Model Refinement

Pavel Afonine

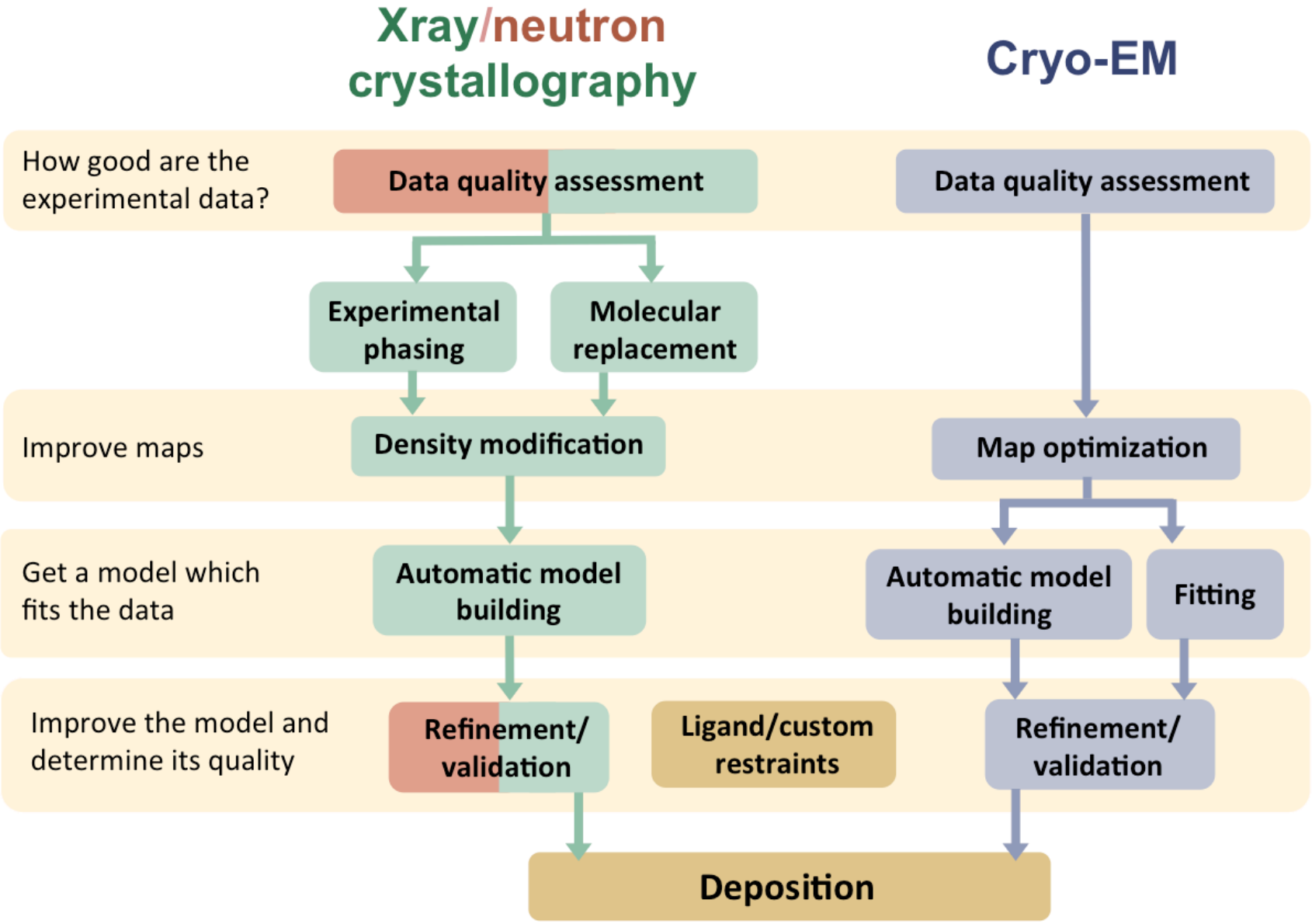
phenix-online.org

lbl.gov

qrefine.com

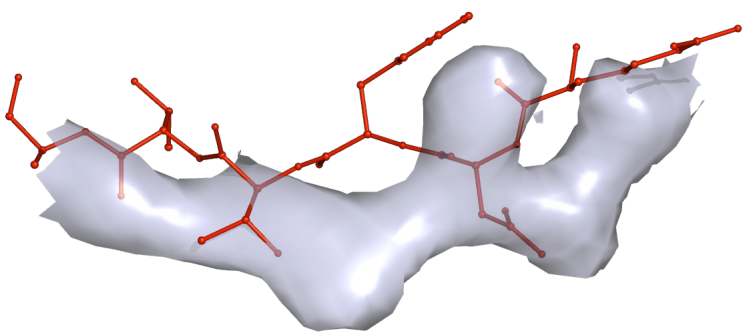
October 14th, 2023
The 80th Annual Pittsburgh Diffraction Conference
Pittsburgh, PA

Phenix: tools for crystallography and cryo-EM

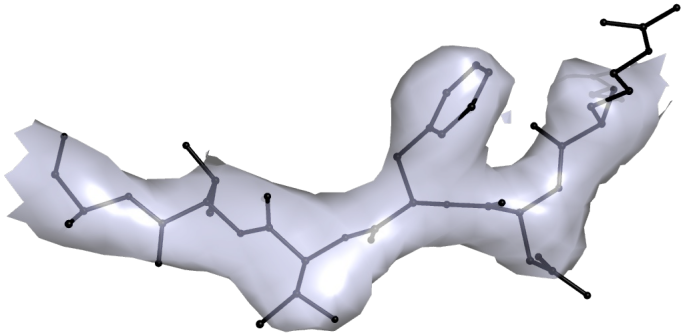


Model refinement in a nutshell

Initial model

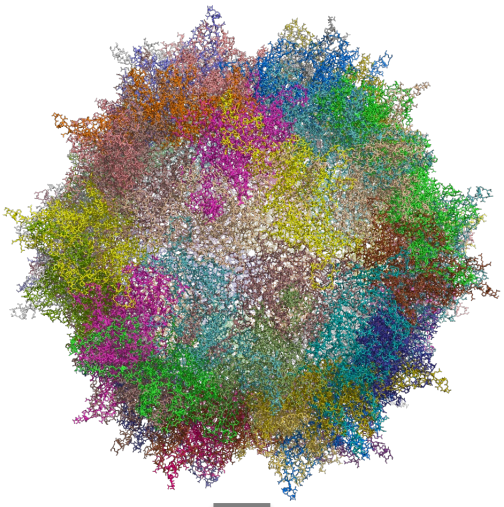


Improved (refined) model

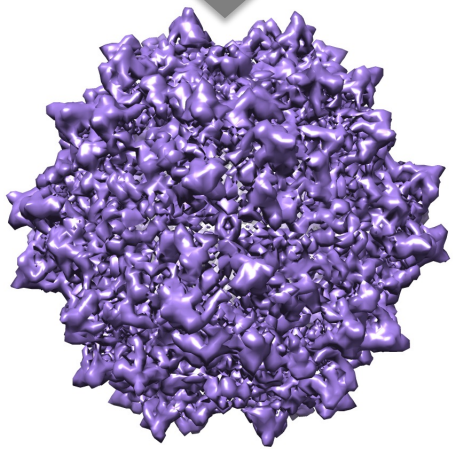


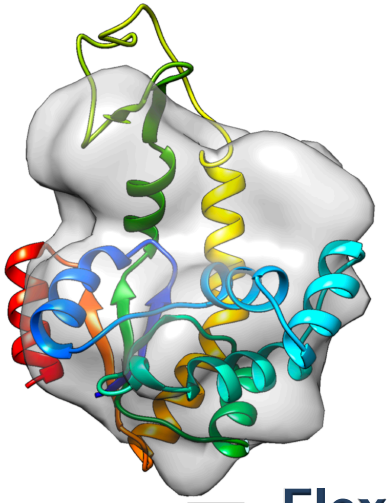
Fit atomic model to experimental data with the help of some *a priori* known information about the model

Not all model-to-data fitting is refinement

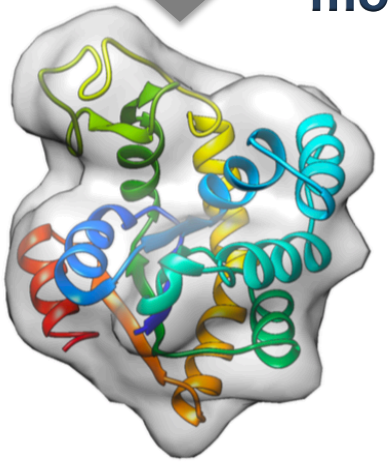


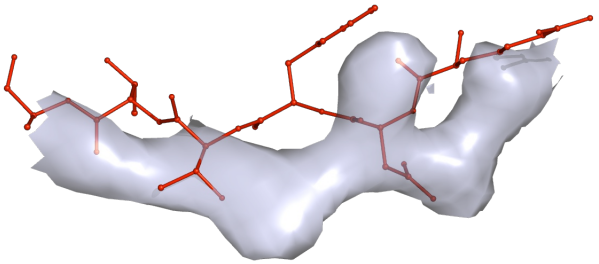
Docking



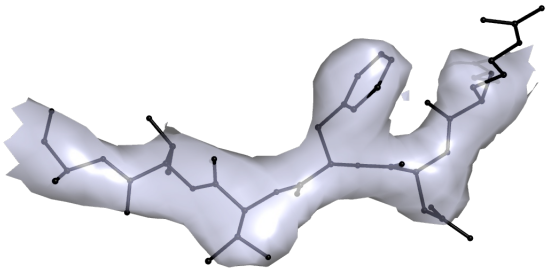


Flexible fitting, morphing

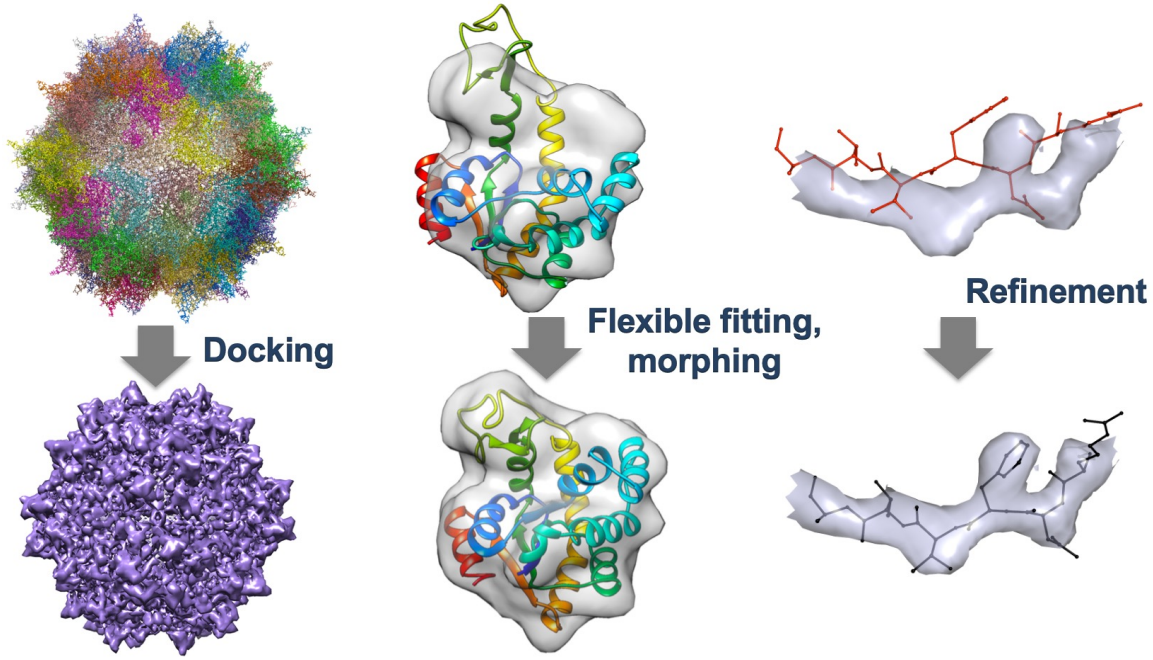




Refinement



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*, ~ 1Å)

Refinement used to be tedious and time consuming

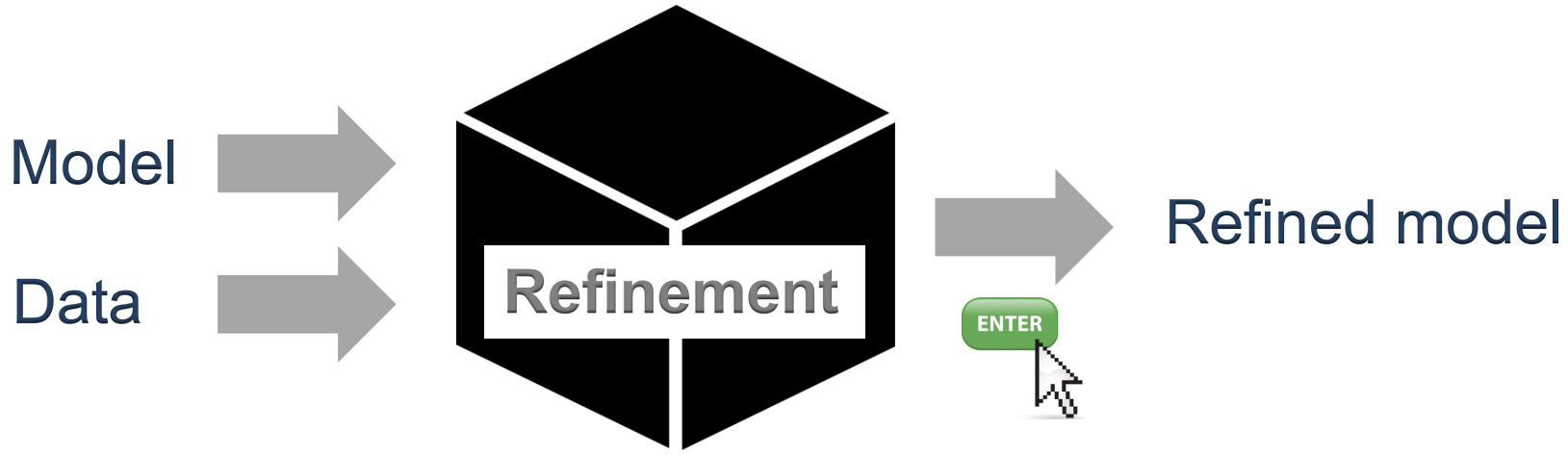
- Familiar with multiple software packages
- Coding knowledge (typically FORTRAN or C)
- Expertise in Unix
- Reading thick books (no Google or ChatGPT!)
  - Anyone remembers 405 pages X-plor book by A. Brunger?
- Don’t expect your questions answered by email within 24 hours

Refinement used to be tedious and time consuming

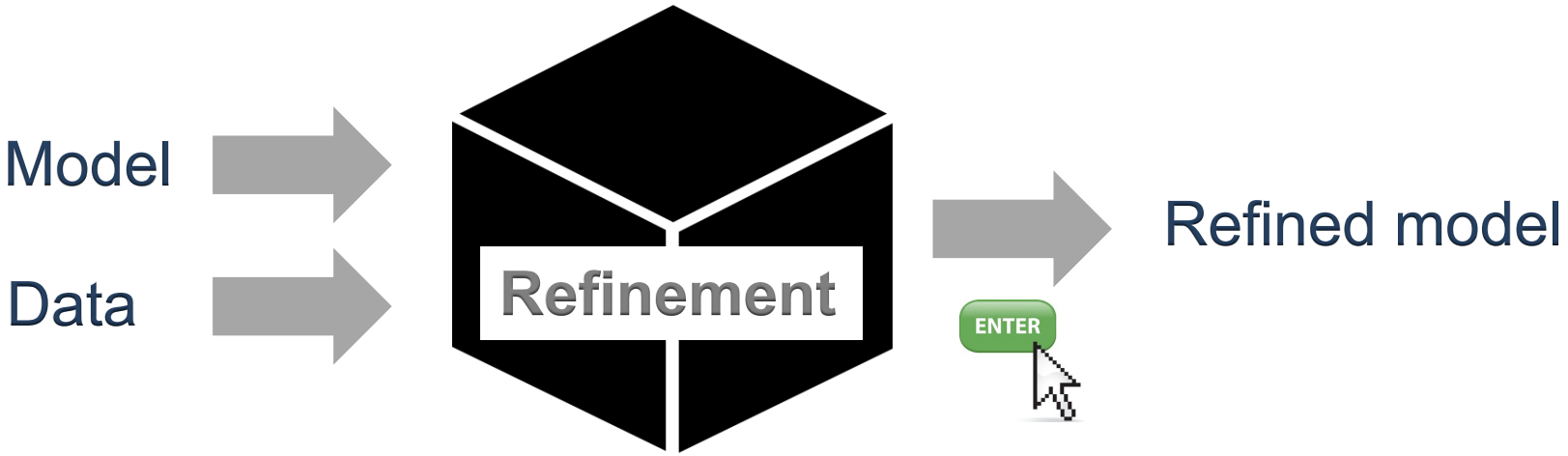
- Many months to complete
  - Spend days on graphics (manual building)
  - Run refinements overnight

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?

Group B vs individual?

Local minima?

ADP?

Clashes?

NCS?

IAS?

tNCS?

Grid search?

Weights?

CDL?

SA?

Minimization?

Rama plot restraints?

f’ & f”?

Hydrogens?

Restraints?

Bulk-Solvent?

Rama-Z?

Anisotropy?

Rigid body?

Twinning?

SS restraints?

NQH flips?

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
  - Twining, 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…

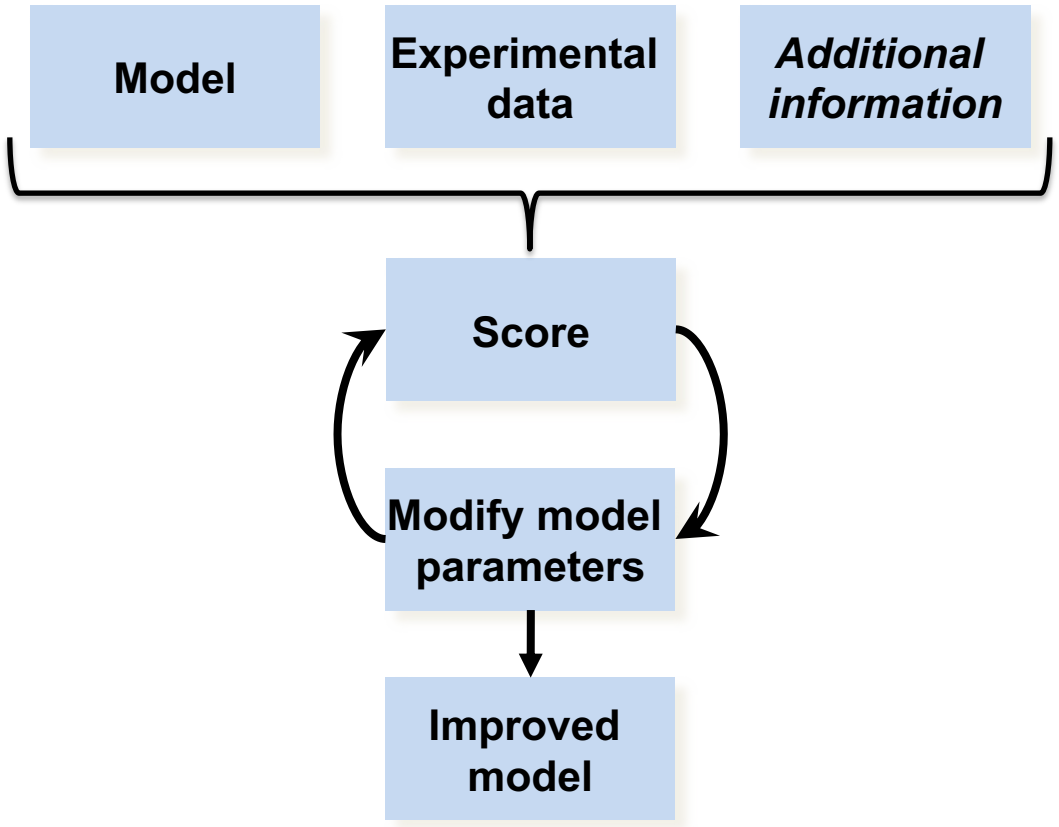
- … refinement worked ?
- … you did it correctly ?
- … the model you got is good enough to publish ?

- Do validation!

Standard validation protocols are designed to answer these questions

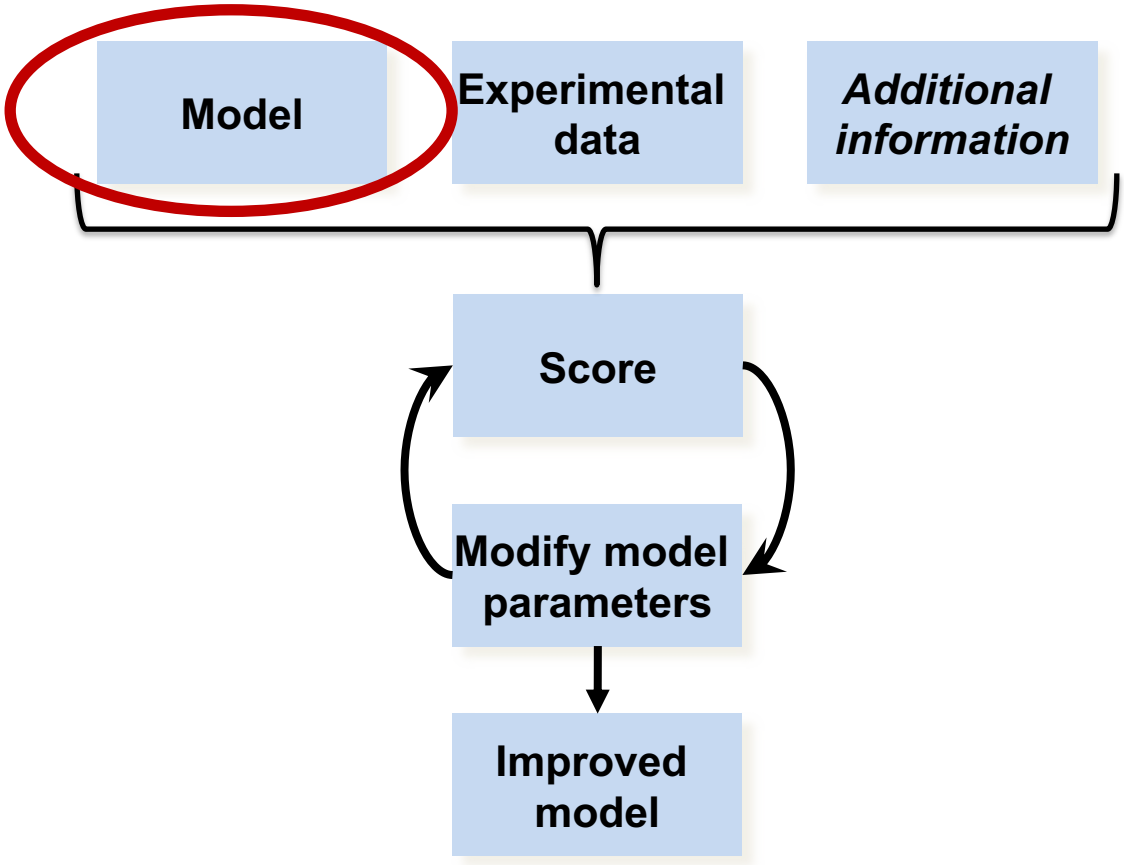
Refinement: a closer look

Model refinement



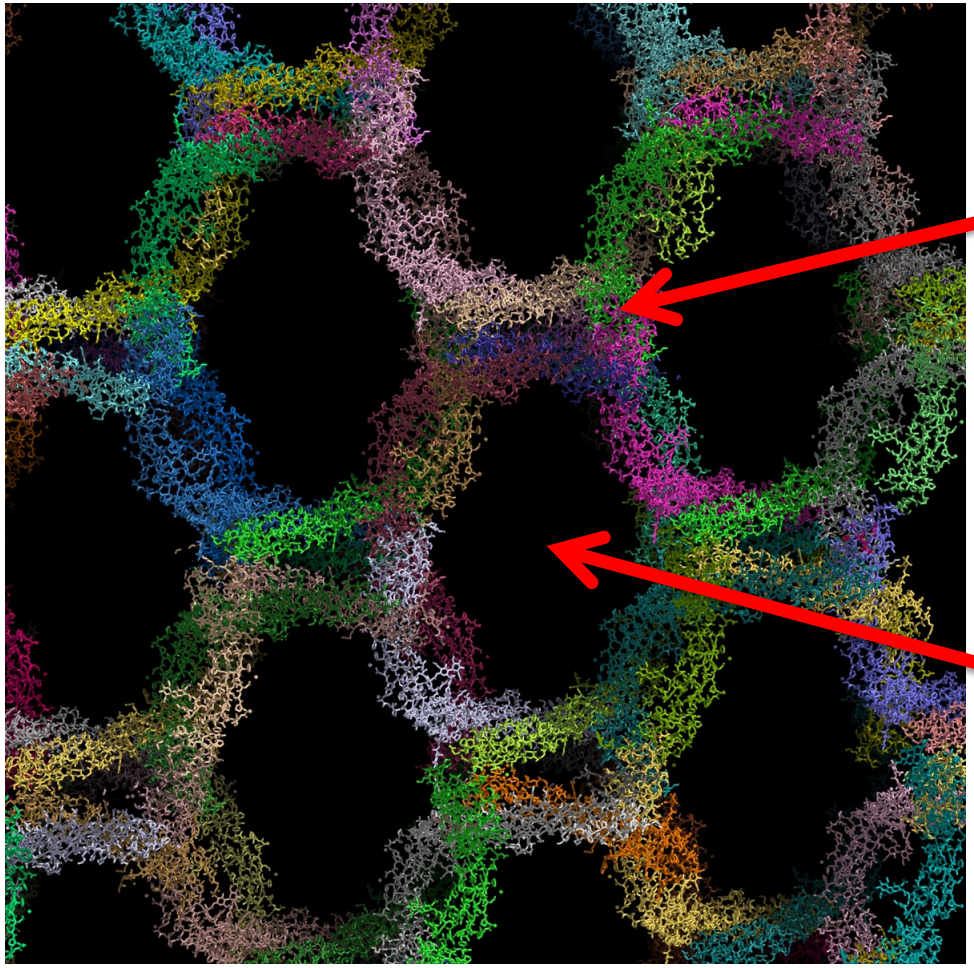
Refinement – optimization process of fitting model parameters to experimental data

Model refinement



Crystal structure model

PDB code: 1QUB

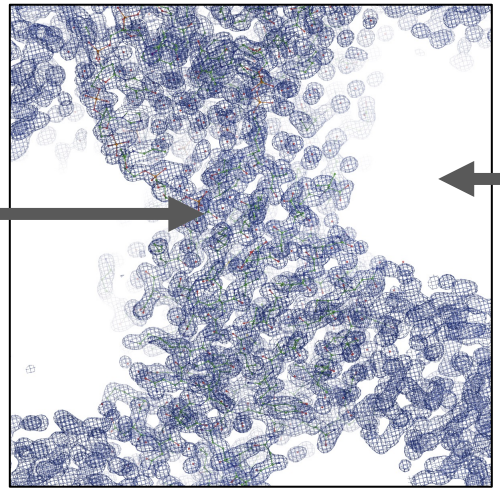


Macro-molecule

Bulk-solvent:
~ 50% of unit cell volume

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

Crystal structure model: structure factors

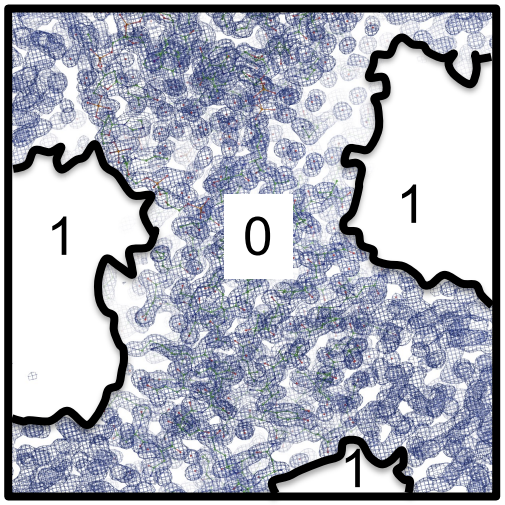


Macromolecule
(atomic model)

Bulk-solvent

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

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_{MASK}= FT(M)
3. Define solvent contribution F_{BULK}:

   $F_{BULK} = k_{MASK} * F_{MASK}$
4. Combine with $F_{CALC(ATOMS)}$

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

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

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)

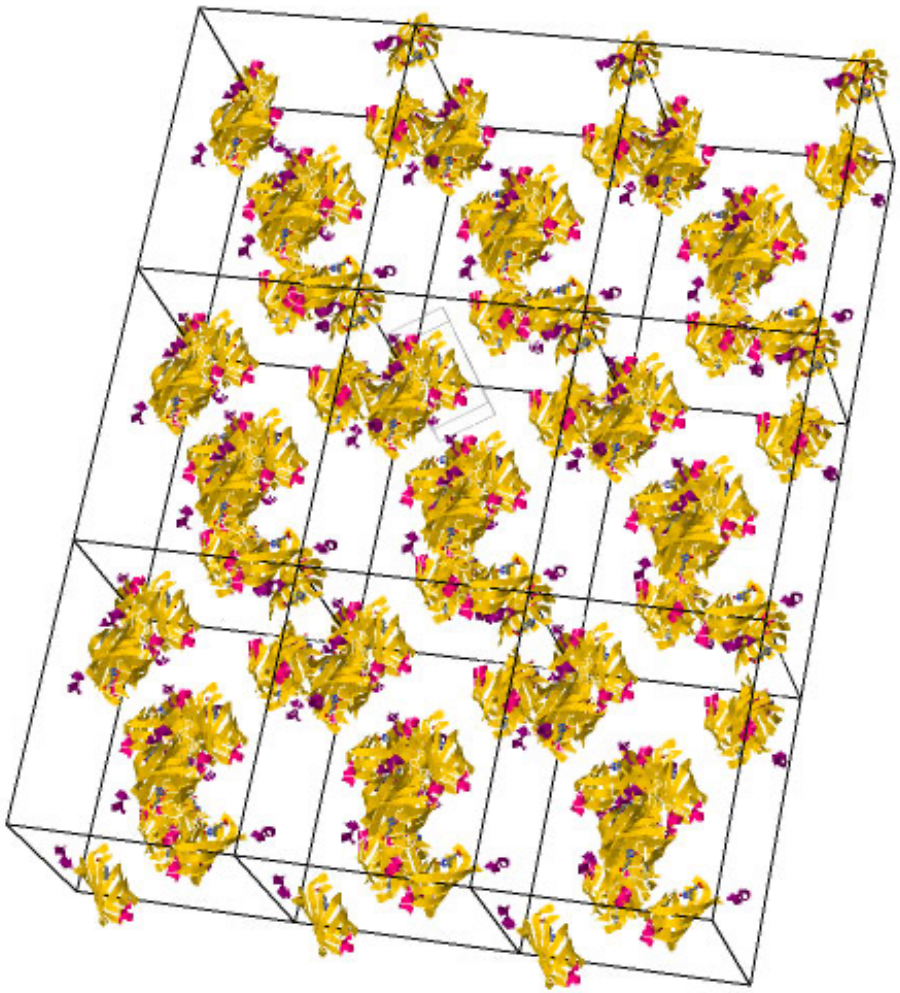
$$\mathbf{F}_{\text{CALC (ATOMS)}}(h,k,l) = \sum_{n=1}^{Natoms} q_n f_n(s) \exp\left(-\frac{B_n s^2}{4}\right) \exp\left(2i\pi \mathbf{r}_n \mathbf{s}\right)$$

Atom type `C`

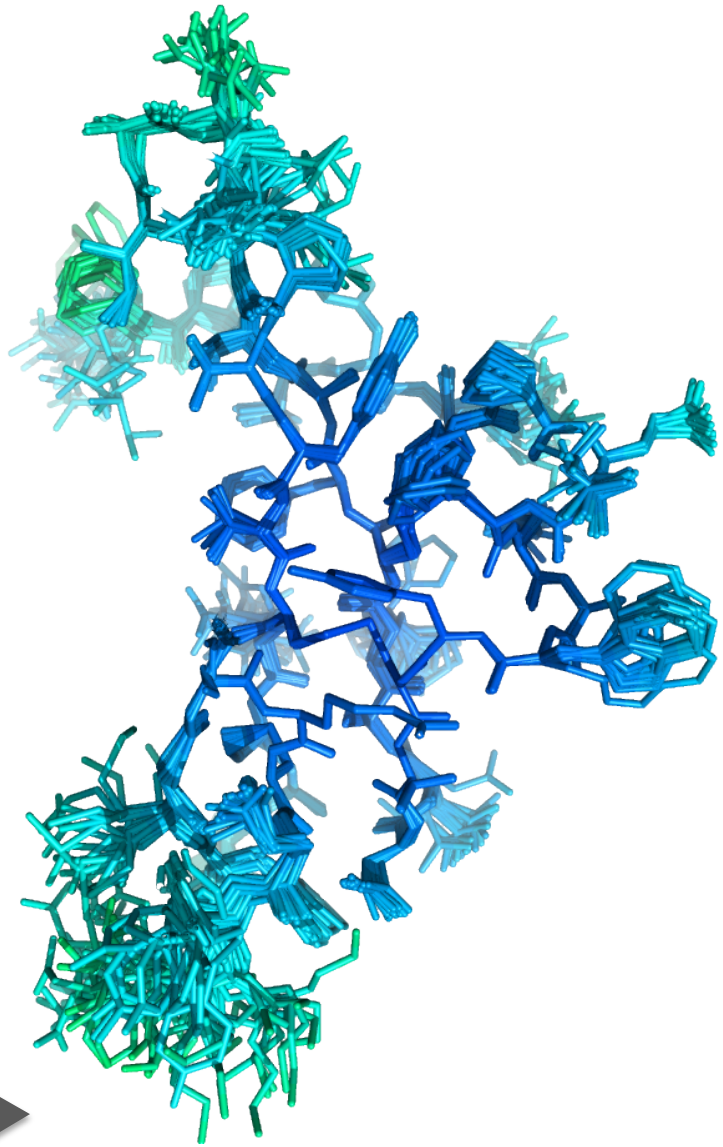
`31.309 29.489 26.044`

Atomic coordinates

Atomic model: disorder

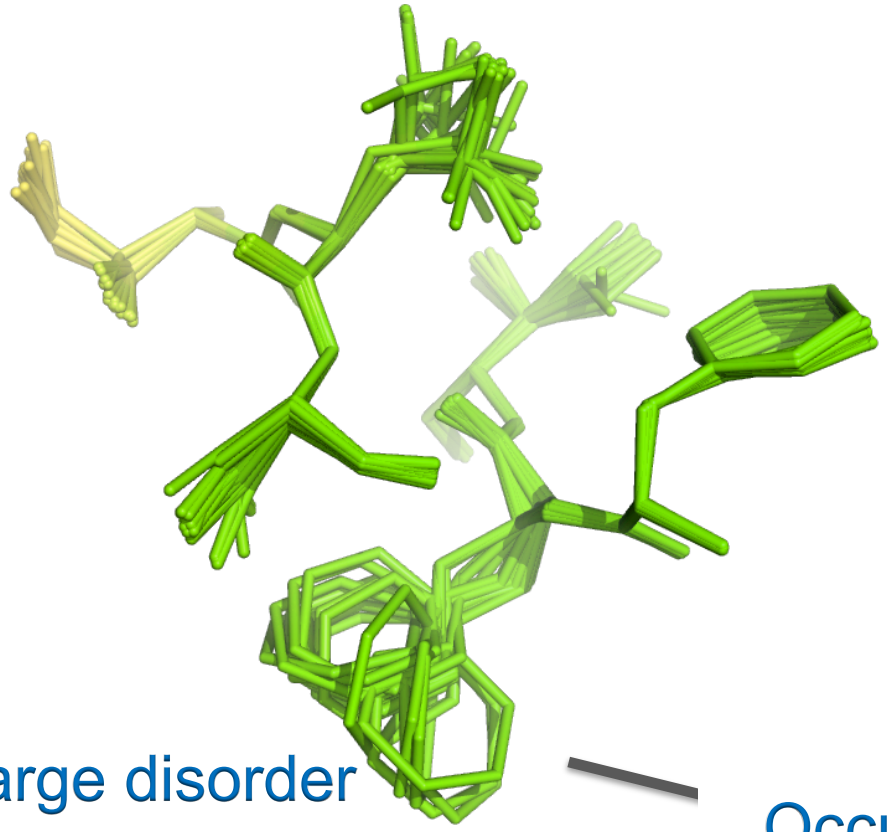


Crystal = many unit cells



Superpose all structures
from each unit cell

Atomic model: disorder



Small disorder

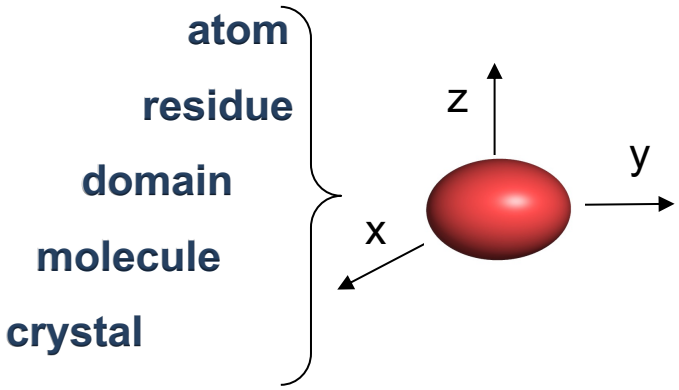
ADP (B-factor)

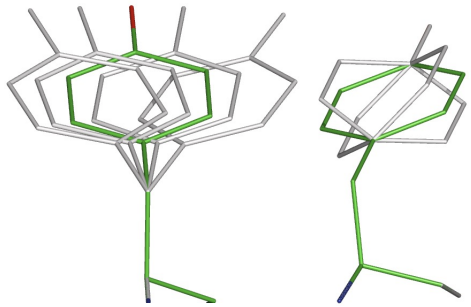
Large disorder

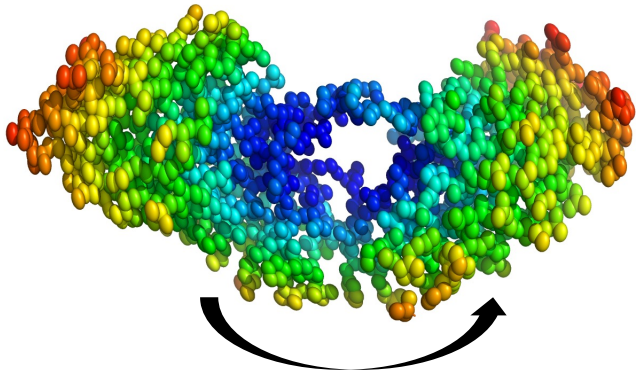
Occupancy

```
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)



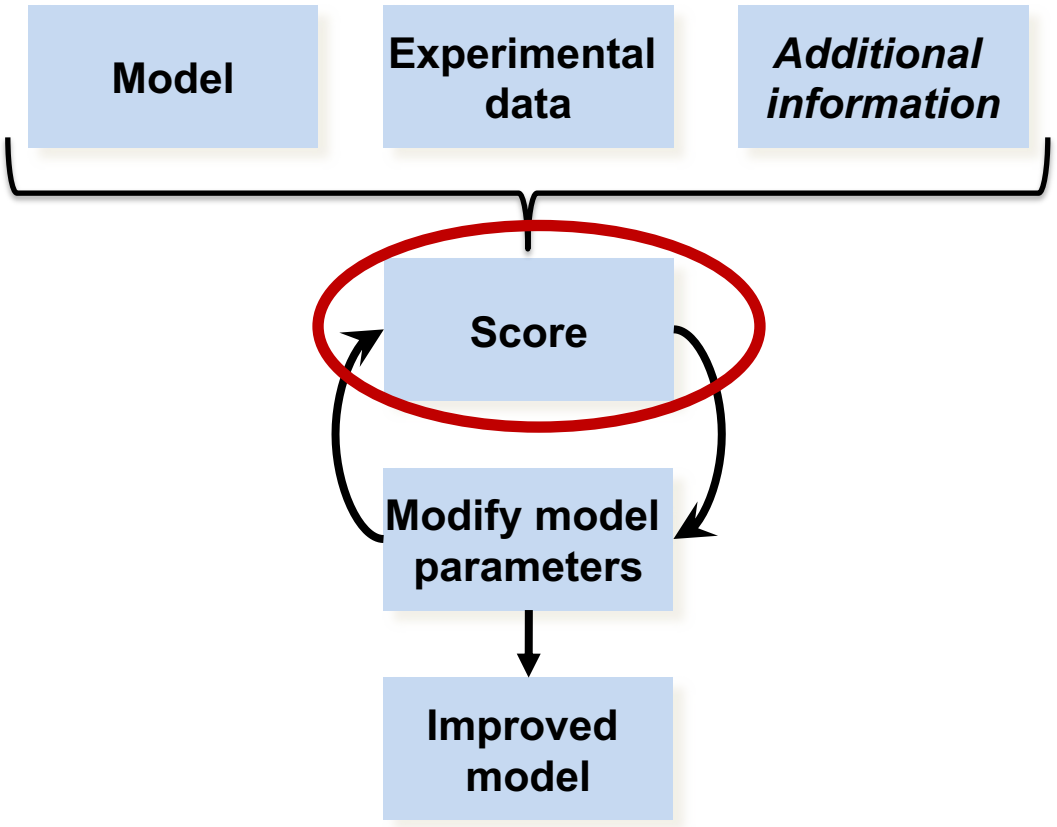




TLS

B_{TOTAL} = sum of individual contributions

Refinement target function (score)



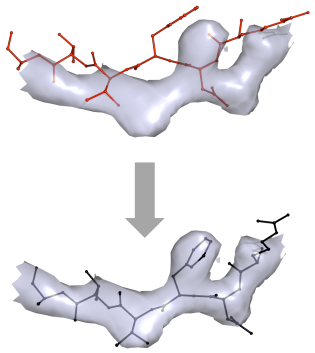
Model refinement

T = **T_{DATA}** + **w** * **$T_{RESTRAINTS}$**

T

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

T_{DATA}



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

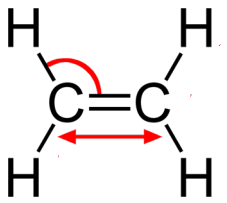
$$\sum_{hkl}\frac{||F_{obs}| - |F_{model}||}{|F_{obs}|}$$

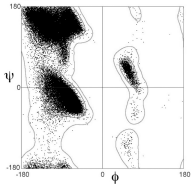
Maximum-Likelihood

w

Dose of extra information

$T_{RESTRAINTS}$





T_{BOND} +

T_{ANGLE} +

$T_{DIHEDRAL}$ +

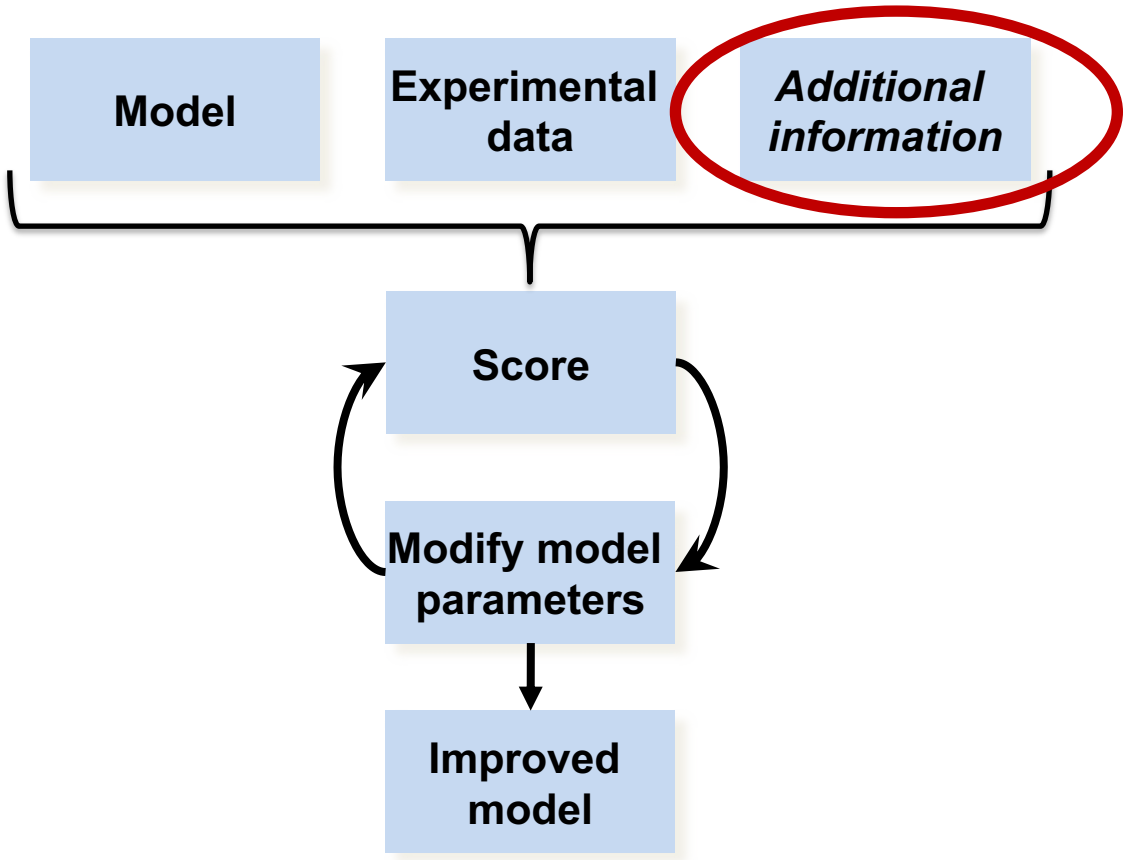
T_{PLANE} +

$T_{REPULSION}$ +

$T_{CHIRALITY}$ +

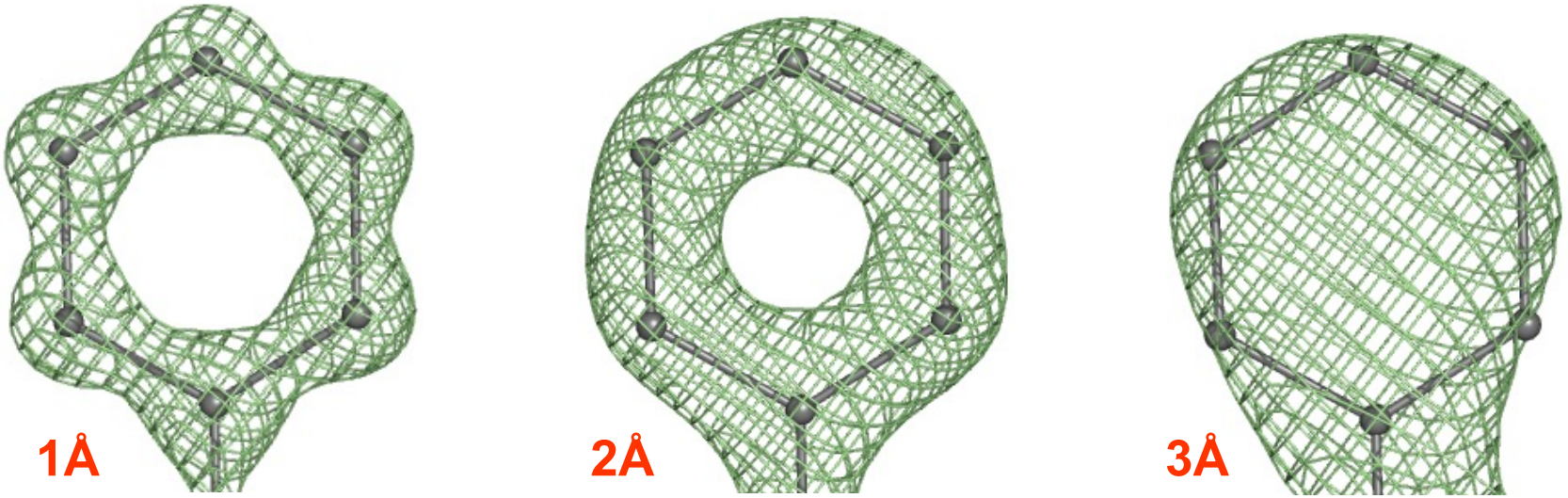
...

Additional information (restraints, constraints)



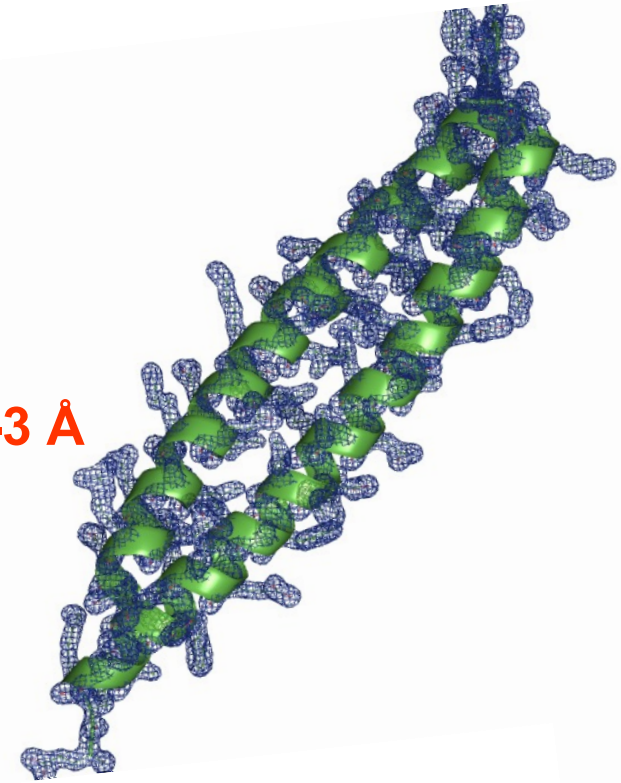
Restraints and constraints

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

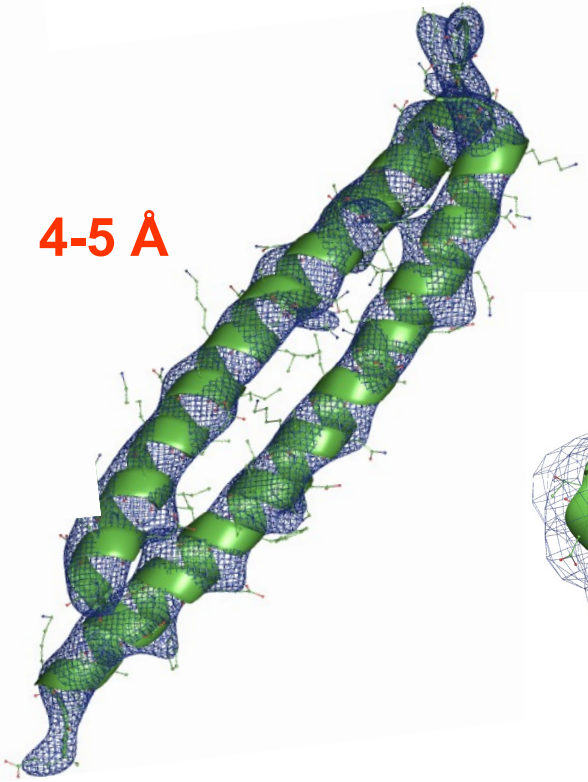


Restraints and constraints

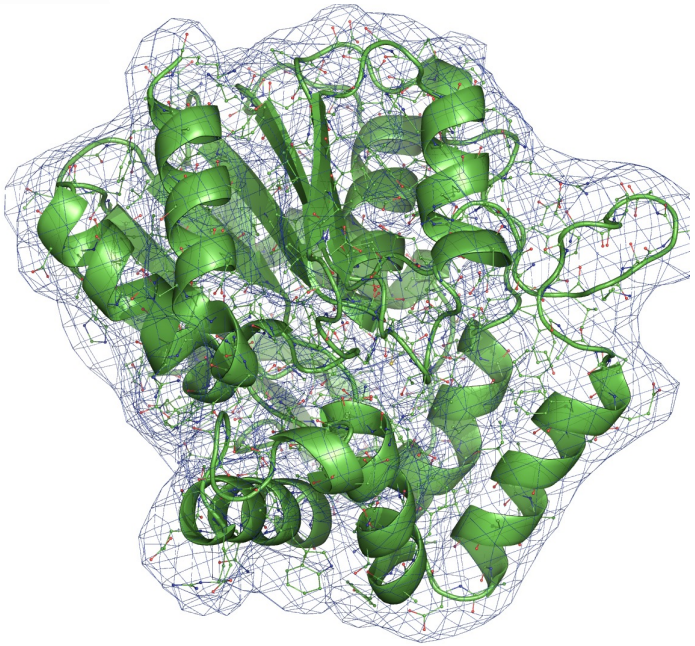
2-3 Å



4-5 Å

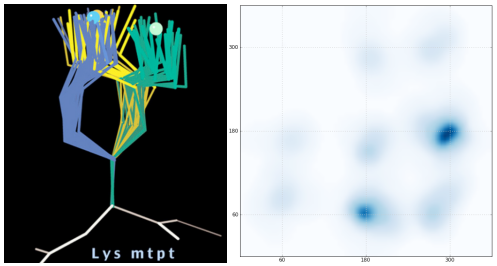


6Å-lower

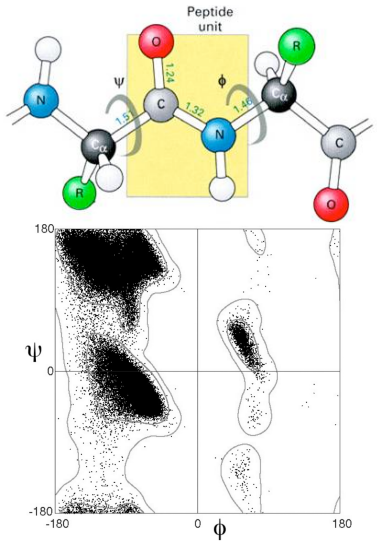


More restraints for low resolution

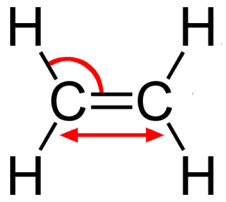
Side chain distributions



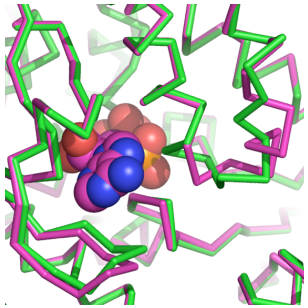
Main chain distributions



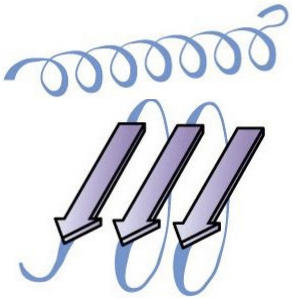
Covalent geometry



Internal symmetry (NCS)



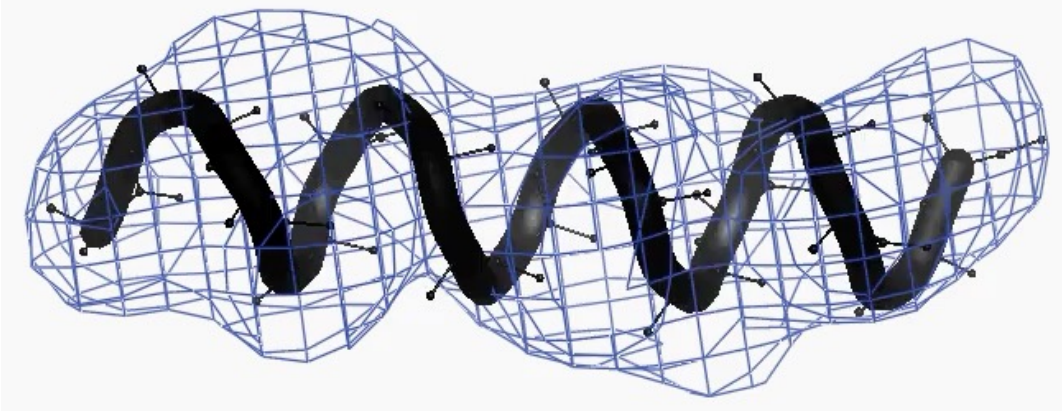
Similar (homologous) structures (reference model restraints)



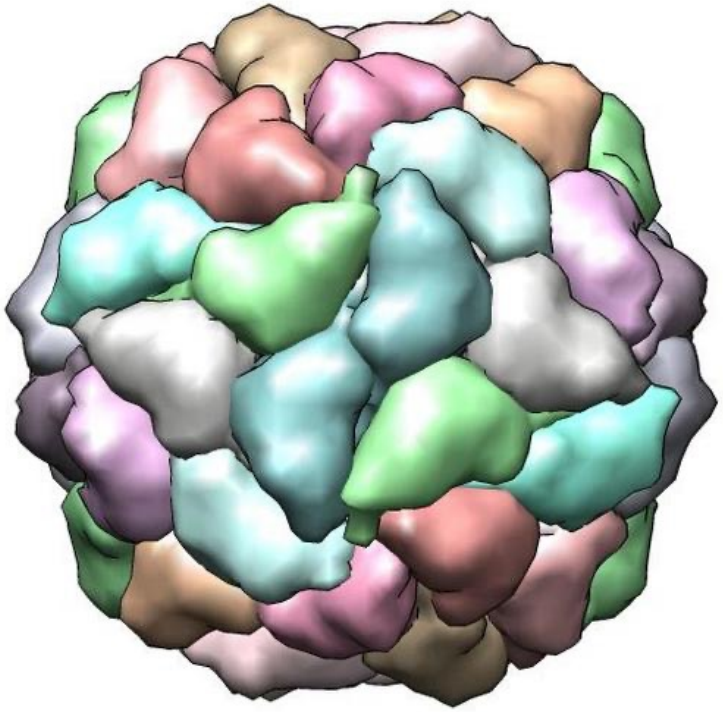
Secondary structure

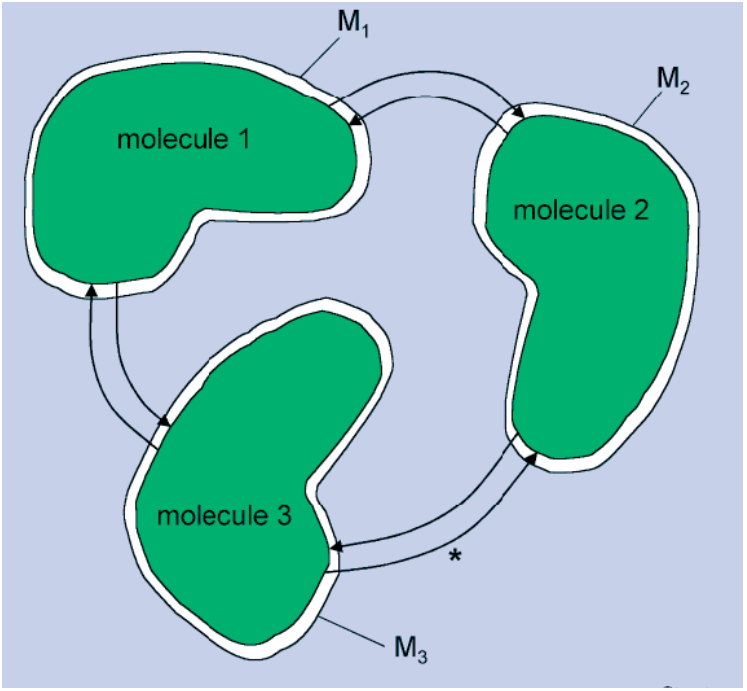
Importance of more restraints at low resolution

- 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



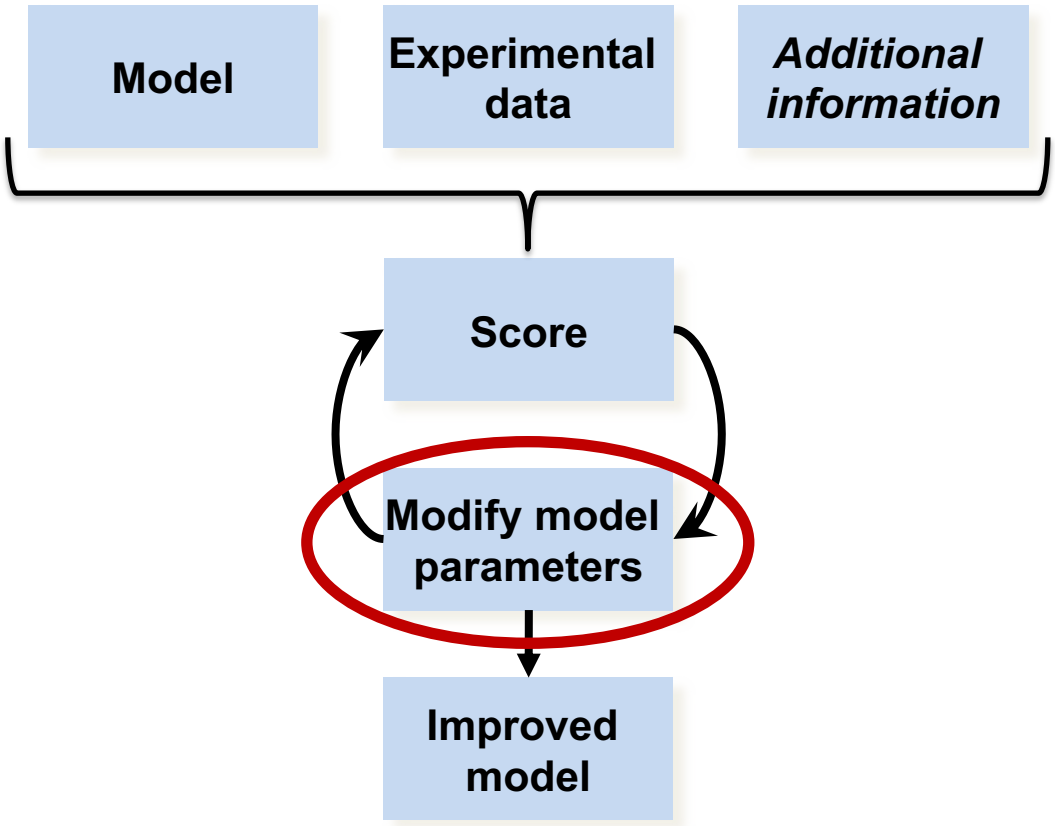


Source: Internet

- **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

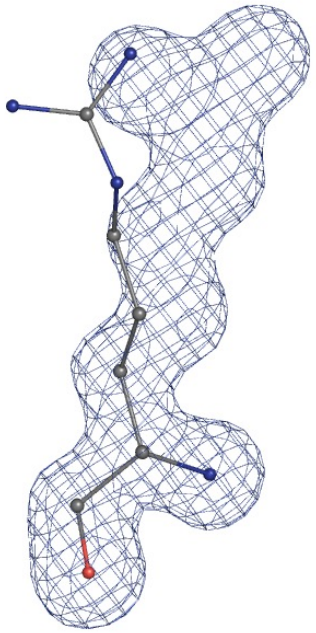

Model
Experimental data
Additional information
Score
Modify model parameters
Improved model

Choices of optimization method

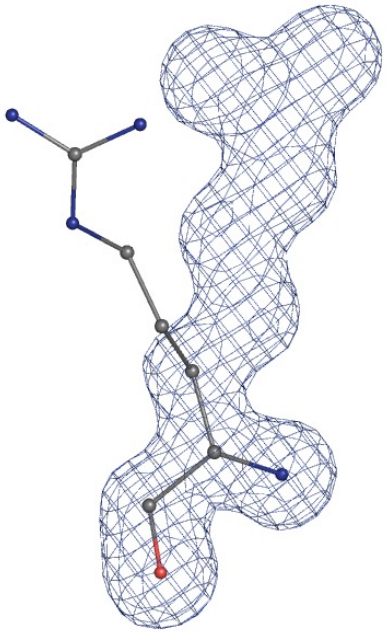
- Gradient-based minimization
- Simulated annealing
- Grid (systematic) searches
- Manual using molecular graphics programs (Coot, Chimera,…

Choice of refinement method and refinement convergence

Minimization

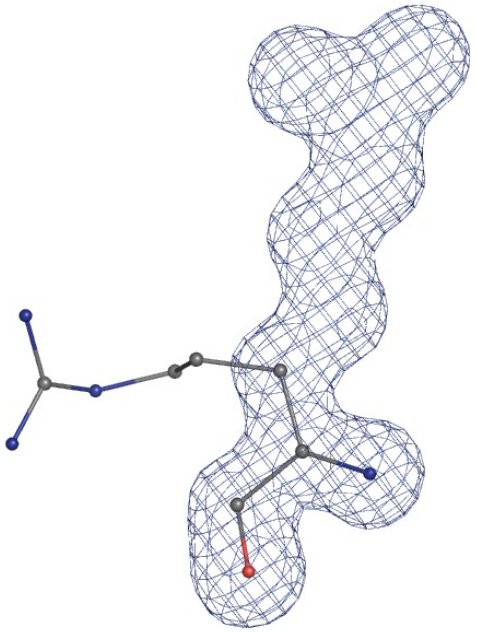


Simulated Annealing



Beyond convergence radius of minimization

Real-space grid search

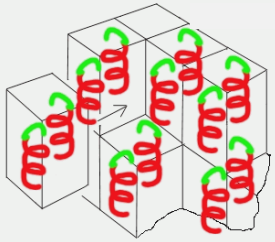


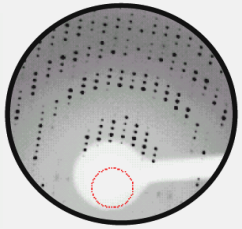
Beyond convergence radius of minimization and SA

Phenix tools for model refinement

Refinement

Crystallography





Initial model

Experimental data

A priori knowledge

Score

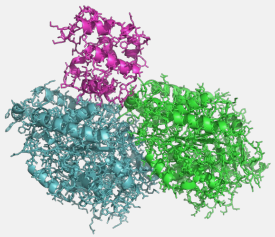
Modify model parameters

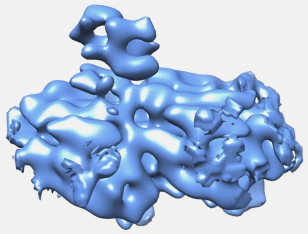
Improved model

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

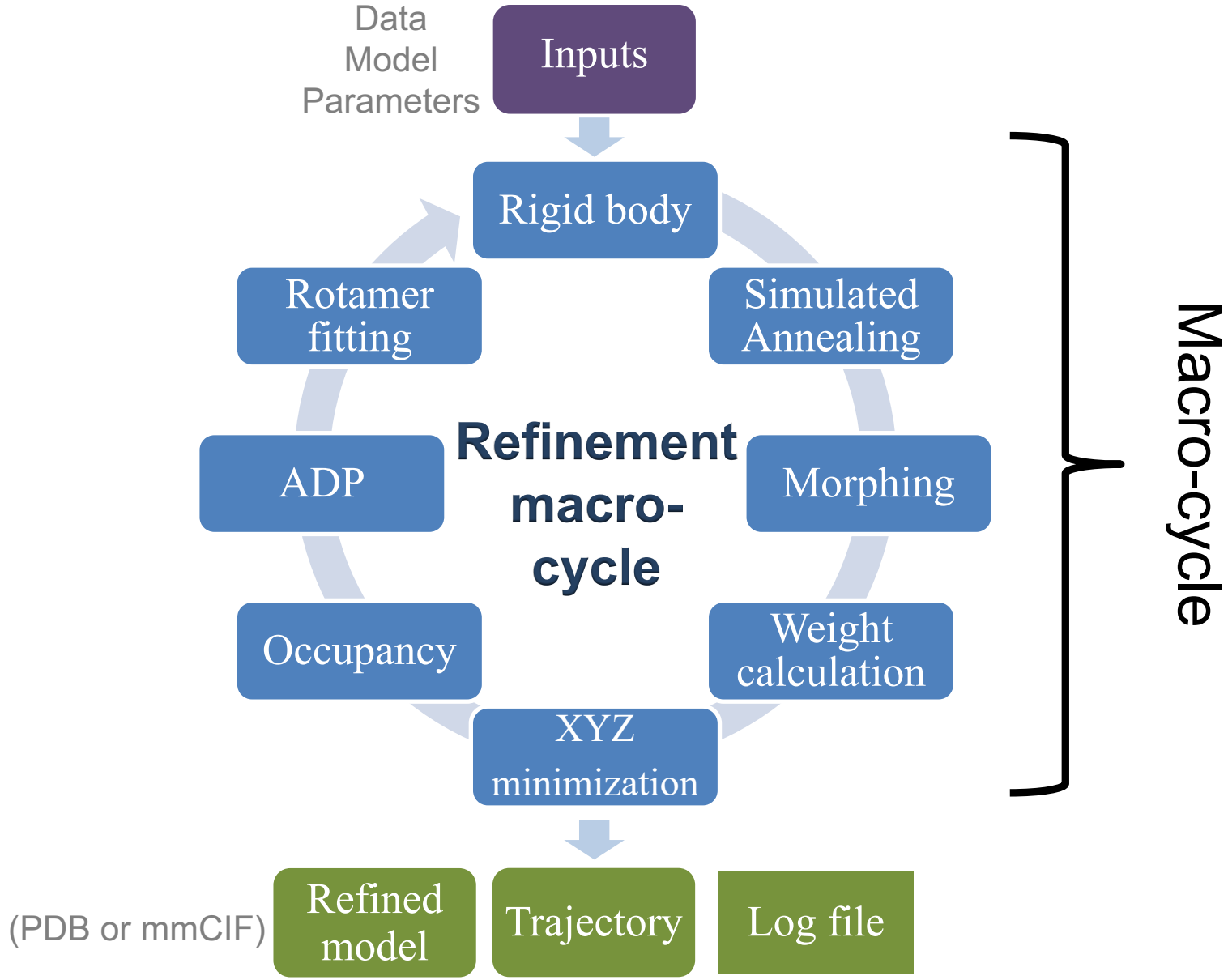
Crystallographic refinement

- 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

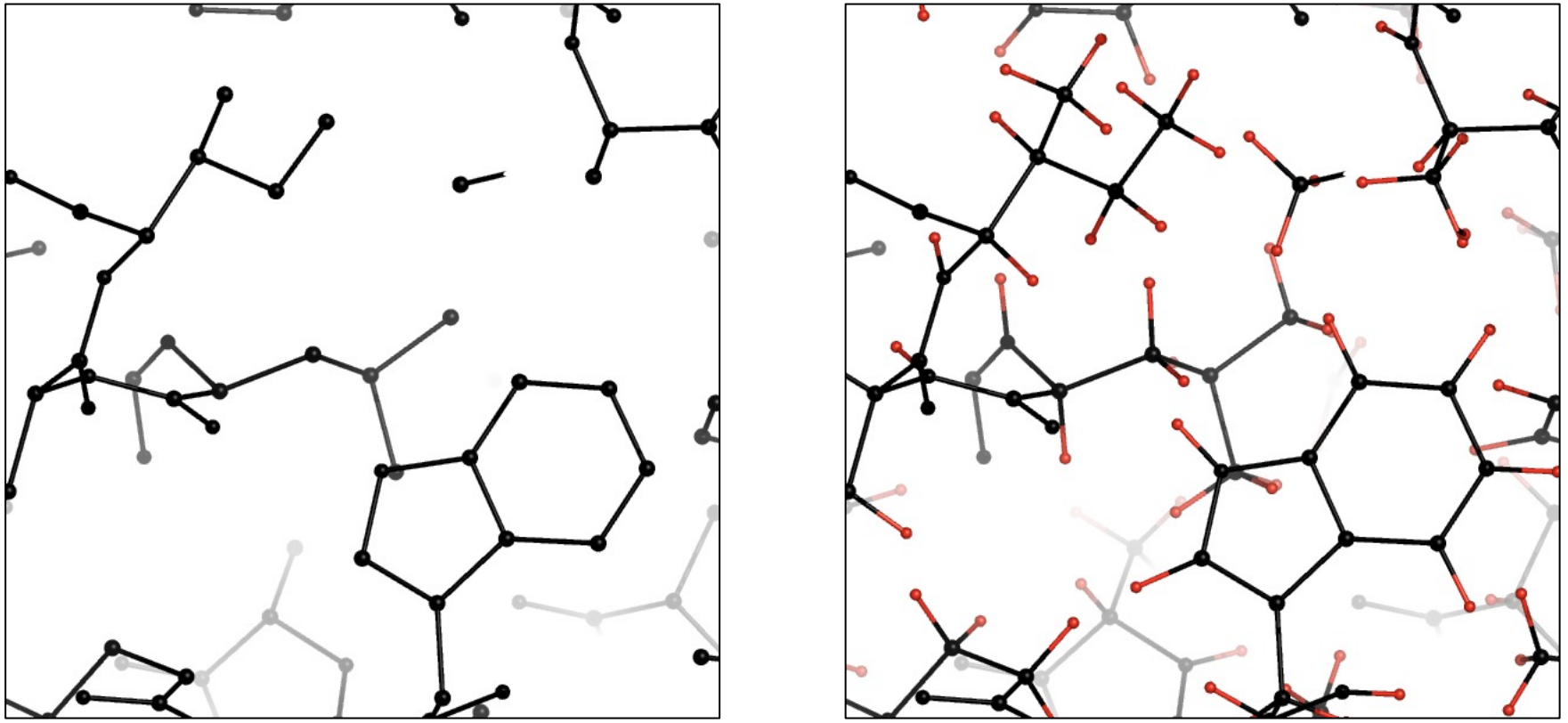
Refinement protocol



Refinement: practical considerations

Use Hydrogen atoms

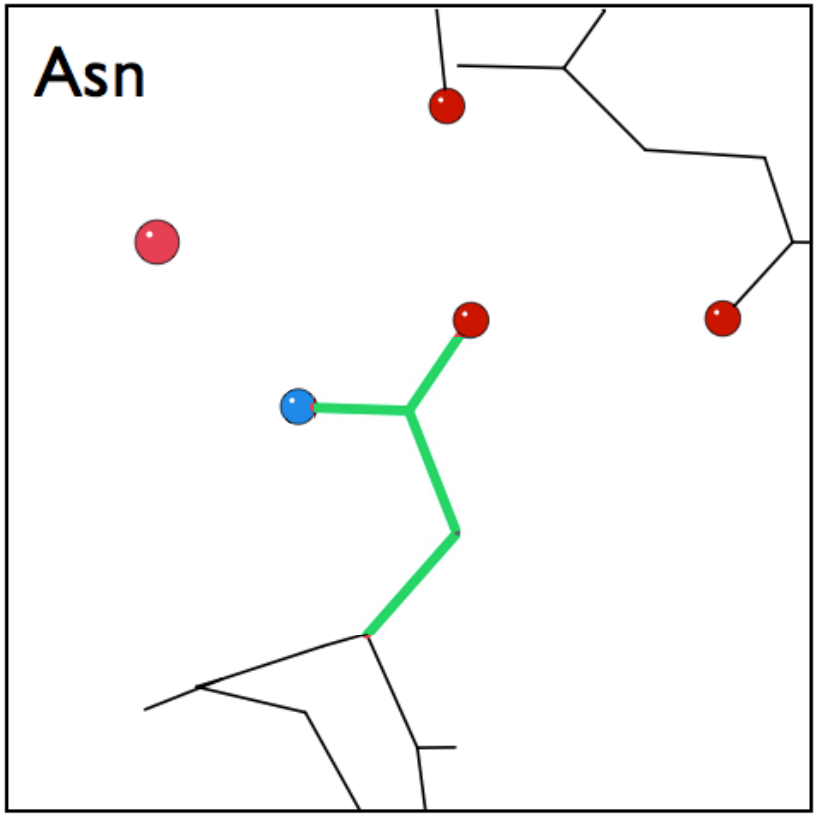
- Half of the atoms in a protein molecule
- Make most interatomic contacts
- Add to model towards the end, data resolution does not matter
- Once added, do not remove before the PDB deposition
- H do contribute to R-factors (expect 0.1-2% drop in R)

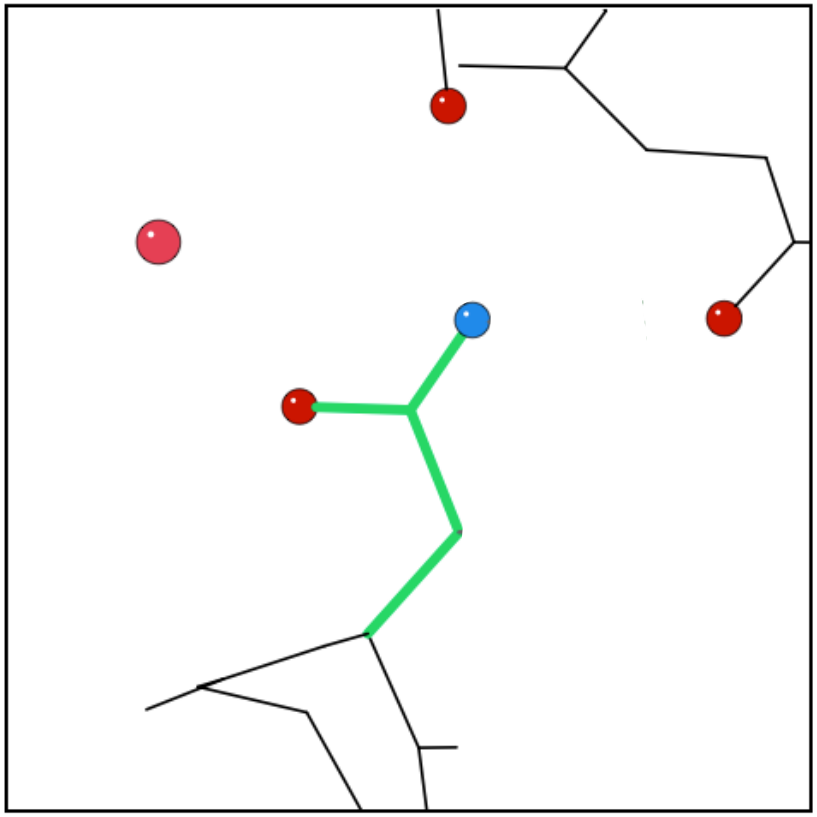


A structure without (left) and with (right) hydrogen atoms

Use Hydrogen atoms

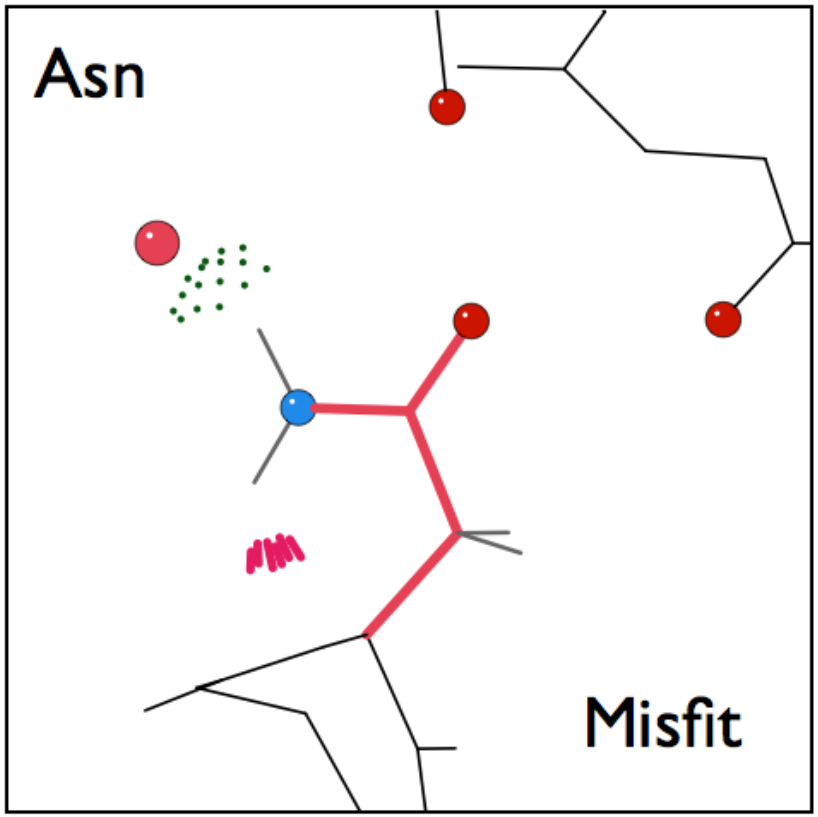
- N/Q/H flips (asparagine/glutamine/histidine)
  - Based on clash analysis
  - Requires H present

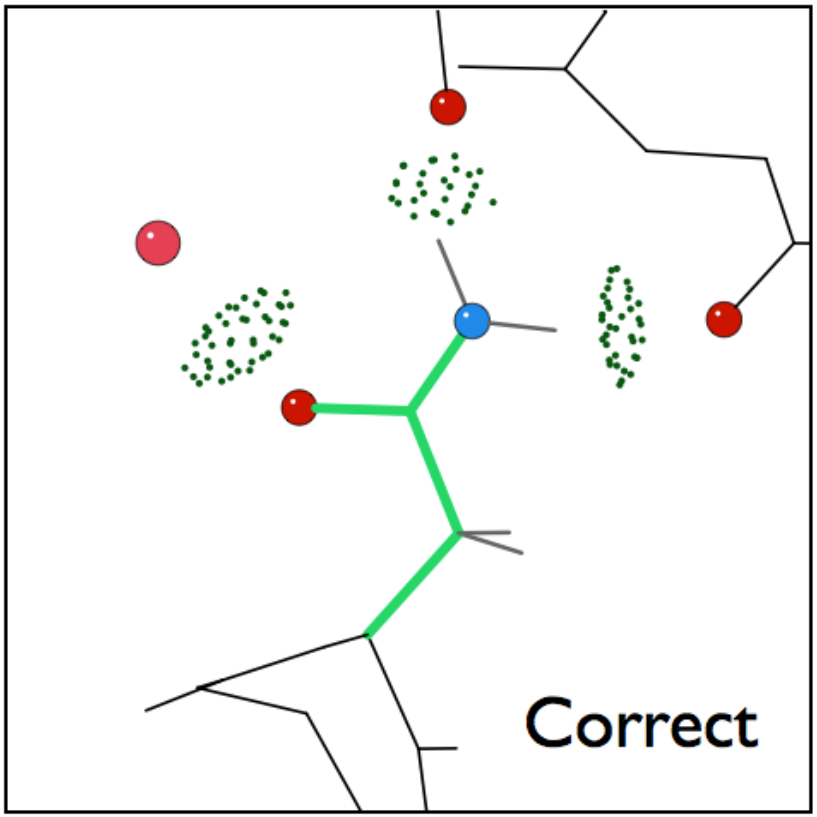




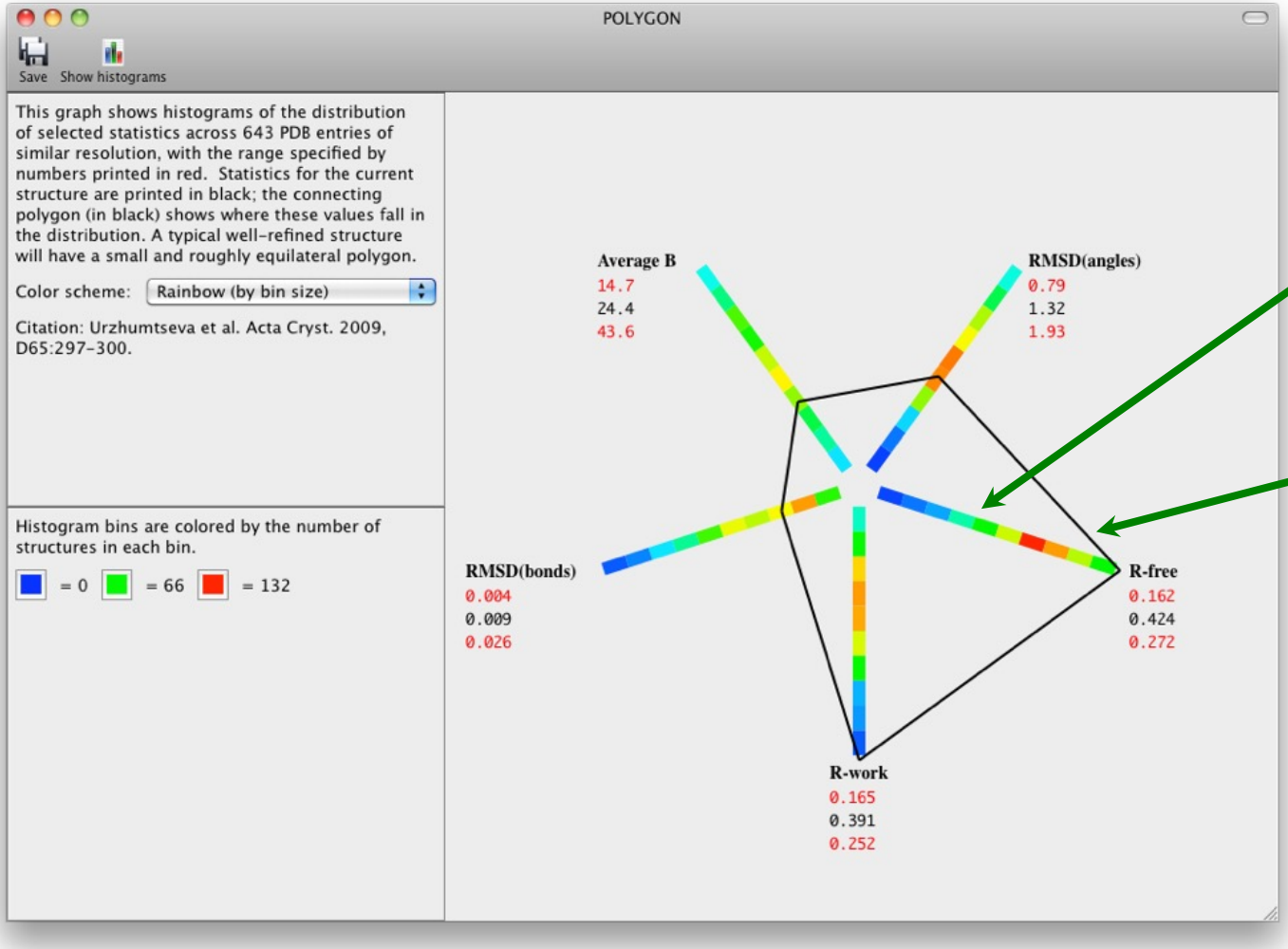
Use Hydrogen atoms

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





Know when to stop



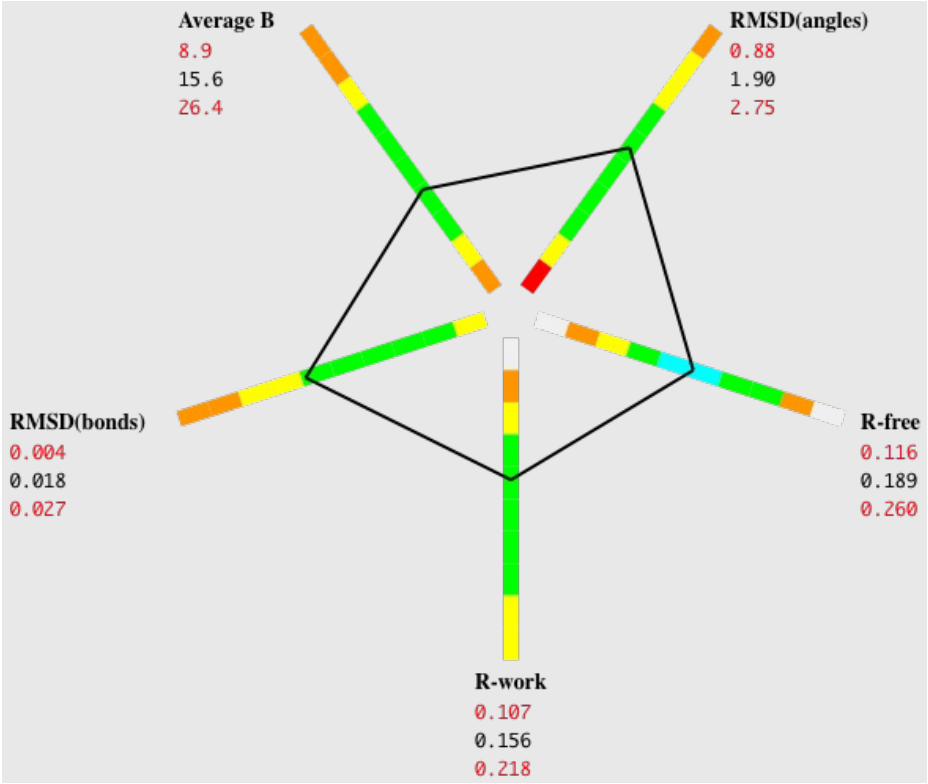
Colored bars are histograms showing distribution of values for structures at similar resolution

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

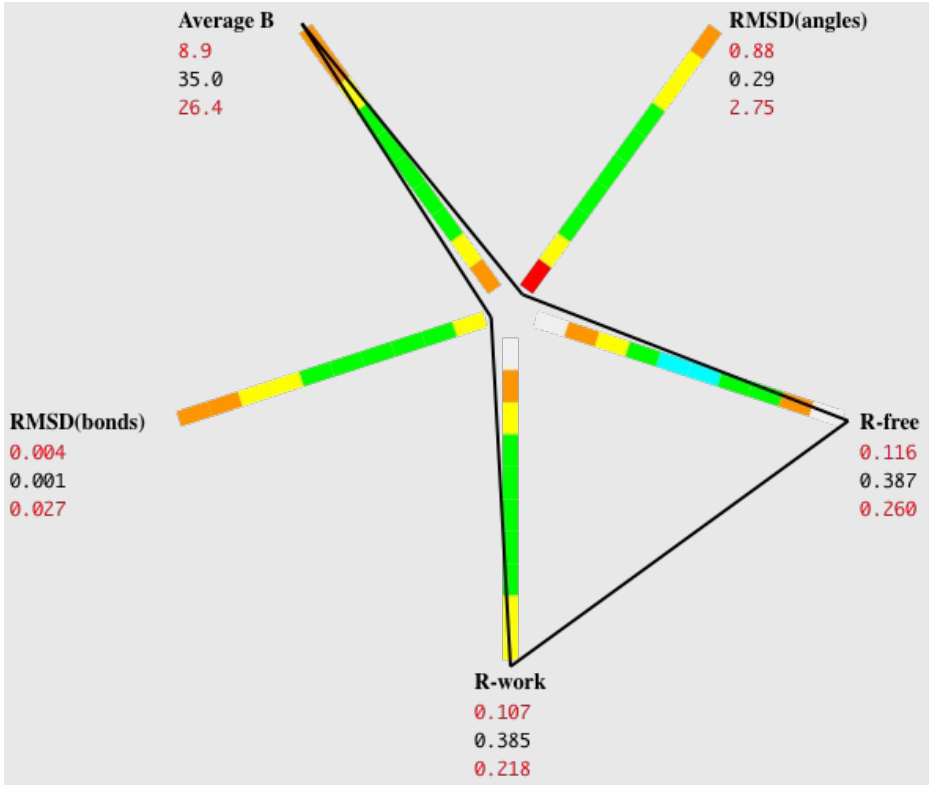
Crystallographic model quality at a glance.
L.Urzhumtseva, P.V.Afonine, P.D.Adams & A.Urzhumtsev. *Acta Cryst*. **D**65, 297-300 (2009)

Know when to stop

Likely overall good model



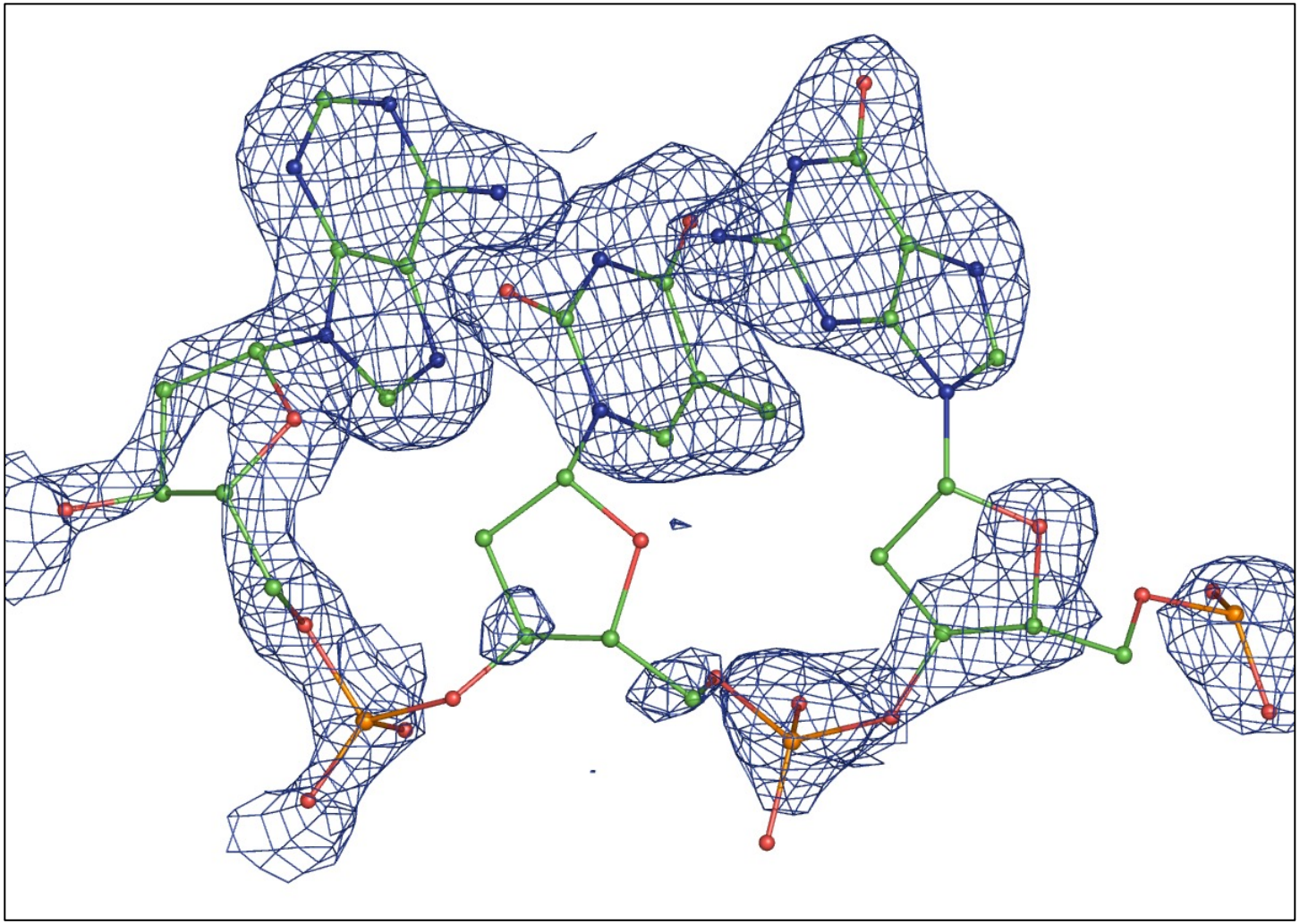
Clearly there are problems



Don’t waste time fixing unfixable

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

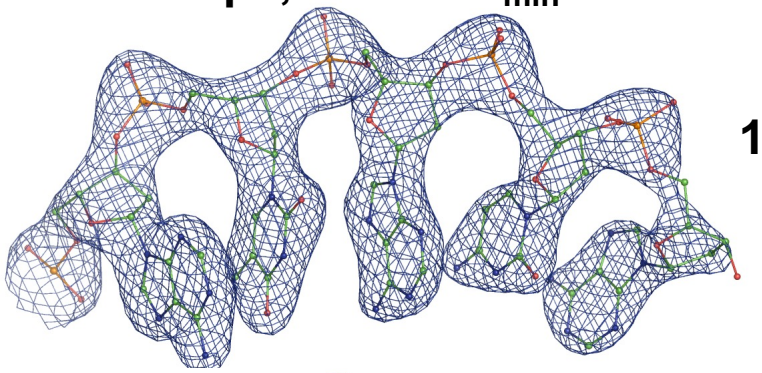
2mFo-DFc , 1σ



Don’t waste time fixing unfixable

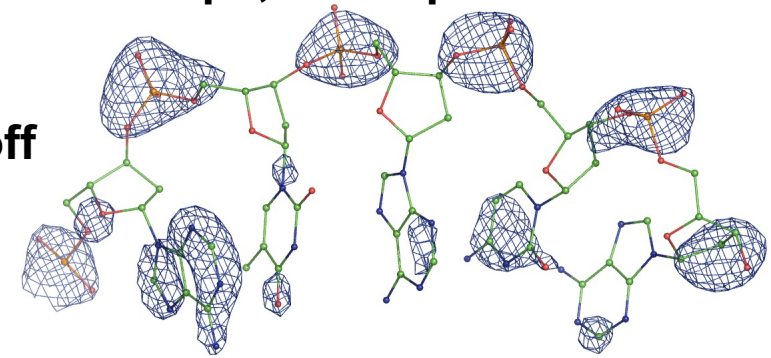
```
Completeness by resolution:
 19.9274 -  3.2441 0.78
  3.2441 -  2.5767 0.99
  2.5767 -  2.2515 1.00
  2.2515 -  2.0459 1.00
  2.0459 -  1.8993 0.99
Overall completeness in dmin-inf: 0.95
```

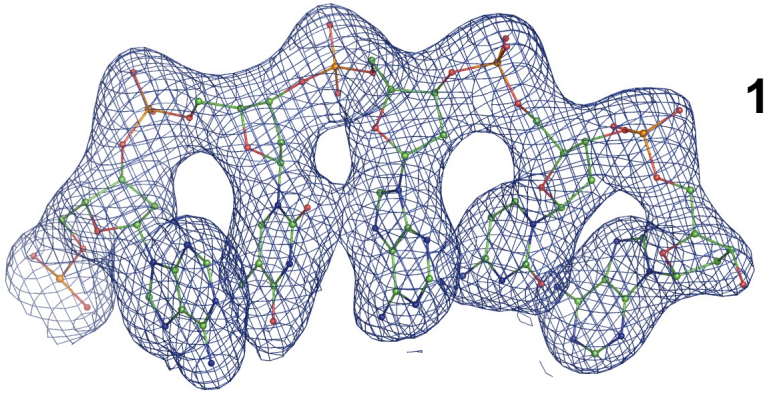
Fcalc maps, full set d_{min}-inf



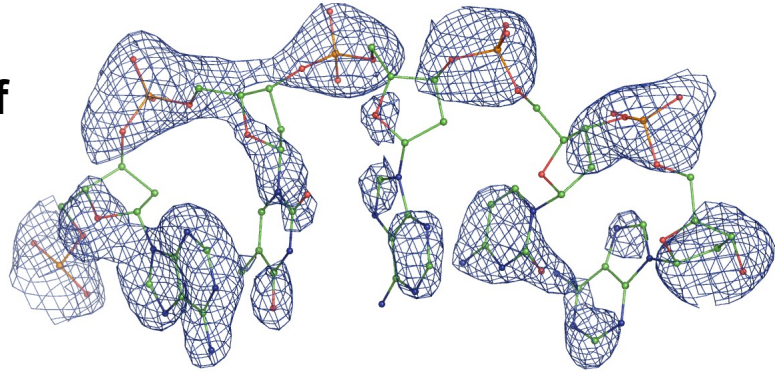
1.5σ map cutoff

Fcalc maps, incomplete set





1σ map cutoff



Data incompleteness distorts maps

Local *vs* Global

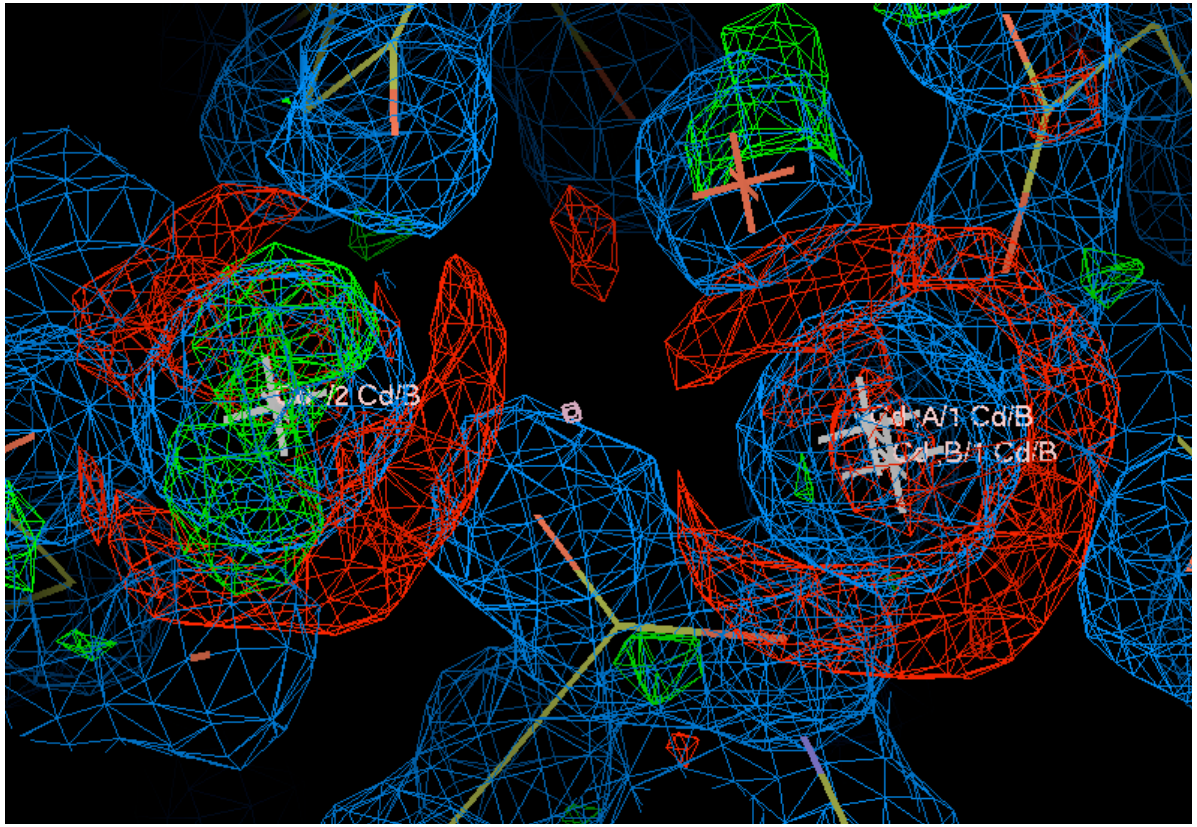
- 2.5Å: R_{WORK}/R_{FREE} = 17.1/21.2% bonds = 0.01Å angles = 1.6°
  - R-factors are great, overall geometry is great, but…

Histogram of deviations from ideal values

Bonds		Angles	
0.000 - 0.035:	2645	0.000 - 9.313:	4208
0.035 - 0.070:	19	9.313 - 18.626:	9
0.070 - 0.106:	13	18.626 - 27.939:	3
0.106 - 0.141:	5	27.939 - 37.252:	4
0.141 - 0.176:	3	37.252 - 46.565:	0
0.176 - 0.211:	0	46.565 - 55.878:	0
0.211 - 0.246:	0	55.878 - 65.191:	2
0.246 - 0.281:	0	65.191 - 74.504:	1
0.281 - 0.317:	2	74.504 - 83.817:	0
0.317 - 0.352:	18	83.817 - 93.130:	8

- Problem with a few atoms, while the rest is ok
  - Poor ligand geometry

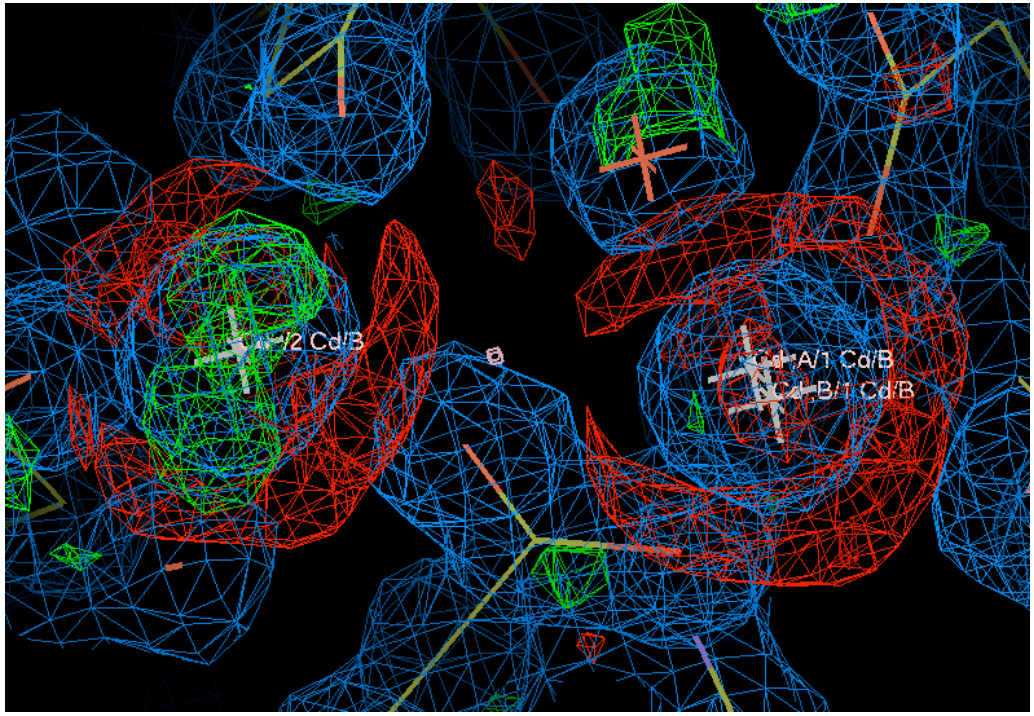
Map and model errors



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

Map and model errors



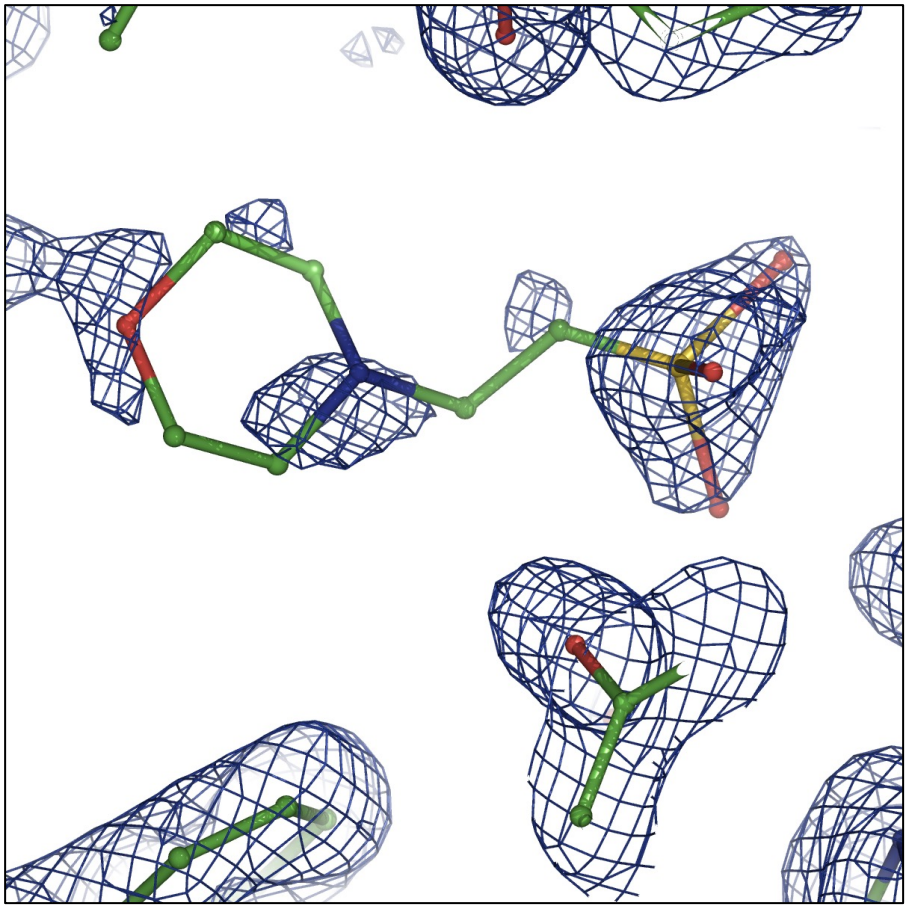
NEW: phenix.oat : try all possibilities, one atom at a time

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

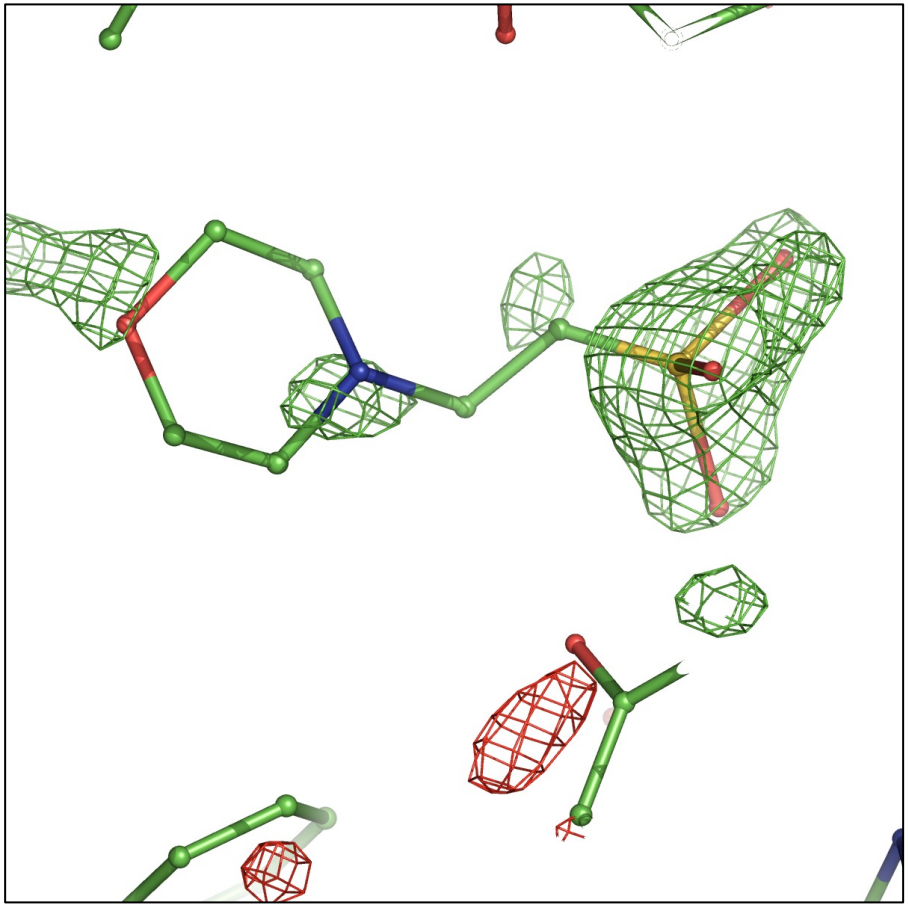
Use the correct map

PDB code: 1ABA, Resolution: 1.45 Å

2mFo-DFc (1σ)



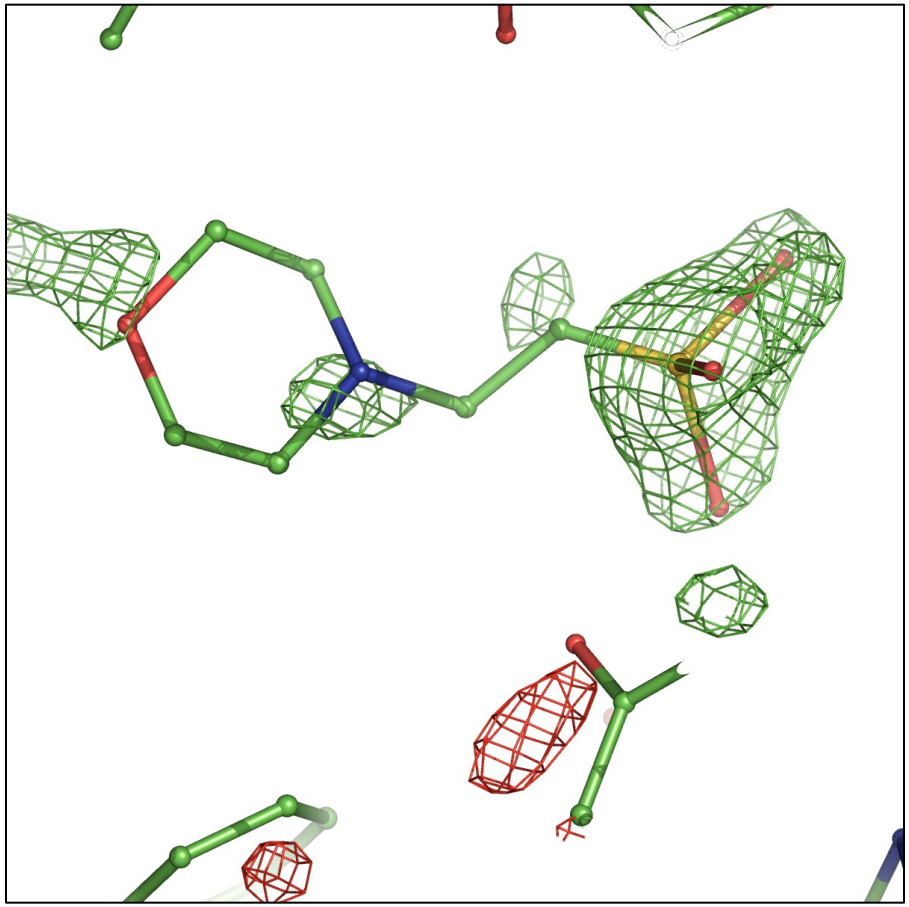
mFo-DFc (±3σ)



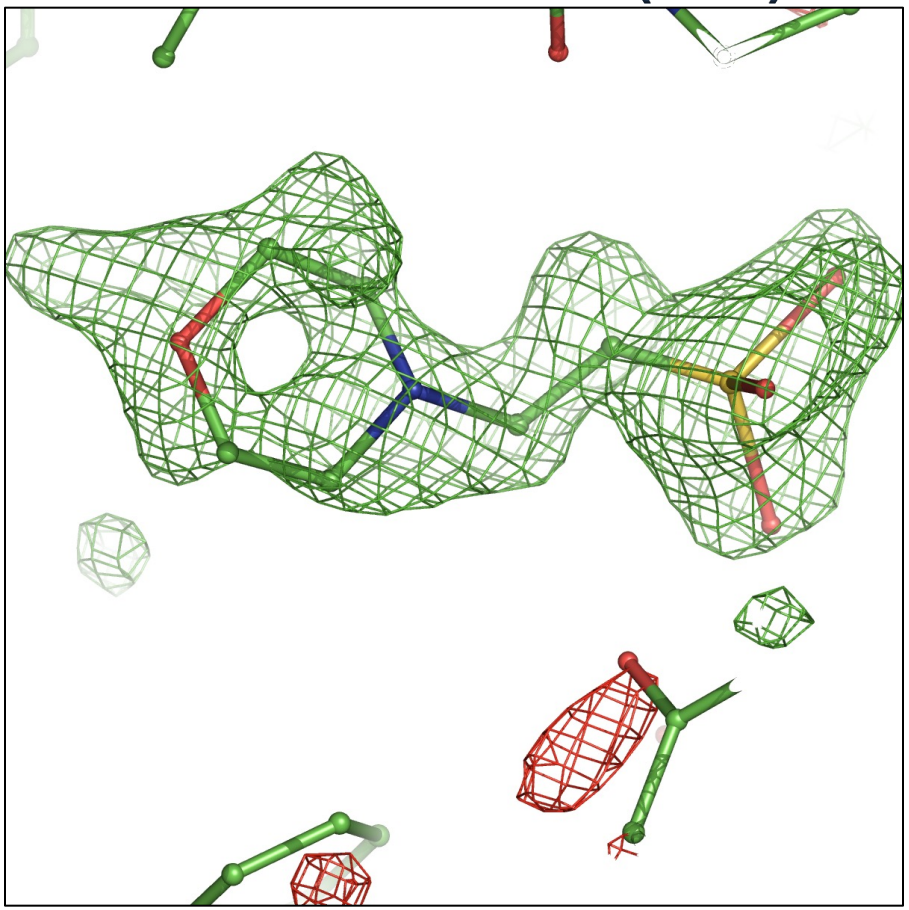
Use the correct map: Polder map

PDB code: 1ABA, Resolution: 1.45 Å

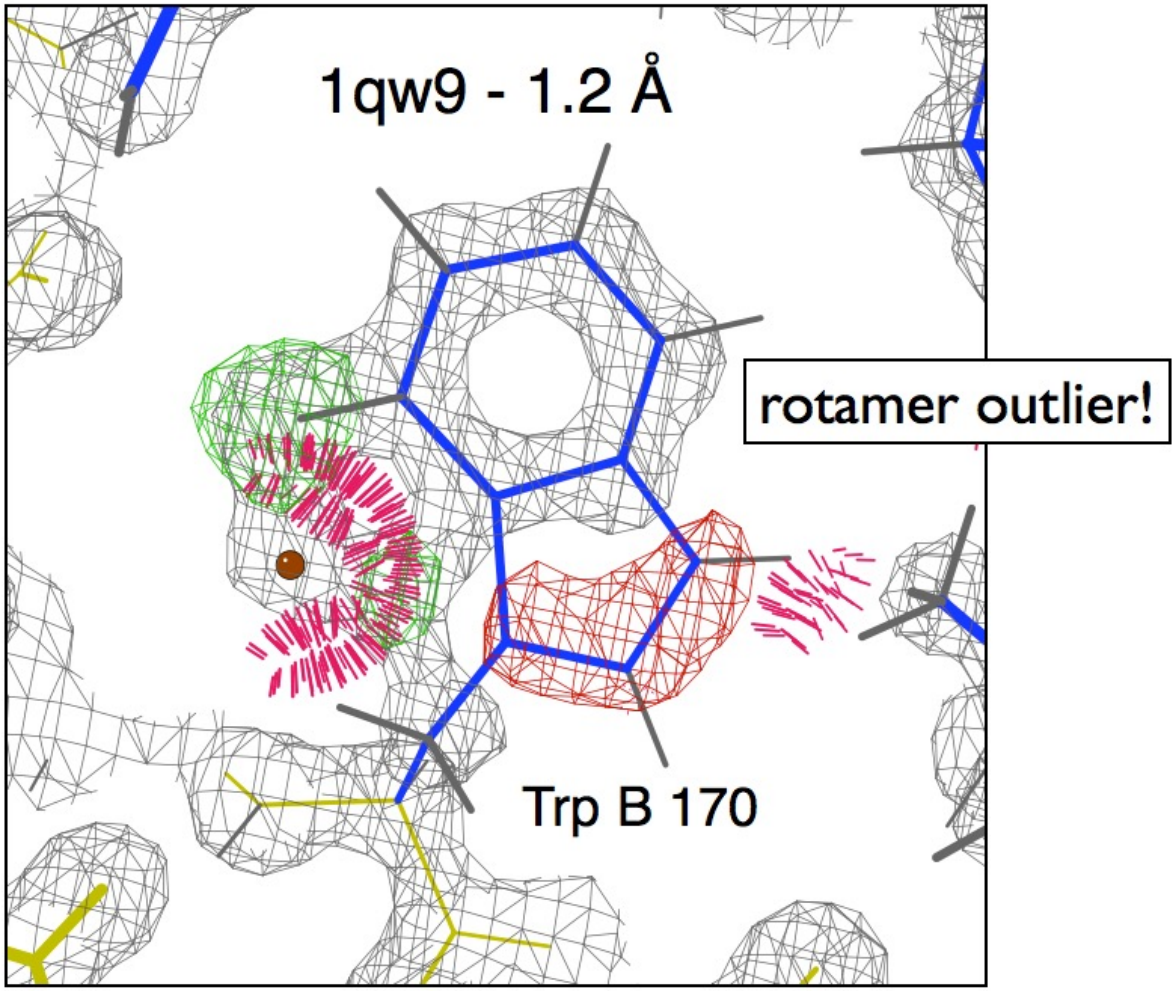
mFo-DFc (±3σ)



Polder mFo-DFc (±3σ)



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)

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

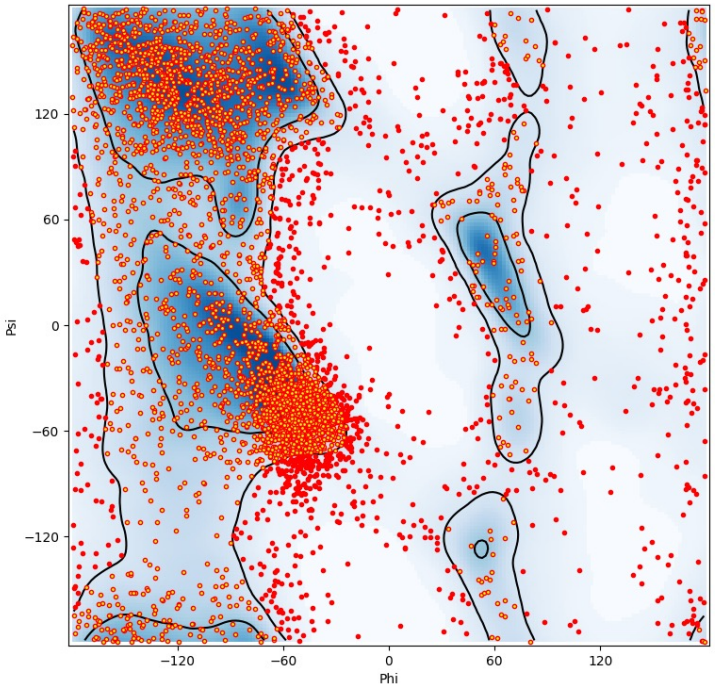
- Likely need at about 3Å and worse
- Better than 3Å: use if needed (preserve good initial model from deterioration)
- Check Ramachandran plot regularly
- Don’t use to fix outliers. Fix outliers first (manually), then use Ramachandran plot restraints to stop re-occurring outliers

Ramachandran plot restraints

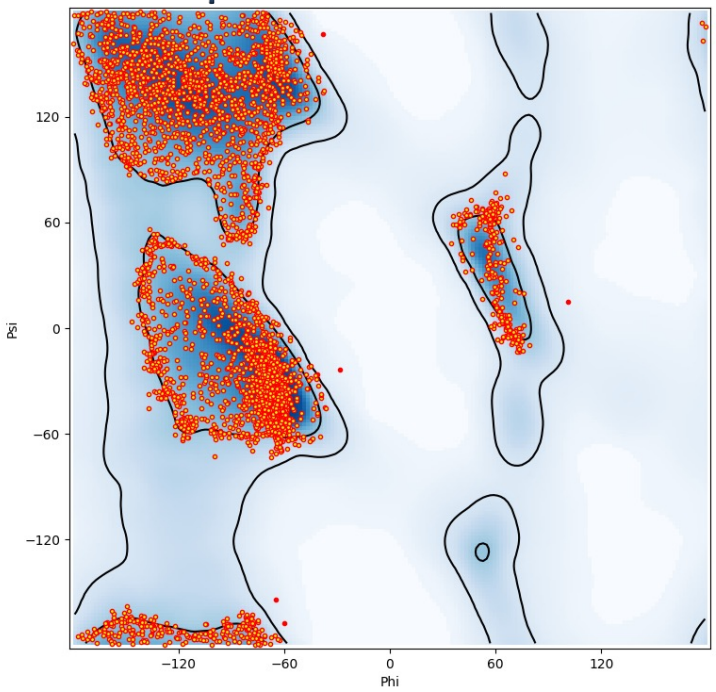
- Ramachandran plot restraints
  - Don’t use to fix outliers. Fix outliers first, then use Ramachandran plot restraints to prevent re-occurring outliers.

PDB code: 5a9z

Original



Refined with Ramachandran plot restraints

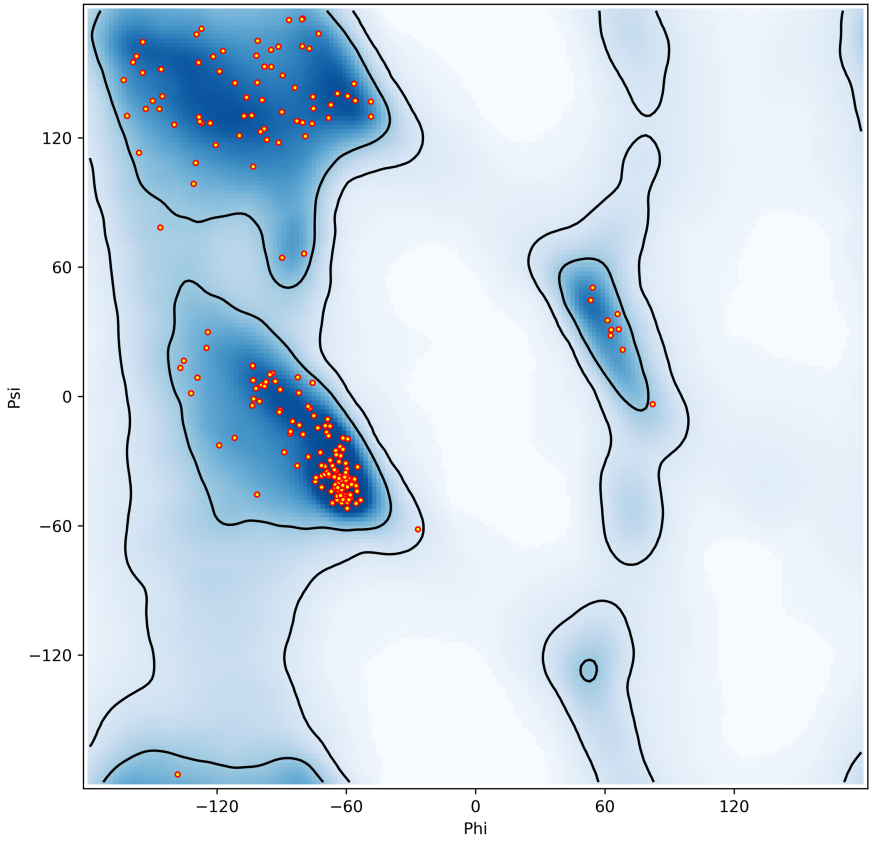


Bad idea to use Ramachandran plot restraints in this case. Fix outliers first!

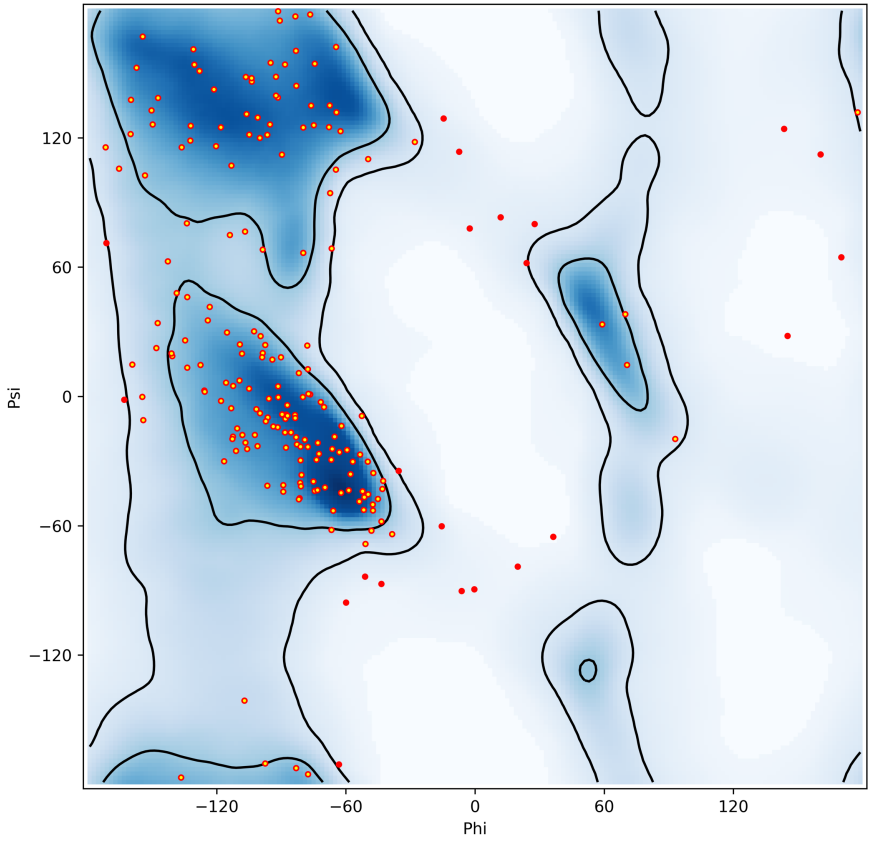
Ramachandran plot restraints

- Ramachandran plot restraints
  - Use to stop outliers from occurring

Before refinement

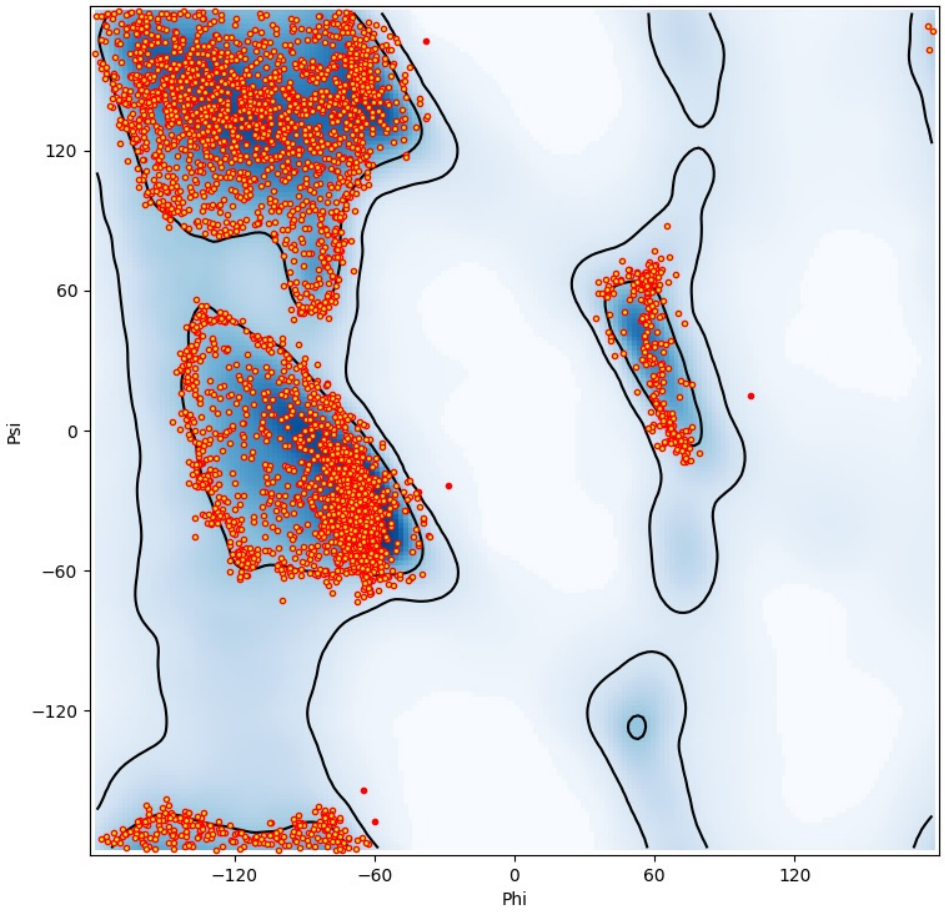


After refinement (No Ramachandran plot restraints)



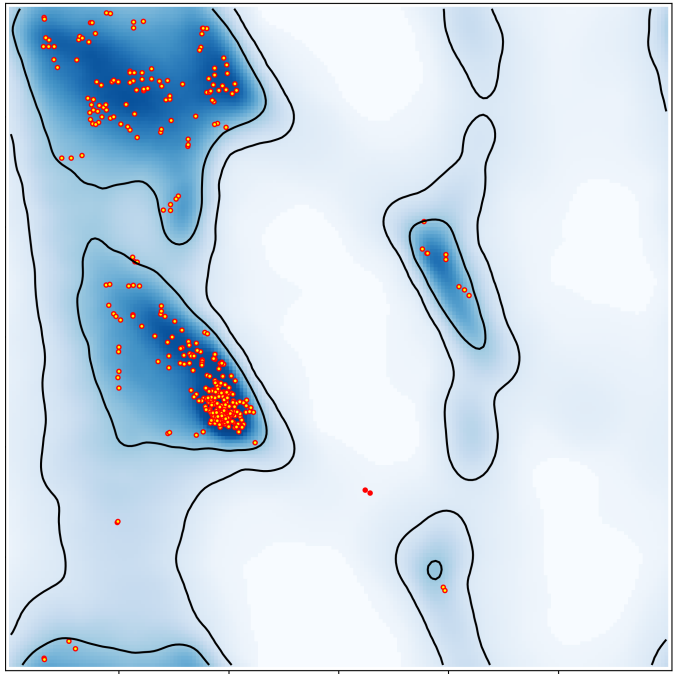
Ramachandran plot restraints

- What is wrong with this plot?



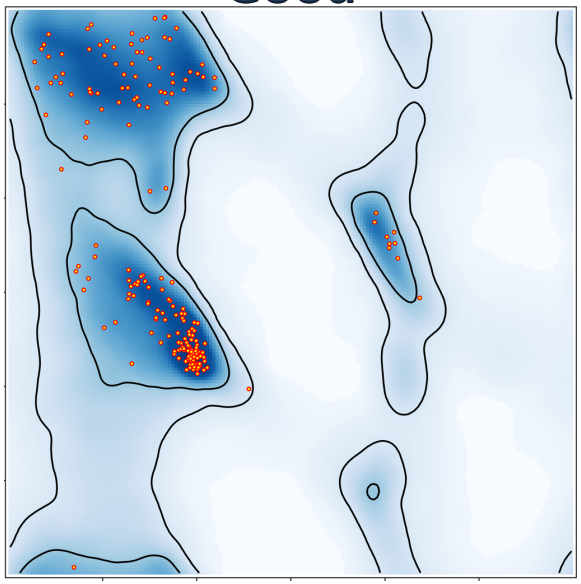
Ramachandran plot restraints

- They are very different from what we expect!

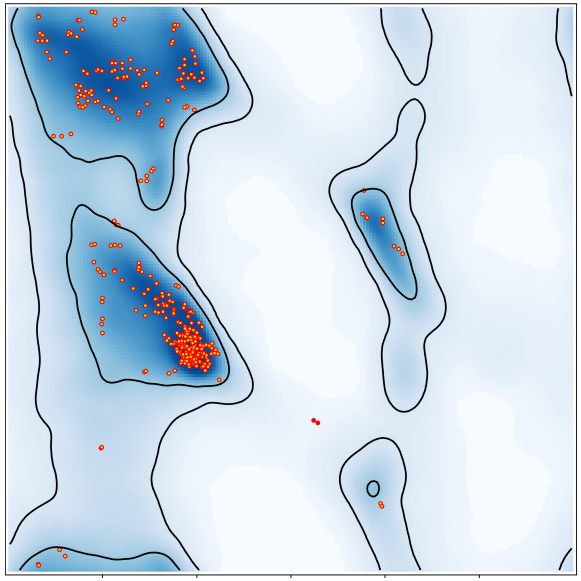


How you can tell good vs bad plot?

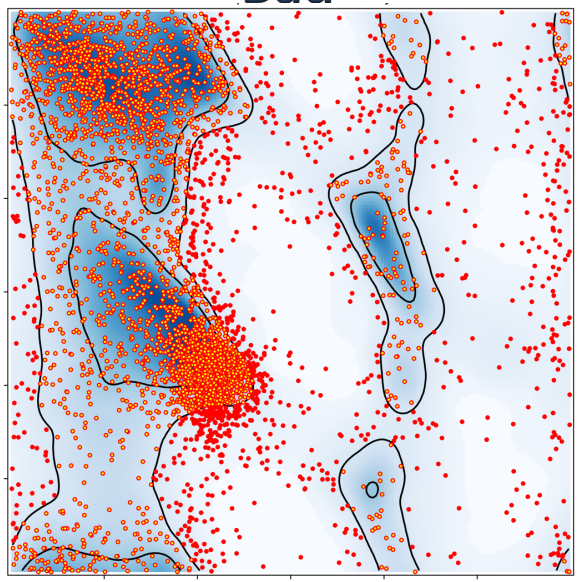
Good



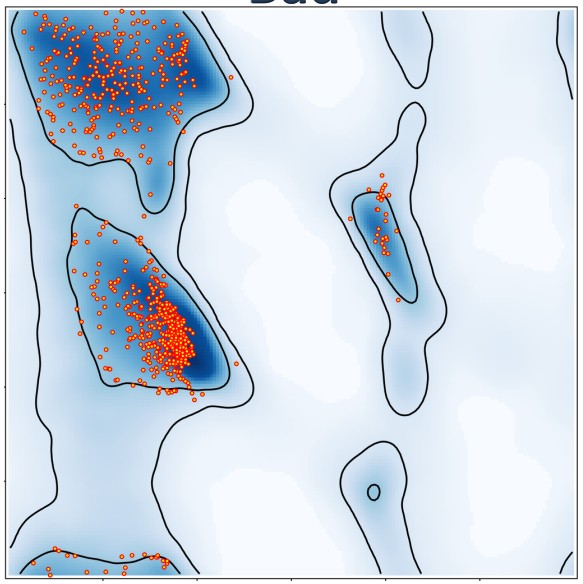
Good



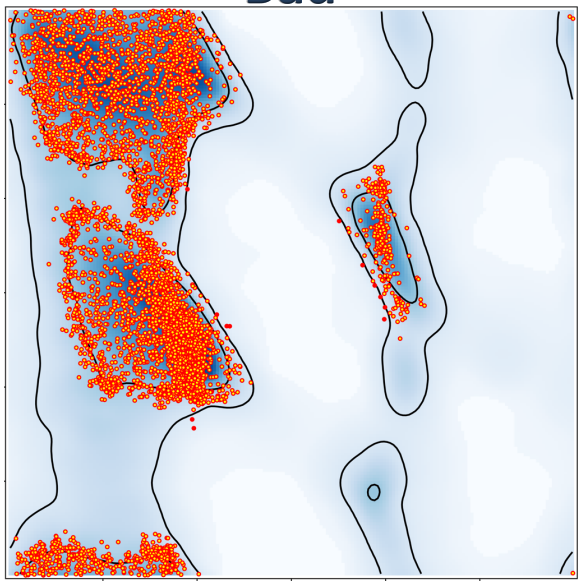
Bad



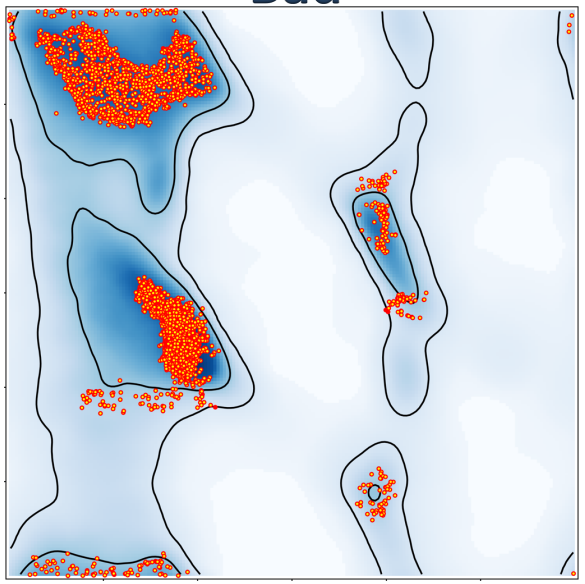
Bad



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

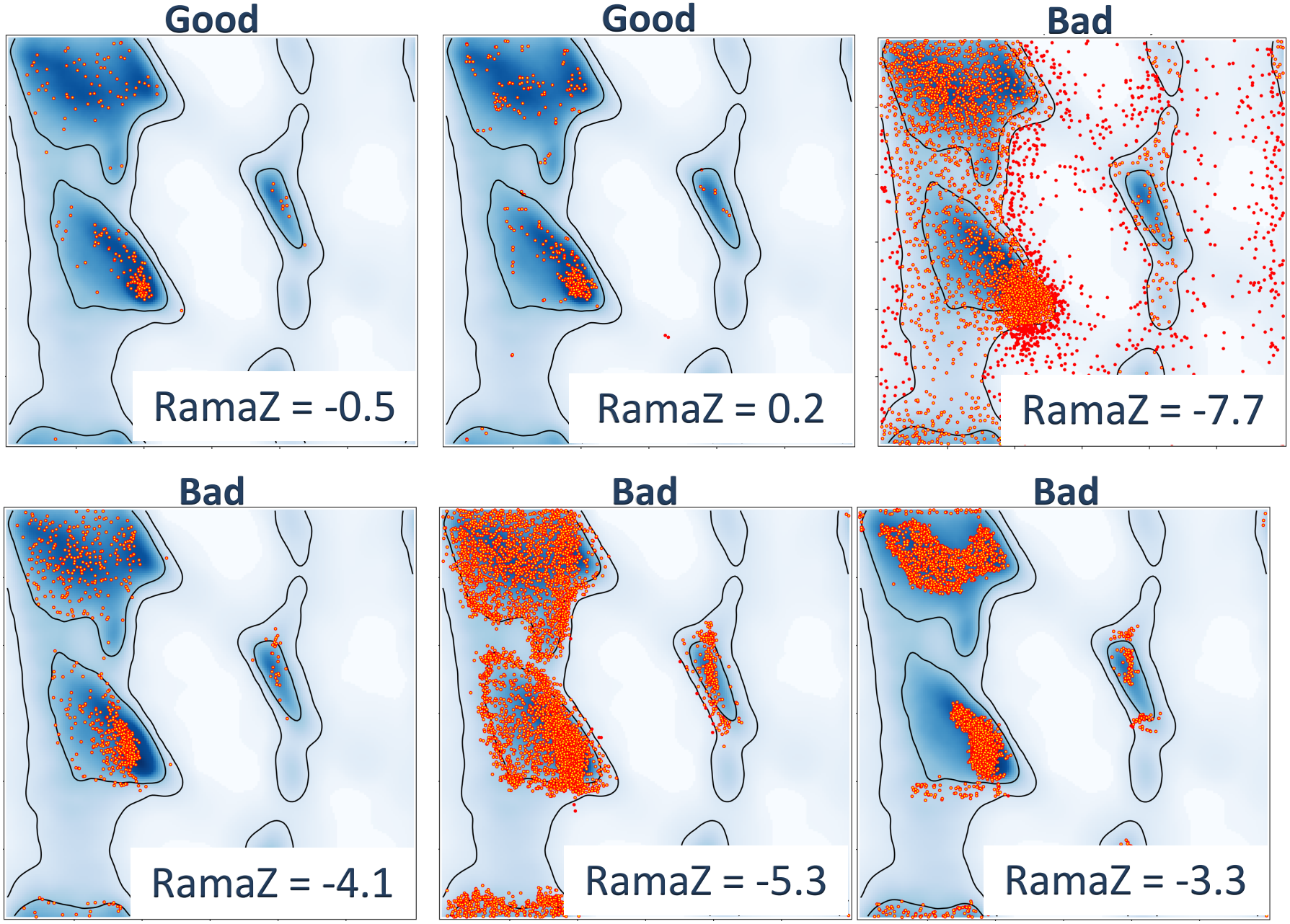
CellPress

Resource

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

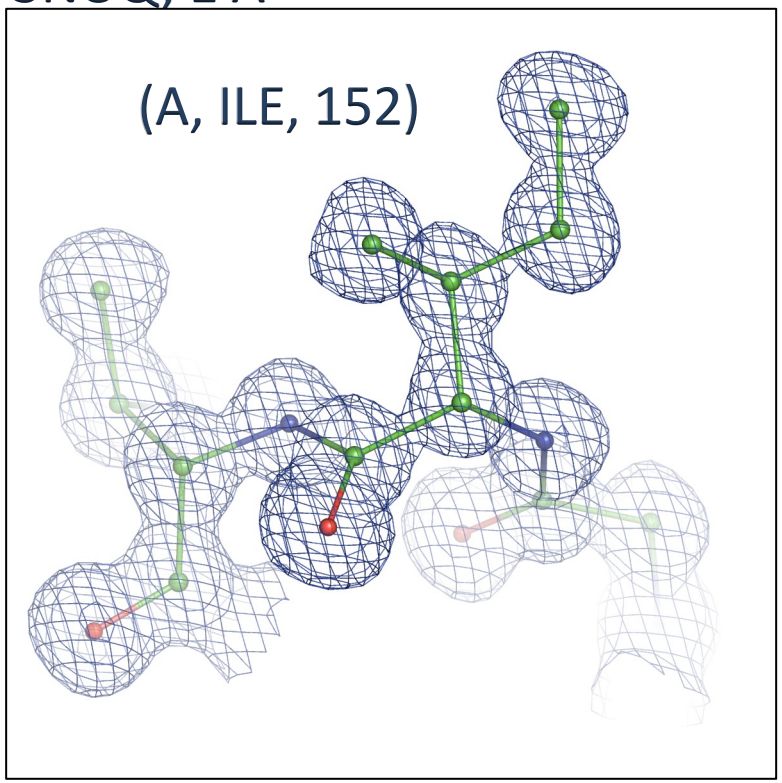
Oleg V. Sobolev,[1,5,*] Pavel V. Afonine,[1] Nigel W. Moriarty,[1] Maarten L. Hekkelman,[2,3] Robbie P. Joosten,[2,3,*] Anastassis Perrakis,[2,3] and Paul D. Adams[1,4]

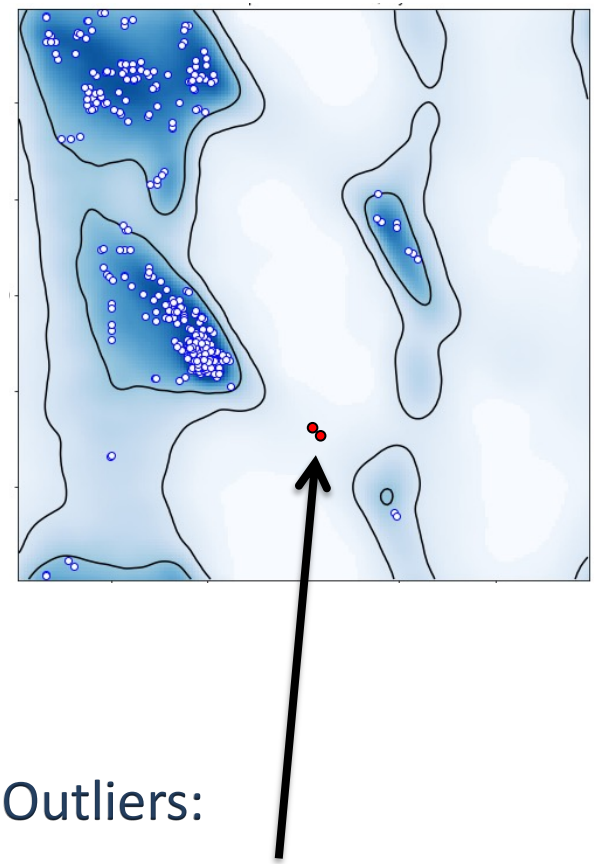
Model validation: *Ramachandran plot Z-score*



An outlier ≠ wrong

3NOQ, 1 Å





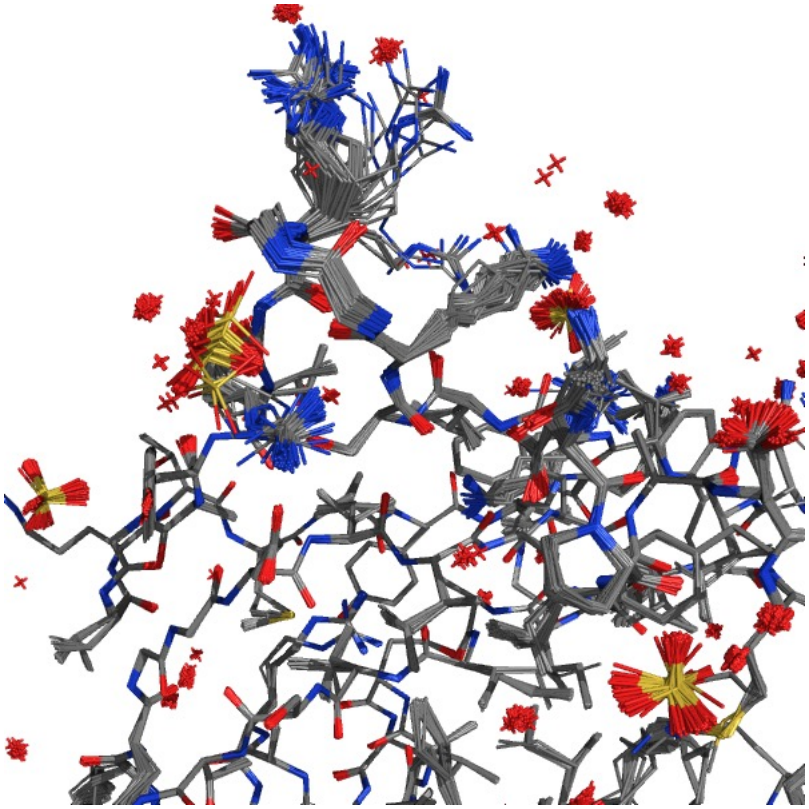
Outliers:

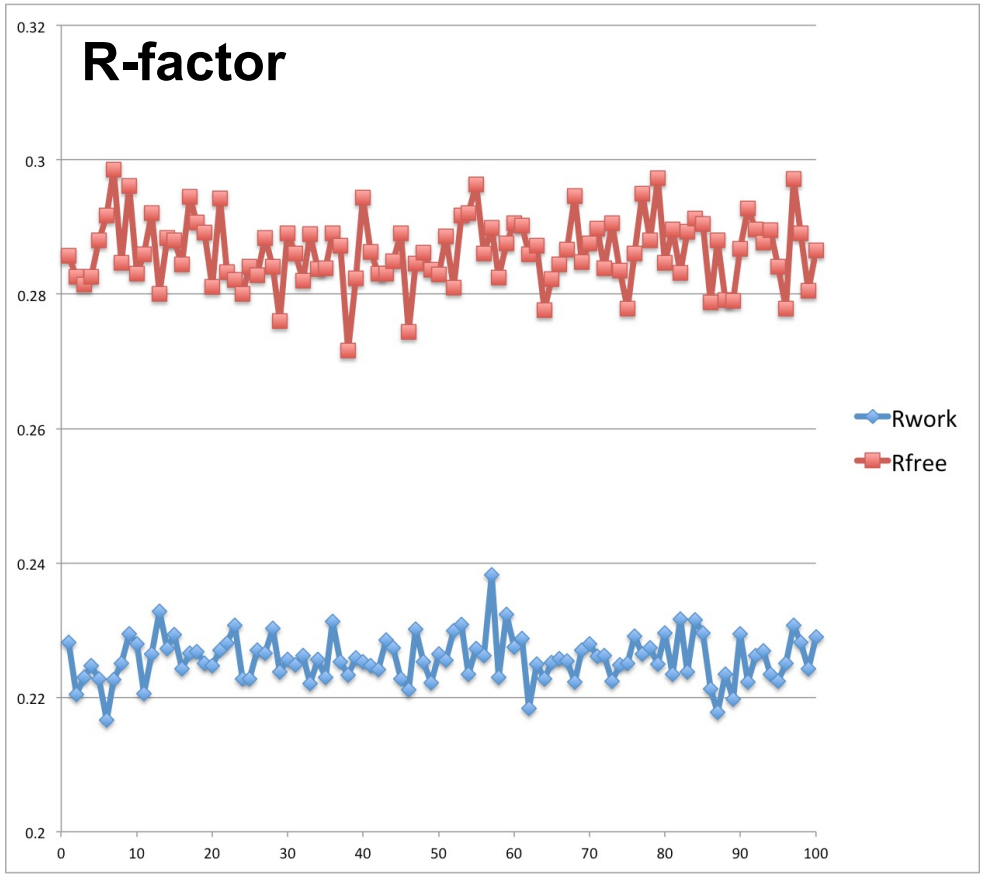
(A, ILE, 152), (B, ILE, 154)

- 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: practical considerations

- Final stages
  - Make the model as complete as possible
  - Build alternative conformations
  - Use Hydrogen atoms (and keep them in the final model!)
  - Add ordered solvent components
- Remember: better model = better map
  - You may see and model your ligands better!

Reading

RESEARCH PAPERS

Acta Cryst. (2018). D**74**, 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). D**68**, 352-367
https://doi.org/10.1107/S0907444912001308
Cited by 2576

OPEN ACCESS

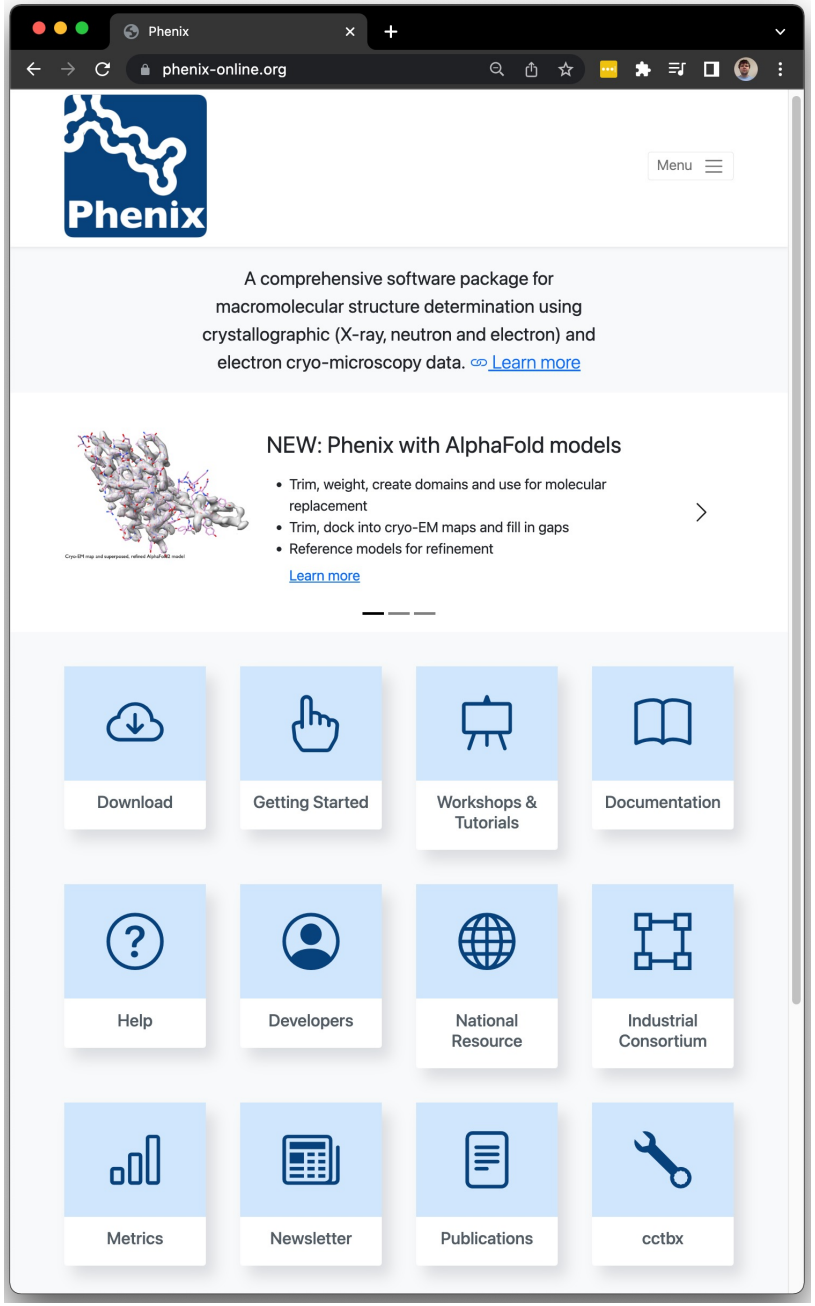
Part of *CCP4 Study Weekend 2011*

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

Phenix resources



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 Phenix Project

Lawrence Berkeley Laboratory

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

Los Alamos National Laboratory
New Mexico Consortium

Tom Terwilliger, Li-Wei Hung

UTHealth

Matt Baker, Corey Hyrc

University of Cambridge

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

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