

Model Refinement

Pavel Afonine



phenix-online.org



lbl.gov



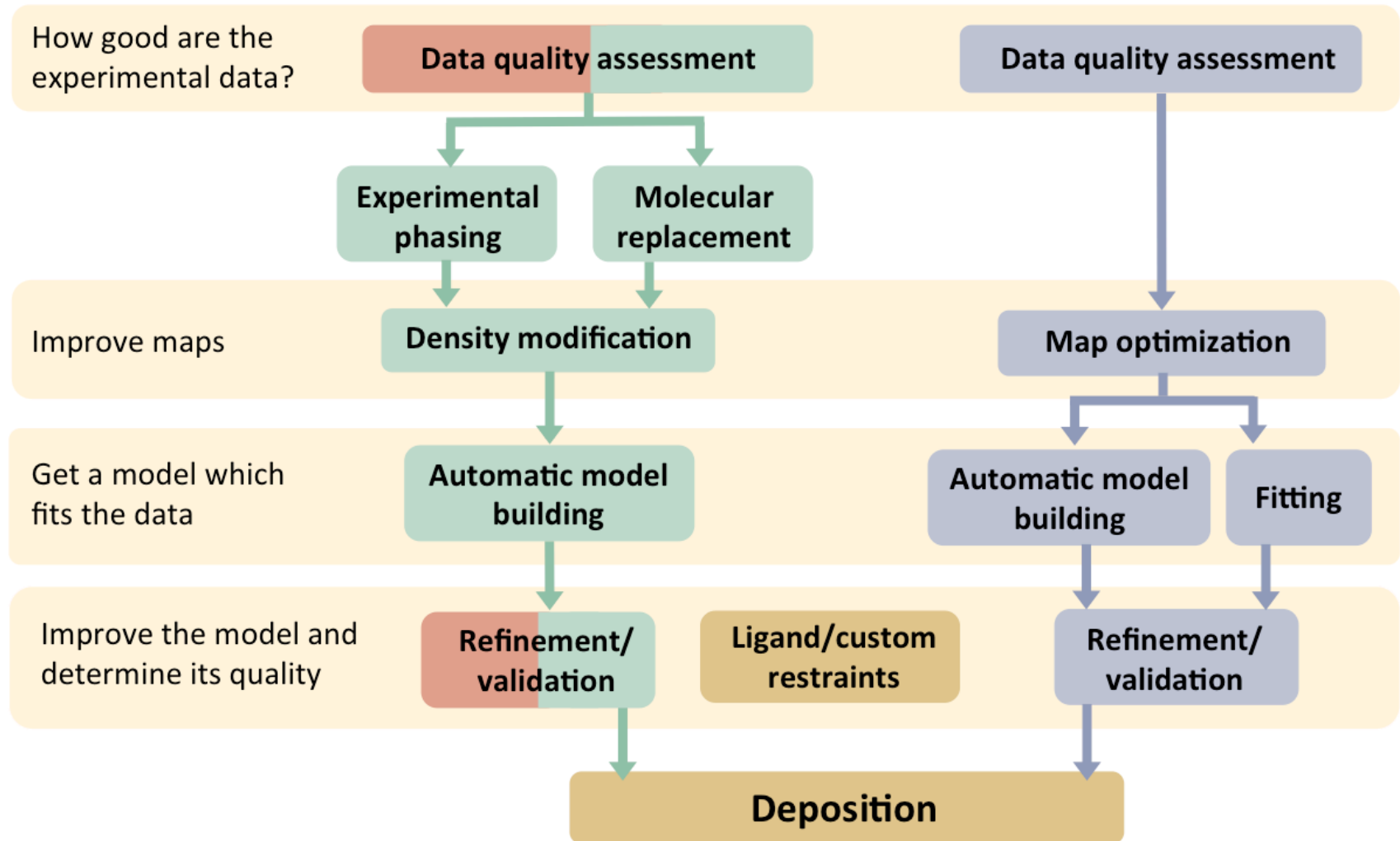
qrefine.com

November 9th 2023
KU

Phenix: tools for crystallography and cryo-EM

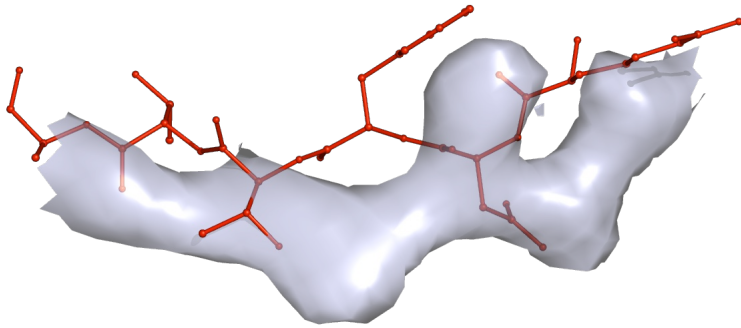
Xray/neutron crystallography

Cryo-EM

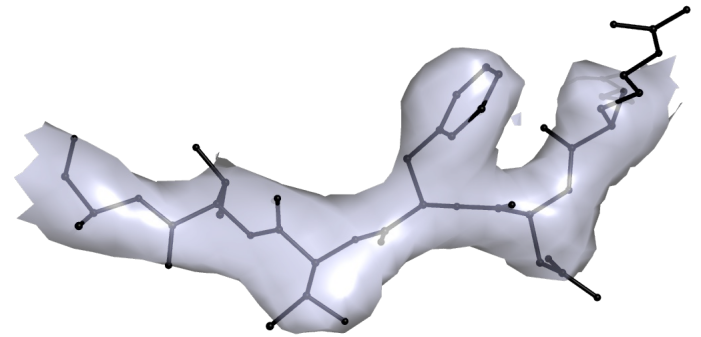


Model refinement in a nutshell

Initial model

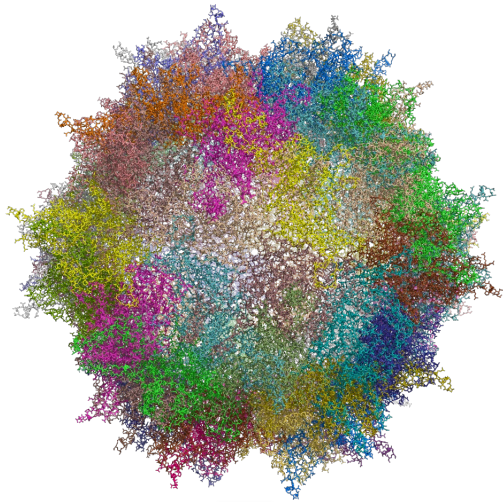


Improved (refined) model

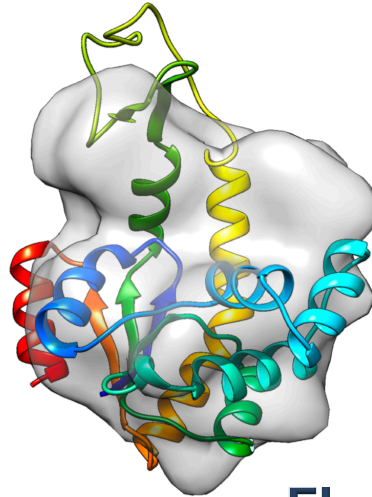
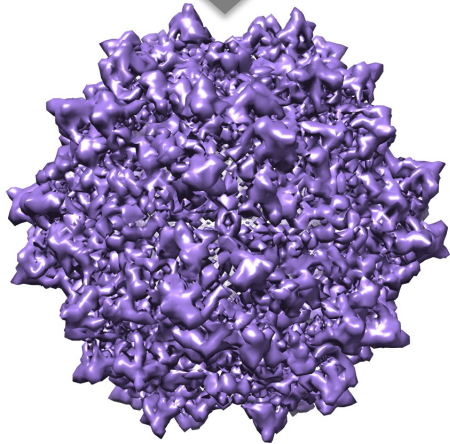


Fit atomic model to experimental data with the help of some *a priori* known information about the model

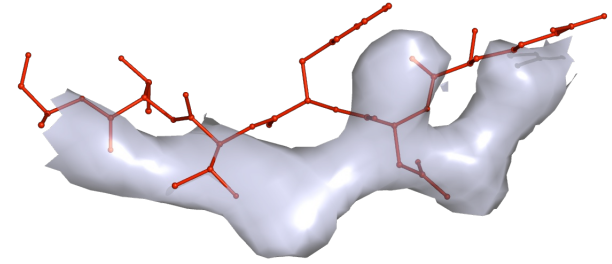
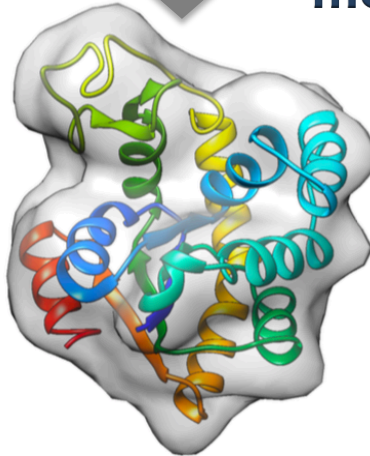
Not all model-to-data fitting is refinement



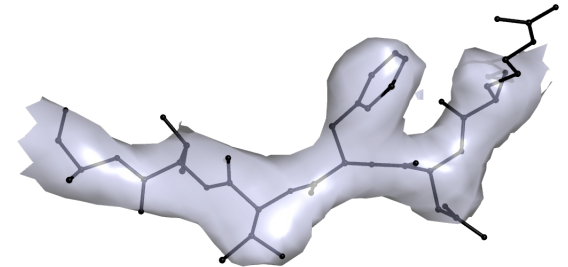
↓
Docking



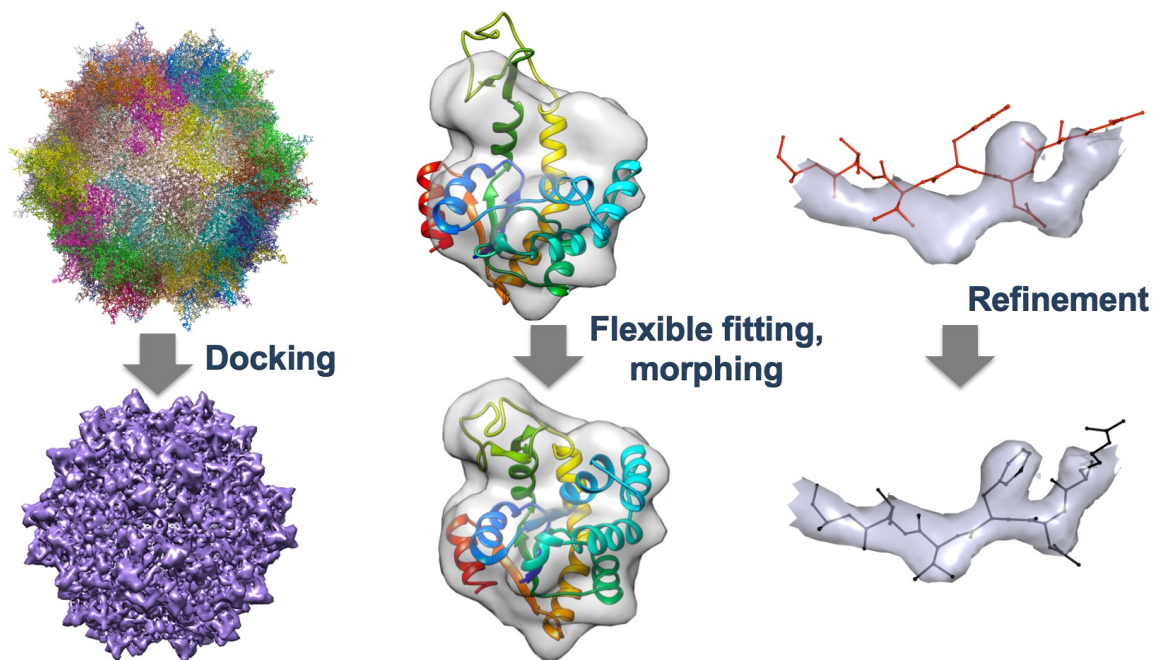
↓
**Flexible fitting,
morphing**



↓
Refinement



Not all model-to-data fitting is refinement



- Docking, flexible fitting, morphing are **not** refinement
- Refinement is to fine-tune an already fine atomic model
- Refinement does only small changes to the model (within *convergence radius of refinement*, $\sim 1\text{\AA}$)

Refinement used to be tedious and time consuming

- Familiar with multiple software packages
- Coding knowledge (typically FORTRAN or C)
- Expertise in Unix
- Reading thick books (no Google or ChatGPT!)
 - Anyone remembers 405 pages X-plor book by A. Brunger?
- Don't expect your questions answered by email within 24 hours

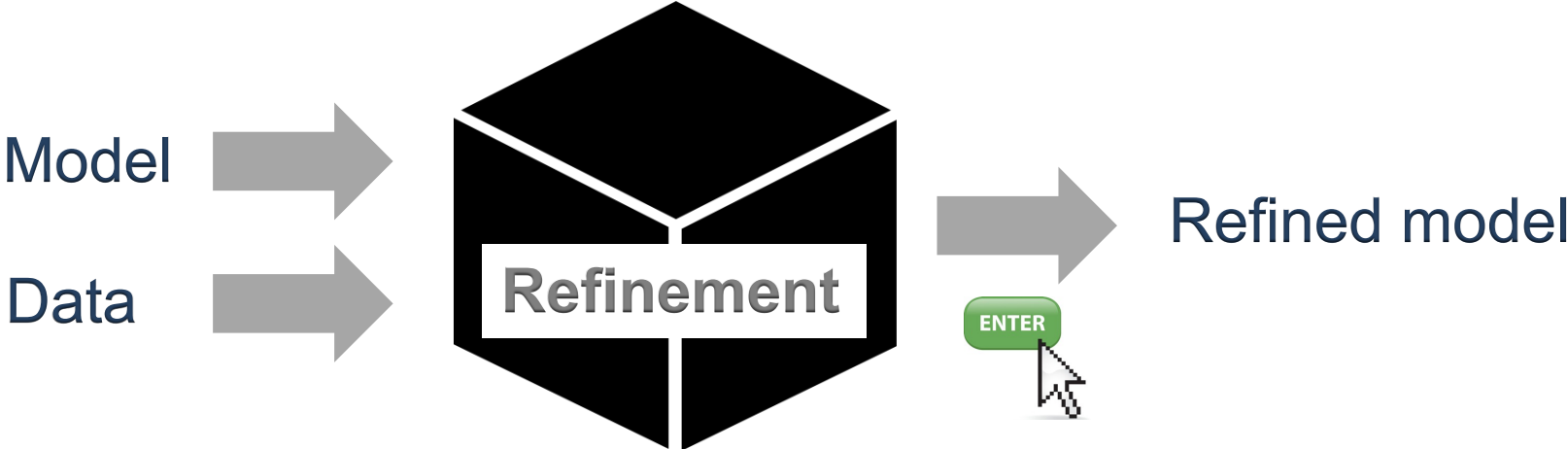
Refinement used to be tedious and time consuming

- Many months to complete
 - Spend days on graphics (manual building)
 - Run refinements overnight

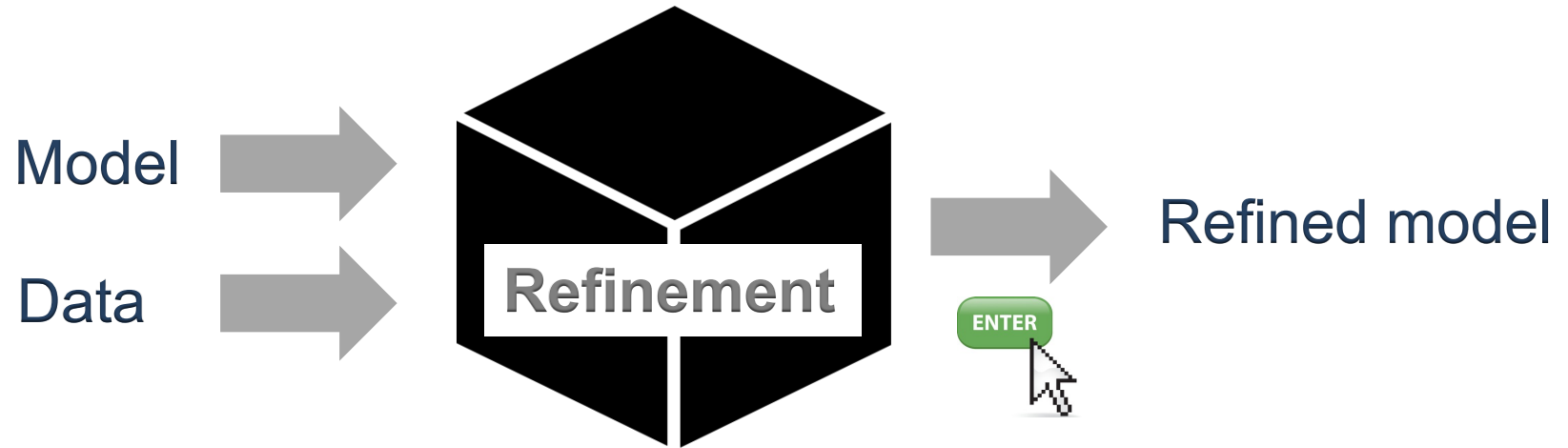


Solving my first structure in 1997

Model refinement: black box



Model refinement: black box



- Does it always work?
- Is it always as easy as *poor model in, better model out*?

Model refinement: black box

- **No.** Because:
 - Refinement parameterization isn't easy (next slide)
 - Default settings suit most common scenario
 - Typical resolution data, model reasonably fits data
 - Less typical situations need customizations
 - Low or high resolution data
 - Incomplete models
 - Final models
 - AlphaFold predicted models
 - Novel ligands

Model refinement: lots of stuff to know...

Reference model?

TLS?

Rotamer fixing?

AltLocs?

ADP?

Group B vs individual?

Local minima?

tNCS?

Clashes?

NCS?

IAS?

Weights?

CDL?

SA?

Grid search?



Minimization?

Rama plot restraints?

f' & f''?

Hydrogens?

Restraints?

Bulk-Solvent?

Rigid body?

Rama-Z?

Anisotropy?

NQH flips?

SS restraints?

Twinning?

Model refinement: black box

- What to do when the 'black box' does not work?
- **Your decision-making** is needed (and it is not always easy!)

Model refinement: decision-making variables

- **Crystal**

- Disorder
- Twinning, tNCS
- Solvent content
- Symmetry

- **Data**

- Resolution
- Errors
- Completeness
- Processing

- **Model**

- Stage
- Source
- Parameterization
- Fit to data

How you know...

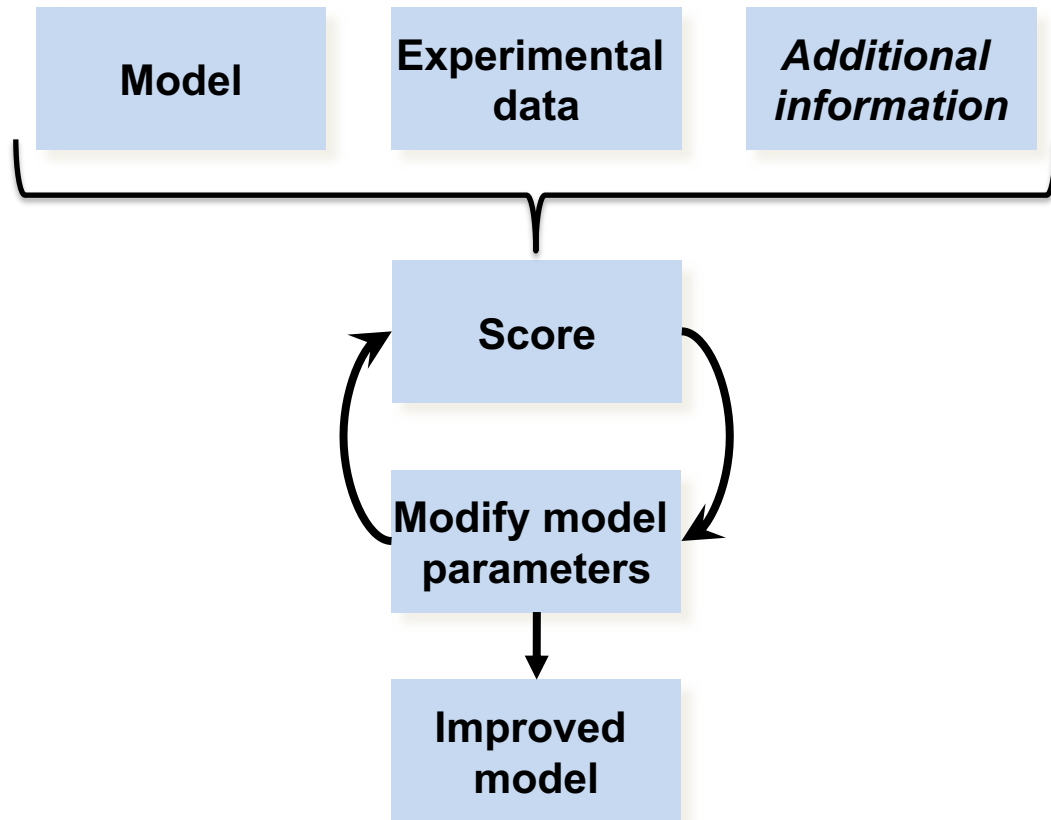
- ... refinement worked ?
- ... you did it correctly ?
- ... the model is good enough to publish ?

How you know...

- ... refinement worked ?
 - ... you did it correctly ?
 - ... the model you got is good enough to publish ?
-
- **Do validation!**
- Standard validation protocols are designed to answer these questions**

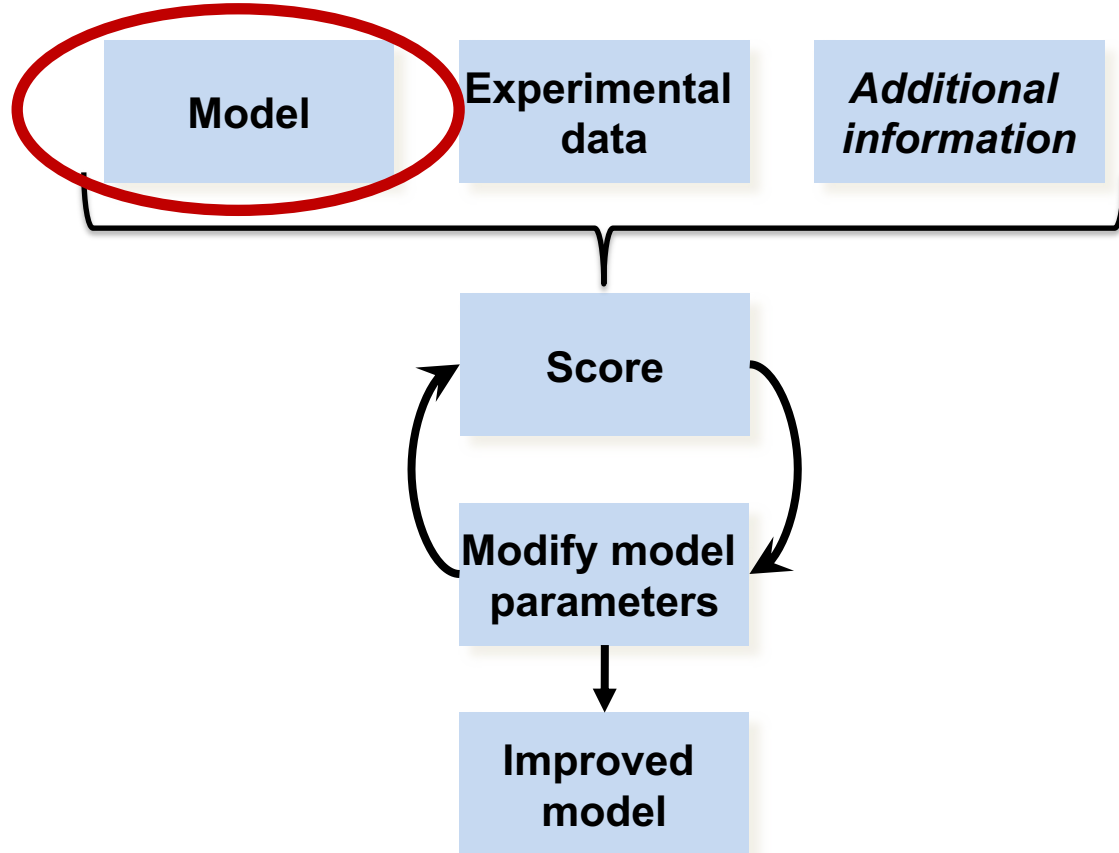
Refinement: a closer look

Model refinement



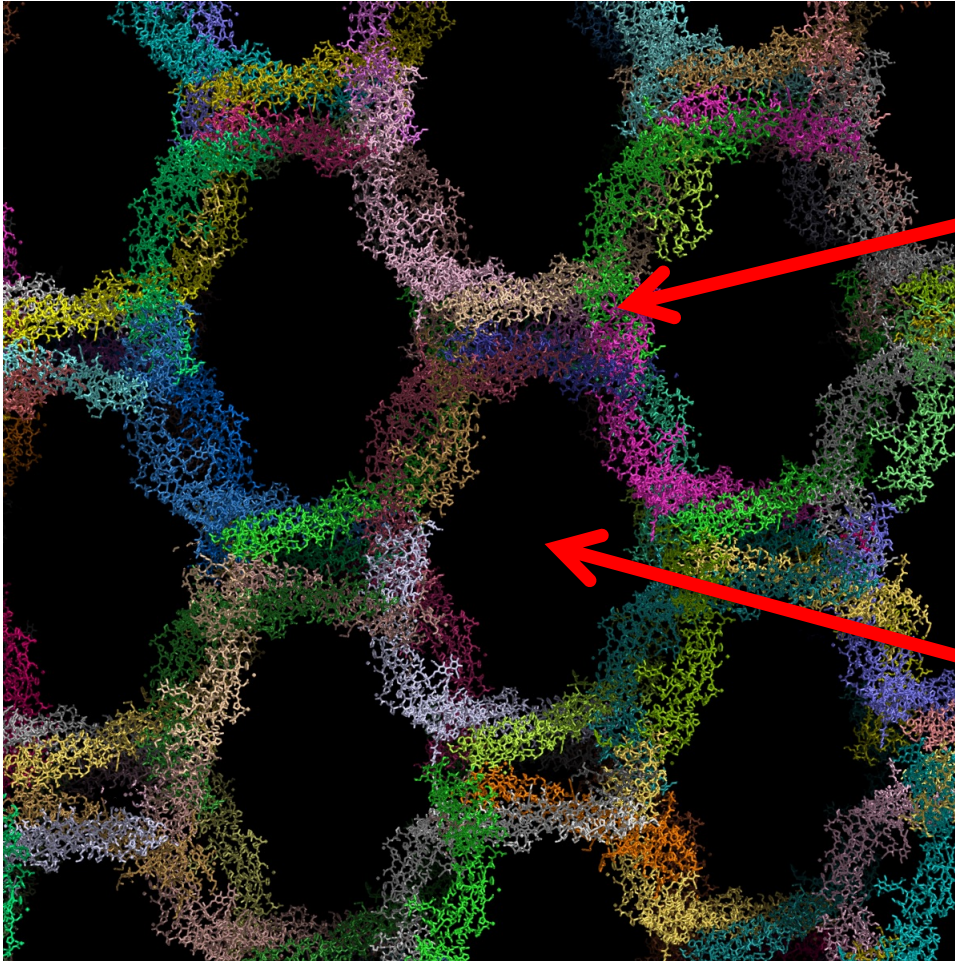
Refinement – optimization process of fitting model parameters to experimental data

Model refinement



Crystal structure model

PDB code: 1QUB

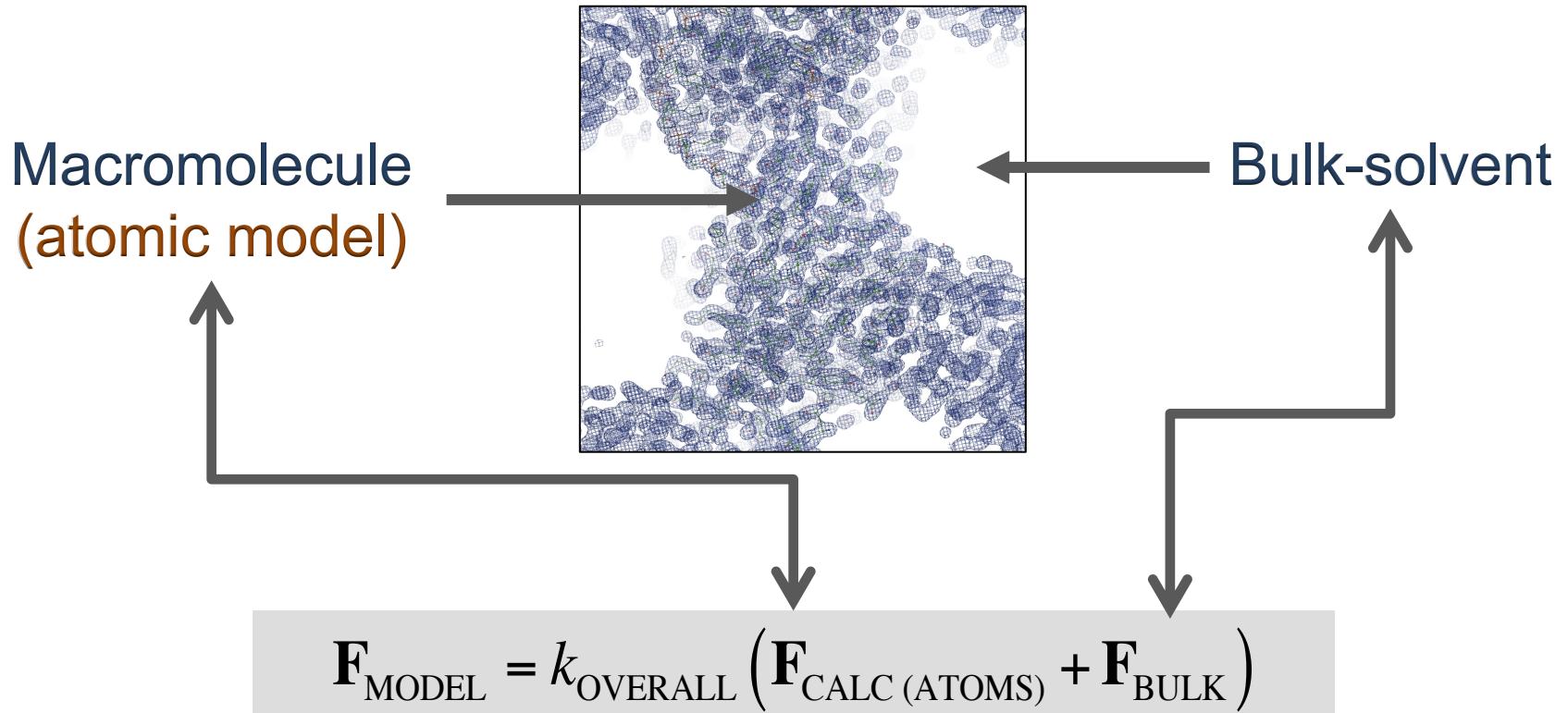


Macro-molecule

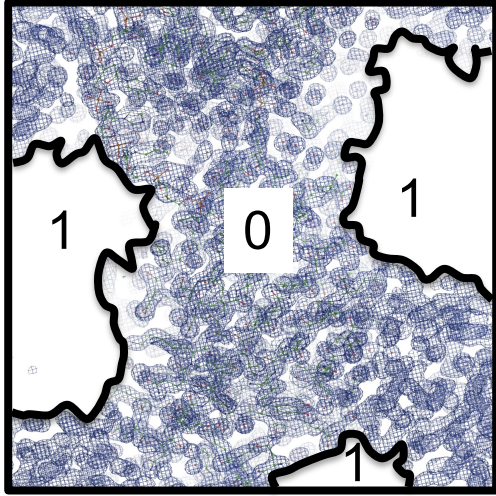
**Bulk-solvent:
~ 50% of unit cell
volume**

Crystal model: $\rho_{\text{crystal}} = \rho_{\text{atoms}} + \rho_{\text{bulk solvent}}$

Crystal structure model: structure factors



Bulk solvent: F_{BULK}



Steps to account for bulk-solvent:

1. Compute solvent mask, M:
0 – inside protein, 1 – outside

2. Structure factors from M:

$$F_{\text{MASK}} = \text{FT}(M)$$

3. Define solvent contribution F_{BULK} :

$$F_{\text{BULK}} = k_{\text{MASK}} * F_{\text{MASK}}$$

4. Combine with $F_{\text{CALC(ATOMS)}}$

Refine k_{MASK} by fitting $|F_{\text{MODEL}}|$ to F_{obs}

$$F_{\text{MODEL}} = k_{\text{OVERALL}} \left(F_{\text{CALC(ATOMS)}} + F_{\text{BULK}} \right)$$

Atomic model

Position

Larger-scale disorder

ATOM	25	CA	PRO A	4	31.309	29.489	26.044	1.00	57.79	C	
ANISOU	25	CA	PRO A	4	8443	7405	6110	2093	-24	-80	C

Local mobility (harmonic vibrations)

$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} \left(\mathbf{F}_{\text{CALC (ATOMS)}} + \mathbf{F}_{\text{BULK}} \right)$$

Occupancy 1.00

57.79 ADP (B-factor)

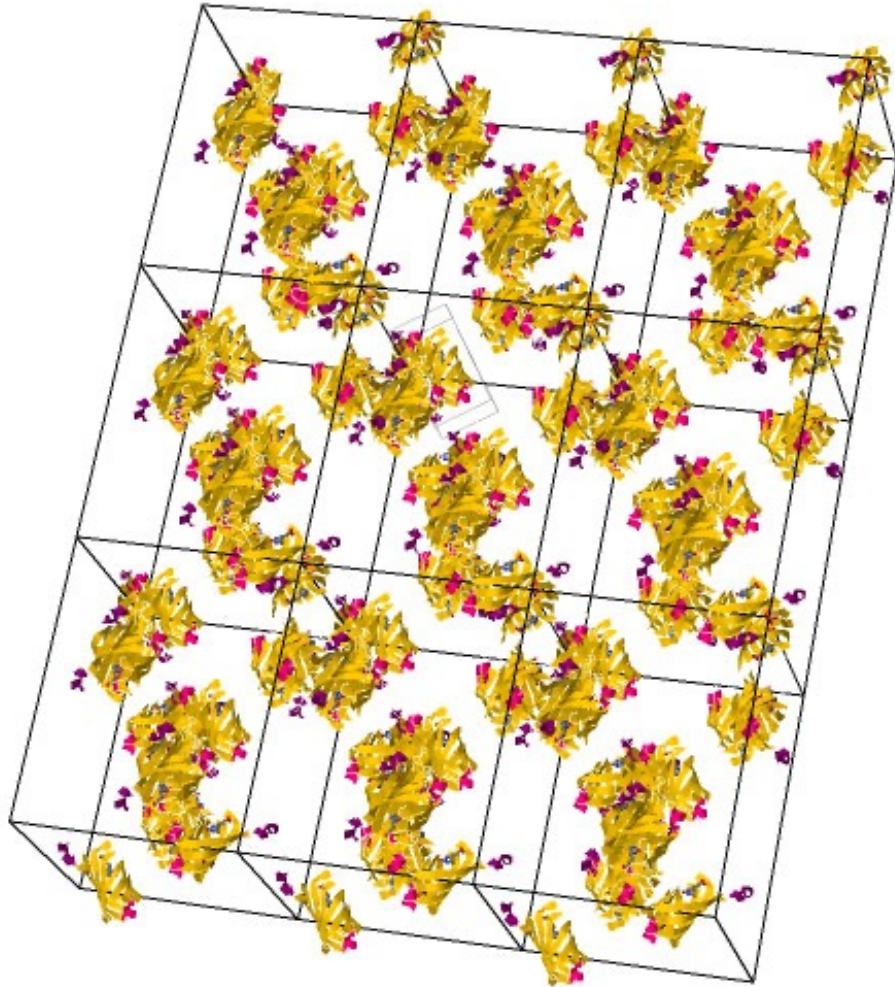
$$\mathbf{F}_{\text{CALC (ATOMS)}}(h, k, l) = \sum_{n=1}^{N_{\text{atoms}}} q_n f_n(s) \exp\left(-\frac{B_n s^2}{4}\right) \exp(2i\pi \mathbf{r}_n \cdot \mathbf{s})$$

Atom type C

31.309 29.489 26.044

Atomic coordinates

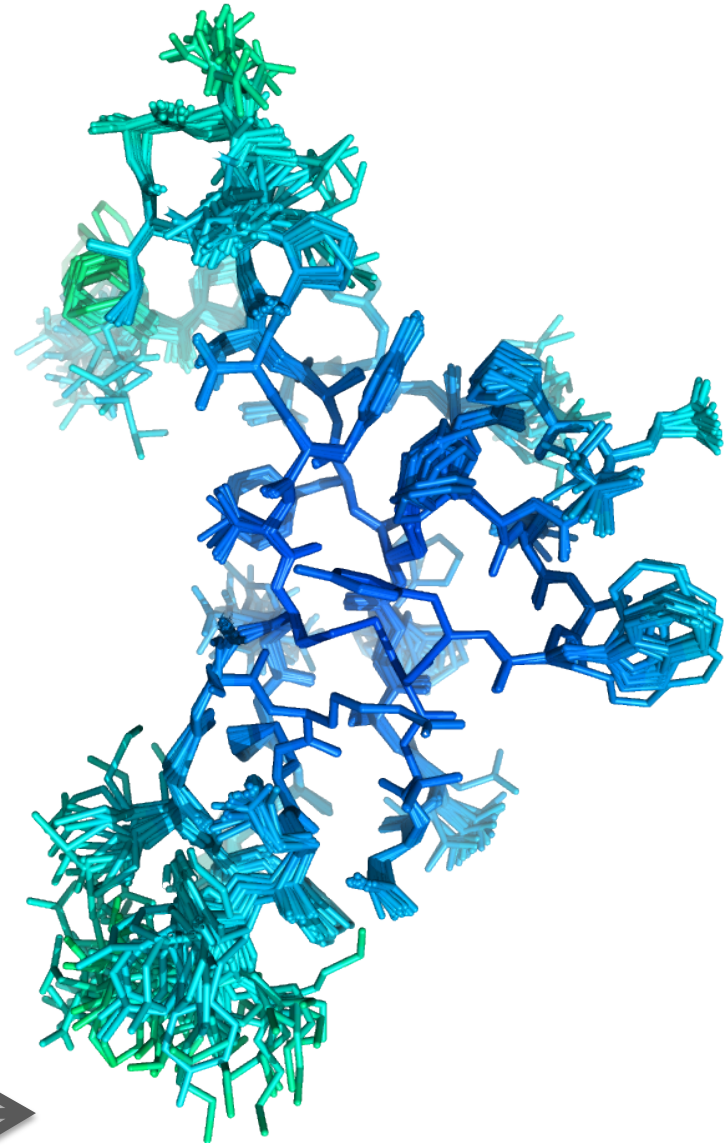
Atomic model: disorder



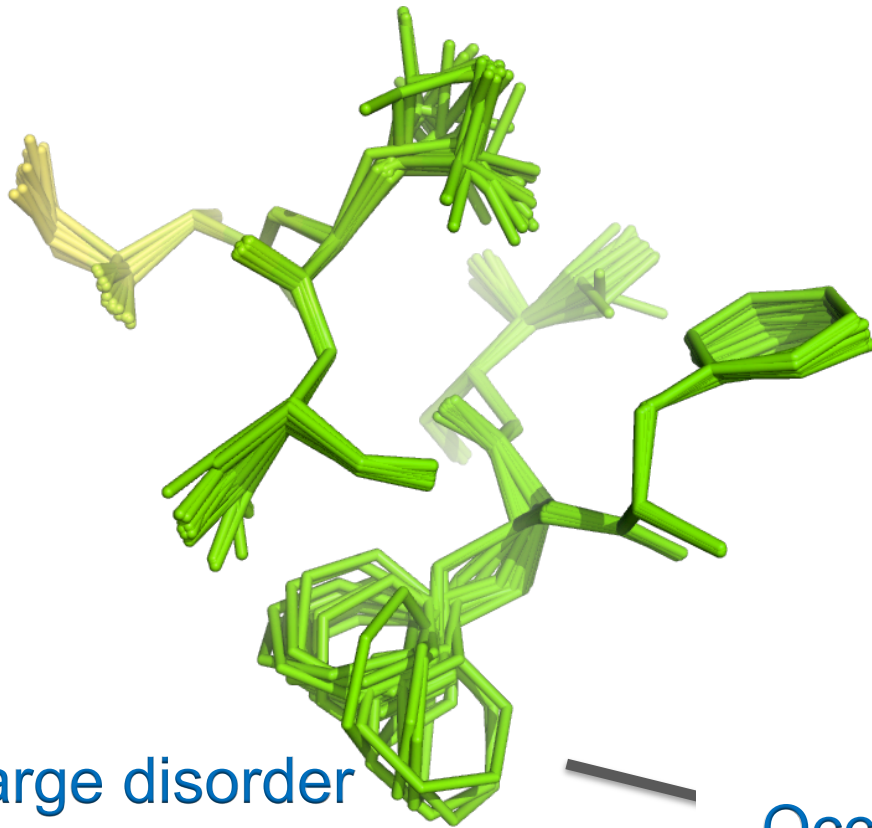
Crystal = many unit cells



Superpose all structures
from each unit cell



Atomic model: disorder



Small disorder

ADP (B-factor)

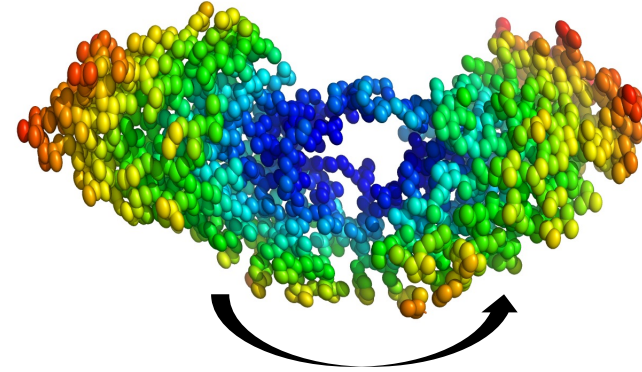
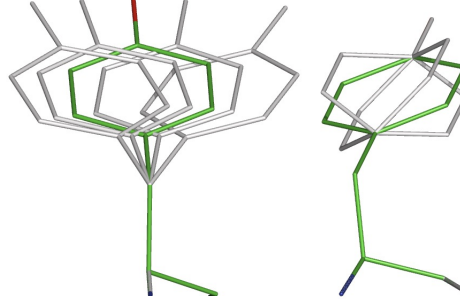
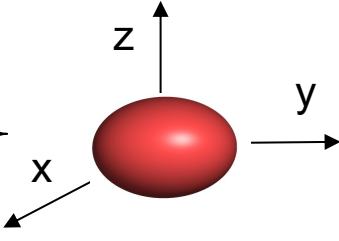
Large disorder

Occupancy

ATOM	25	CA	PRO	A	4	31.309	29.489	26.044	1.00	57.79	C	
ANISOU	25	CA	PRO	A	4	8443	7405	6110	2093	-24	-80	C

Atomic Displacement Parameters (ADP, B-factors)

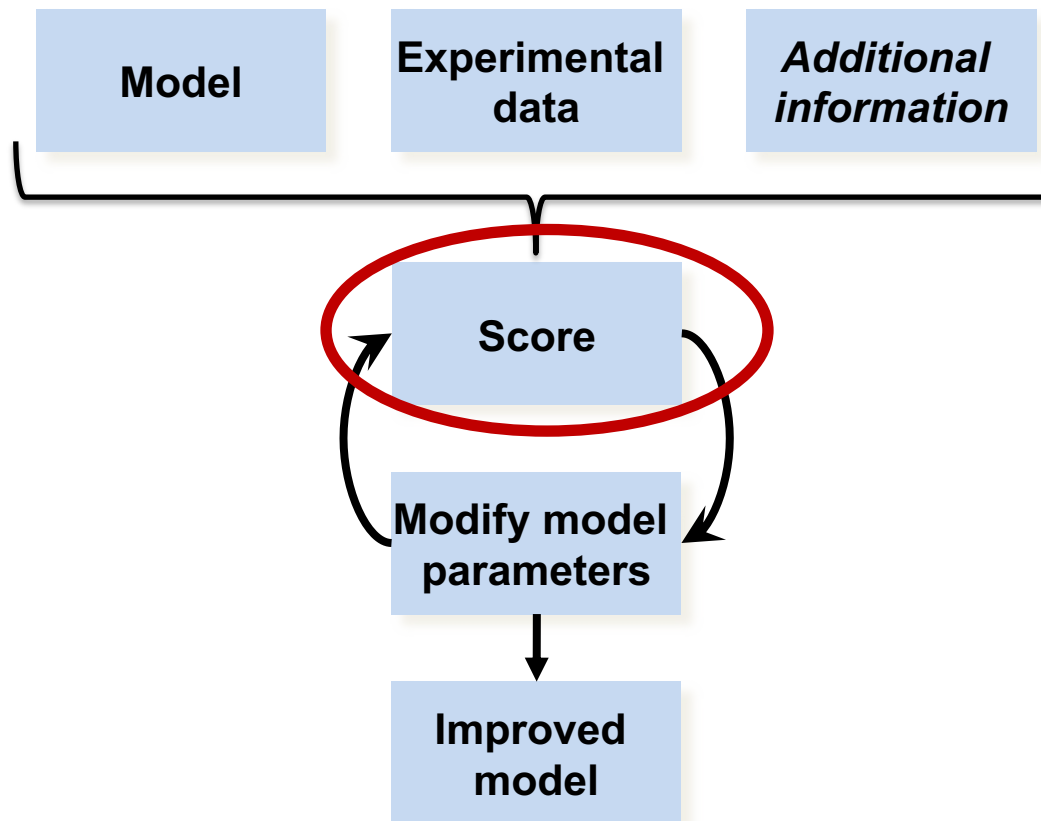
atom
residue
domain
molecule
crystal



TLS

$B_{\text{TOTAL}} = \text{sum of individual contributions}$

Refinement target function (score)



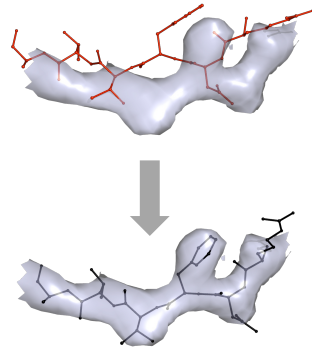
Model refinement

T

Consensus
between model-
to-data fit and
extra information
about the model

=

T_{DATA}



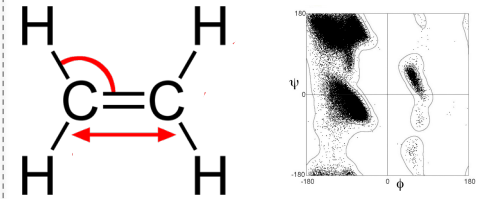
+

W

Dose of extra
information

*

T_{RESTRAINTS}



$$\sum_{hkl} (F_{obs} - F_{model})^2$$

$$\sum_{hkl} \frac{||F_{obs}| - |F_{model}||}{|F_{obs}|}$$

Maximum-Likelihood

T_{BOND} +

T_{ANGLE} +

T_{DIHEDRAL} +

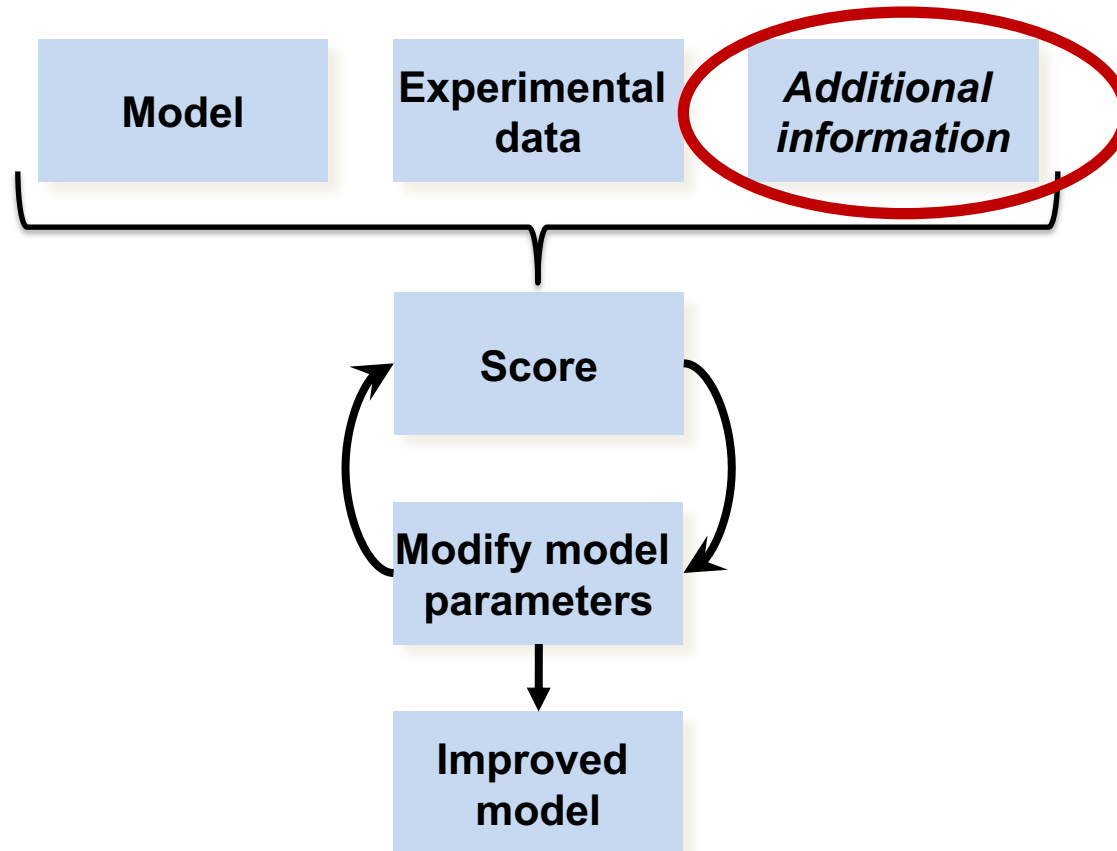
T_{PLANE} +

T_{REPULSION} +

T_{CHIRALITY} +

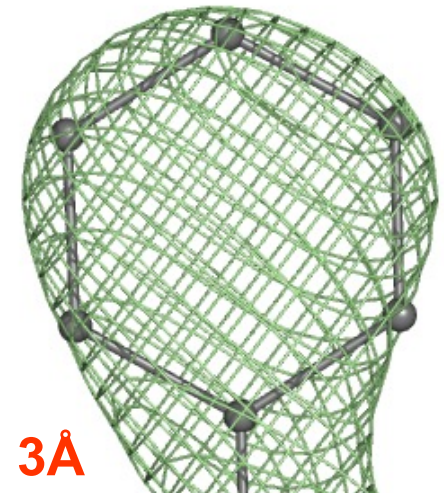
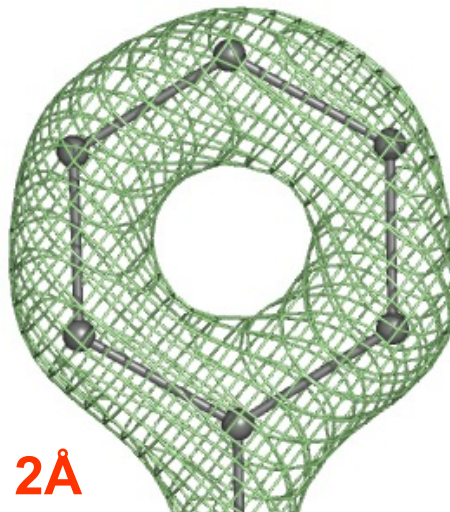
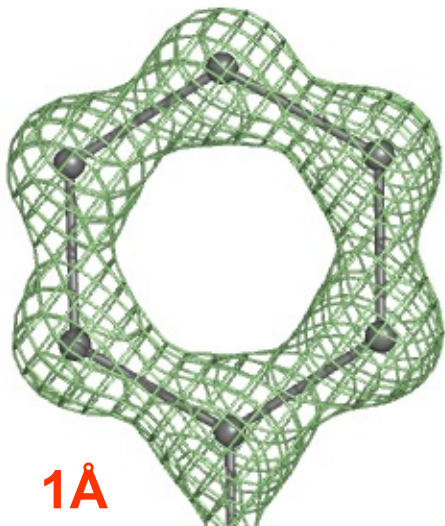
...

Additional information (restraints, constraints)



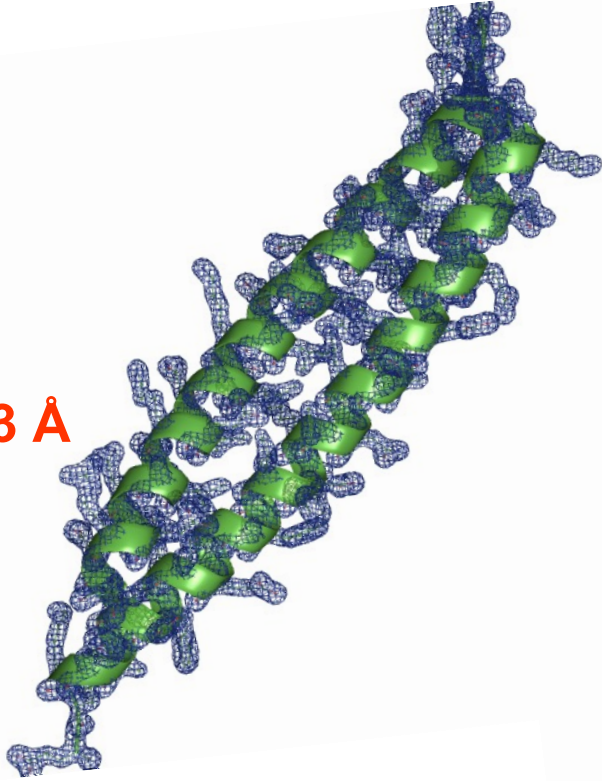
Restraints and constraints

- Why?
 - Experimental data are not perfect:
 - Finite resolution
 - Contains errors
 - Typically less than model parameters (overfitting)
 - Phases are approximate
- Effect of resolution

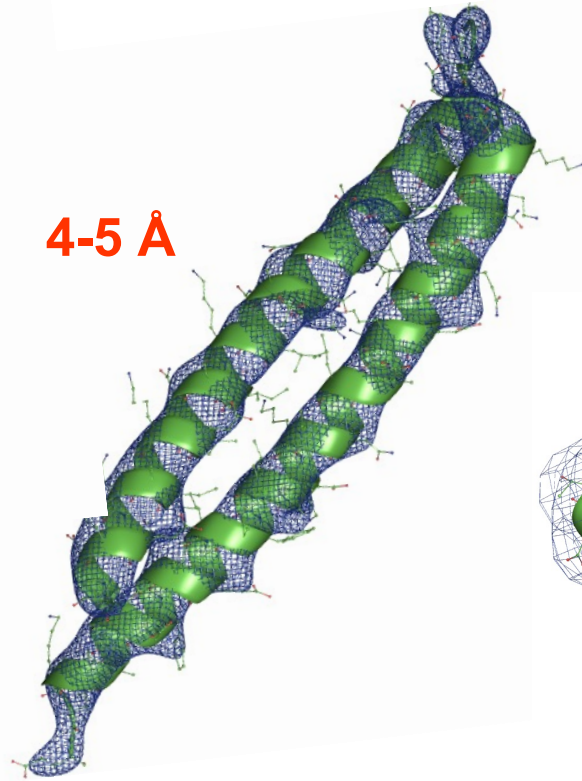


Restraints and constraints

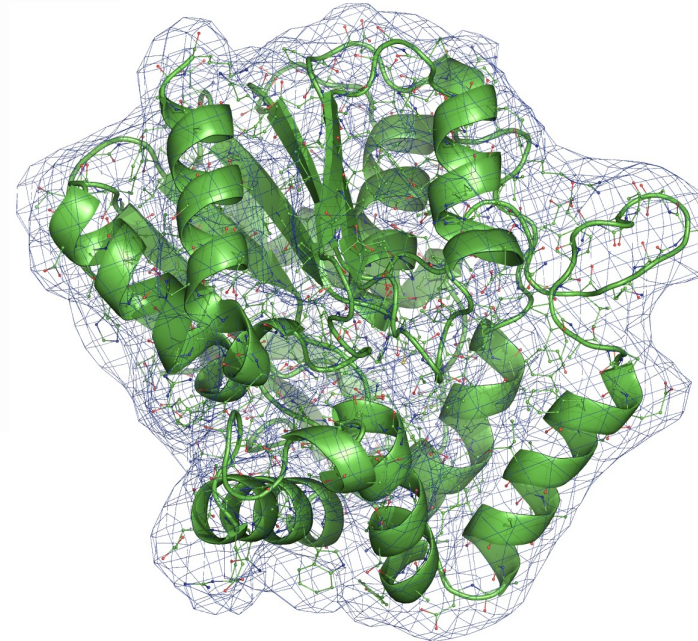
2-3 Å



4-5 Å

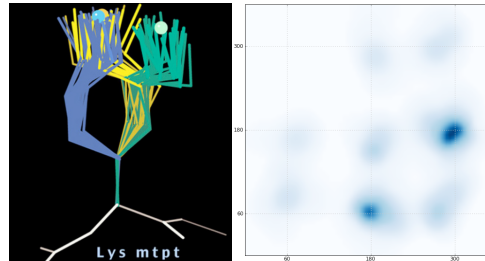


6Å-lower

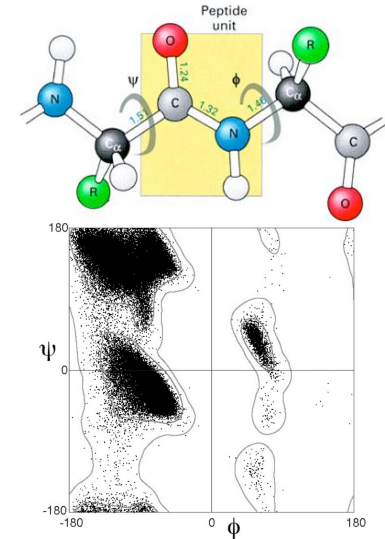


More restraints for low resolution

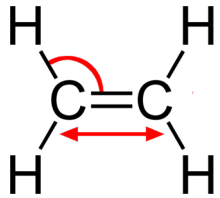
Side chain distributions



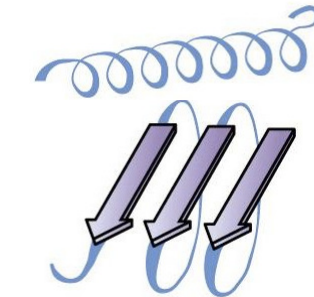
Main chain distributions



Covalent geometry

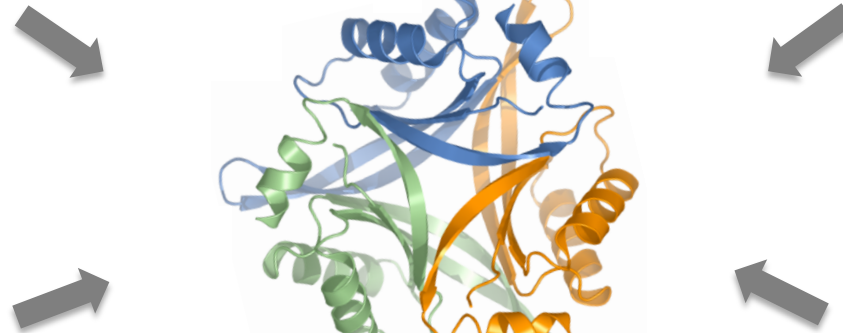


Internal symmetry (NCS)



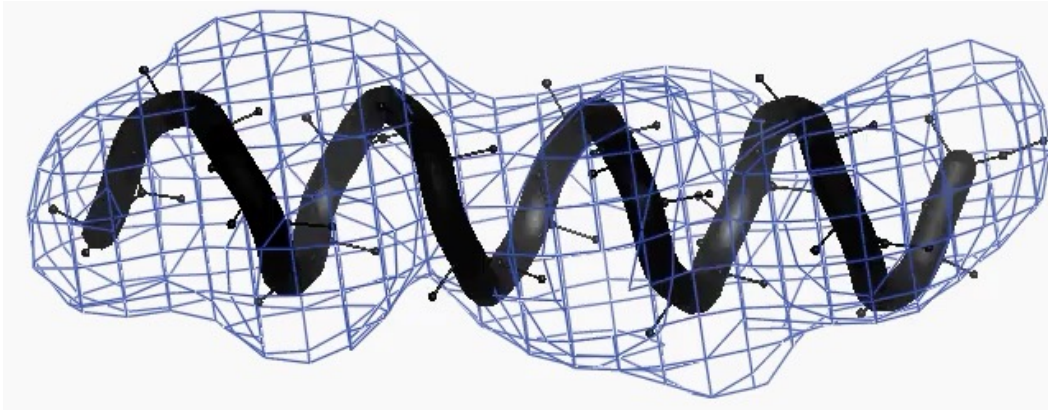
Secondary structure

Similar (homologous) structures (reference model restraints)

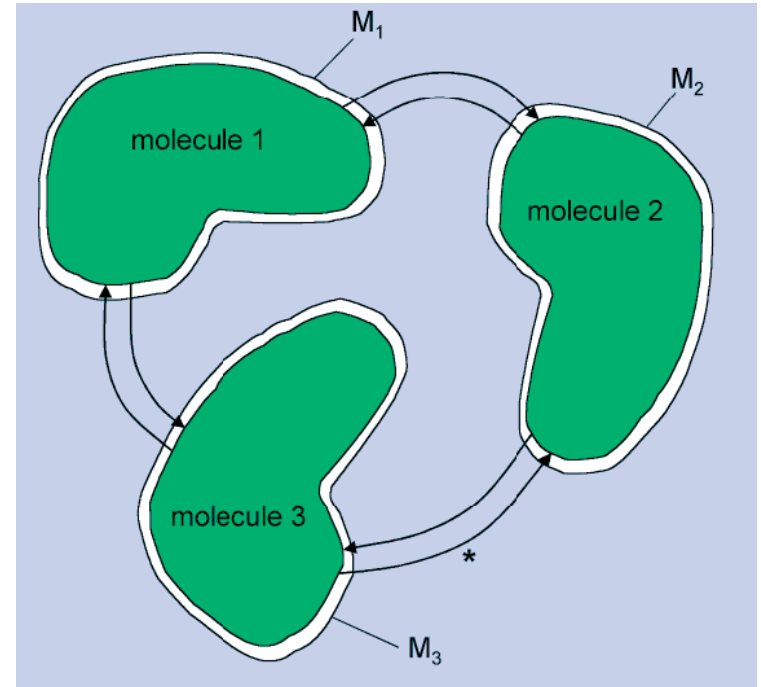
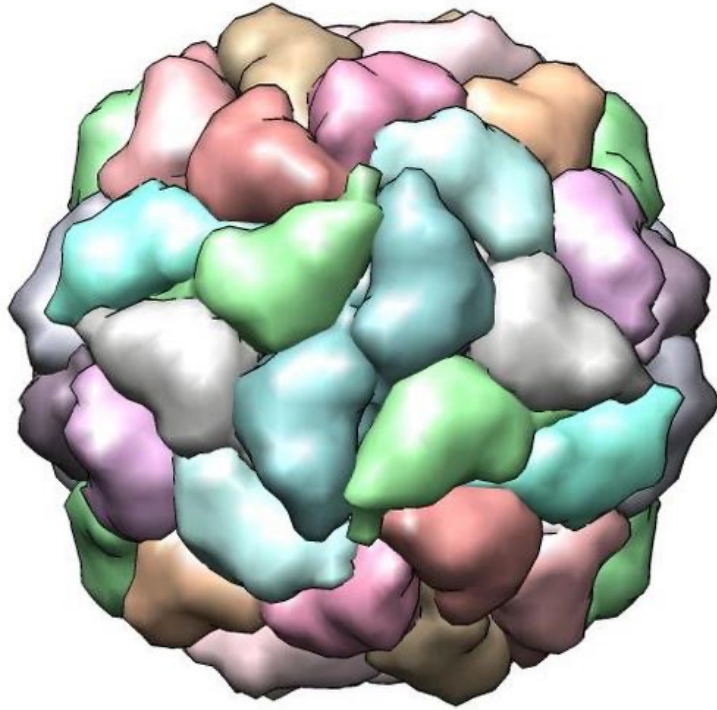


Importance of more restraints at low resolution

- Toy example: refinement of a perfect α -helix into low-res map
 - Standard restraints on covalent geometry isn't sufficient
 - Model geometry deteriorates as result of refinement



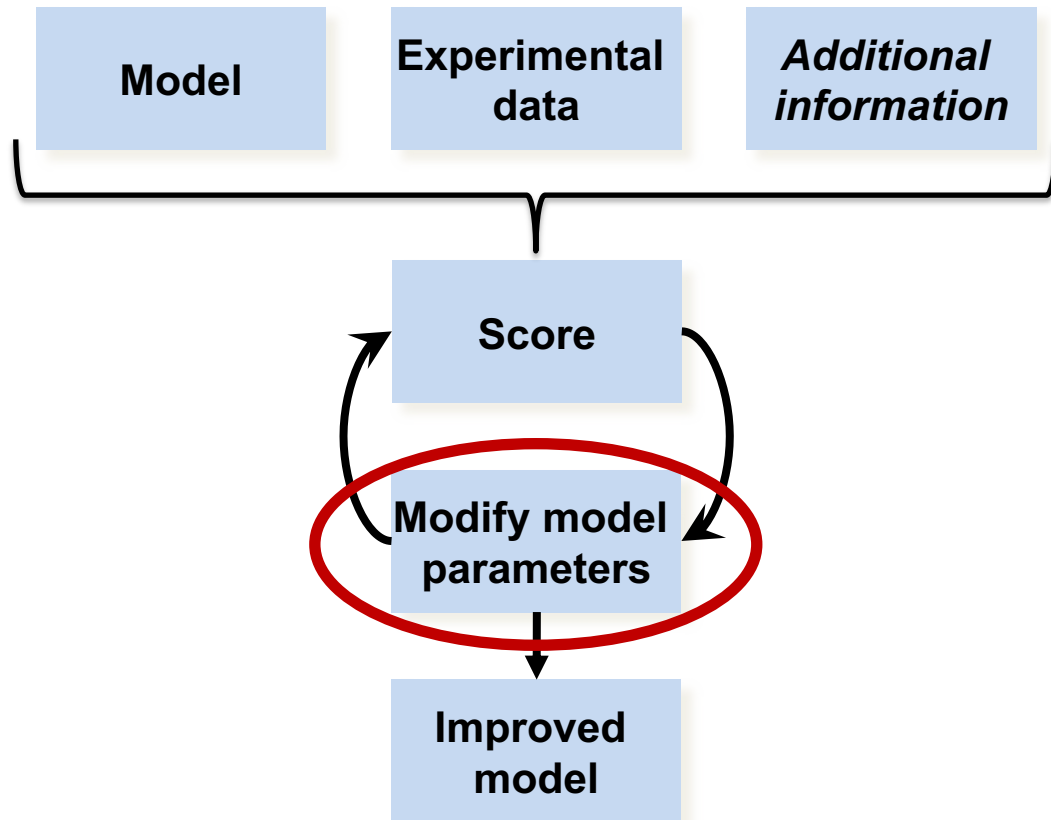
NCS (internal symmetry): constraints vs restraints



Source: Internet

- **Constraints:** molecules 1, 2 and 3 are required to be **identical**
- **Restraints:** molecules 1, 2 and 3 are required to be **similar** but not necessarily identical

Refinement

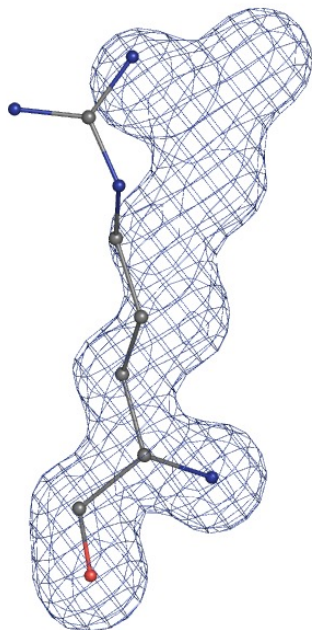


Choices of optimization method

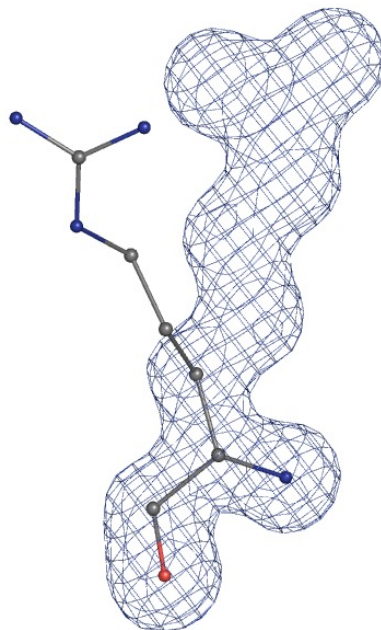
- Gradient-based minimization
- Simulated annealing
- Grid (systematic) searches
- Manual using molecular graphics programs (Coot, Chimera,...

Choice of refinement method and refinement convergence

Minimization

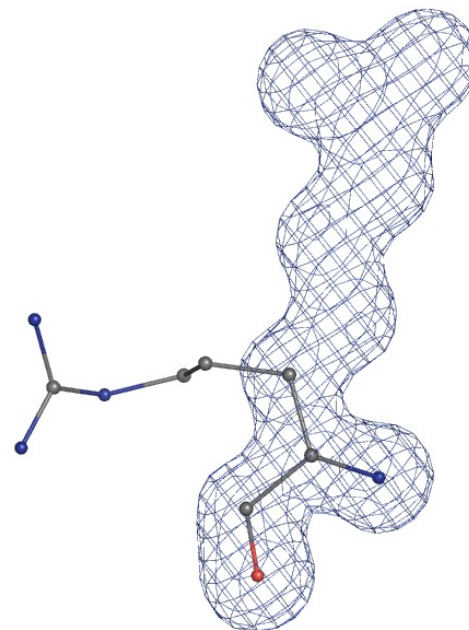


Simulated Annealing



**Beyond
convergence radius
of minimization**

Real-space grid search

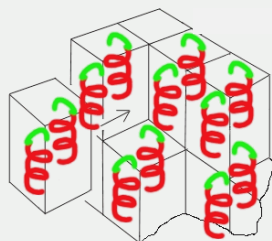


**Beyond convergence
radius of
minimization and SA**

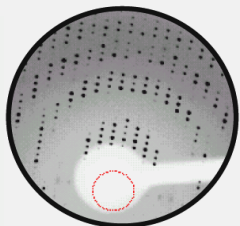
Phenix tools for model refinement

Refinement

Crystallography



Initial model



Experimental data

A priori knowledge

Score

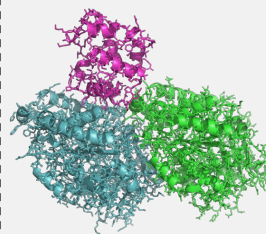
Modify model parameters

Improved model

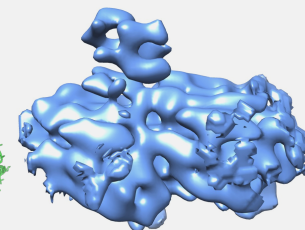
phenix.refine

Available since 2005

Cryo-EM



Initial model



Experimental data

A priori knowledge

Score

Modify model parameters

Improved model

phenix.real_space_refine

Available since 2013

Atomic model refinement: crystallography vs cryo-EM

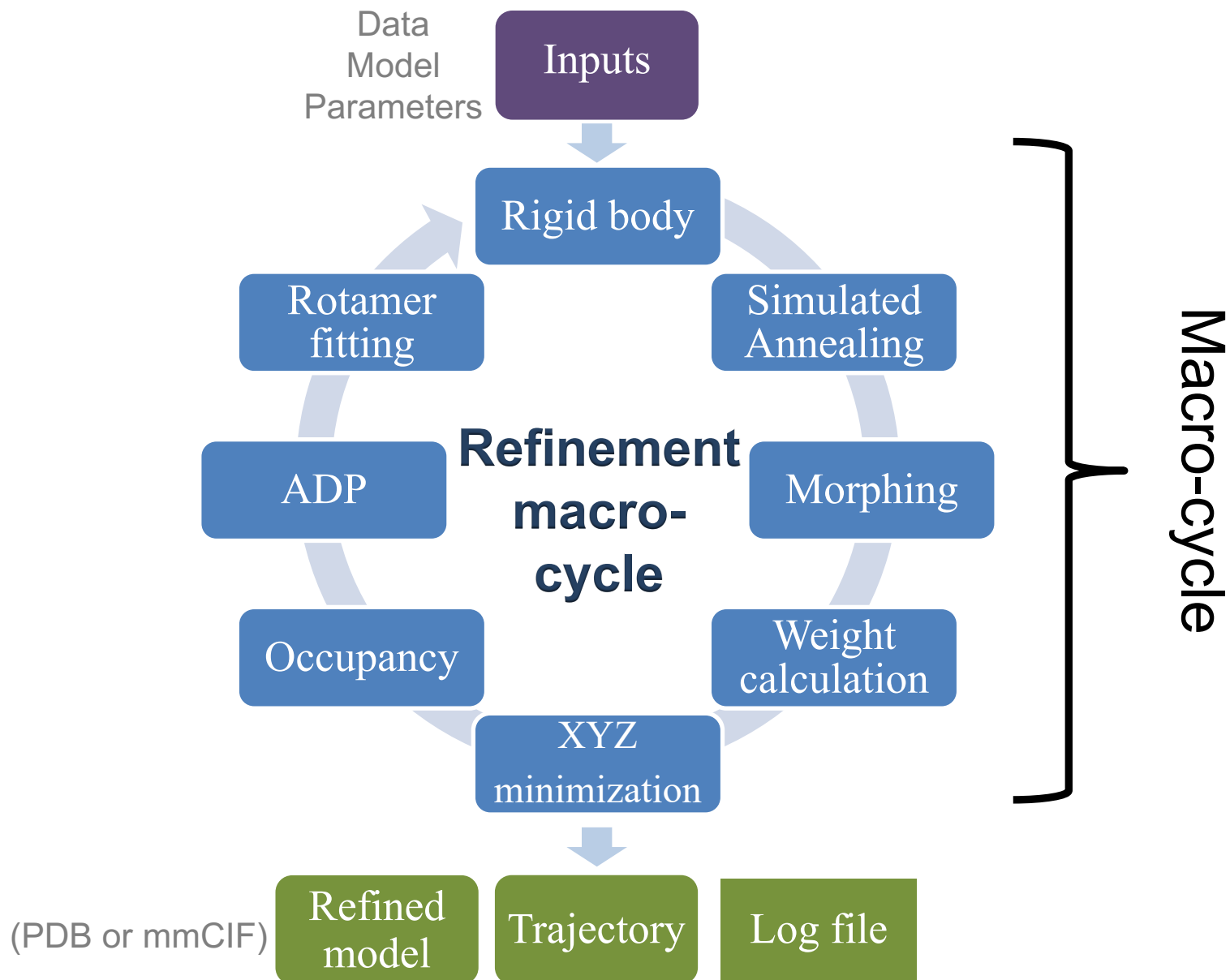
Crystallographic refinement

- Improving model improves map
 - (2mFo-DFc, Model phase), (mFo-DFc, Model phase)
 - Better model leads to better map
 - Better map leads to more model built
 - Improving model in one place lets build more model elsewhere in the unit cell
 - Refine all model parameters (XYZ, B) from start to end of structure solution
 - Build solvent (ordered water) early
- Experimental data never changed
- Data / restraints weight is global and time expensive to find best value
- Whole model needs to be refined

Cryo-EM refinement

- Changing model does not change map
 - Build solvent (water) last
 - Get as complete and accurate model as possible before refining B factors and occupancies
- Experimental data changes a lot during the process (filtering, boxing, using maps with implied symmetry or not, etc.)
 - What map to use in refinement?
 - Refined B factors depend on map used
- Data / restraints weight can be local and is always optimal
- Boxed parts of the model can be refined

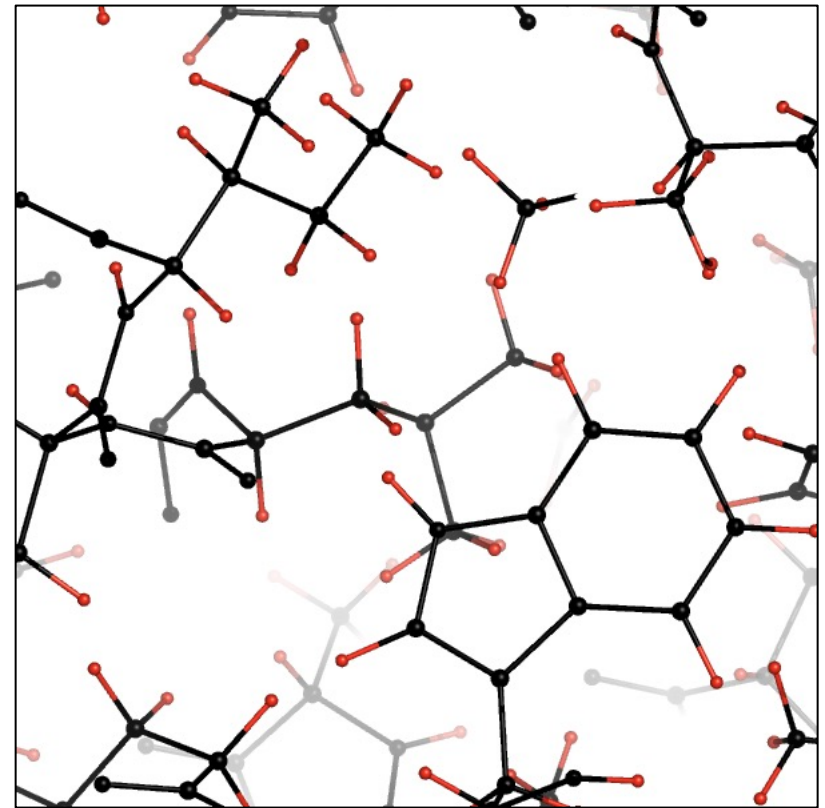
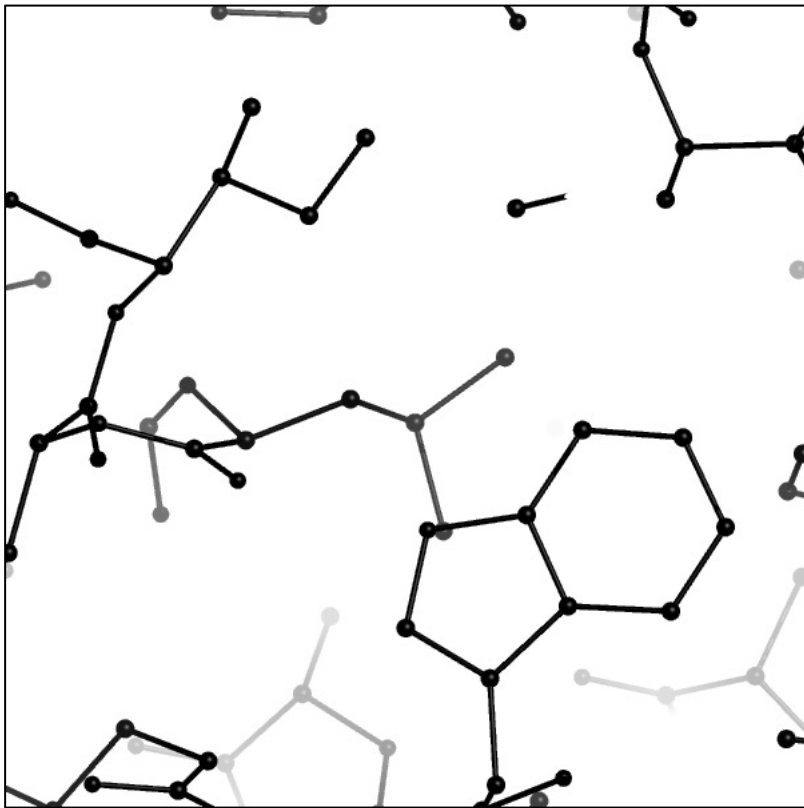
Refinement protocol



Refinement: practical considerations

Use Hydrogen atoms

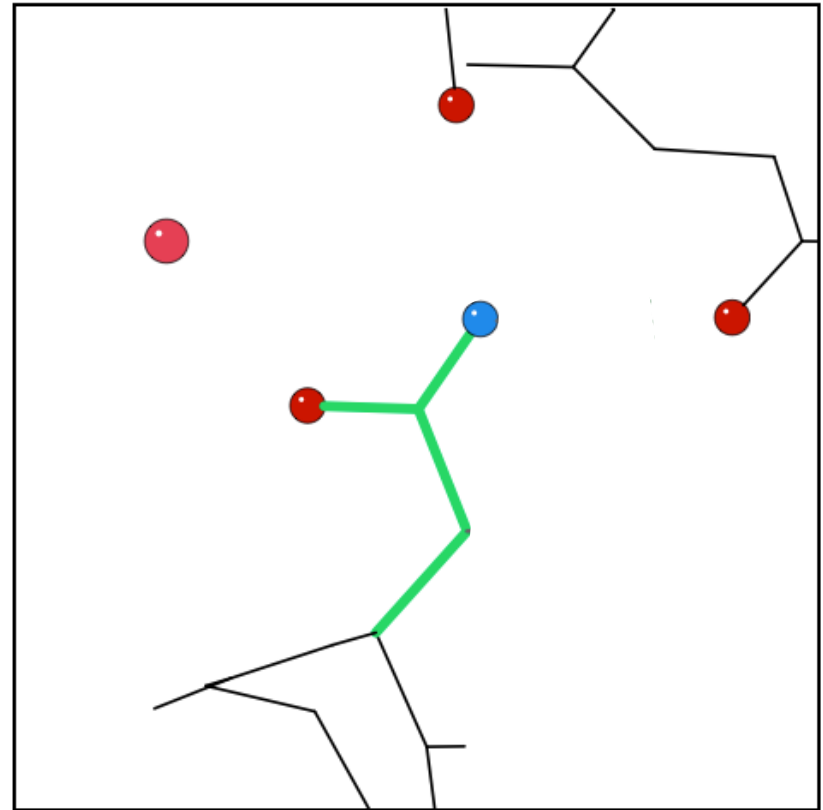
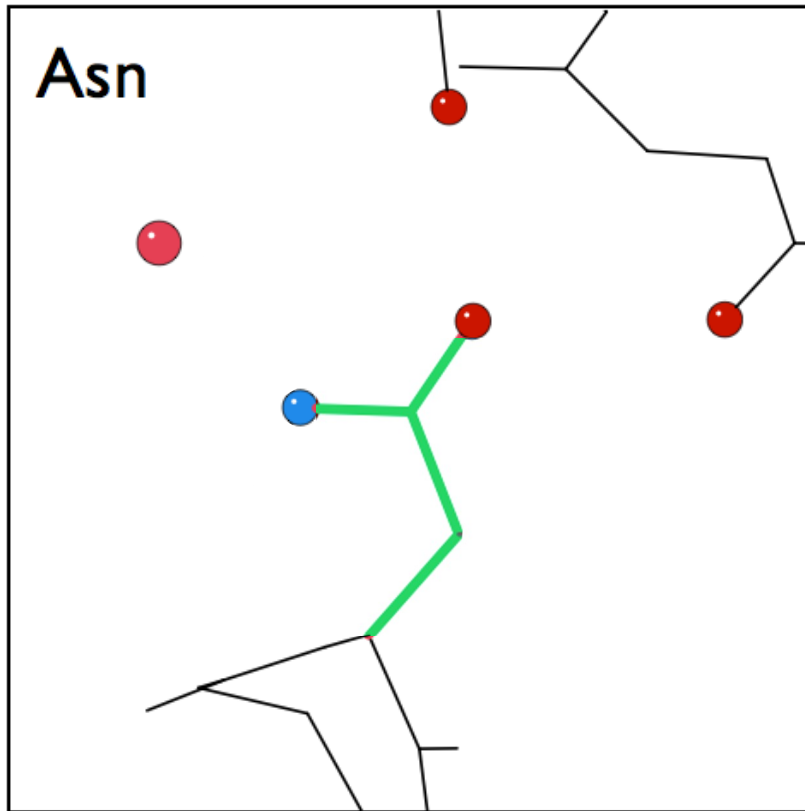
- Half of the atoms in a protein molecule
- Make most interatomic contacts
- Add to model towards the end, data resolution does not matter
- Once added, do not remove before the PDB deposition
- H do contribute to R-factors (expect 0.1-2% drop in R)



A structure without (left) and with (right) hydrogen atoms

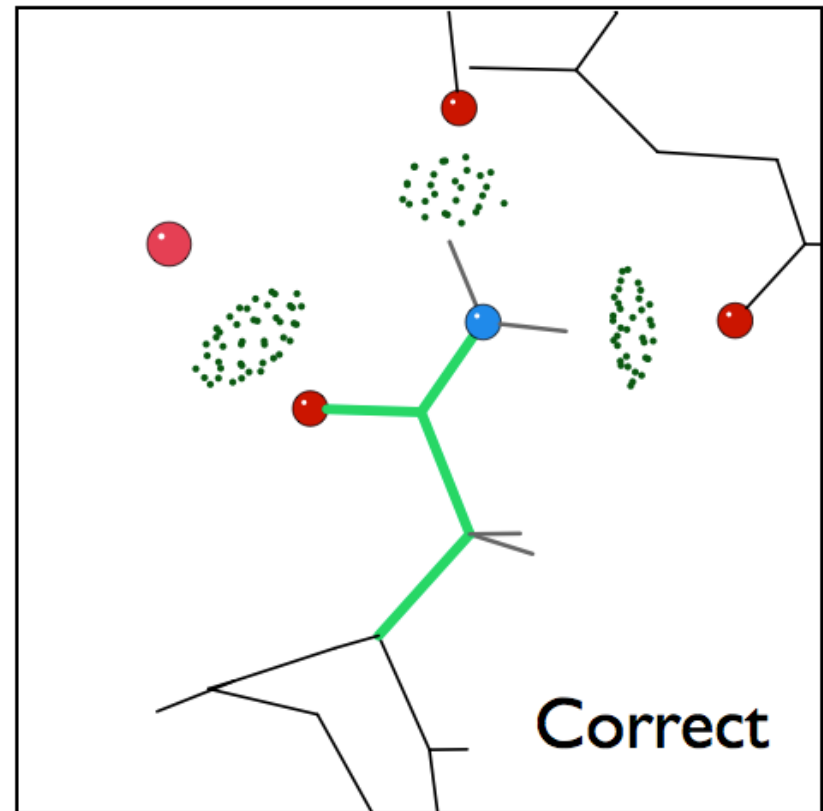
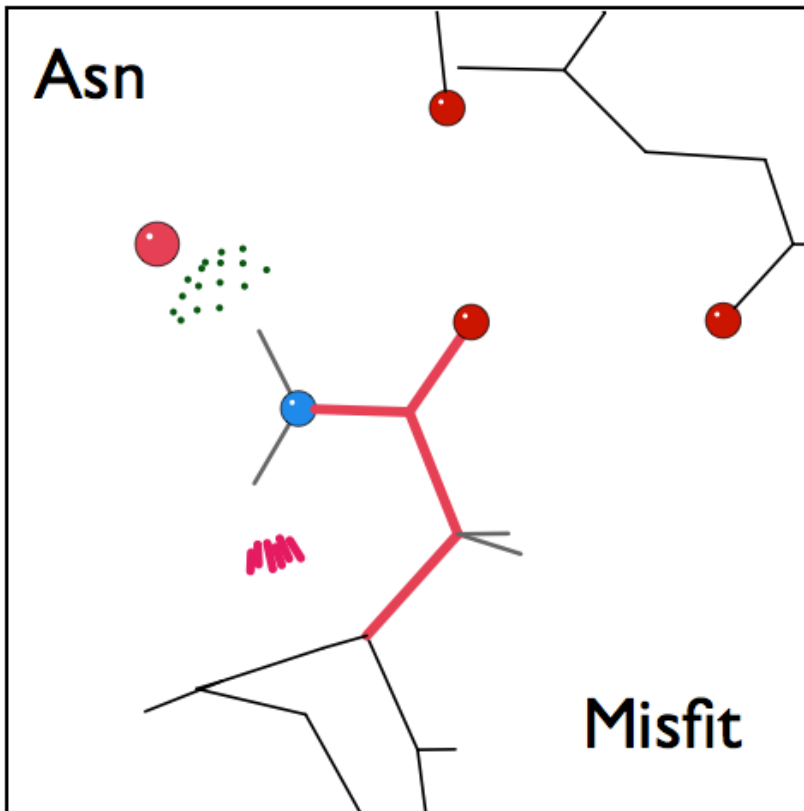
Use Hydrogen atoms

- N/Q/H flips (asparagine/glutamine/histidine)
 - Based on clash analysis
 - Requires H present

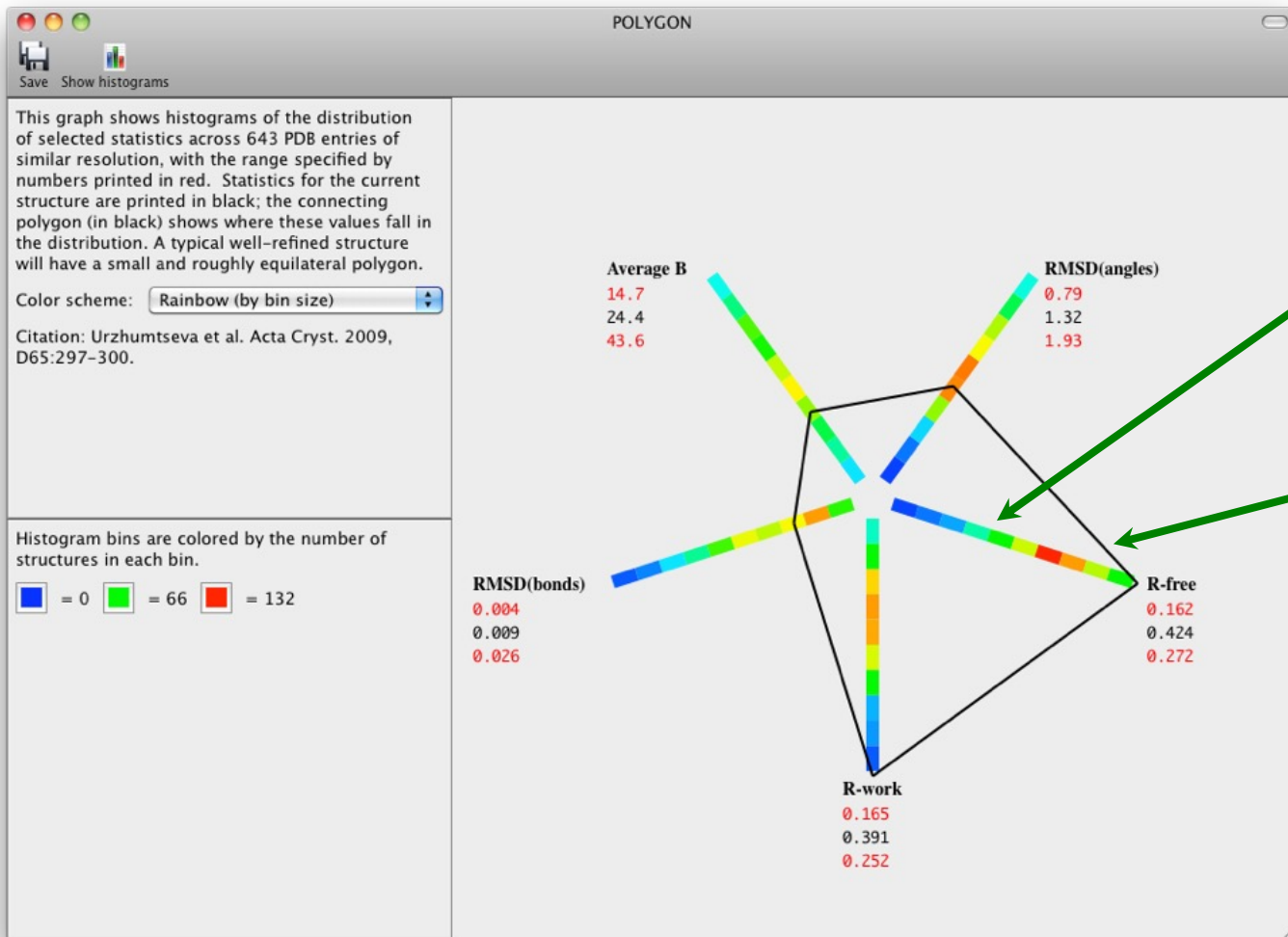


Use Hydrogen atoms

- N/Q/H flips
 - Based on clash analysis
 - Requires H present



Know when to stop



Colored bars are histograms showing distribution of values for structures at similar resolution

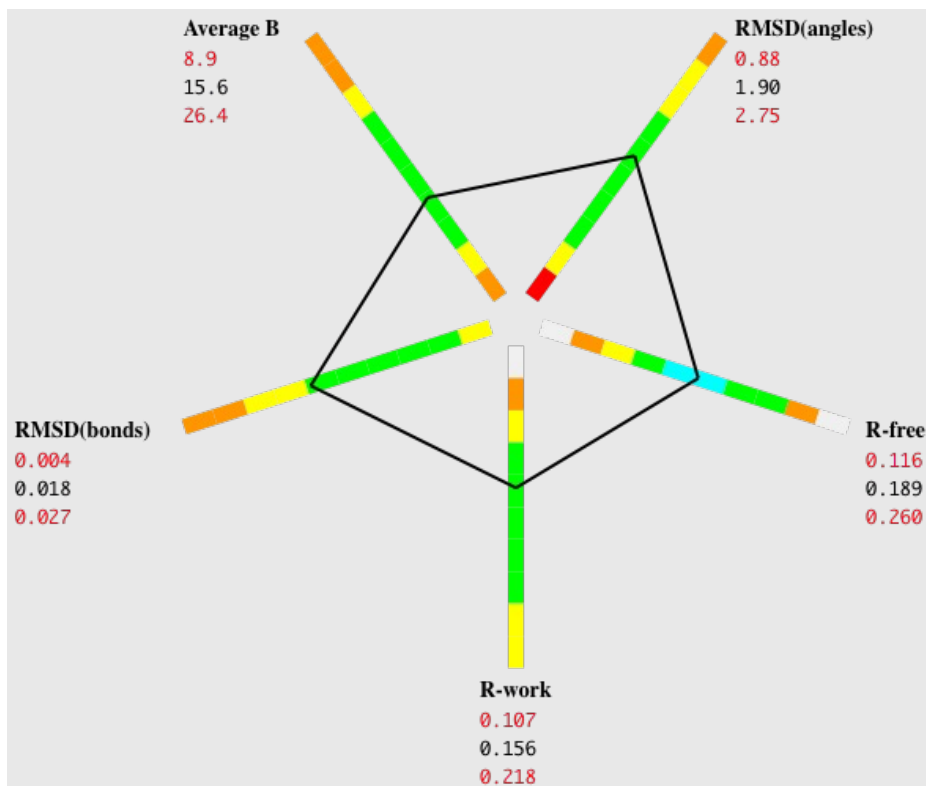
The black polygon shows where the statistics for the user's structure fall in each histogram

Crystallographic model quality at a glance.

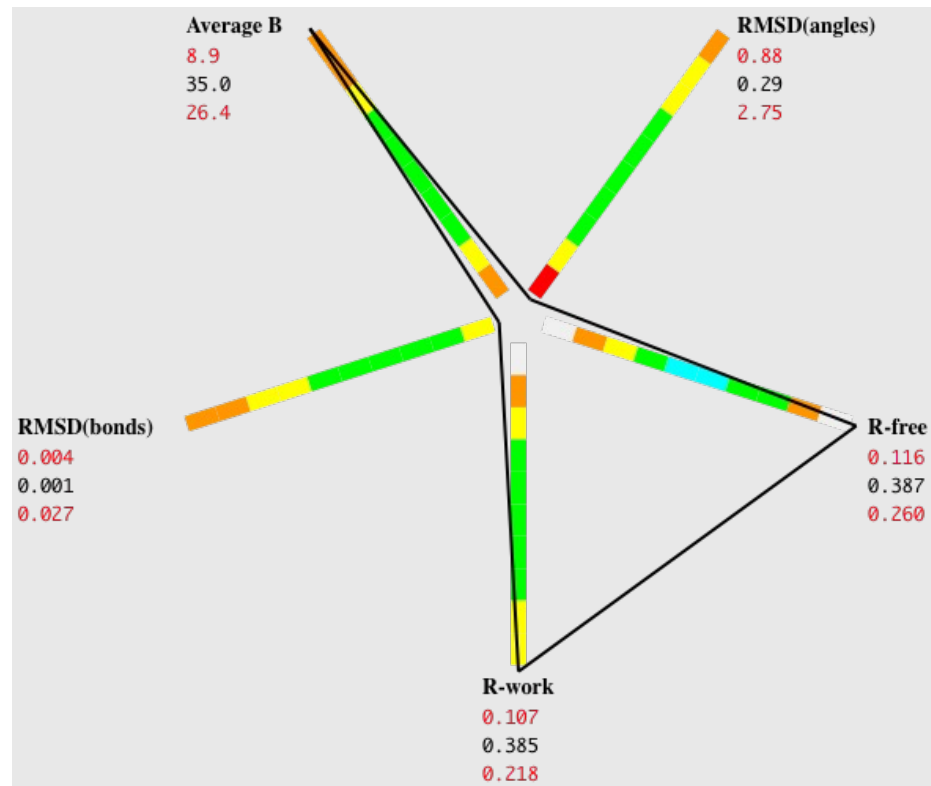
L.Urzhumtseva, P.V.Afonine, P.D.Adams & A.Urzhumtsev. *Acta Cryst.* D65, 297-300 (2009)

Know when to stop

Likely overall good model



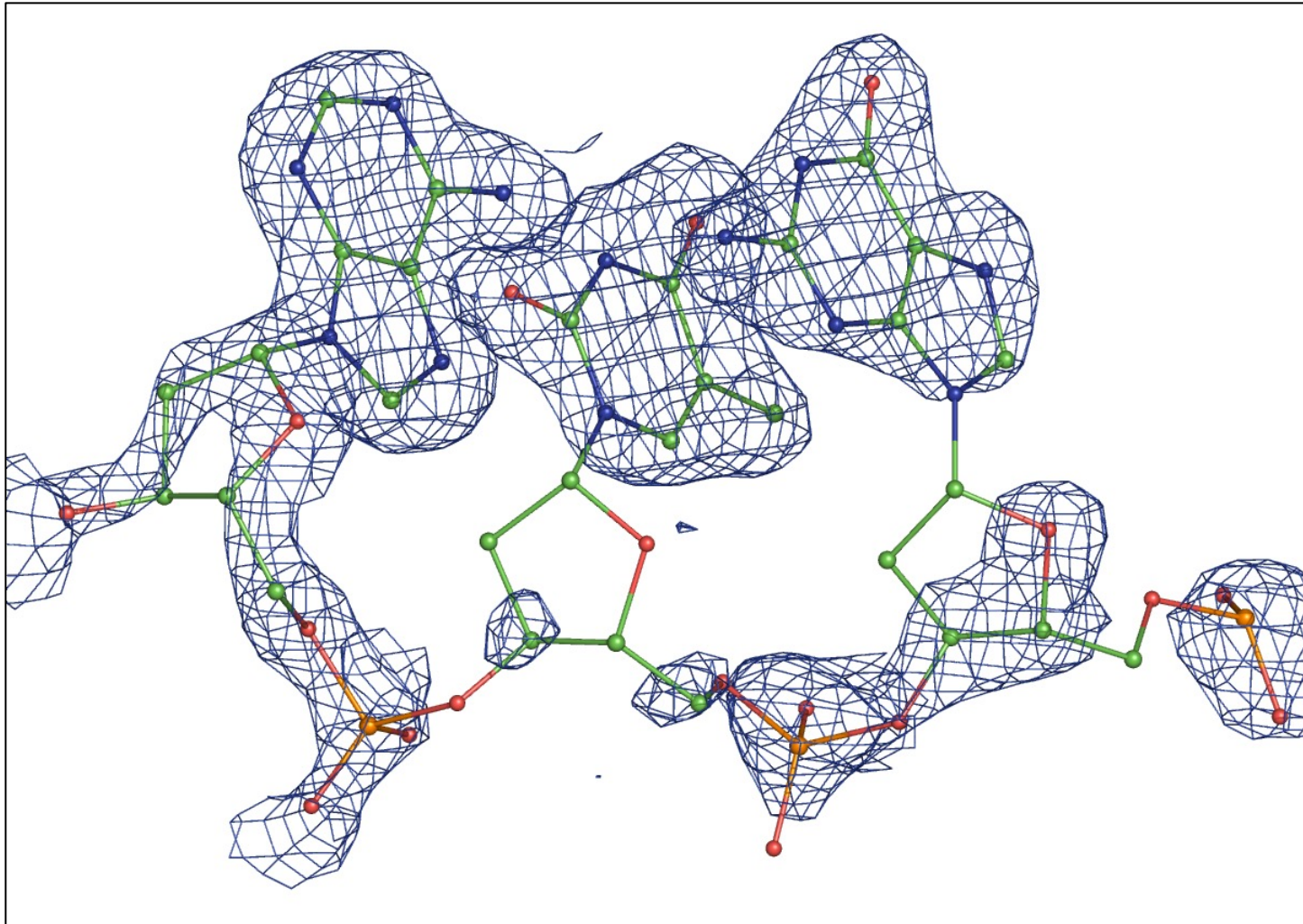
Clearly there are problems



Don't waste time fixing unfixable

PDB code: 1NH2, resolution 1.9Å, showing E6-E8

2mFo-DFc , 1σ



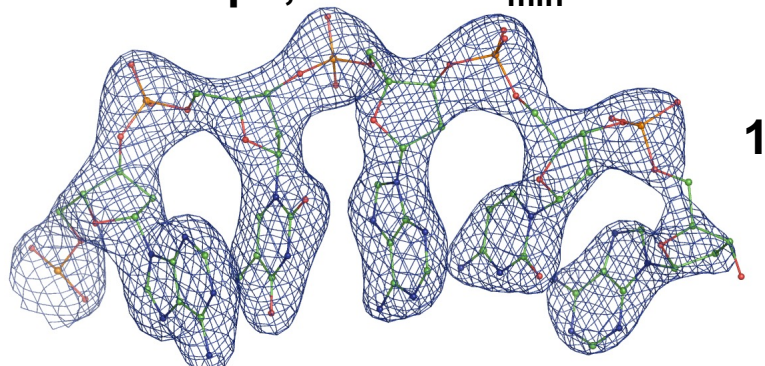
Don't waste time fixing unfixable

Completeness by resolution:

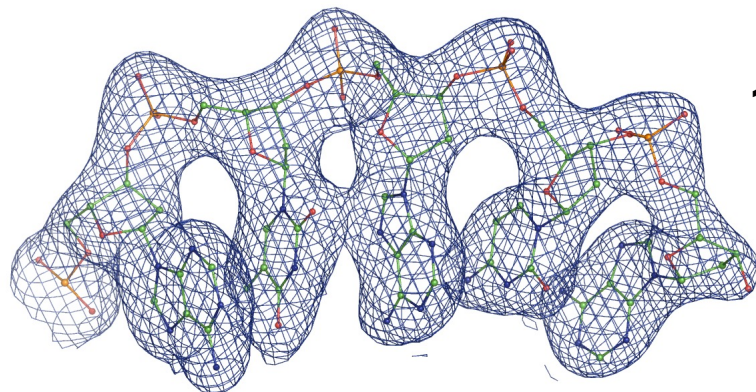
19.9274	-	3.2441	0.78
3.2441	-	2.5767	0.99
2.5767	-	2.2515	1.00
2.2515	-	2.0459	1.00
2.0459	-	1.8993	0.99

Overall completeness in d_{\min} -inf: 0.95

Fcalc maps, full set d_{\min} -inf

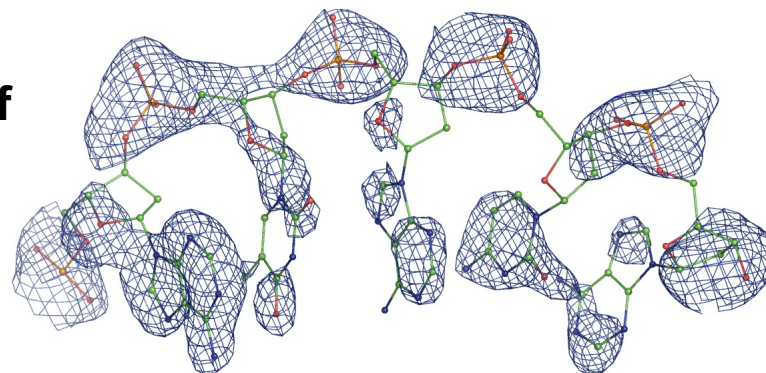
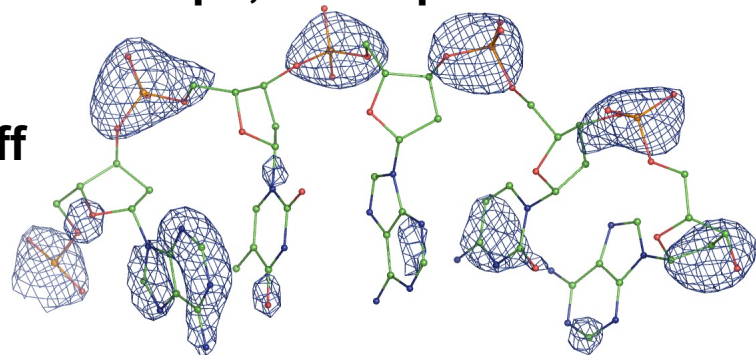


1.5 σ map cutoff



1 σ map cutoff

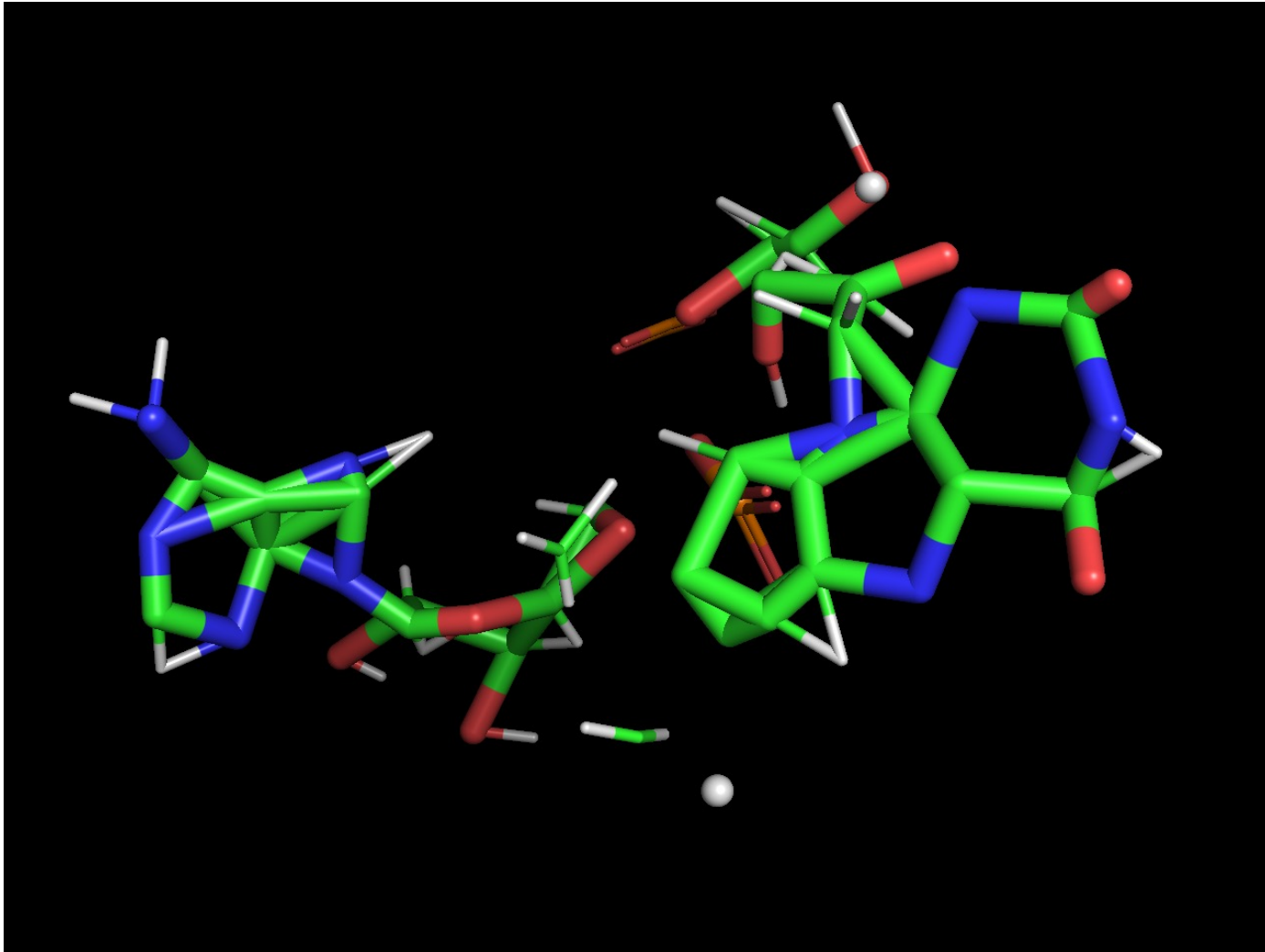
Fcalc maps, incomplete set



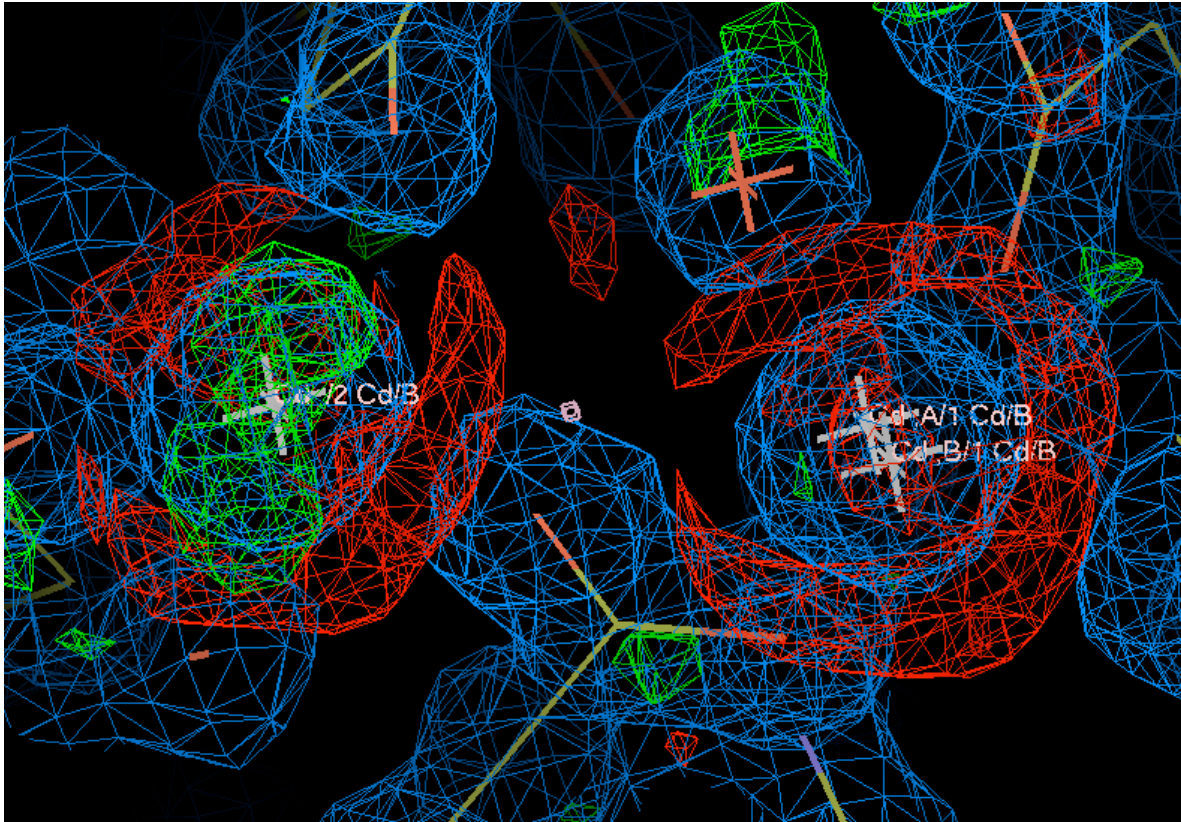
Data incompleteness distorts maps

Local vs Global

- $R_{\text{WORK}}/R_{\text{FREE}}$, bond/angle RMSDs etc do not report on local errors



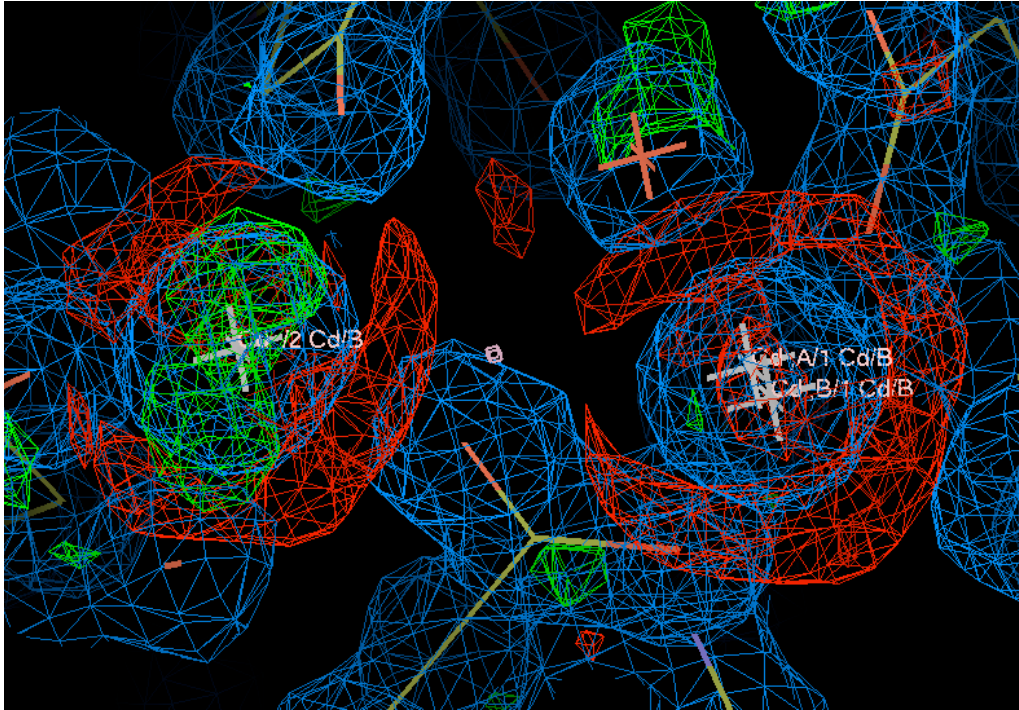
Map and model errors



Reasons for +ve/-ve density:

- Suboptimal xyz, occupancy, ADP, anomalous f' & f'' , charge
- Refinement has not reached convergence
- Wrong atom (ion)
- Suboptimal ADP (B-factor) type: isotropic vs anisotropic
- **NEW** phenix.oat is the new tool to help with this

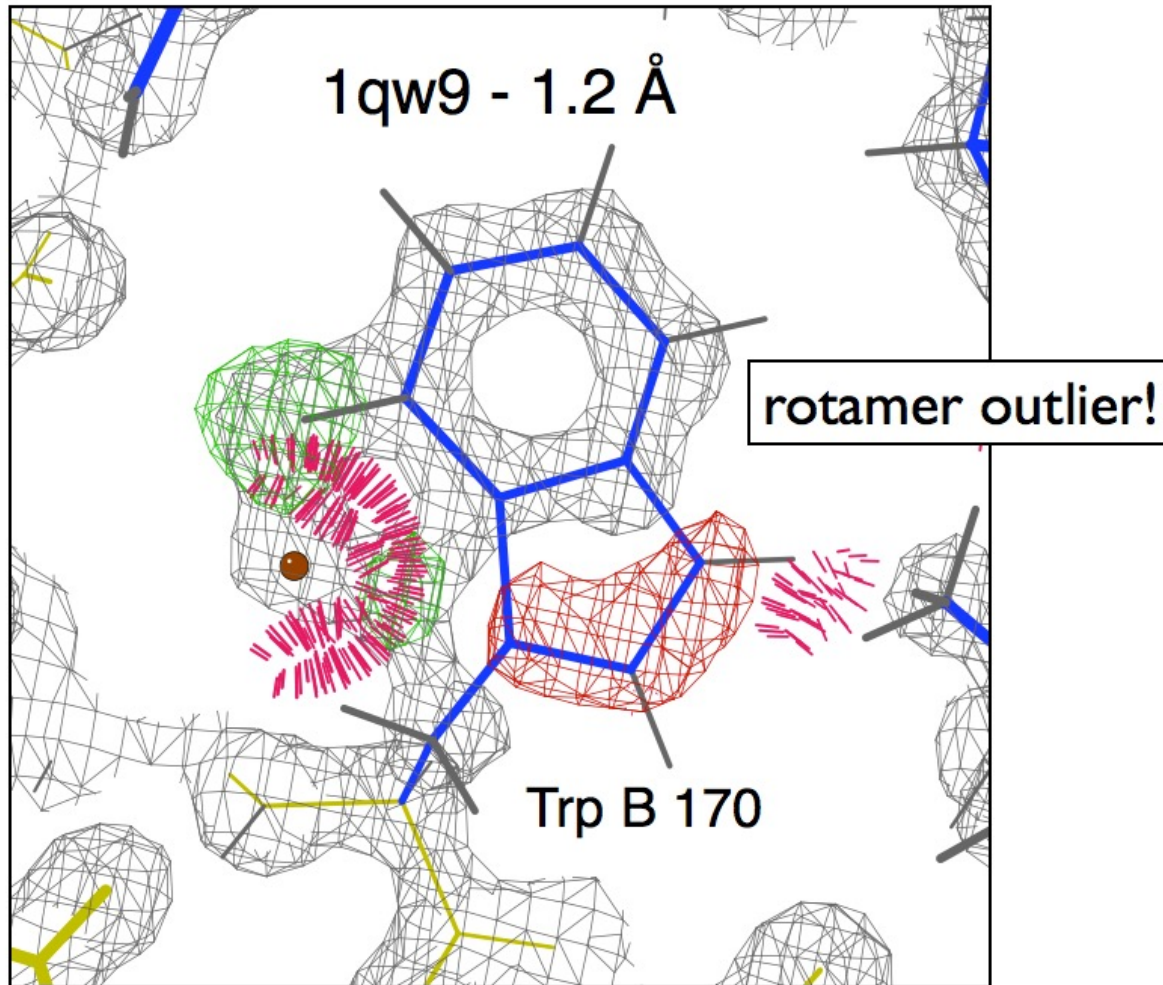
Map and model errors



NEW: phenix.oat : try all possibilities, one atom at a time

```
phenix.oat model.pdb data.mtz selection="chain A and resseq 123  
and name CD"
```

Not all modeling errors can be fixed by refinement



- Sadly, manual validation is still required

Low resolution (3Å or worse)

- Use:
 - Ramachandran plot restraints
 - Secondary structure restraints
 - Reference model restraints (if quality homology model is available)
 - NCS (restraints or constraints)

NCS (Non-crystallographic symmetry)

- Constraints vs restraints
 - Constraints:
 - 4-5 Å or worse
 - Highly symmetric molecules
 - Restraints:
 - 2-4 Å
- Torsion vs Cartesian NCS
 - Torsion are preferable in most cases
- Symmetry related copies:
 - Can be found automatically as part of refinement
 - Can be specified manually
 - Automatic determination relies on model quality
 - Always check automatically detected NCS copies

Secondary structure (SS) restraints

- Always use at 3Å and worse
- Better than 3Å: use if needed
- Require SS annotation
- SS annotation must be accurate
 - Errors in SS annotation will propagate into refined model
- Secondary structure (SS) annotation
 - SS information
 - HELIX/SHEET records in PDB file or equivalent in mmCIF
 - *Phenix* generated parameter files
 - Tools to create SS annotation
 - Command line (*phenix.secondary_structure_restraints*)
 - *Phenix* GUI
 - Quality of SS annotation:
 - Depends on quality of input model (GIGO)
 - No software can annotate SS fully reliably and correctly
 - Manual validation and editing almost always required

Aggressive optimization methods

- Simulated annealing (SA)
- Model morphing
 - Only use if model has gross errors (correction requires large movements)
 - Do not use if model is relatively good and only needs small corrections

Ramachandran plot restraints

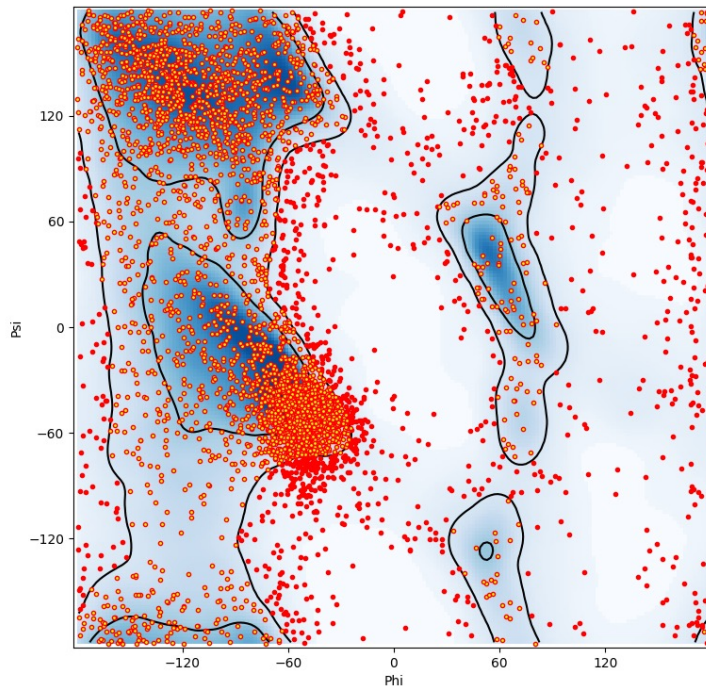
- Likely need at about 3Å and worse
- Better than 3Å: use if needed (preserve good initial model from deterioration)
- Check Ramachandran plot regularly
- Don't use to fix outliers. Fix outliers first (manually), then use Ramachandran plot restraints to stop re-occurring outliers

Ramachandran plot restraints

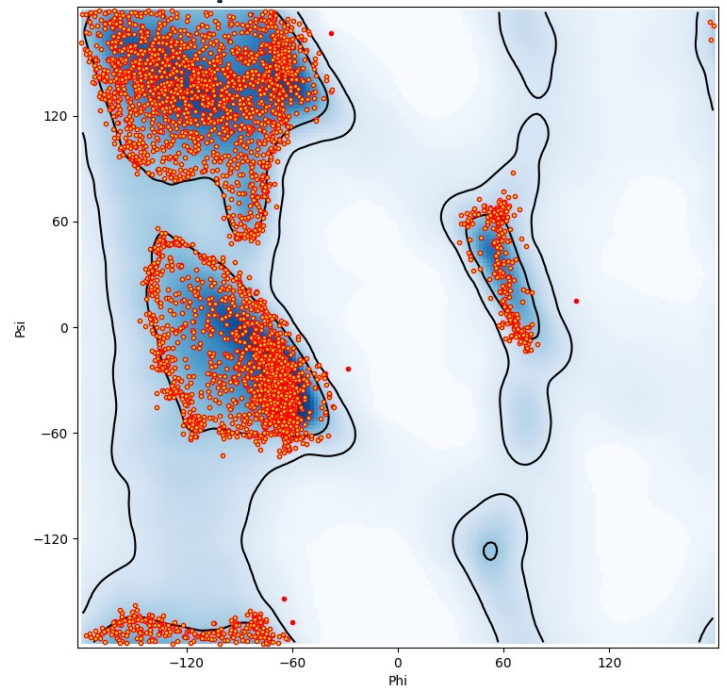
- Ramachandran plot restraints
 - Don't use to fix outliers. Fix outliers first, then use Ramachandran plot restraints to prevent re-occurring outliers.

PDB code: 5a9z

Original



Refined with Ramachandran plot restraints

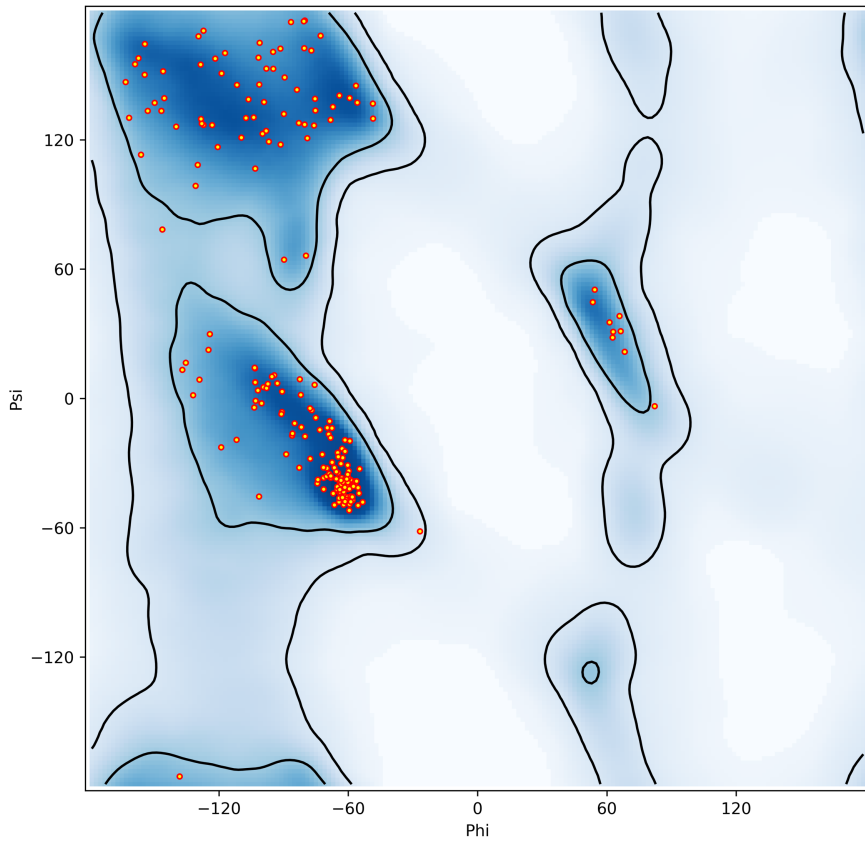


Bad idea to use Ramachandran plot restraints in this case. Fix outliers first!

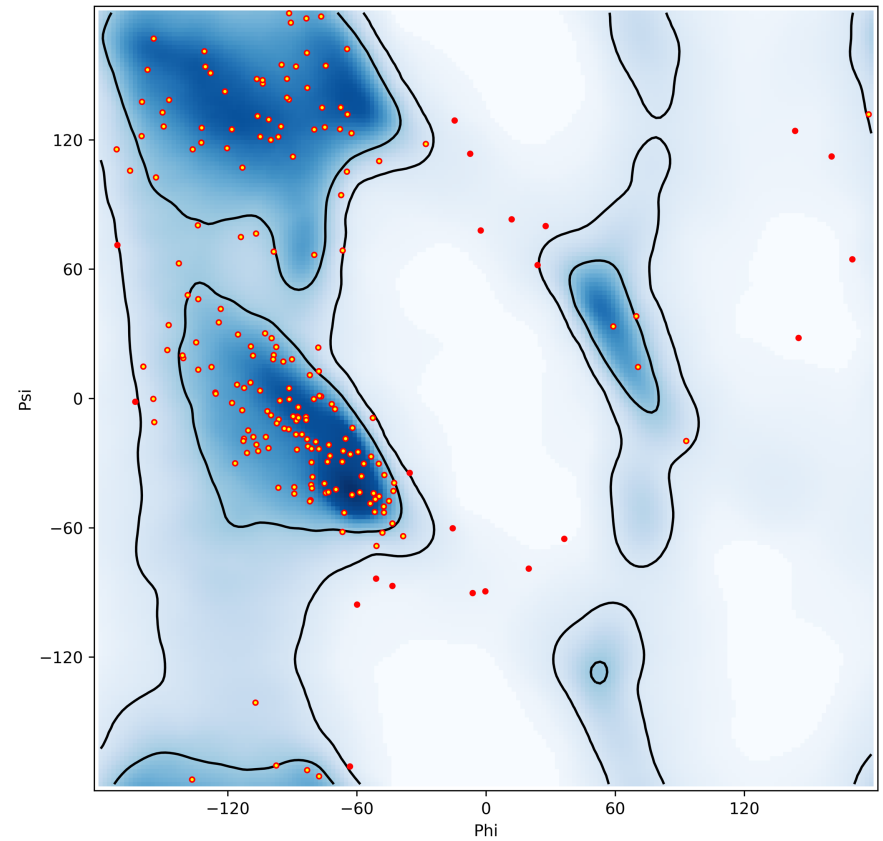
Ramachandran plot restraints

- Ramachandran plot restraints
 - Use to stop outliers from occurring

Before refinement

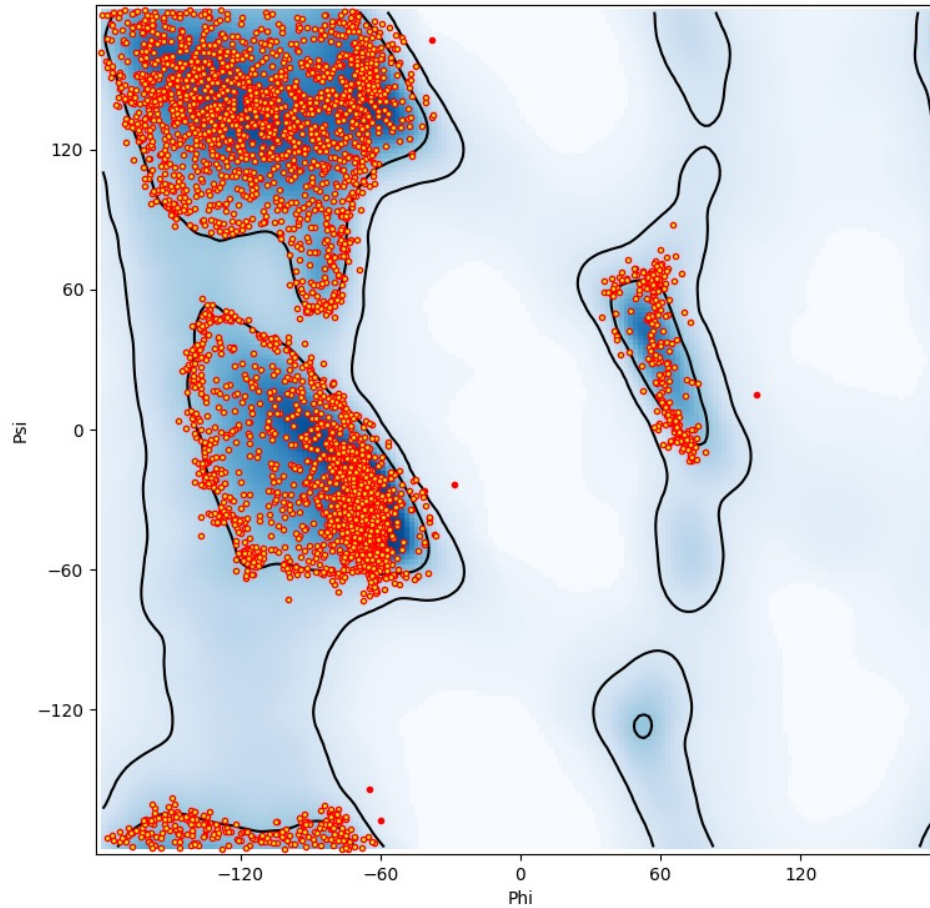


After refinement (No Ramachandran plot restraints)



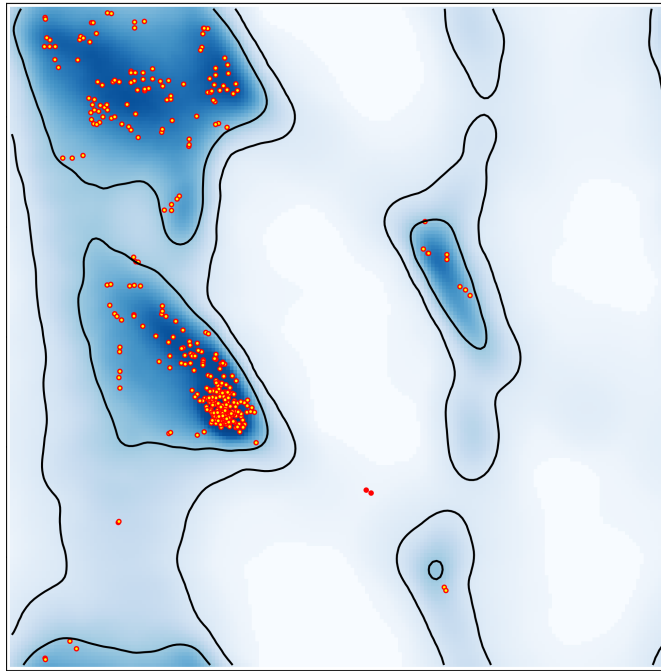
Ramachandran plot restraints

- What is wrong with this plot?



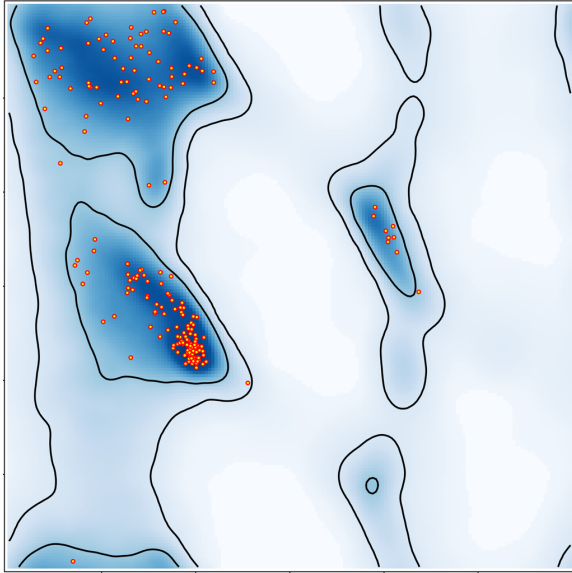
Ramachandran plot restraints

- They are very different from what we expect!

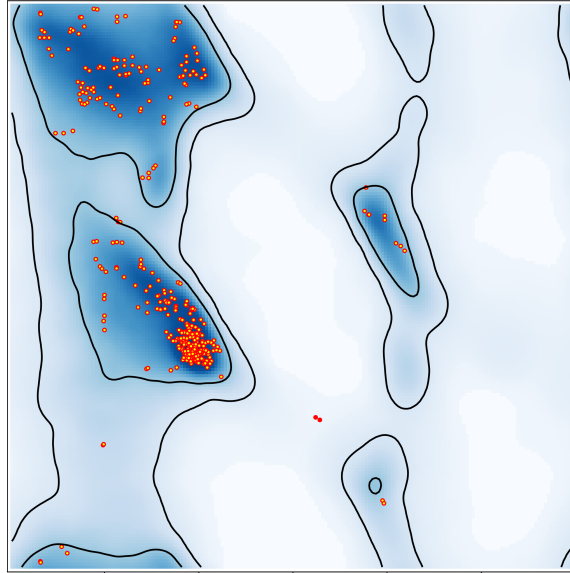


How you can tell good vs bad plot?

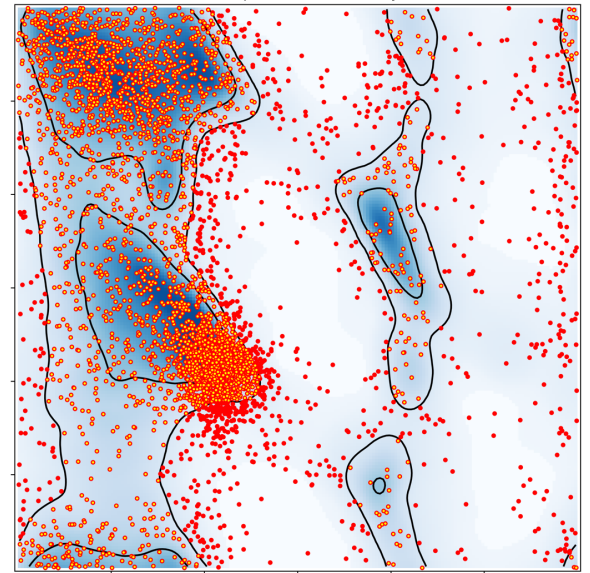
Good



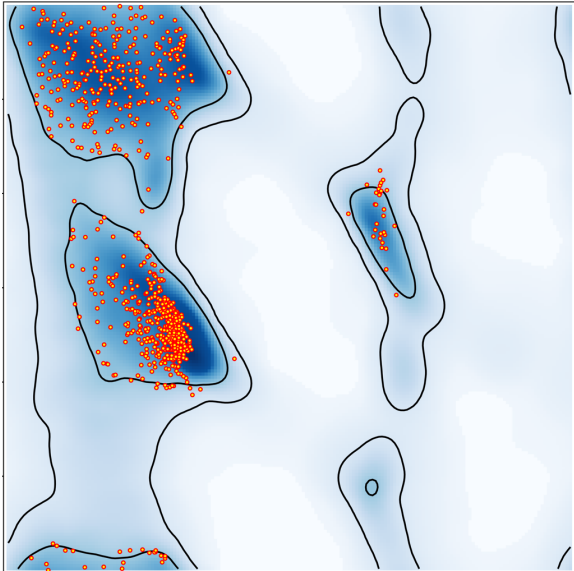
Good



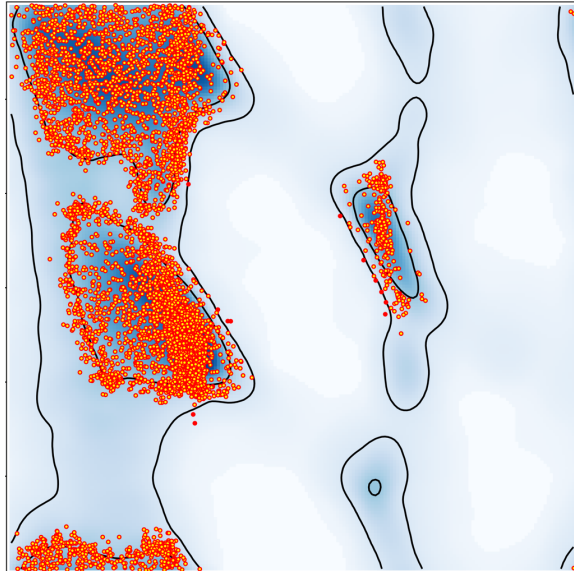
Bad



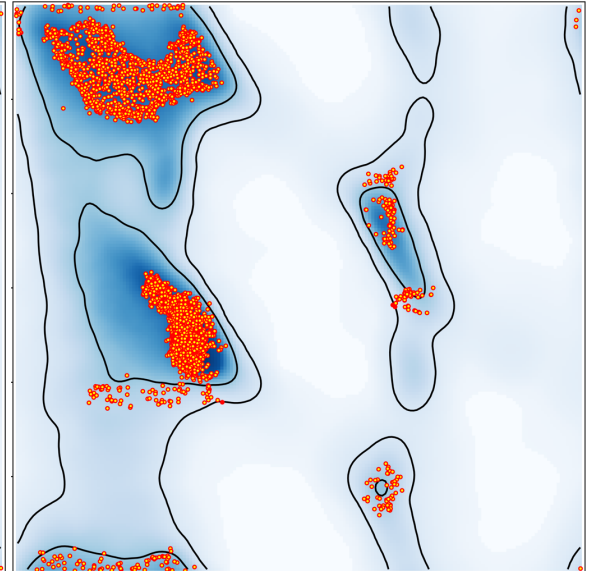
Bad



Bad



Bad



Ramachandran plot Z-score

CABIOS

Vol. 13 no. 4 1997
Pages 425–430

Objectively judging the quality of a protein structure from a Ramachandran plot

Rob W.W.Hooft, Chris Sander and Gerrit Vriend

- Good at spotting odd plots
- One number, simple criteria:
 - Poor: $|Z| > 3$ Suspicious: $2 < |Z| < 3$ Good: $|Z| < 2$

Structure

 CellPress

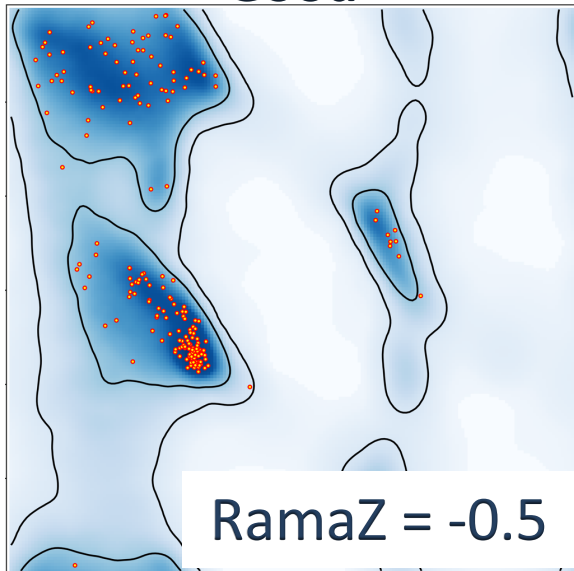
Resource

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

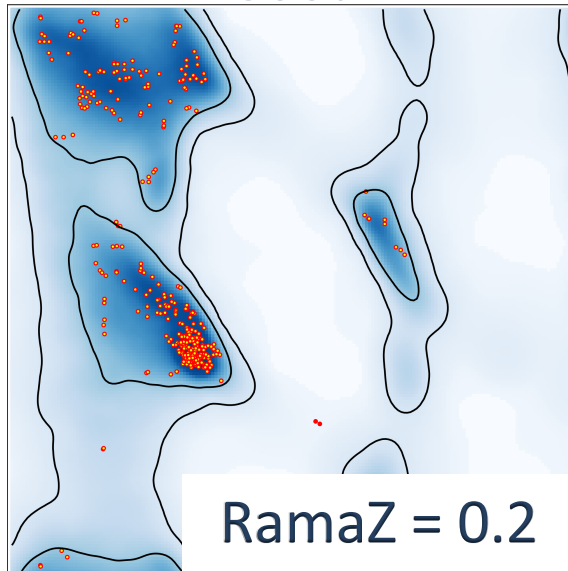
Oleg V. Sobolev,^{1,5,*} Pavel V. Afonine,¹ Nigel W. Moriarty,¹ Maarten L. Hekkelman,^{2,3} Robbie P. Joosten,^{2,3,*} Anastassis Perrakis,^{2,3} and Paul D. Adams^{1,4}

Model validation: *Ramachandran plot Z-score*

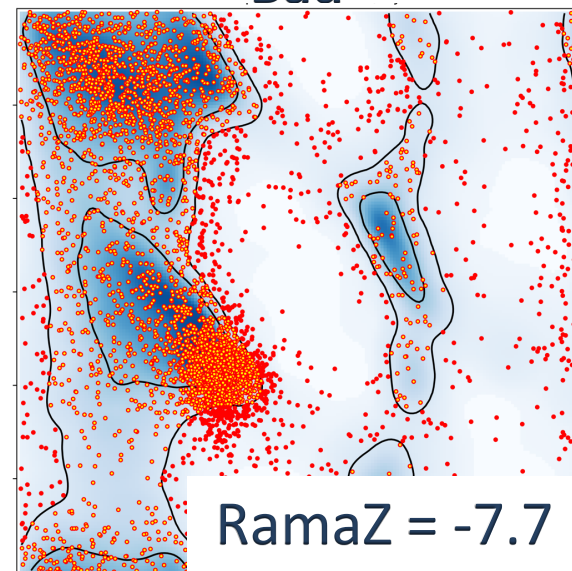
Good



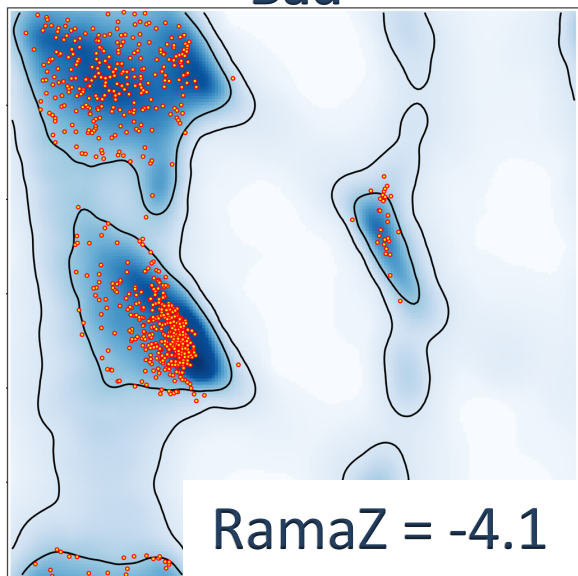
Good



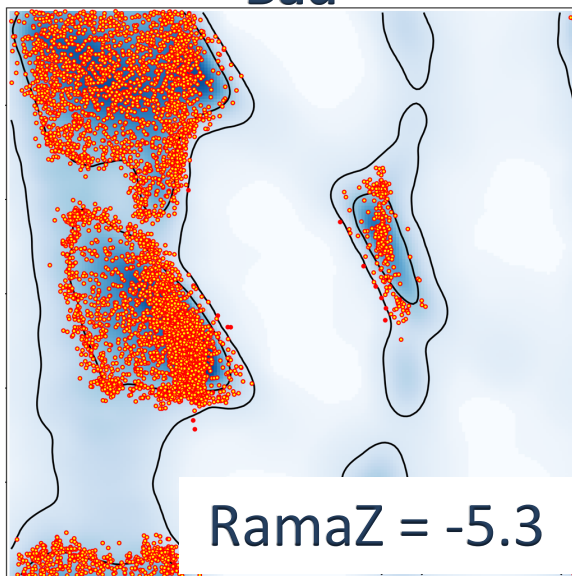
Bad



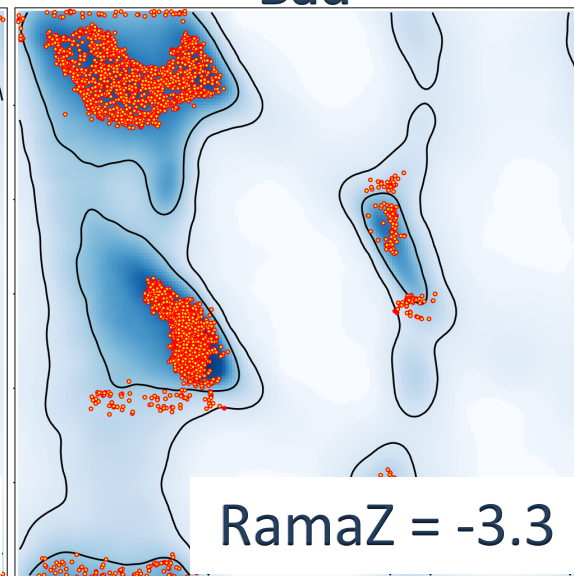
Bad



Bad

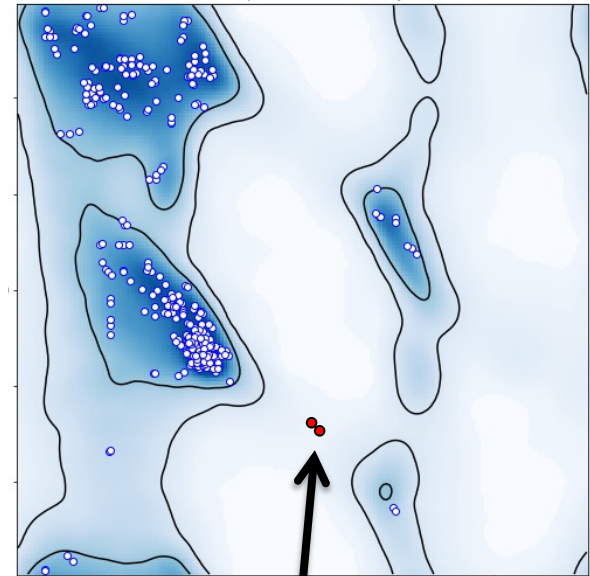
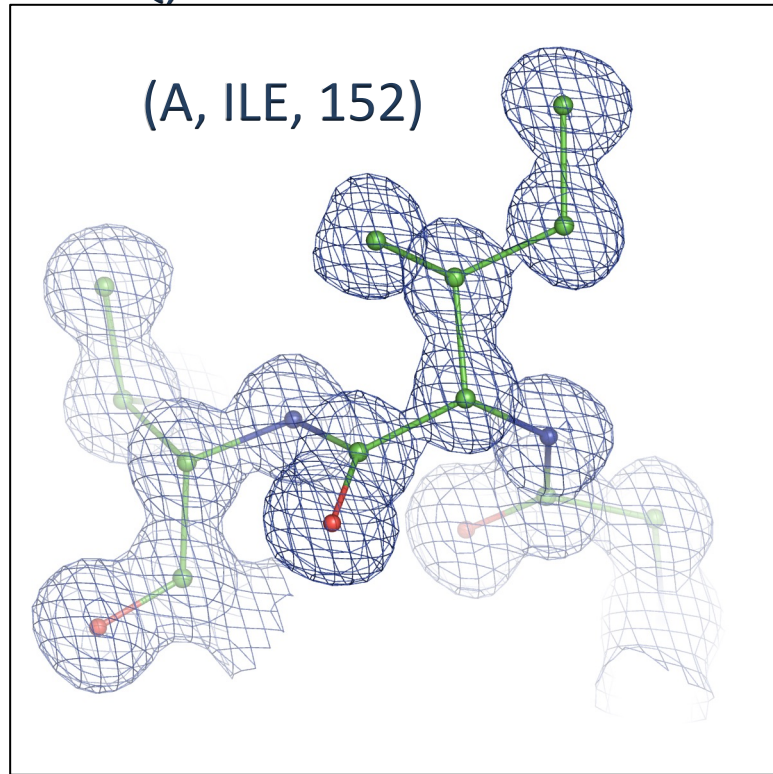


Bad



An outlier \neq wrong

3NOQ, 1 Å



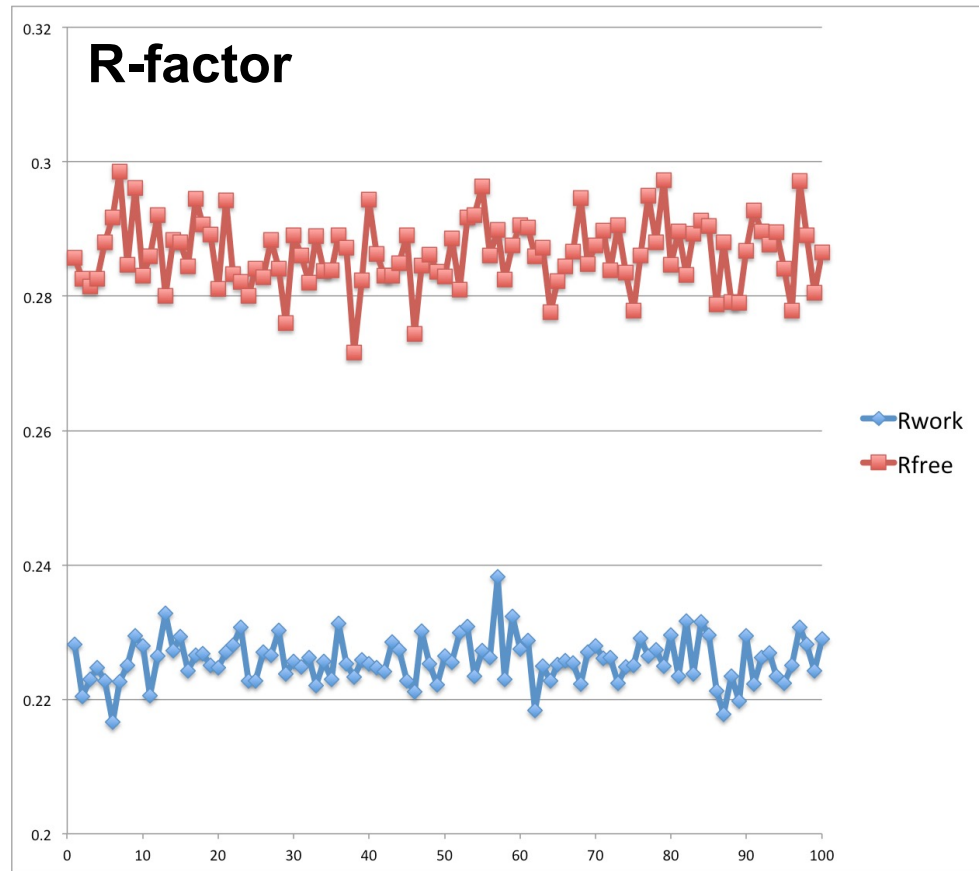
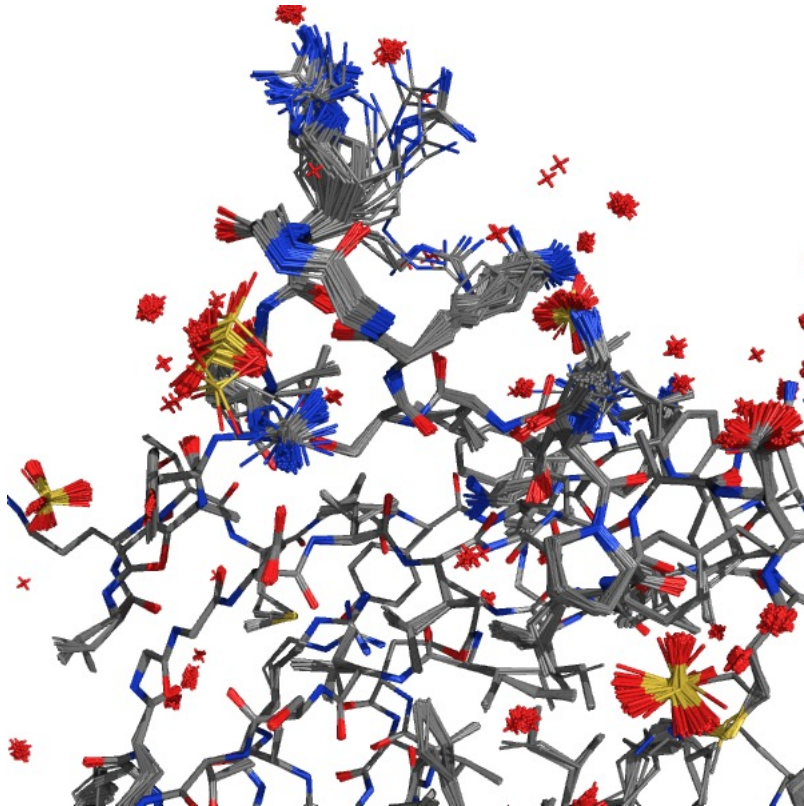
Outliers:

(A, ILE, 152), (B, ILE, 154)

- All outliers need to be explained (supported by the data)

Estimating and using uncertainty

100 identical refinement runs each one starting with slightly perturbed model



Refinement run

Reading

RESEARCH PAPERS

Acta Cryst. (2018). **D74**, 531-544
<https://doi.org/10.1107/S2059798318006551>

Cited by **672**

Part of *CCP-EM Spring Symposium 2017*



Real-space refinement in *PHENIX* for cryo-EM and crystallography

P. V. Afonine^{}, B. K. Poon^{}, R. J. Read^{}, O. V. Sobolev^{}, T. C. Terwilliger^{}, A. Urzhumtsev and P. D. Adams^{}

RESEARCH PAPERS

Acta Cryst. (2012). **D68**, 352-367
<https://doi.org/10.1107/S0907444912001308>

Cited by **2576**

Part of *CCP4 Study Weekend 2011*



Towards automated crystallographic structure refinement with *phenix.refine*

P. V. Afonine^{}, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty^{}, M. Mustyakimov, T. C. Terwilliger^{}, A. Urzhumtsev, P. H. Zwart^{} and P. D. Adams^{}

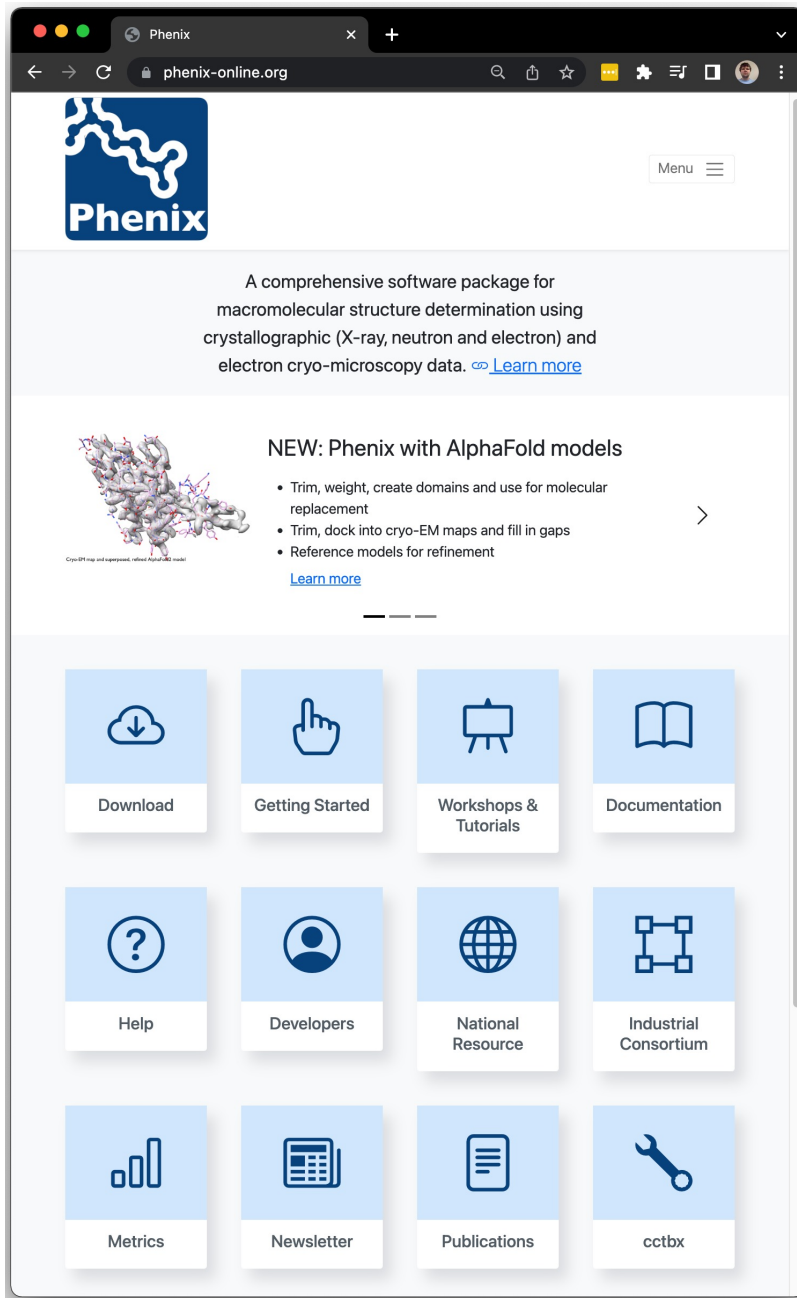
phenix.refine is a program within the *PHENIX* package that supports crystallographic structure



OPEN  ACCESS



Phenix resources



The screenshot shows the Phenix website homepage. At the top left is the Phenix logo, a blue square with a white molecular structure and the word "Phenix" below it. To the right of the logo is a "Menu" button with a hamburger icon. Below the logo is a grey banner with the text: "A comprehensive software package for macromolecular structure determination using crystallographic (X-ray, neutron and electron) and electron cryo-microscopy data. [Learn more](#)".

Below the banner is a section titled "NEW: Phenix with AlphaFold models" with a small image of a protein structure. To the right of the image is a list of bullet points: "Trim, weight, create domains and use for molecular replacement", "Trim, dock into cryo-EM maps and fill in gaps", and "Reference models for refinement". Below the list is a "Learn more" link.

Below the "NEW" section is a grid of 12 blue buttons with white icons and text labels:

- Download (cloud with arrow icon)
- Getting Started (hand cursor icon)
- Workshops & Tutorials (presentation board icon)
- Documentation (open book icon)
- Help (question mark icon)
- Developers (person icon)
- National Resource (globe icon)
- Industrial Consortium (network of nodes icon)
- Metrics (bar chart icon)
- Newsletter (calendar icon)
- Publications (document icon)
- cctbx (wrench icon)

Phenix paper

Video tutorials

Documentation

Relevant papers

Bi-annual newsletters

Slides from workshops

User support

- **Feedback, questions, help**

Mailing list (anyone signed up):

phenixbb@phenix-online.org

Bug reports (developers only):

bugs@phenix-online.org

Ask for help (developers only):

help@phenix-online.org

- **Reporting a bug or asking for help:**

- We can't help you if you don't help us to understand your problem
- Make sure the problem still exist using the latest *Phenix* version
- Send us all inputs (files, non-default parameters) and tell us steps that lead to the problem
- All data sent to us is kept confidentially

The Project



Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine,
Dorothee Liebschner, Nigel
Moriarty, Billy Poon,
Christopher Schlicksup,
Oleg Sobolev



University of Cambridge

Randy Read, Airlie McCoy,
Tristan Croll, Claudia Millán Nebot,
Rob Oeffner



Los Alamos National Laboratory New Mexico Consortium

Tom Terwilliger, Li-Wei Hung



UTHealth

Matt Baker, Corey Hyrc



Duke University

Jane & David Richardson,
Christopher Williams,
Vincent Chen



An NIH/NIGMS funded
Program Project

Liebschner D, *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. Acta Cryst. 2019 **D75**:861–877