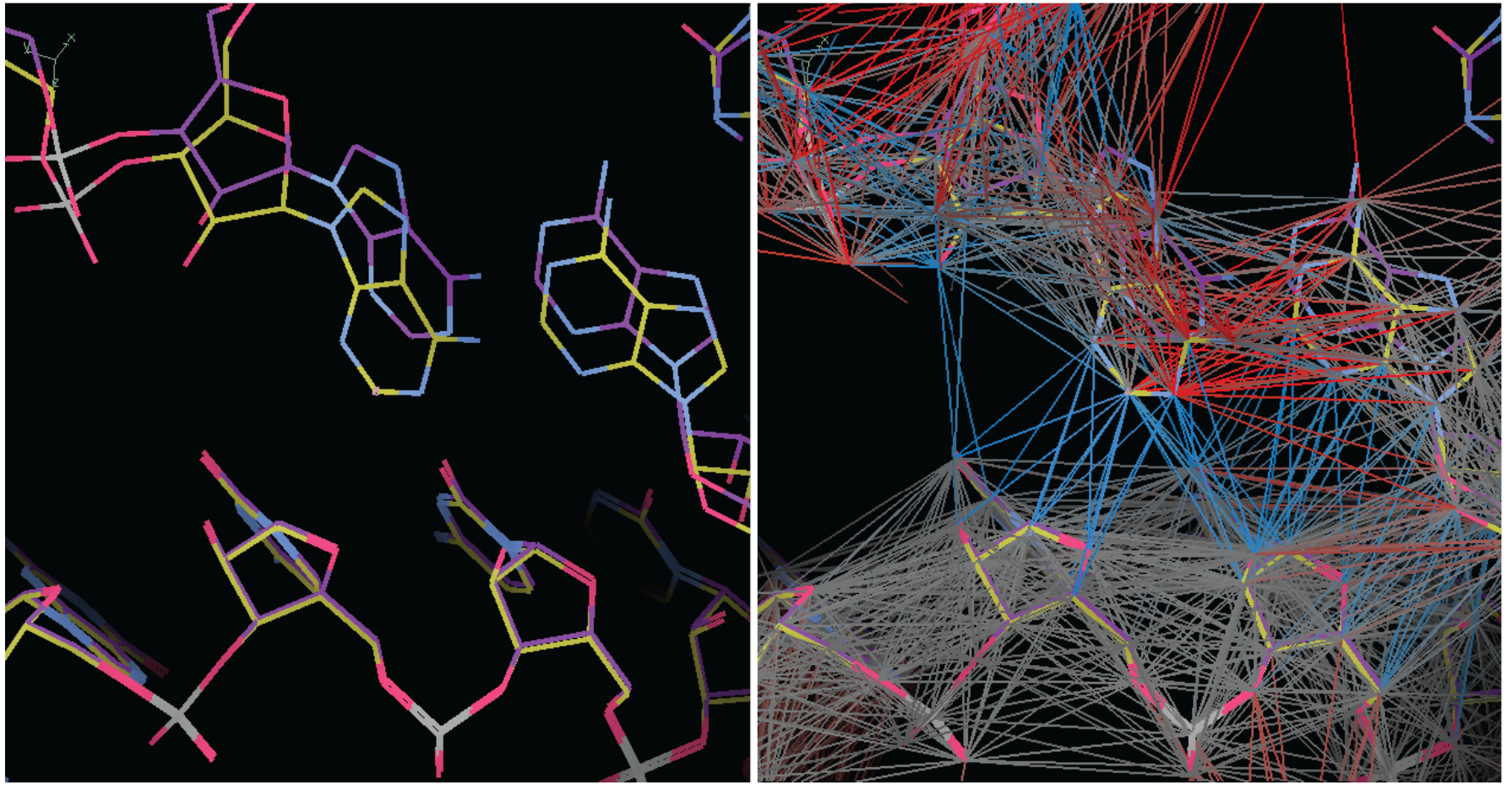
# External restraints from homologous structures



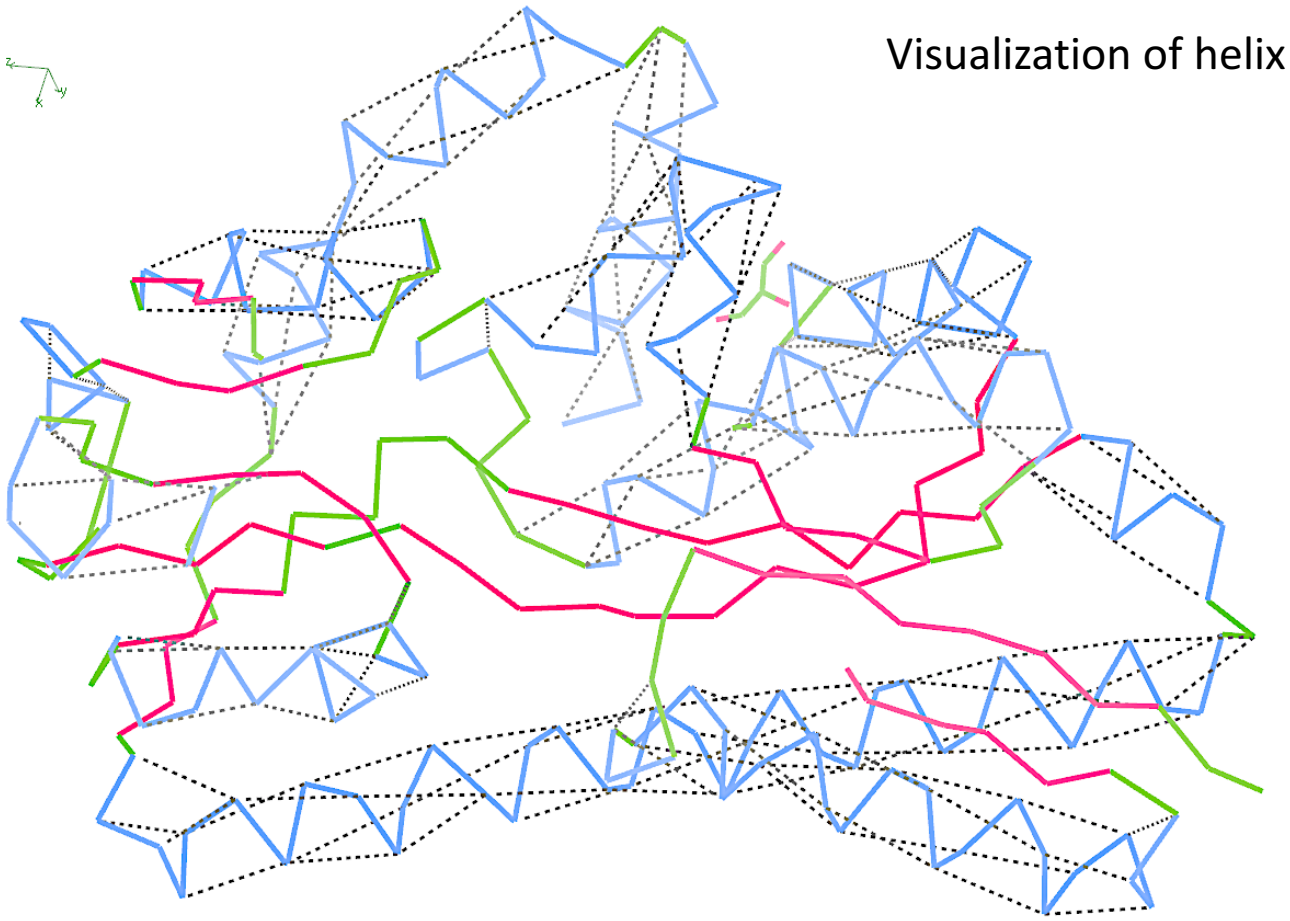
Target + Reference = Refined

# External restraints from homologous structures

All restraints can be visualised and applied in Coot:

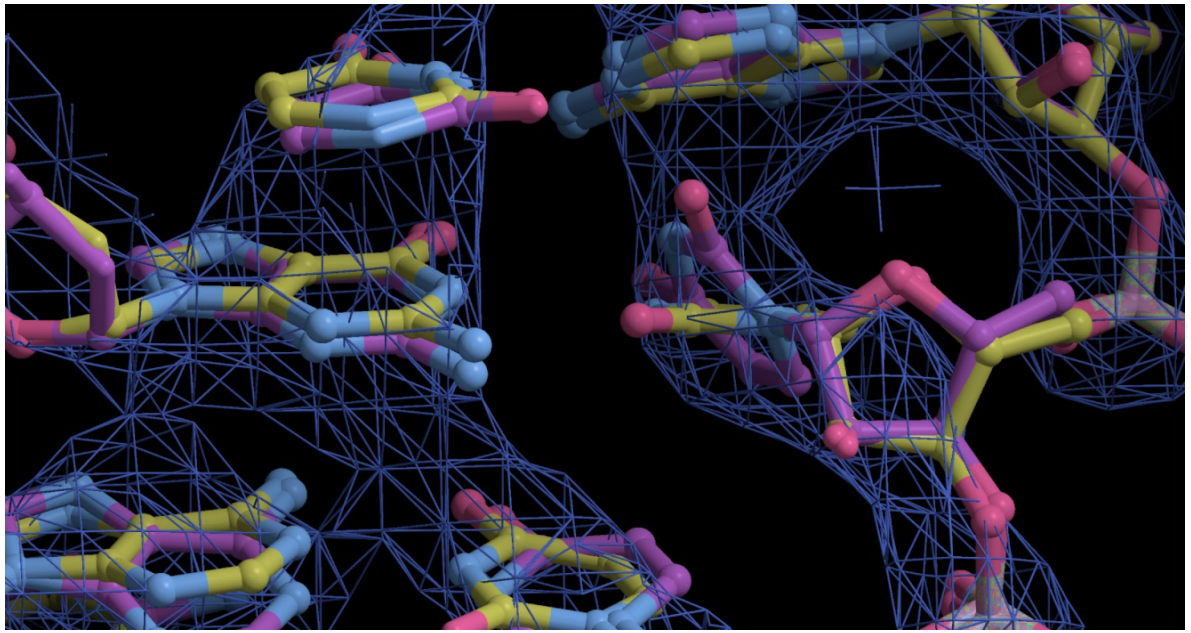


# Secondary structure restraints



Visualization of helix restraints in Coot

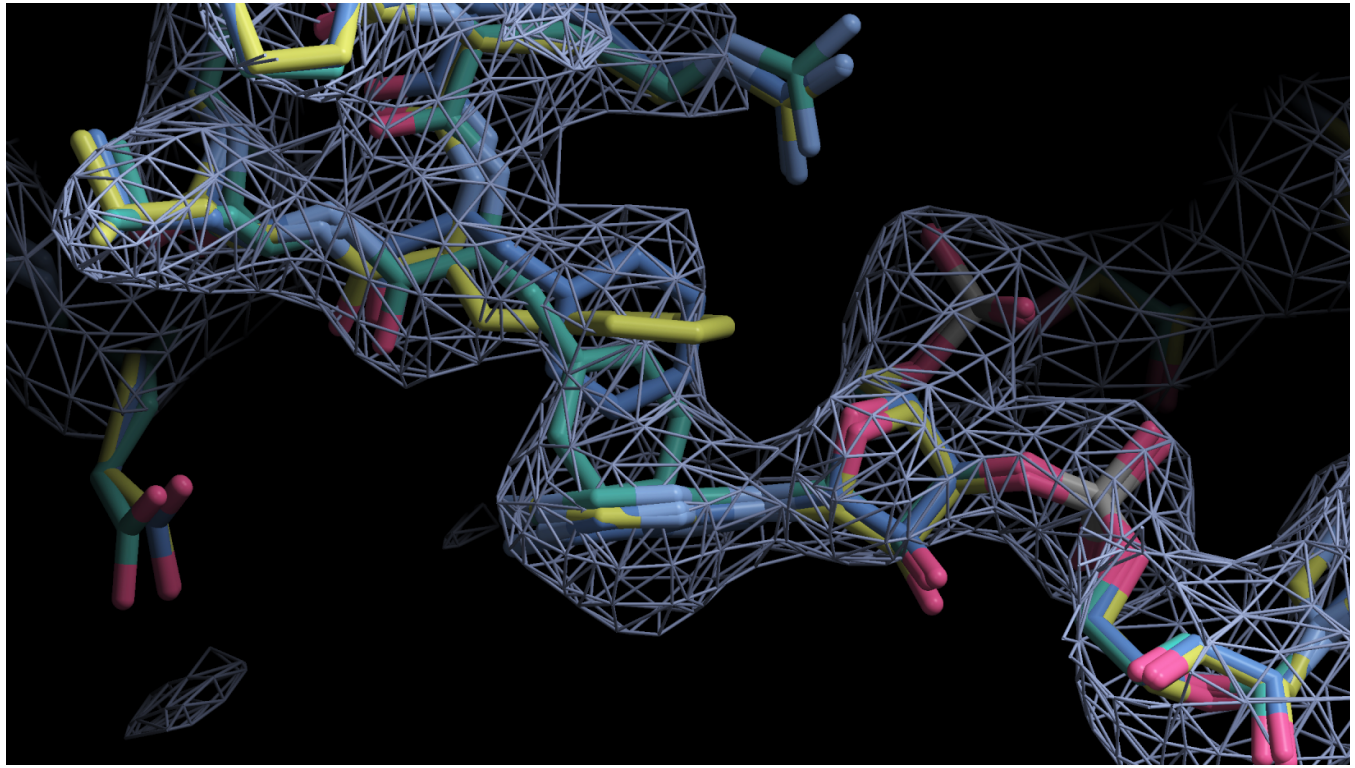
# Base-pair restraints



Purple = without restraints
Yellow = with base pair restraints

# Parallel-plane restraints

- Identifies and maintains sets of atoms that should be in parallel planes
- Nucleic acid bases
- Also amino acid sidechains (Trp, Tyr, His, Arg, Lys, Asn, Gln)
- And base:amino acid sidechain interactions



Green = before refinement; blue = refinement without stacking restraints;
Yellow = after refinement with stacking restraints

# Options

1. REFMAC (Fourier/reciprocal-space refinement)
2. Phenix.real_space_refine
3. Rosetta

# REFMAC

- **Key developer**: Garib Murshudov (MRC-LMB)
- **Basic premise**: Macromolecular refinement using maximum likelihood and elements of Bayesian statistics
- **Availability**: CCPEM / CCP4
- **References**: Brown et al (2015) Acta Cryst D, 71: 136-153
- **Tutorial**: http://i2pc.cnb.csic.es/download/spring_course_2016/ccpem_refmac_tutorial.pdf

$$f_{tot} = wf_{data} + f_{geom}$$

likelihood of the data

probability of the model

$f_{data} = -\log[P(obs;model)]$

$f_{geom} = -\log[P(model)]$

w : relative weighting

# Phenix.real_space_refine

- **Key developer**: Paul Adams / Pavel Afonine (Lawrence Berkeley National Laboratory)
- **Basic premise**: Refines a model into a map in real space.
- **Availability**: Phenix
- **References**: Afonine, et al. Computational Crystallography Newsletter (2013). Volume 4, Part 2, 43-44.
- **Tutorial**: https://www.youtube.com/watch?v=shmBHtyUdCc

# Features available in REFMAC and Phenix

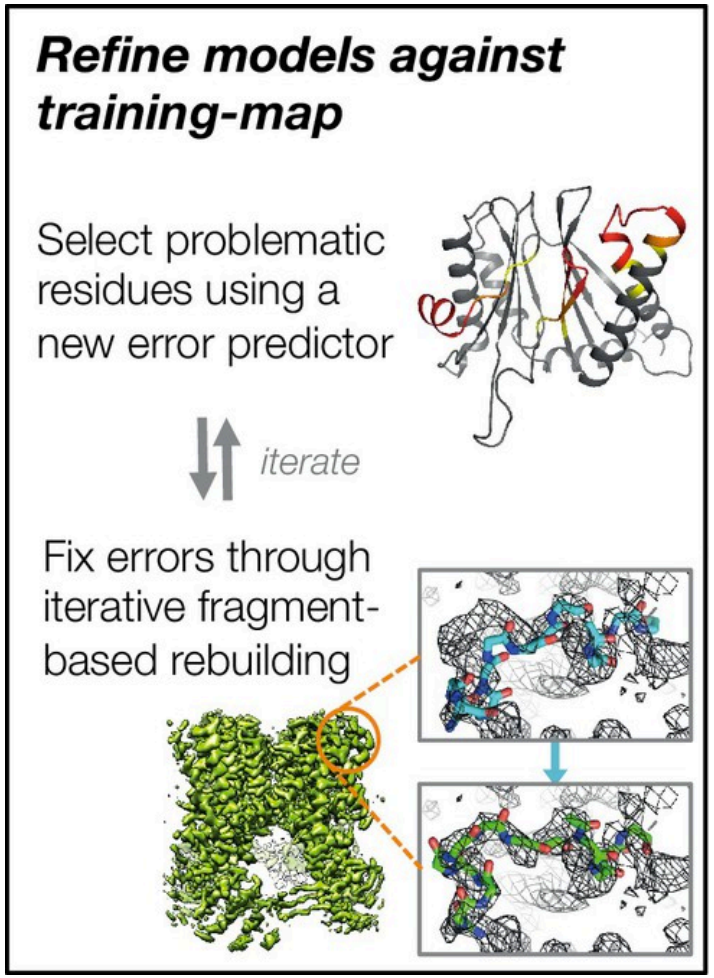
| Options | REFMAC | Phenix |
|---|---|---|
| GUI / command line | both | both |
| Morphing | (available in Coot) | ✔ |
| Rigid-body refinement | ✔ | ✔ |
| Simulated annealing | ✗ | ✔ |
| Jelly-body refinement | ✔ | ✗ |
| B-factor refinement | ✔ | ✔ (reciprocal space) |
| Composite map refinement | ✔ | ✗ |
| Reference structure restraints | ✔ | ✔ |
| Secondary structure restraints | ✔ | ✔ |
| Nucleic acid restraints | ✔ | ✔ |
| Symmetry restraints / constraints | both | both |
| Ramachandran restraints | ✗ | ✔ |
| Rotamer restraints | ✗ | ✔ |
| Ligand restraint handling | ✔ | ✔ |

# Rosetta (model refinement for cryo-EM)

- **Key developer**: Frank DiMaio (University of Washington)
- **Basic premise**: rebuilds models using iterative fragment-based sampling
- **Availability**: through the Rosetta software package
- **References**: DiMaio et al (2015) Nat. Methods 12:361-5 ; Wang et al (2016) eLife, 5:e17219
- **Tutorial:** https://faculty.washington.edu/dimaio/wordpress/software/

# Rosetta refinement : stage 1

Stage 1



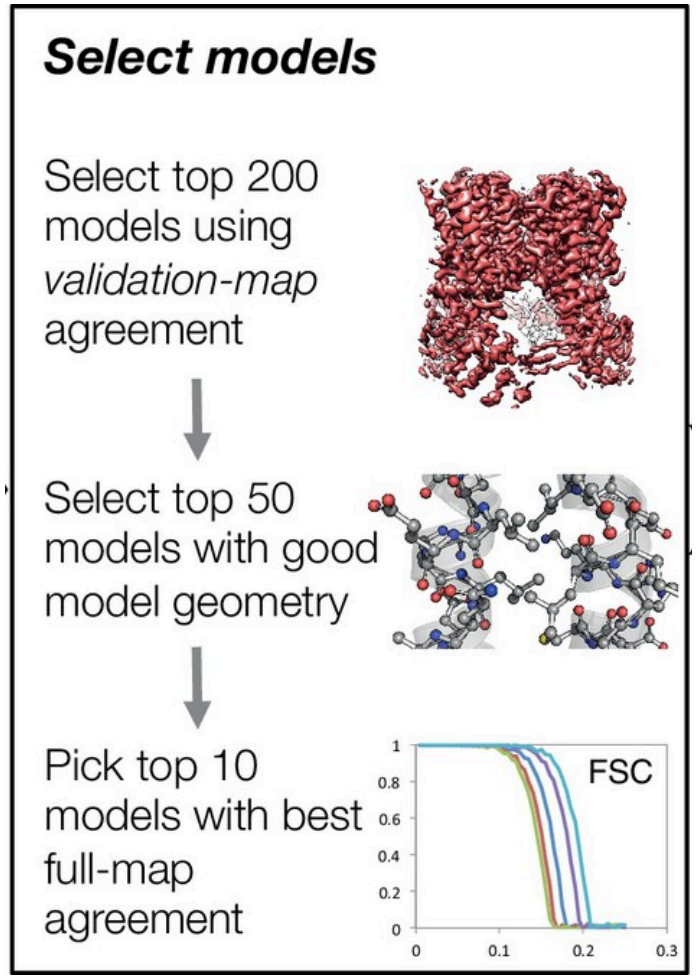

- Hand-built models usually fit the density well
- but are incorrect geometrically (the model is strained)
- Identify problematic residues by assessing local model-strain & local agreement to density
- Rebuild problematic regions using iterative fragment-based rebuilding followed by all-atom refinement
- This rebuilding happens in just **one** of the half maps

# Rosetta refinement : fragment-based backbone rebuilding

- Backbone fragments are collected from the PDB
- Superposed onto the current model
    1. First, use 25 x 17-residue fragments
    2. And then use 25 x 9-residue fragments
- Each fragment is optimized to fit the density
- At each position, the fragment with best fit to the density that has an r.m.s. of less than 0.5 Å over the terminal residues is selected
- Backbone atomic positions from the selected fragment then replace the corresponding backbone in the current model
- The backbone geometry at the stitching site is regularized

# Rosetta refinement : stage 2

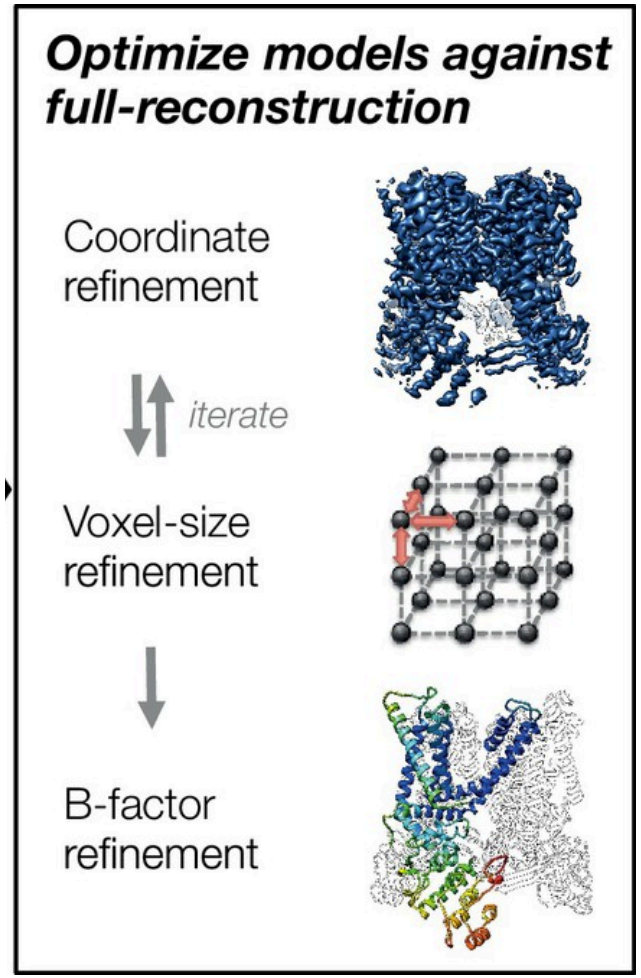
Stage 2




- Identify the top 200 stereo-chemically correct models with best agreement to an independent half-map (this prevents overfitting)
- Select top 50 models with good geometry
- Picks top 10 models with best agreement to the final full map

# Rosetta refinement : stage 3

Stage 3

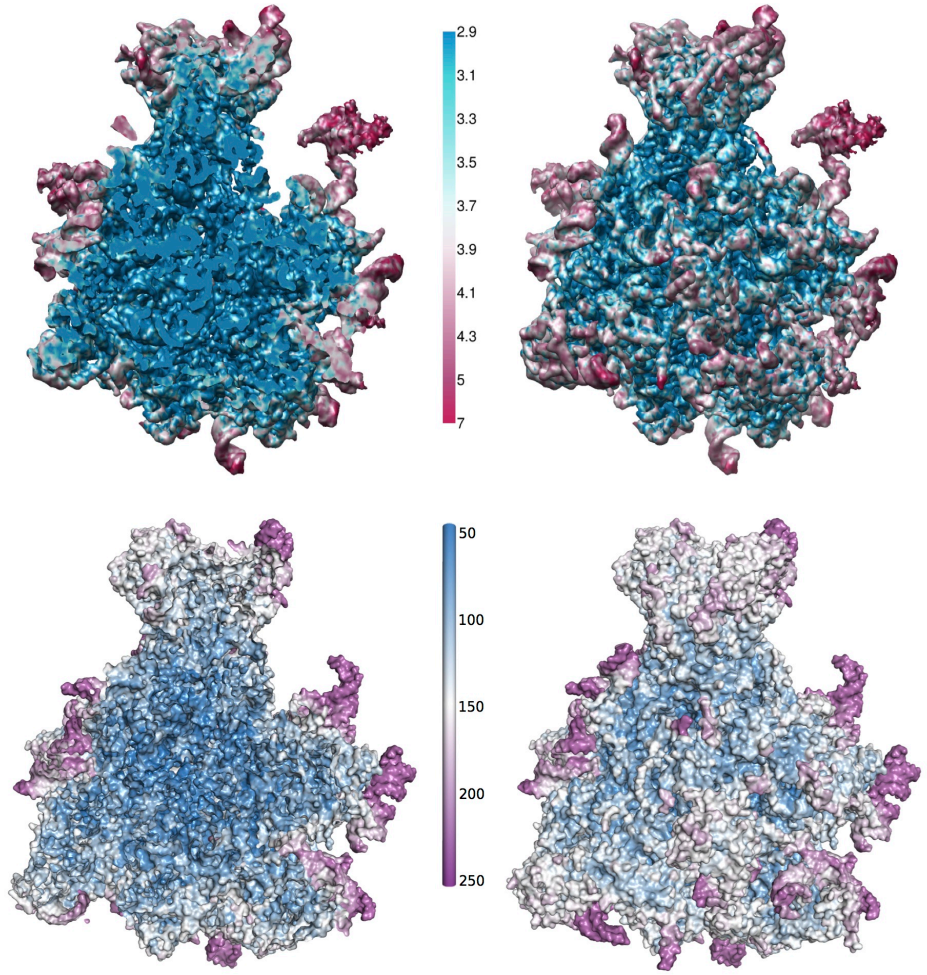



- Models are further optimized to the full-reconstruction
- This refinement uses a weight optimally scaled between experimental data and the forcefield using the 'validation' half map
- Can perform a magnification refinement – if the magnification on your microscope is poorly calibrated
- Performs a B-factor refinement

# Refinement : B-factors

B-factors are usually interpreted as a measurement of the amount of motion that an atom experiences

Should correlate with local resolution




sliced view

surface view

# Validation

"with great resolution comes great responsibility"

# Model validation

1. **Does the model agree with the map?**
   - Global measure of how well the model fits the map
   - Local measures of fit
   - B-factor calculations

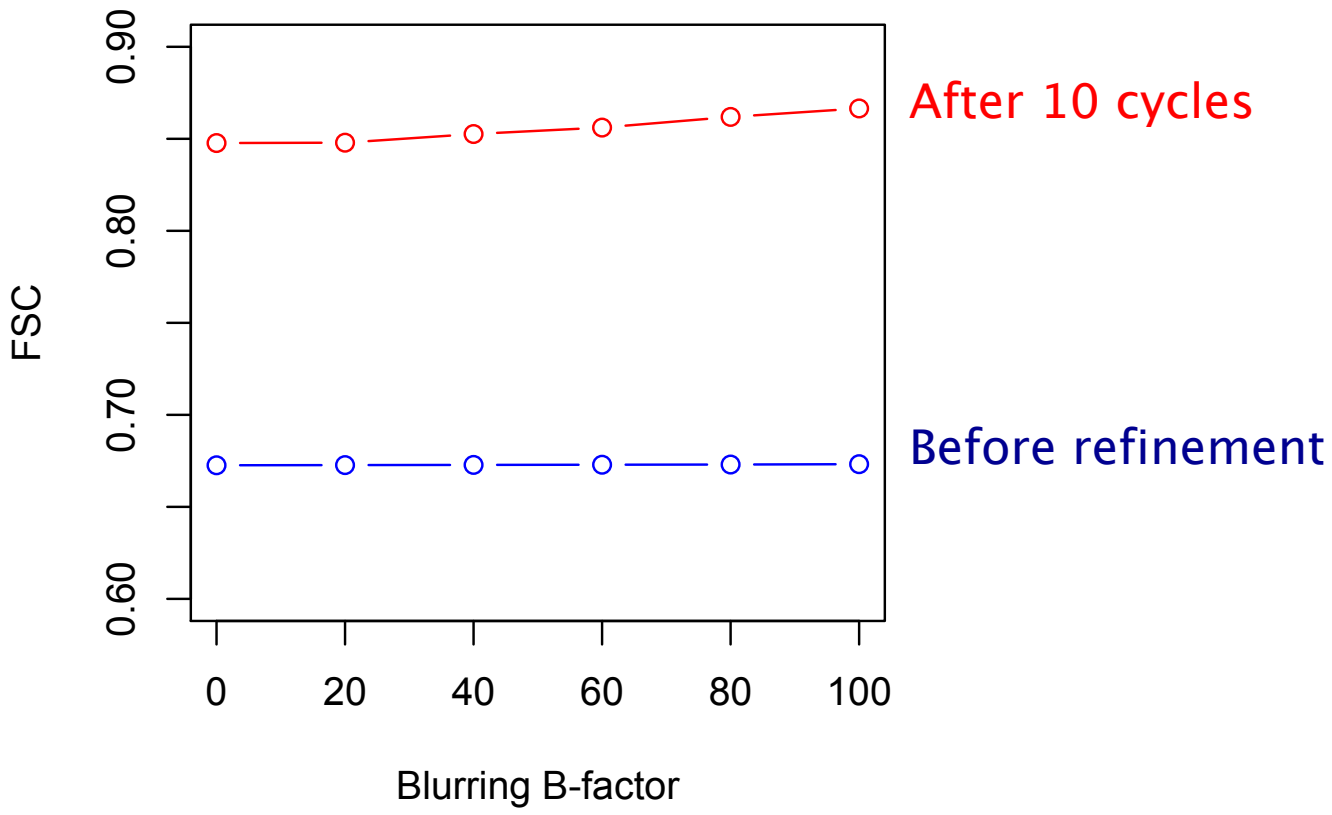
1. **Have we overfitted the data?**
   - Cross-validation

2. **Does the model look like other macromolecules and what we know of chemistry?**
   - Consistency of 3D structure with 1D sequence
   - Deviations from ideal values (bonds, angles, etc)
   - Non-bonded clashing atoms
   - Stereochemistry (Ramachandran plot)
   - Rotamers

# Global measure of fit : FSCaverage

- FSCaverage (as used in REFMAC) is largely independent of Bfactor.
- FSC is calculated over resolution shells. If the shells are sufficiently narrow the weights are roughly the same within each shell.

$$FSC_{average} = \frac{\sum_{i=1}^{N_{shell}} N_i FSC_i}{\sum_{i=1}^{N_{shell}} N_i}$$

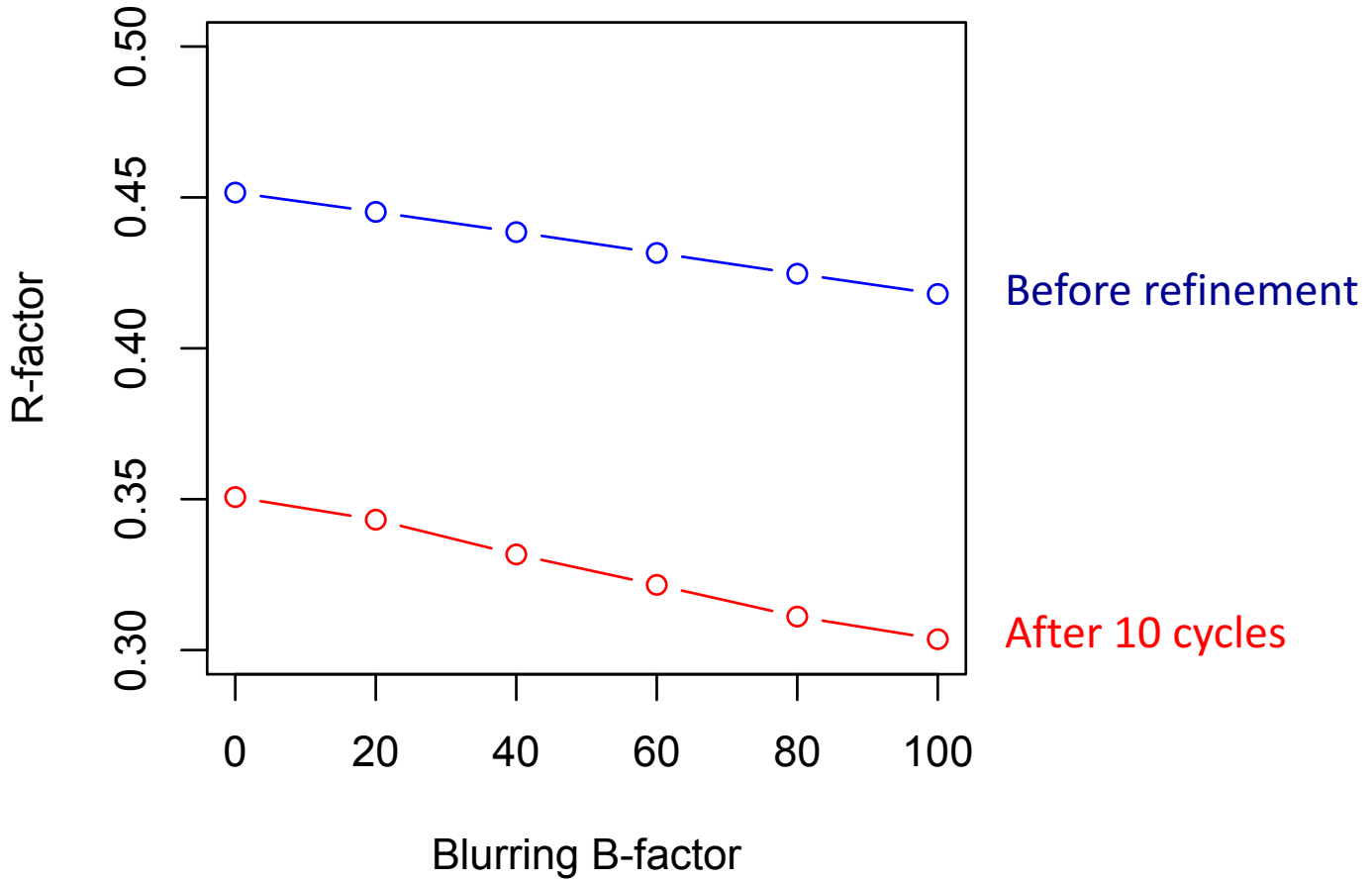


# Rfactor is an inappropriate measure of fit

The crystallographic Rfactor is inappropriate for monitoring fit-to-density as it can be artificially lowered by changing the B-factor

$$R_f = \frac{\sum_{\mathbf{h}} w_{\mathbf{h}} ||\mathbf{F}_{1\mathbf{h}}| - |\mathbf{F}_{2\mathbf{h}}||}{\sum_{\mathbf{h}} w_{\mathbf{h}} |\mathbf{F}_{1\mathbf{h}}|}$$

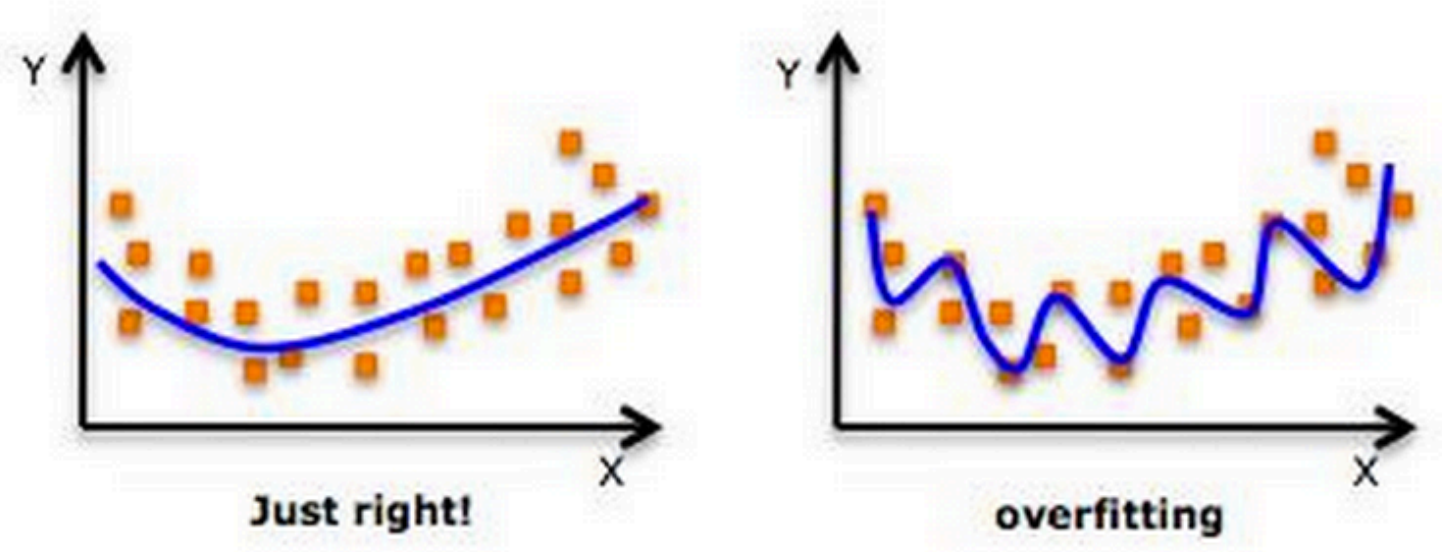
$e^{-Bs^2/4}$



# Local measure of fit

- Per-residue correlation coefficient
- Can be calculated using many different programs:
  - phenix.real_space_correlation
  - Rosetta (-denstools::perres)
  - score_smoc.py (CCP-EM, overlapping residue windows)

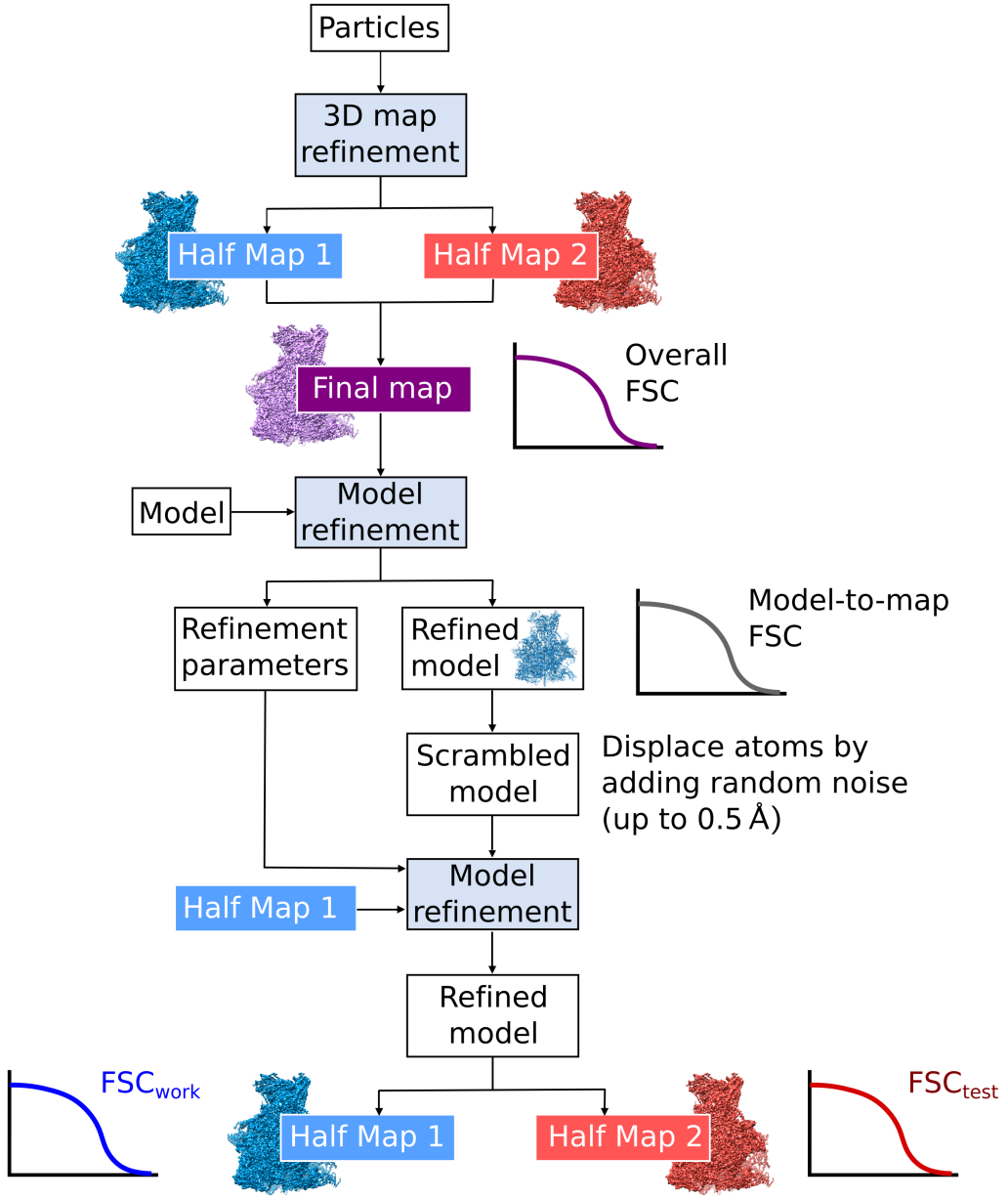
# Overfitting




**What leads to overfitting?**

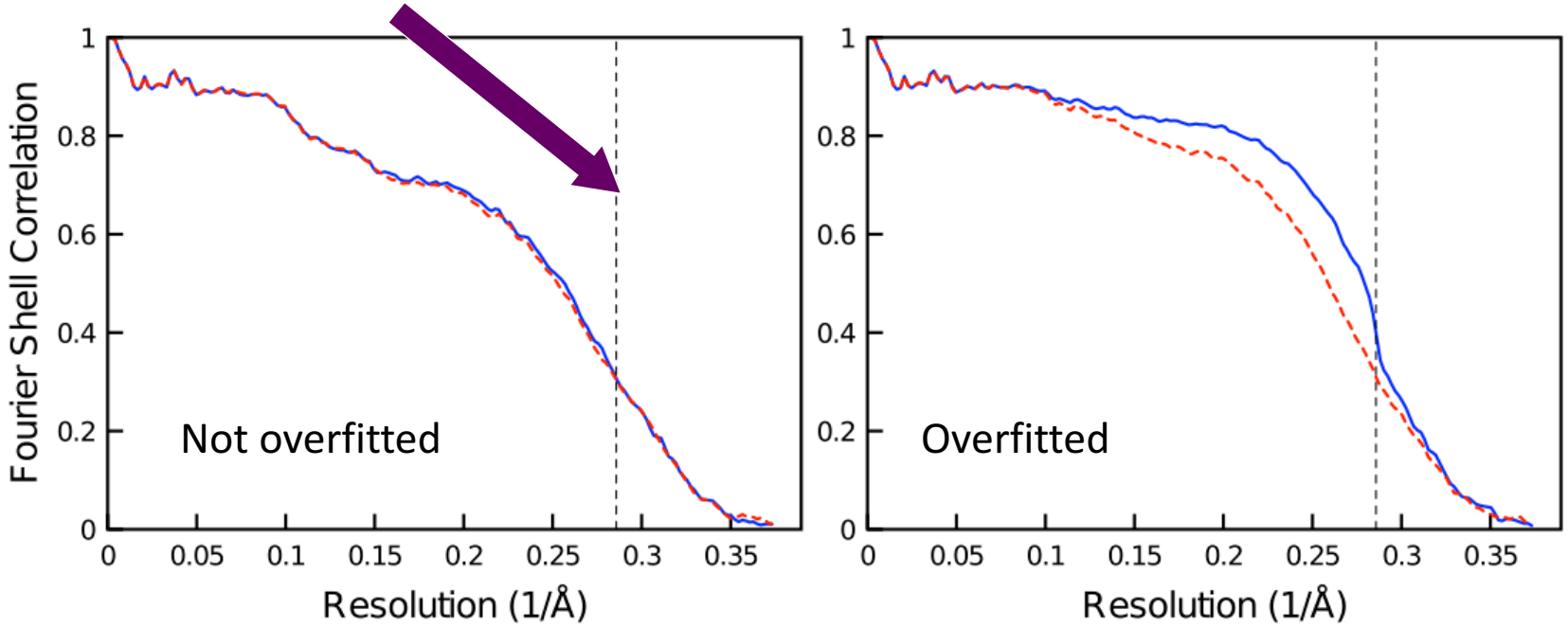
1. Insufficient data (low resolution, partial occupancy)
2. Ignoring data (cutting by resolution)
3. Sub-optimal parameterization
4. Bad weights
5. Excess of imagination

# Overfitting (half map validation)



# Overfitting (half map validation)

Resolution limit must be defined in refinement

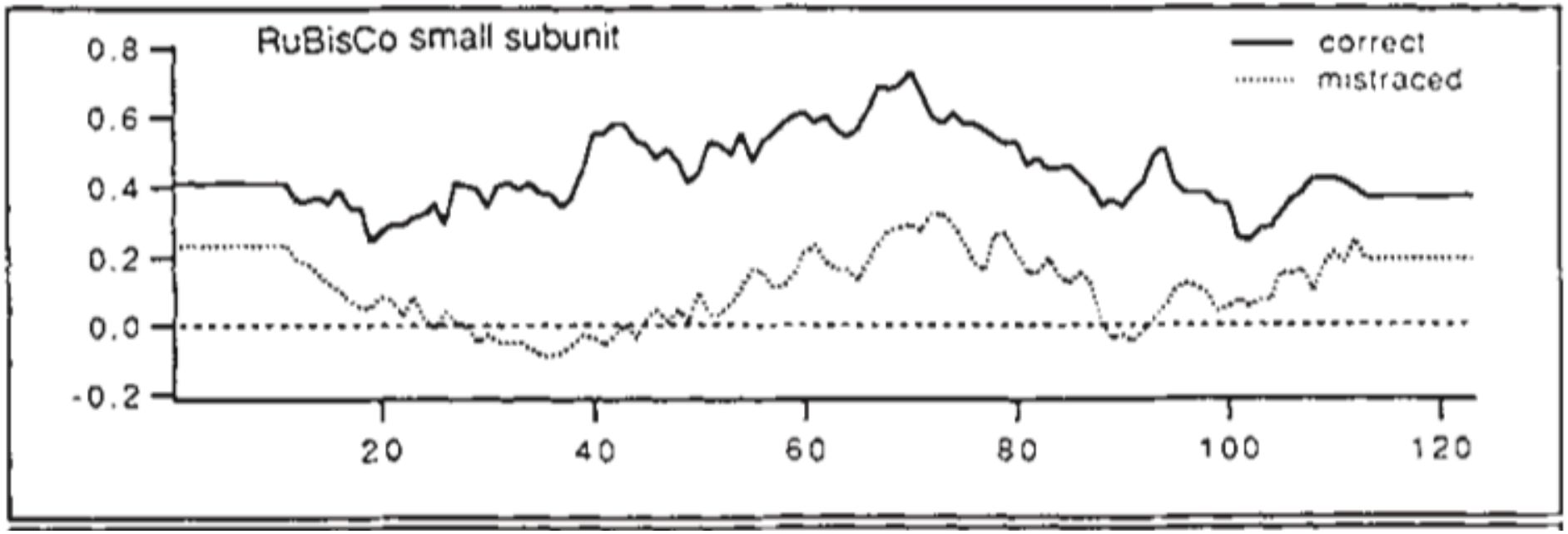



- Both FSC curves should overlap
- Smooth transition beyond resolution cutoff applied during refinement

- Sharp fall beyond resolution used for refinement (loss of predictive power)
- Better fitting to the map the model was refined against (fitting to noise)

# Verify3D

- **Key developers**: David Eisenberg (UCLA)
- **Basic premise**: Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D)
- **Availability**: http://services.mbi.ucla.edu/Verify_3D/
- **References**: (1992) Nature, 356, 6364:83-85

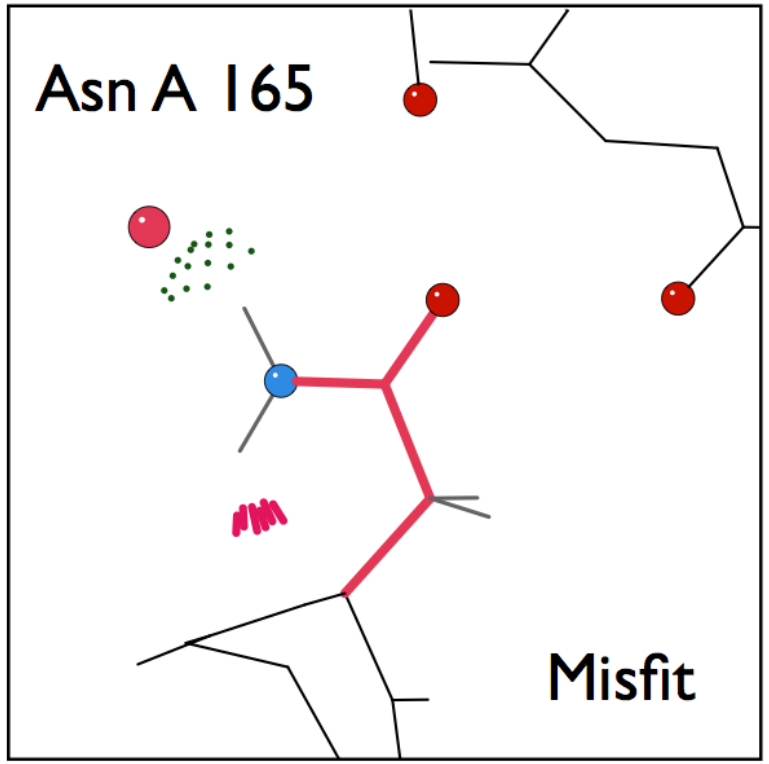


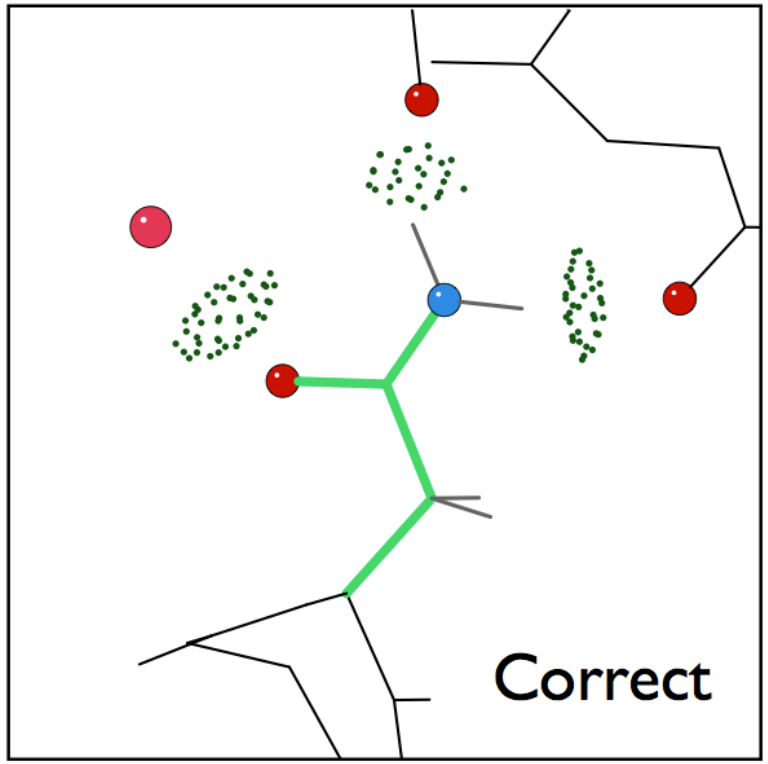
# MolProbity

- **Key developers**: David and Jane Richardson (Duke University)
- **Basic premise**: Evaluates model quality.
- **Availability**: online (http://molprobity.biochem.duke.edu/) and through the Phenix distribution (phenix.molprobity)
- **References**: Chen et al. (2010). Acta Cryst. D66:12-21.

# MolProbity

- Don’t wait until the end of model building before running MolProbity
- By analyzing hydrogen-bonding networks, can automatically detect and correct flipped N/Q/H residues



# MolProbity

What information does MolProbity provide?

PDB : 5K12, EMDB: EMD-8194, resolution = 1.8 Å

| | | | | |
|---|---|---|---|---|
| All-Atom Contacts | Clashscore, all atoms: | 19.75 | | 26th percentile* (N=837, 1.80Å ± 0.25Å) |
| | Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms. | | | |
| Protein Geometry | Poor rotamers | 12 | 0.82% | Goal: <0.3% |
| | Favored rotamers | 1440 | 98.36% | Goal: >98% |
| | Ramachandran outliers | 0 | 0.00% | Goal: <0.05% |
| | Ramachandran favored | 1674 | 96.21% | Goal: >98% |
| | MolProbity score^ | 2.05 | | 59th percentile* (N=11444, 1.80Å ± 0.25Å) |
| | Cβ deviations >0.25Å | 0 | 0.00% | Goal: 0 |
| | Bad bonds: | 0 / 14070 | 0.00% | Goal: 0% |
| | Bad angles: | 0 / 18996 | 0.00% | Goal: <0.1% |
| Peptide Omegas | Cis Prolines: | 6 / 66 | 9.09% | Expected: ≤1 per chain, or ≤5% |

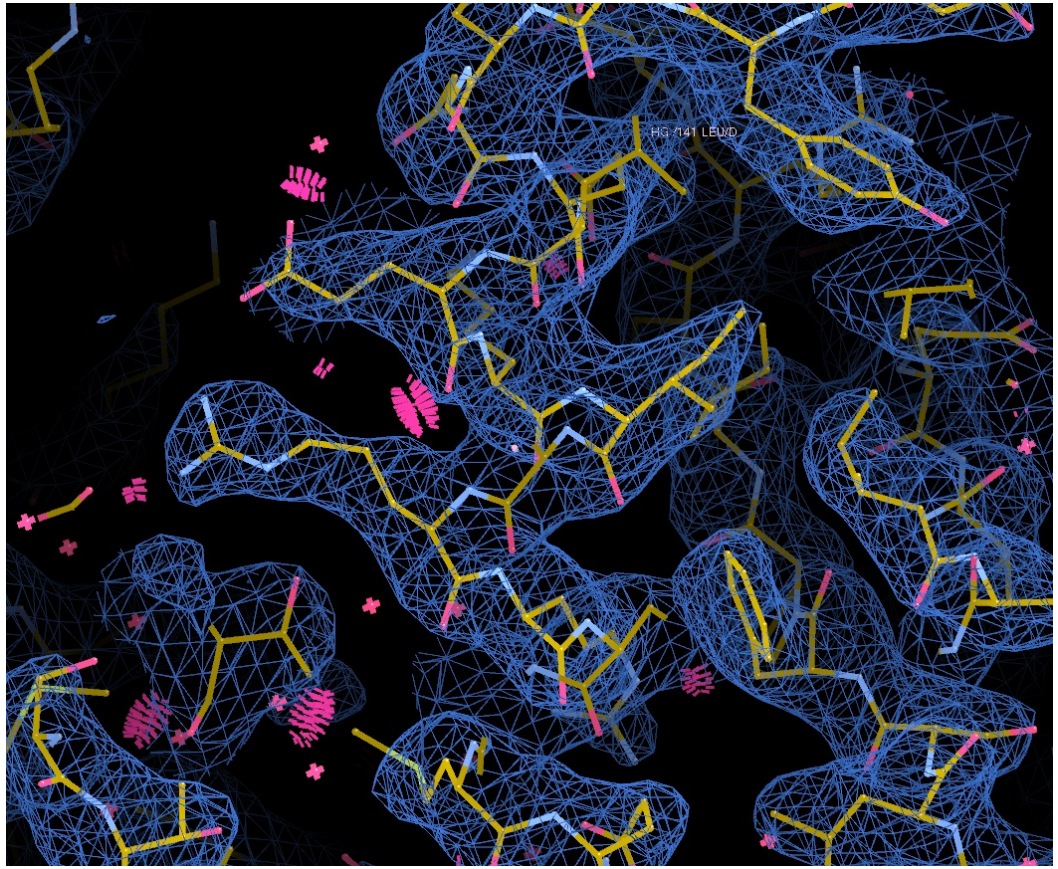
In the two column results, the left column gives the raw count, right column gives the percentage.

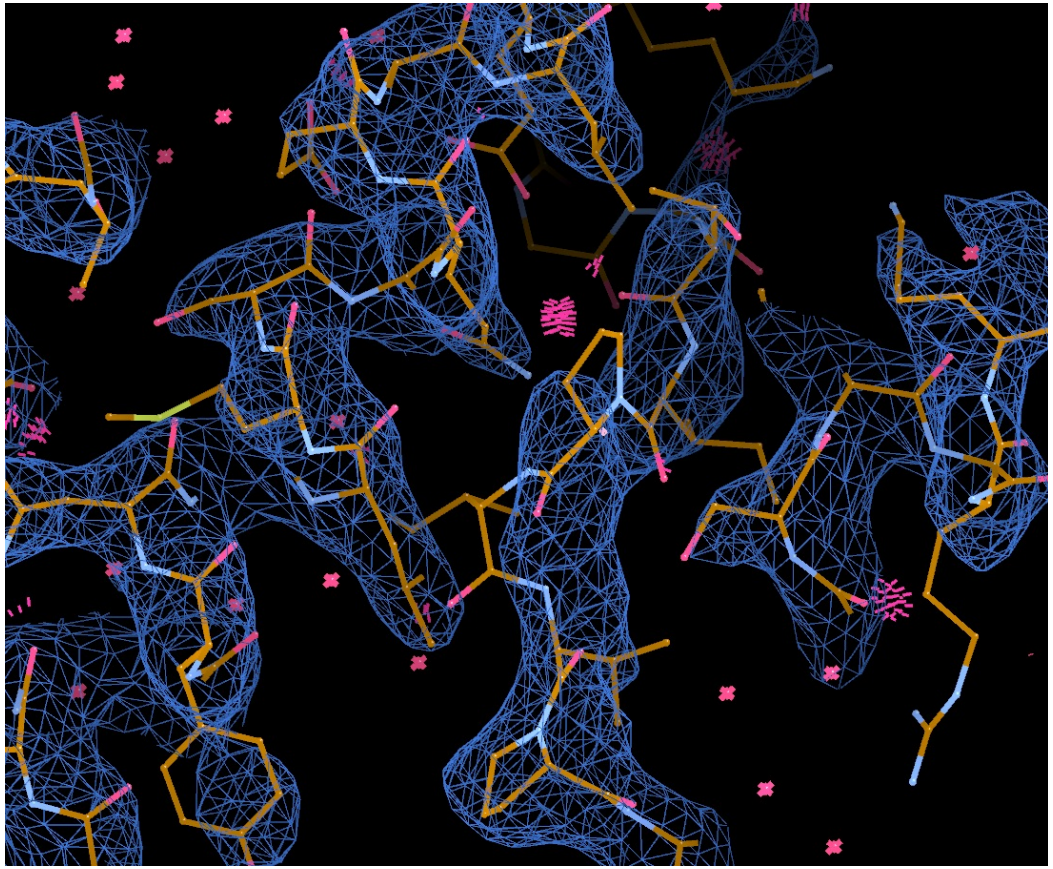
* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst. For clashscore the comparative set of structures was selected in 2004, for MolProbity score in 2006.

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

# MolProbity + Coot

- Validation and model building are not separate entities





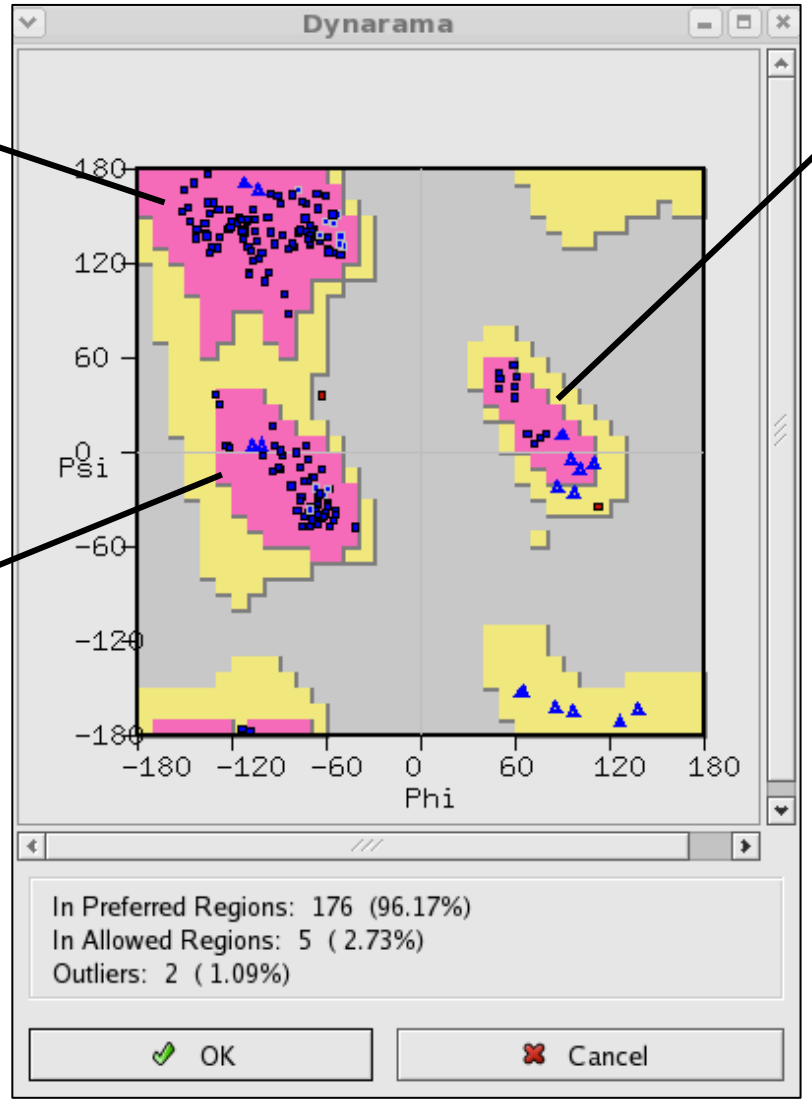
clash

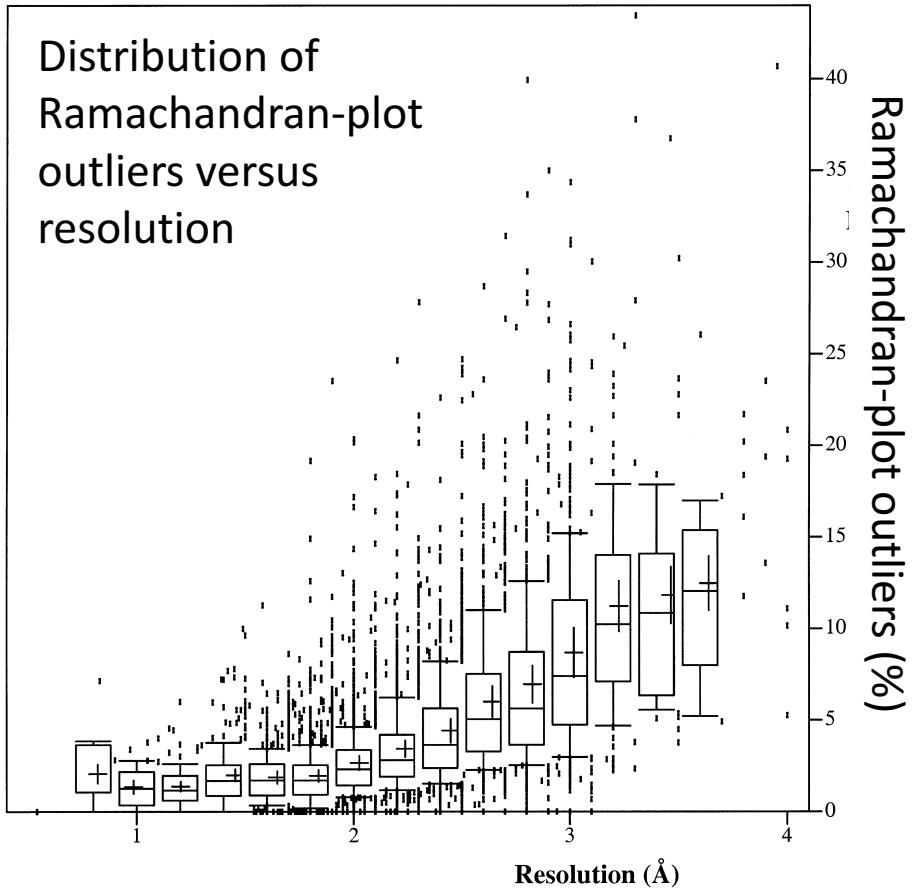
# Ramachandran plot

β-strand

left-handed α-helix

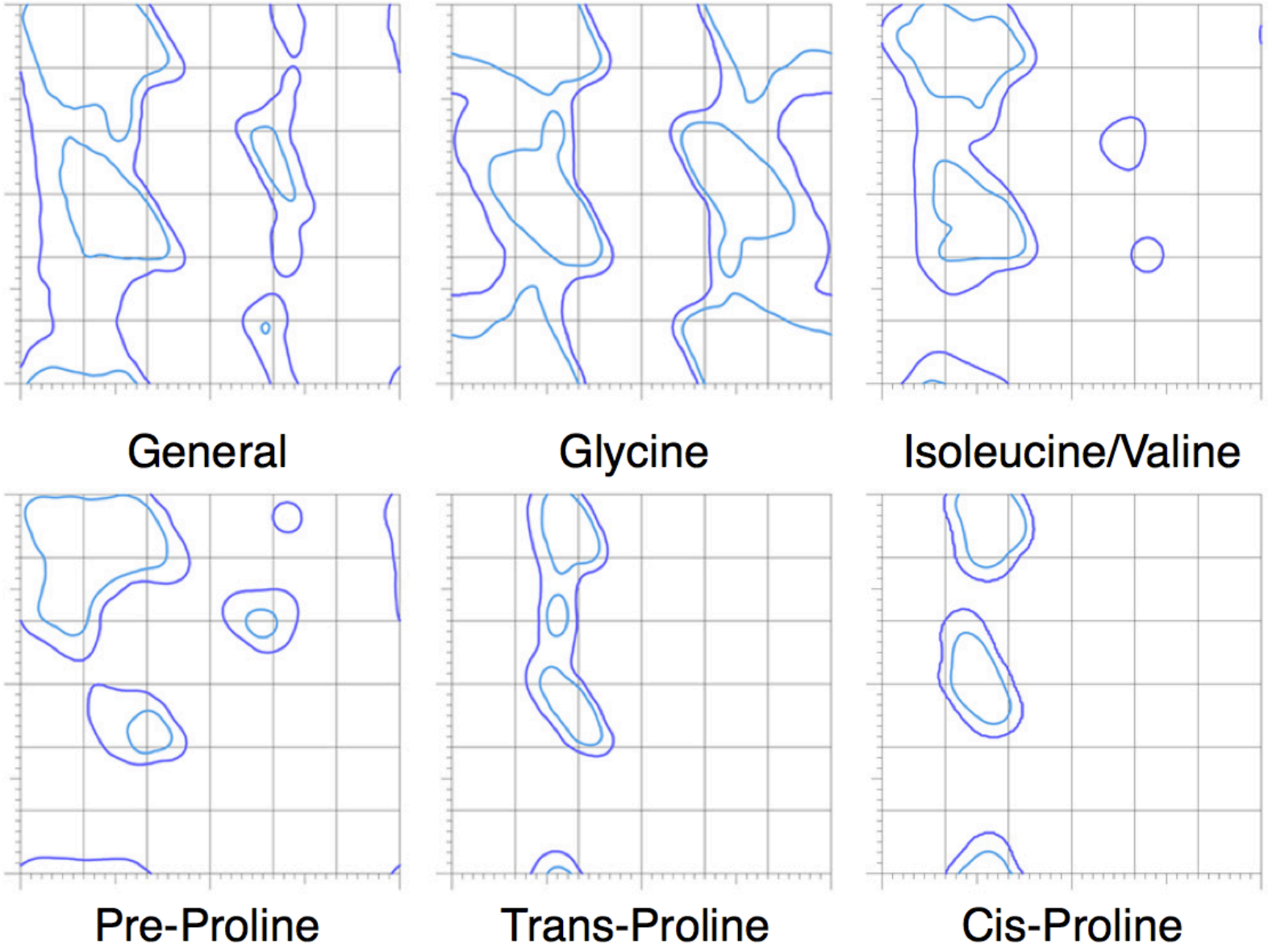
Right-handed α-helix/$3_{10}$



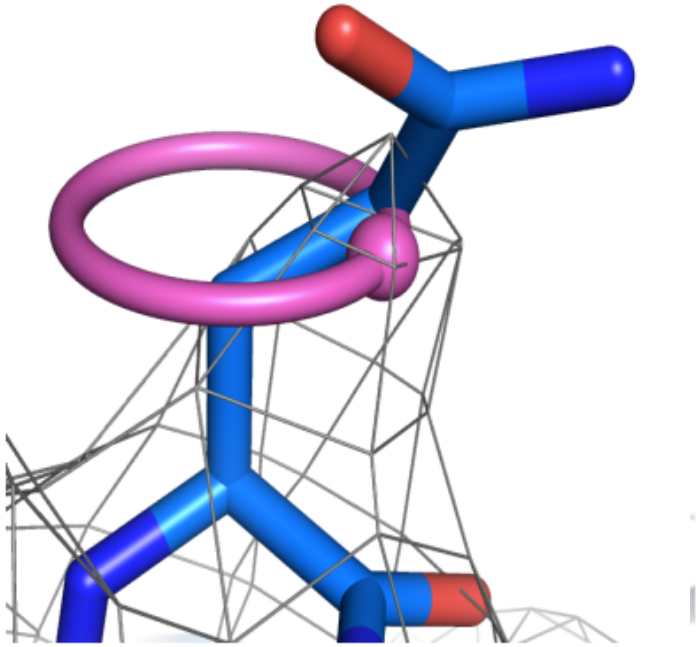

NOTE: not everything flagged as an outlier is wrong - check

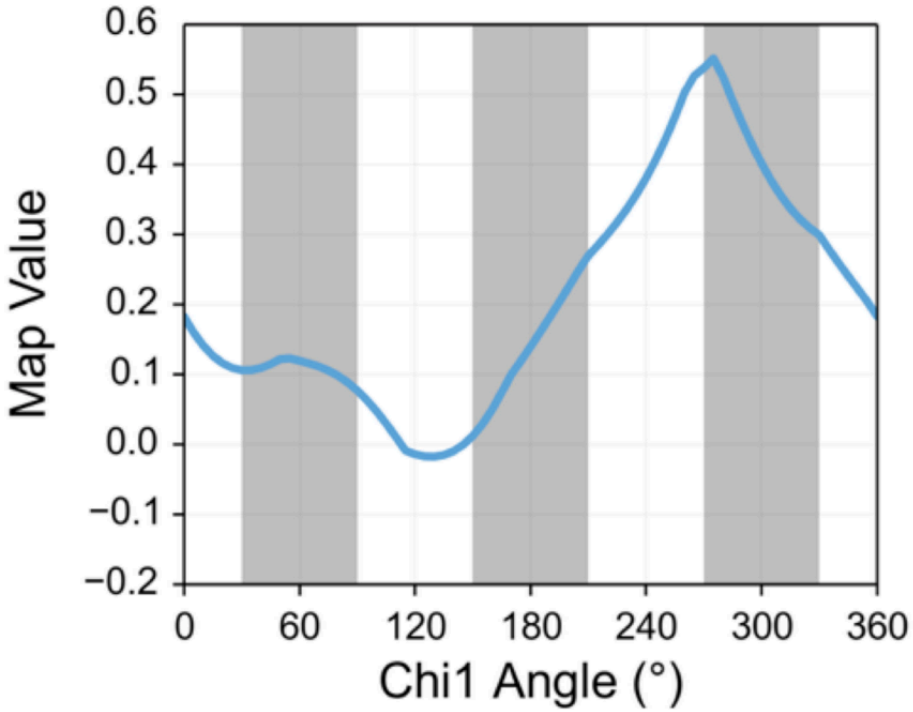
# Ramachandran plots



# EMRinger

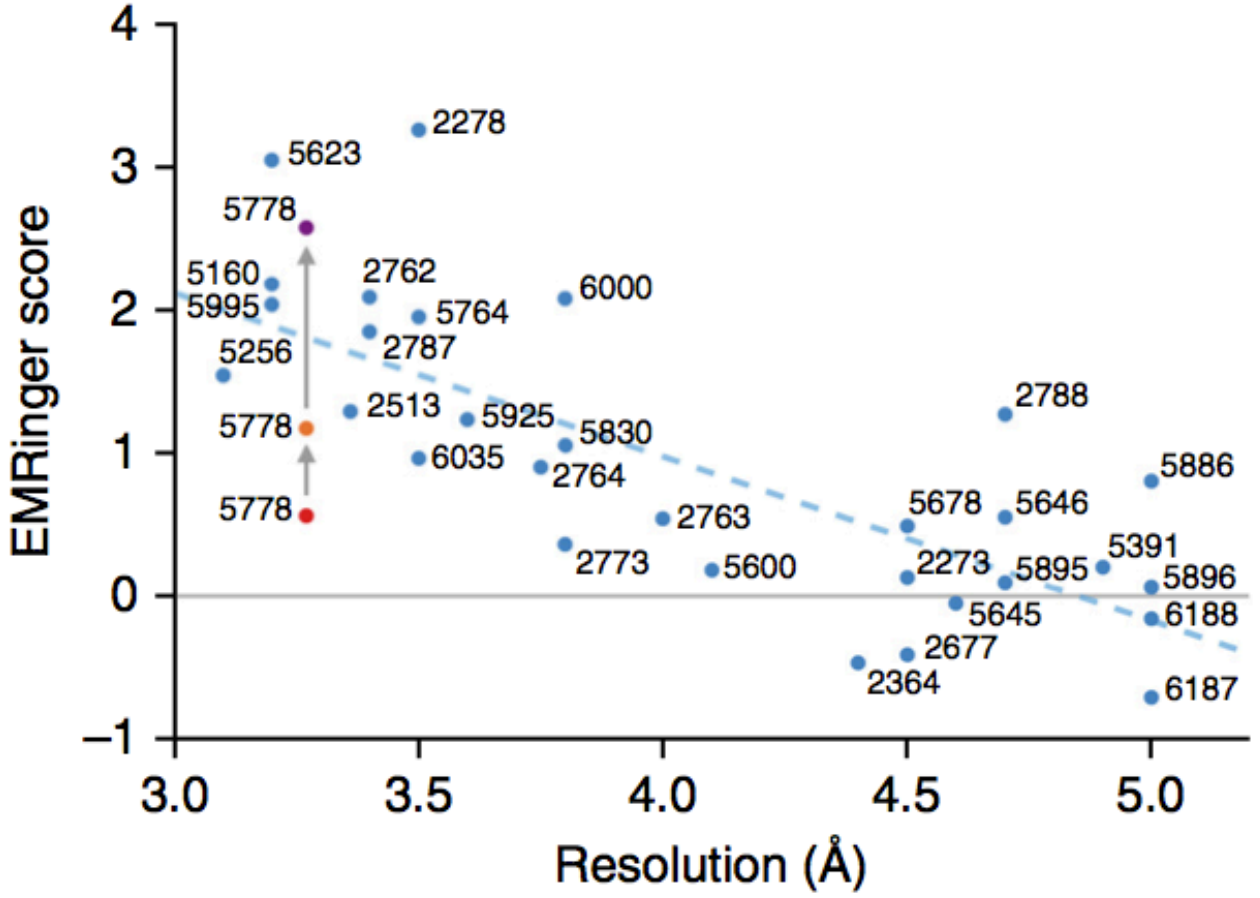
- **Key developer**: Ben Barad/James Fraser (UCSF)
- **Basic premise**: Rotates C-gamma atom around the χ1 angle of a side chain, interpolating the density value in the map as it rotates
- **Availability**: phenix.emringer
- **References**: Barad et al (2015) Nat. Methods, 12:943-946





# EMRinger

- EMRinger score is correlated with resolution
- Gives an idea of what a good score should be



# PDB Validation Reports

Generate a validation report : https://validate-rcsb-1.wwpdb.org/validservice/

WORLDWIDE PDB PROTEIN DATA BANK

EMDataBank Unified Data Resource for 3DEM

Full wwPDB/EMDataBank EM Map/Model Validation Report (i)

Jun 8, 2017 – 06:18 AM EDT

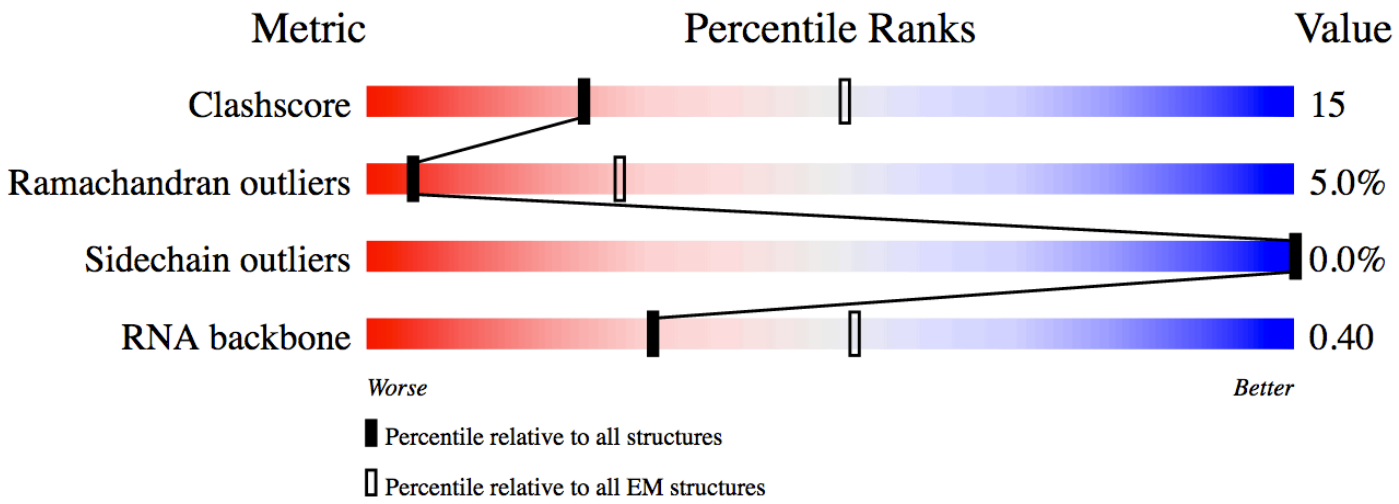
This is a Full wwPDB/EMDataBank EM Map/Model Validation Report.

This report is produced by the standalone wwPDB validation server.
**The structure in question has not been deposited to the wwPDB.**
**This report should not be submitted to journals.**

We welcome your comments at *validation@mail.wwpdb.org*
A user guide is available at
http://wwpdb.org/validation/2016/EMValidationReportHelp
with specific help available everywhere you see the (i) symbol.

PRELIMINARY VALIDATION REPORT

| | | |
|---|---|---|
| MolProbity | : | 4.02b-467 |
| Mogul | : | 1.7.2 (RC1), CSD as538be (2017) |
| Percentile statistics | : | 20161228.v01 (using entries in the PDB archive December 28th 2016) |
| Ideal geometry (proteins) | : | Engh & Huber (2001) |
| Ideal geometry (DNA, RNA) | : | Parkinson et. al. (ll996) |
| Validation Pipeline (wwPDB-VP) | : | rb-20029077 |




**Help**: https://www.wwpdb.org/validation/2016/EMValidationReportHelp

# Presenting validation statistics

## “Table 1”

| | Dataset 1 | Dataset 2 |
|---|---|---|
| **Data Collection** | | |
| Pixel size (Å) | 1.34 | 1.06 |
| Defocus range (μm) | -1.5 to -3.5 | -0.5 to -3.5 |
| Voltage (kV) | 300 | 300 |
| Electron dose (e- $Å^{-2}$) | 25 | 39 |
| | **39S intermediate with folded rRNA** | **39S intermediate with unfolded rRNA** |
| **Map reconstruction** | | |
| Particles | 134,685 | 379,869 |
| Resolution (Å) | 3.06 | 3.03 |
| Map sharpening B-factor ($Å^2$) | -85.0 | -95.0 |
| **Model composition** | | |
| Non-hydrogen atoms | 99,025 | 90,747 |
| Protein residues | 8,230 | 8,135 |
| RNA bases | 1,497 | 1,148 |
| Ligands ($Zn^{2+}$/$Mg^{2+}$) | 3/93 | 3/49 |
| **Fit to map** | | |
| Correlation coefficient (entire box) | 0.76 | 0.77 |
| Correlation coefficient (around atoms) | 0.77 | 0.79 |
| Fourier shell correlation (entire box) | 0.76 | 0.77 |
| Fourier shell correlation (around atoms) | 0.82 | 0.83 |
| **Protein geometry** | | |
| Molprobity score | 1.66 | 1.70 |
| All-atom clashscore | 5.73 | 5.94 |
| EMRinger score | 3.69 | 3.85 |
| RMSD deviation bonds (Å) | 0.010 | 0.016 |
| RMSD deviation angles (°) | 1.02 | 1.27 |
| Favored rotamers (%) | 95.7 | 95.0 |
| Poor rotamers (%) | 0.48 | 0.90 |
| Ramachandran favored (%) | 94.9 | 94.4 |
| Ramachandran outliers (%) | 0.02 | 0.06 |
| **Nucleic acid geometry** | | |
| Poor sugar puckers (%) | 1.14 | 1.39 |
| Bad backbone conformations (%) | 26.3 | 27.7 |

# Deposition


https://deposit-pdbe.wwpdb.org/deposition

- Deposit at the same time as EM maps.
- Recommended depositions:
  - Postprocessed map
  - Both half maps
  - Any masks applied during processing
  - Any map that has been modified in any way (excluding blurring/sharpening)
  - Model