Model Refinement

Pavel Afonine

phenix-online.org

lbl.gov

qrefine.com

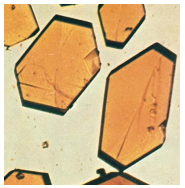
Duke University, NC

September 12th 2024

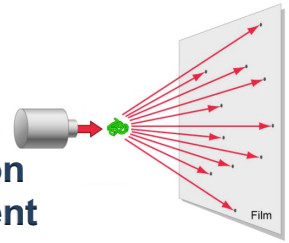
Solving structure by crystallography

Pure sample

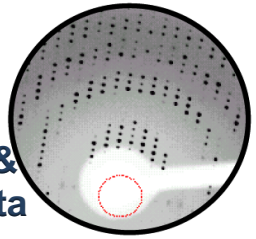
Crystals



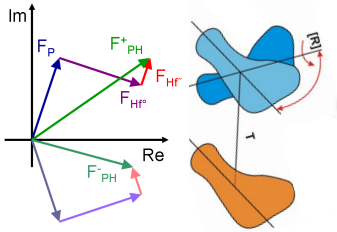
Diffraction experiment



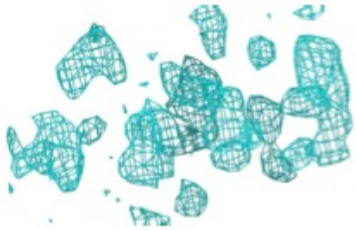
Process & check data



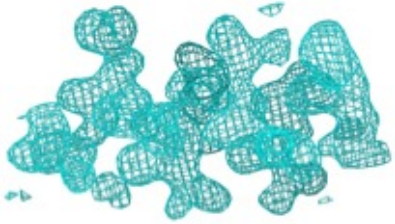
Solve phase problem



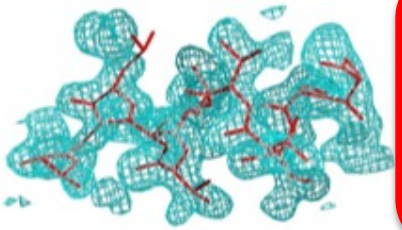
Initial map



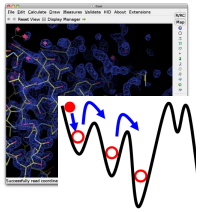
Improved map



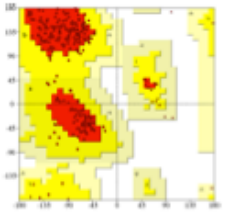
Build model



Refine model



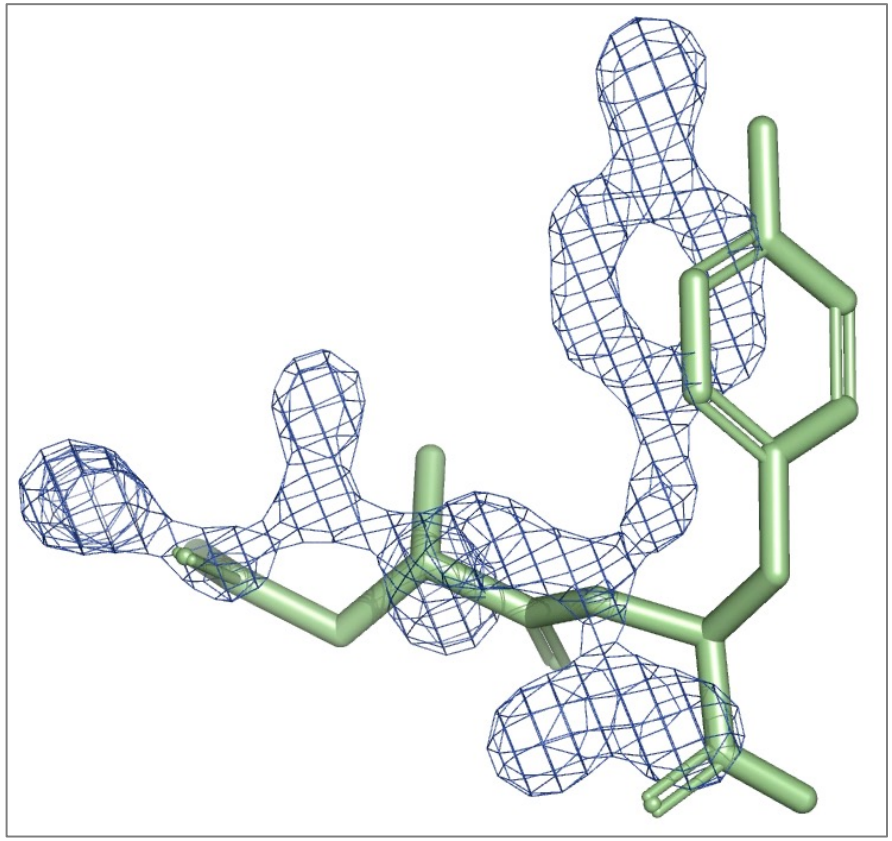
Validate model

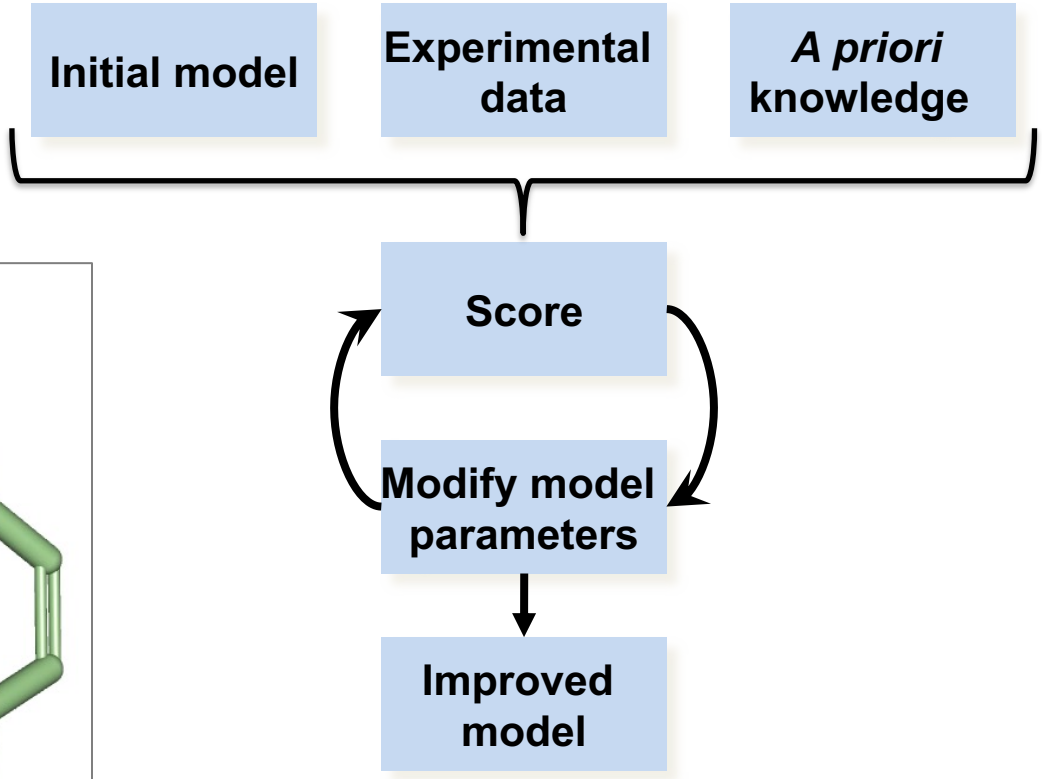


Publish & deposit

- Process is not as ‘linear’ as shown
- Each step has numerous sub-steps
- Crystals may not grow or exhibit pathologies
- Stuck solving phase problem

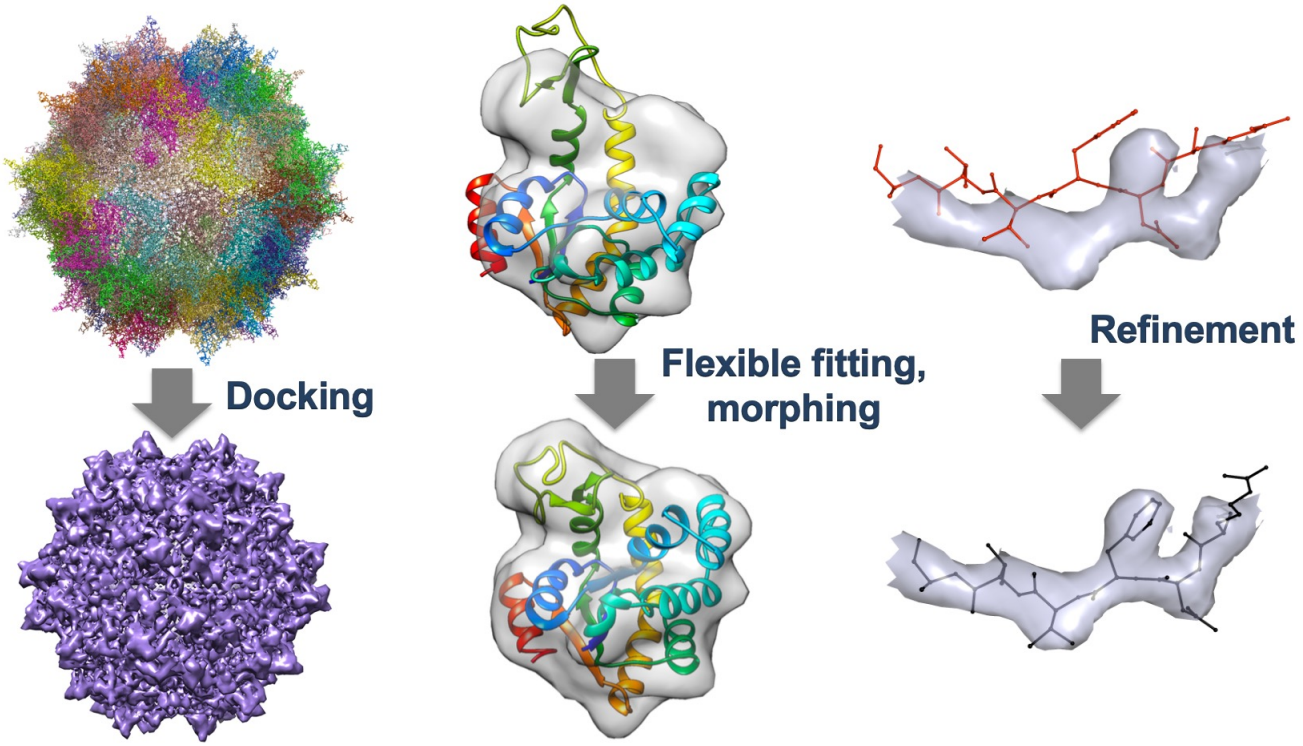
Model refinement





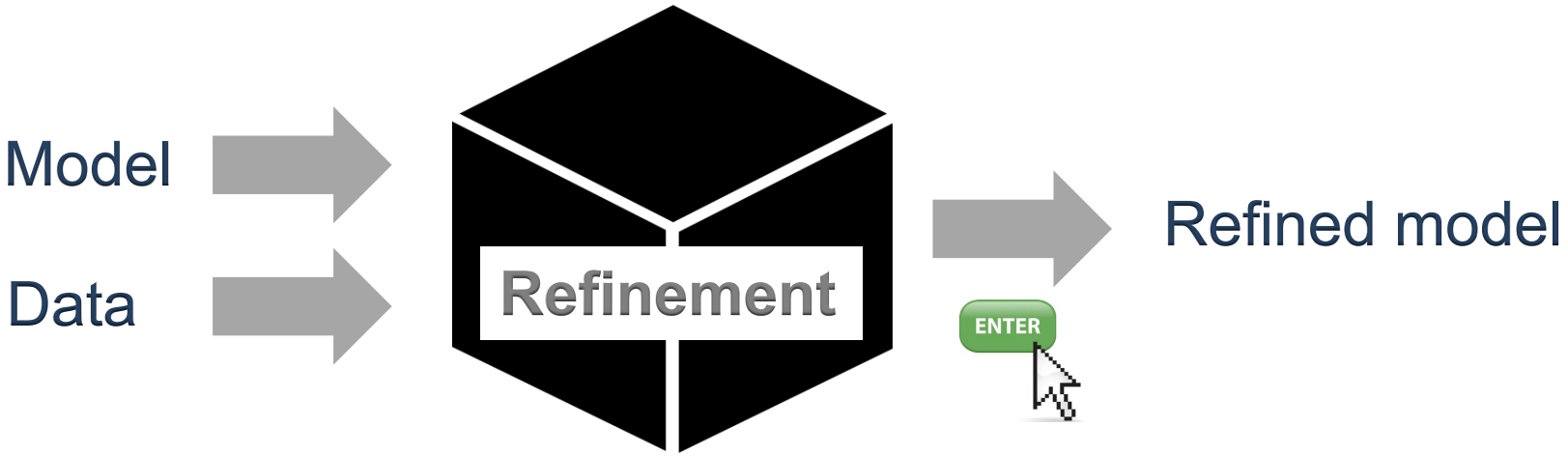
Optimization process of fitting atomic model parameters to experimental data

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)

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
  - 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: lot 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?

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!)

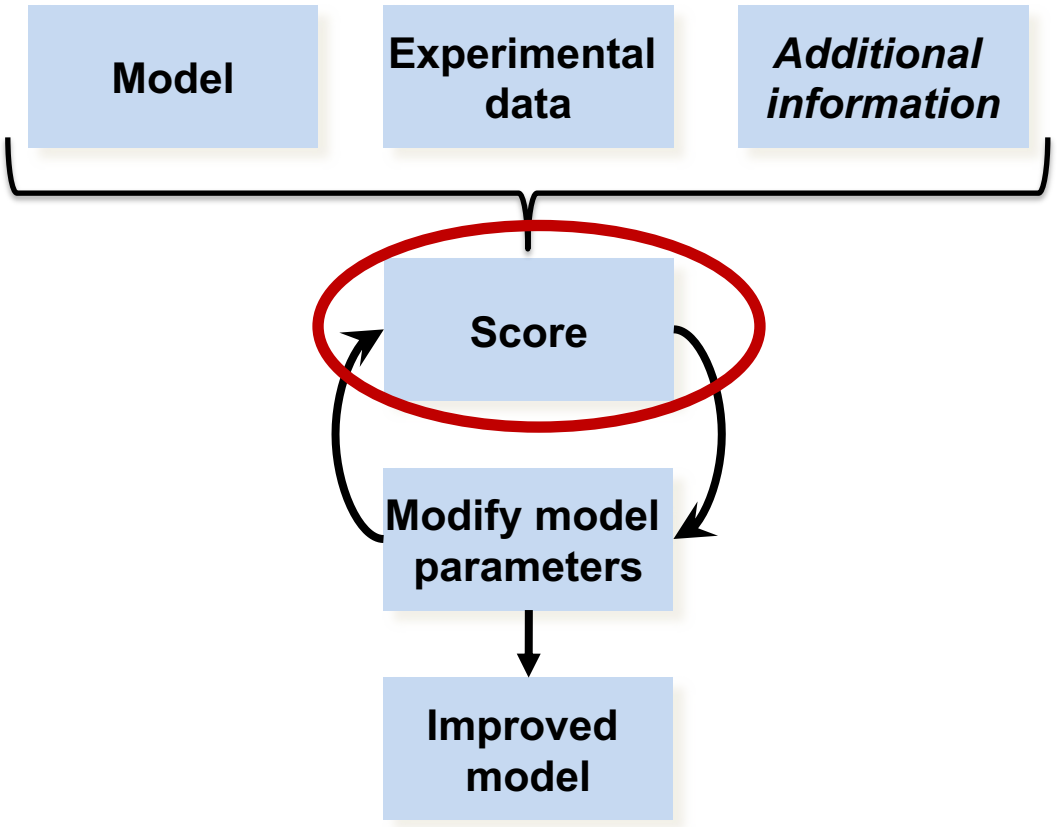
How you know…

- … 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

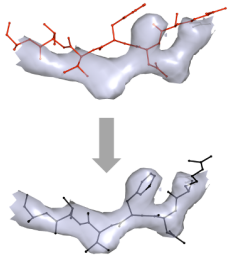
Refinement target function (score)

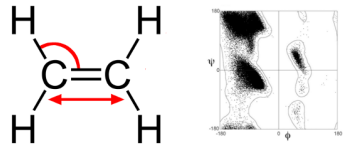


Restraints and data resolution

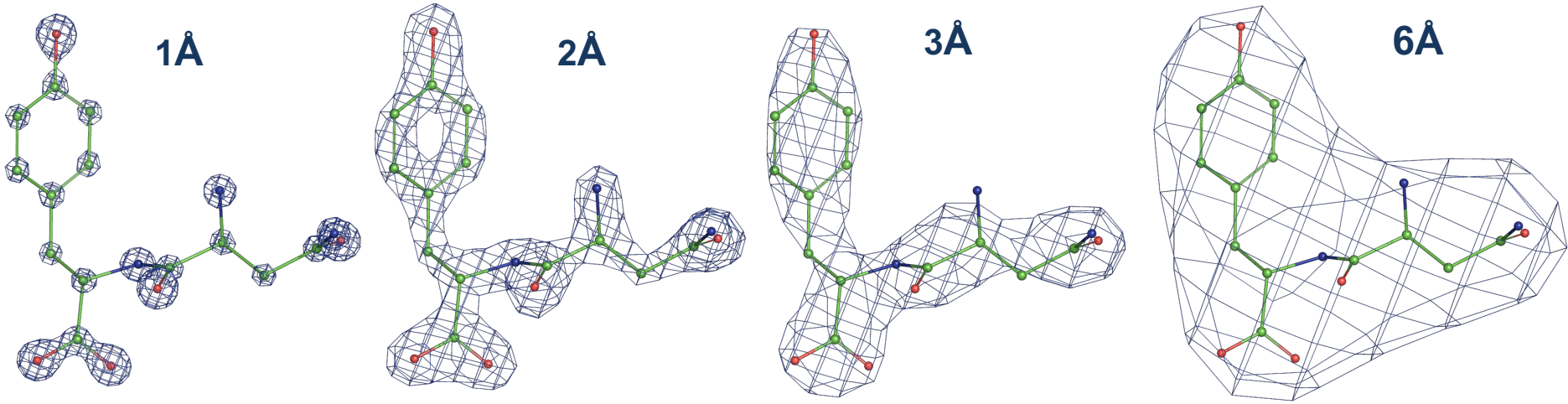
$T = T_{DATA} + w * T_{RESTRAINTS}$

Optimize consensus between model-to-data fit and... common sense



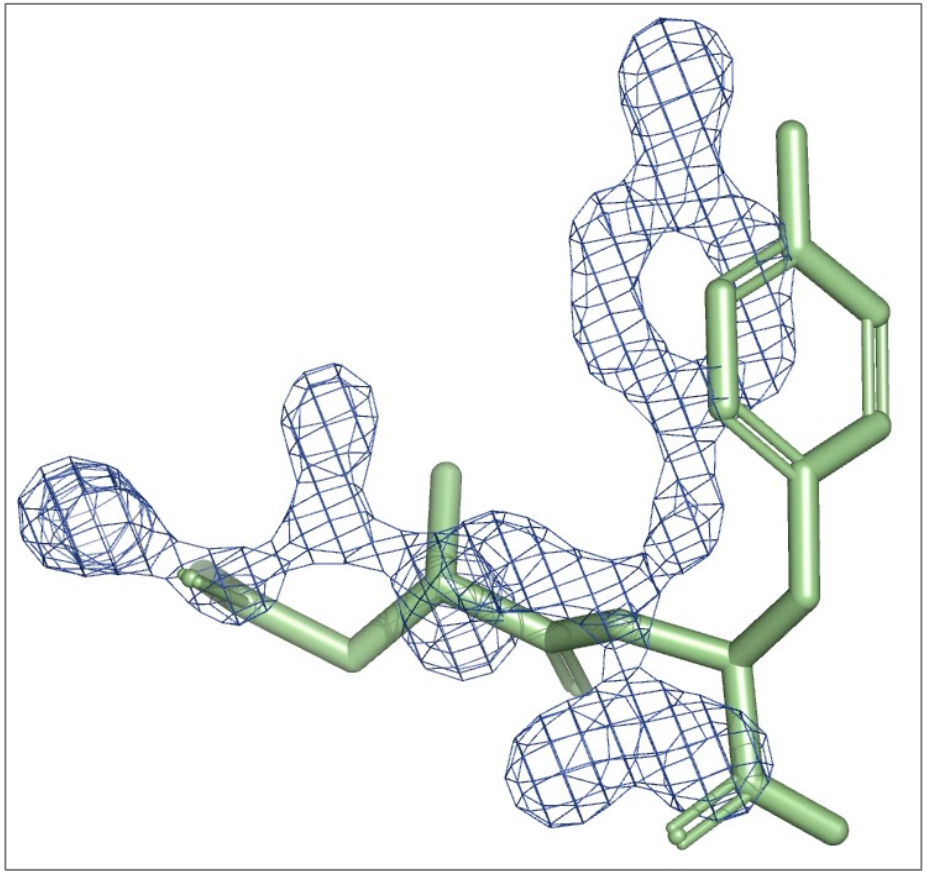


Bonds, angles, planes, torsions, chirality, non-bonded repulsion

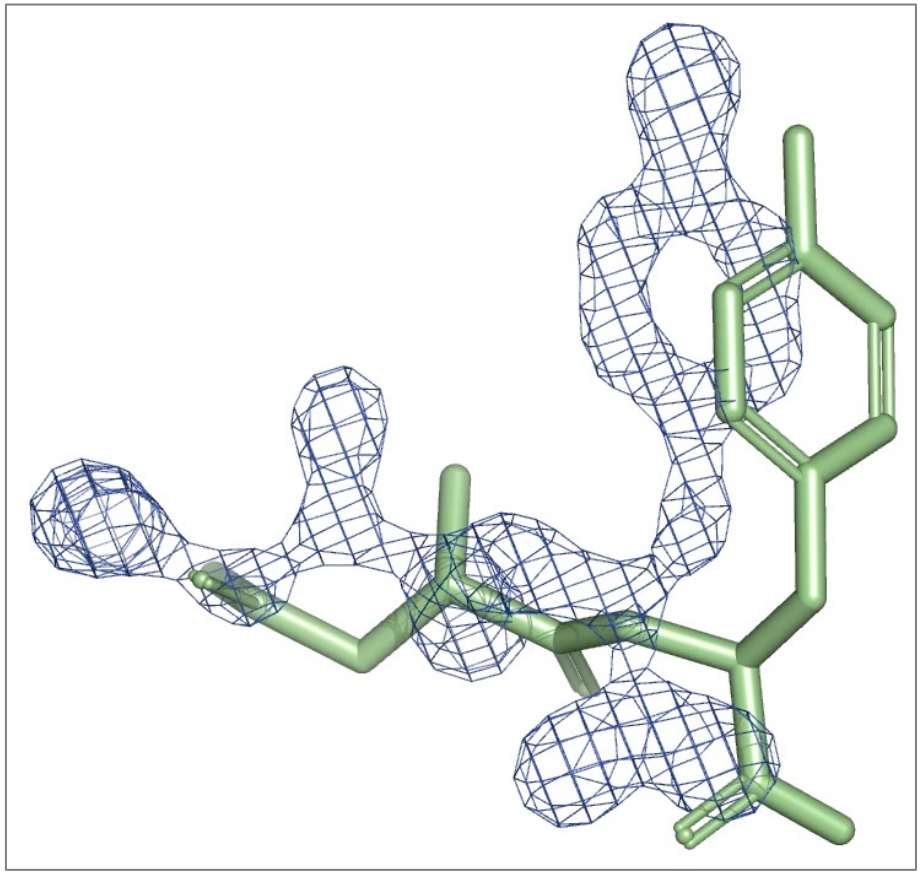


Model refinement with *vs* no restraints

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



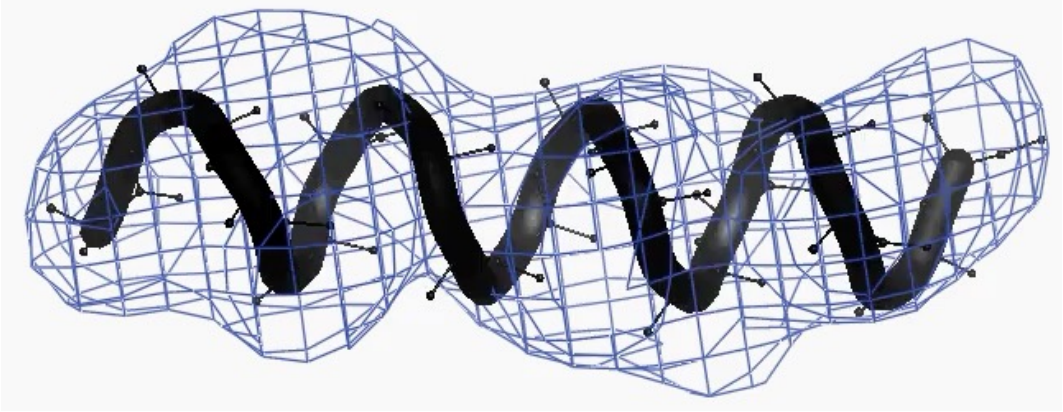
Using restraints



No restraints

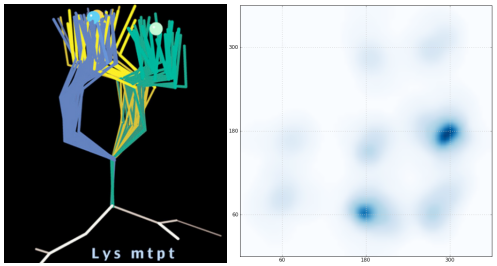
Model refinement with insufficient restraints

- Refinement of a perfect α-helix into low-res map
  - Using simplistic (standard) restraints on covalent geometry
    - Model geometry deteriorates as result of refinement

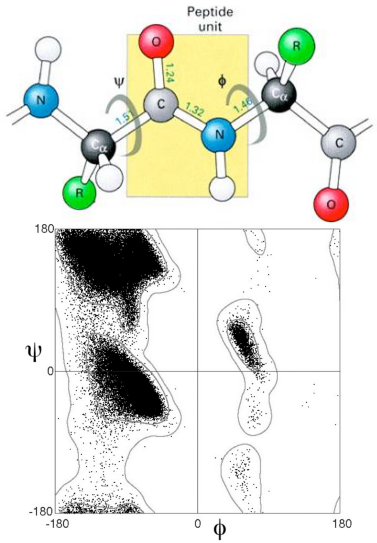


More restraints for low resolution

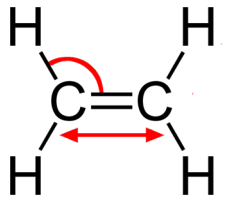
Side chain distributions



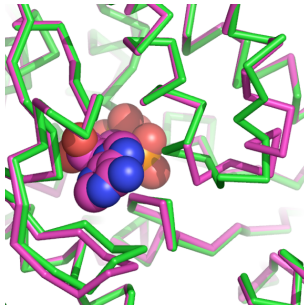
Main chain distributions



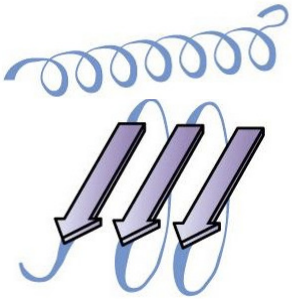
Covalent geometry



Internal symmetry (NCS)



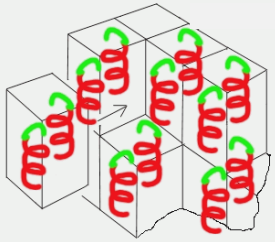
Similar (homologous) structures (reference model restraints)

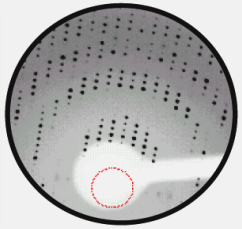


Secondary structure

Refinement

Crystallography





Initial model

Experimental data

A priori knowledge

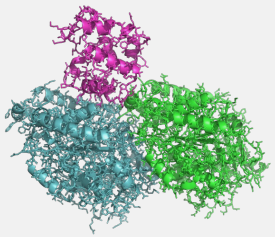
Score

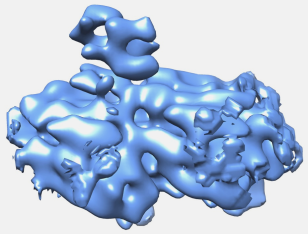
Modify model parameters

Improved model

phenix.refine

Cryo-EM





Initial model

Experimental data

A priori knowledge

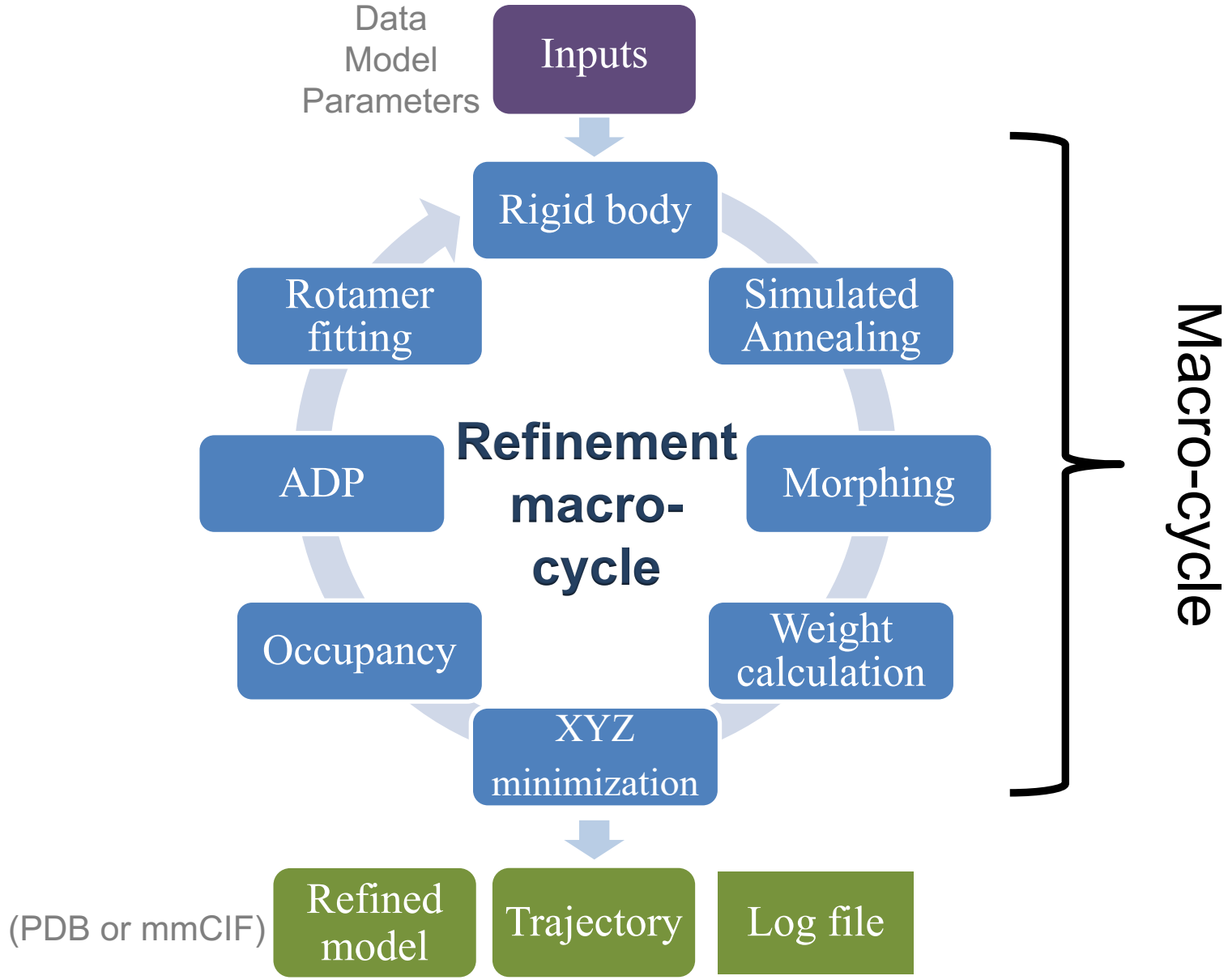
Score

Modify model parameters

Improved model

phenix.real_space_refine

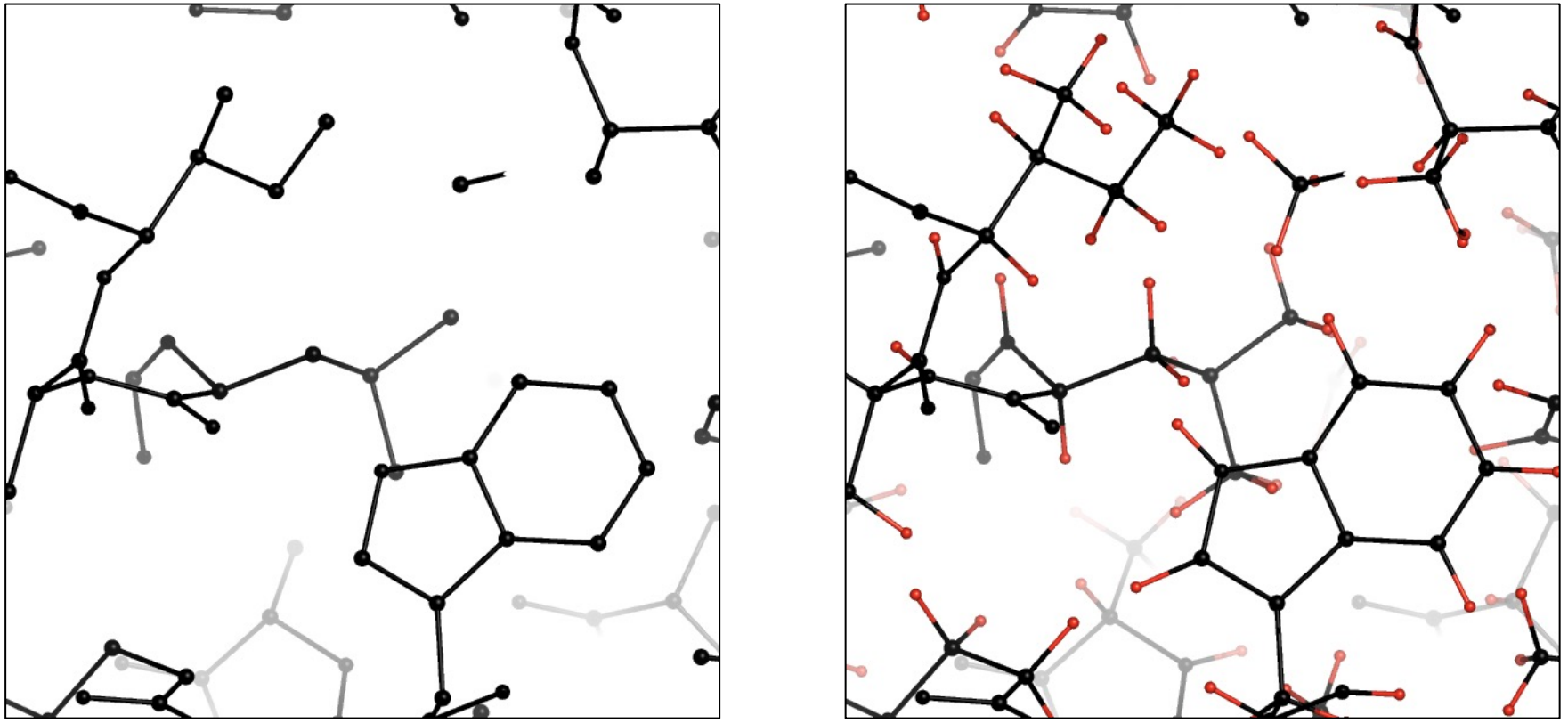
Refinement protocol



Refinement: practical considerations

Use Hydrogen atoms

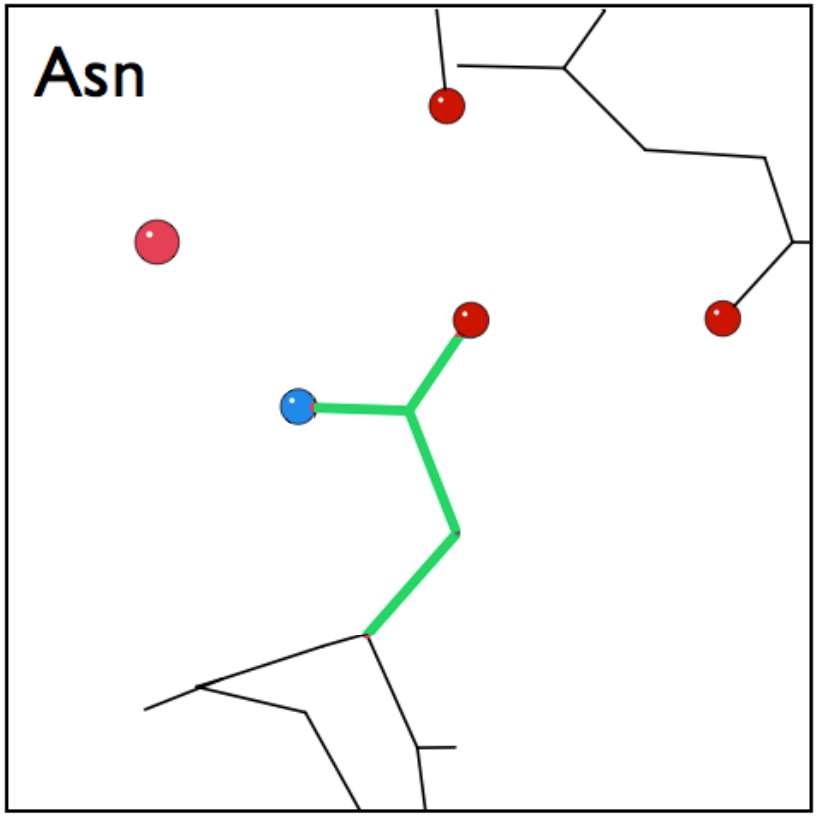
- Half of the atoms in a protein molecule
- Make most interatomic contacts
- Add to model towards the end, data resolution does not matter
- Once added, do not remove before the PDB deposition
- H do contribute to R-factors (expect 0.1-2% drop in R)

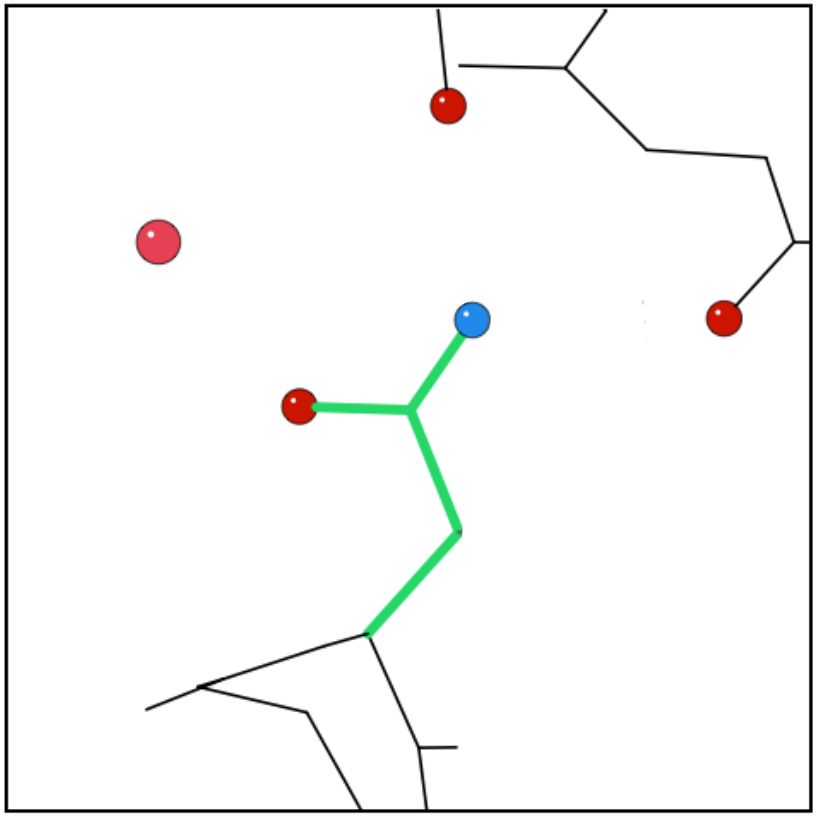


A structure without (left) and with (right) hydrogen atoms

Use Hydrogen atoms

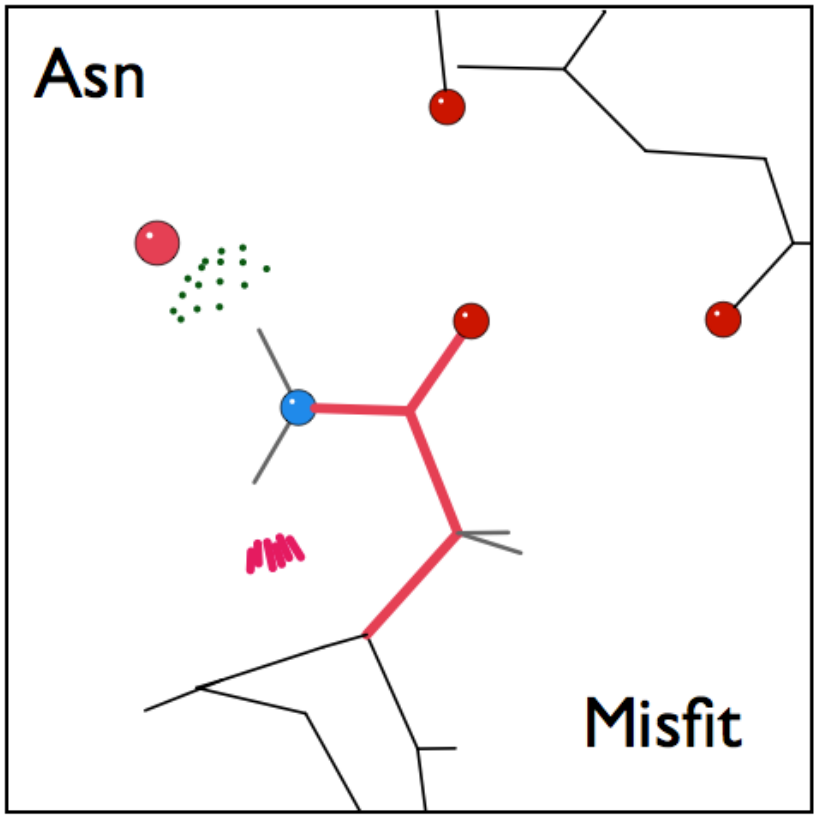
- N/Q/H flips (asparagine/glutamine/histidine)
  - Based on clash analysis
  - Requires H present

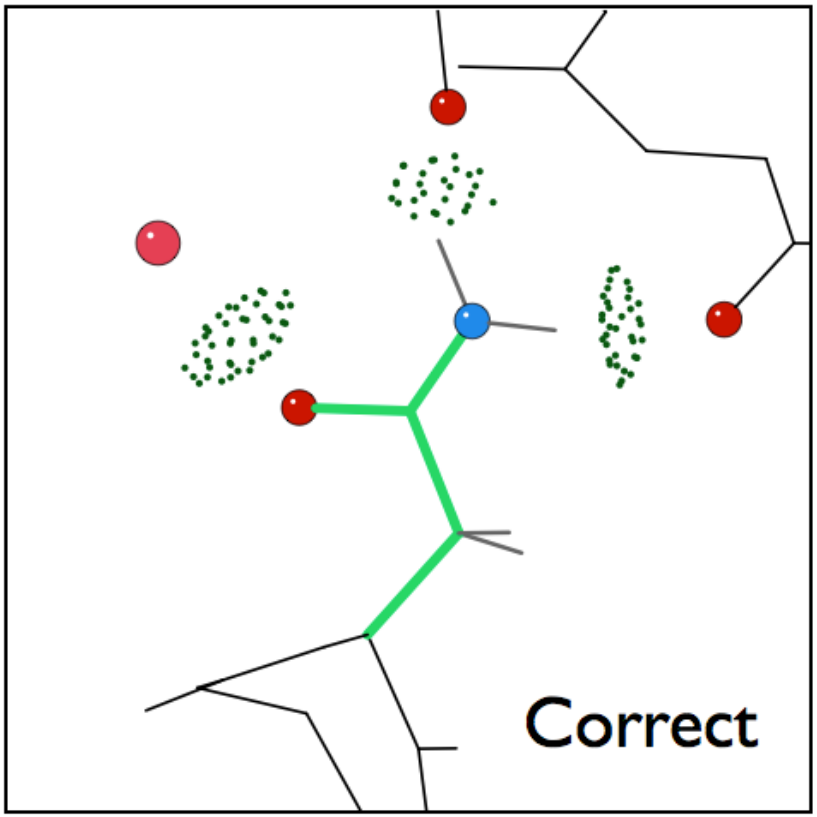




Use Hydrogen atoms

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

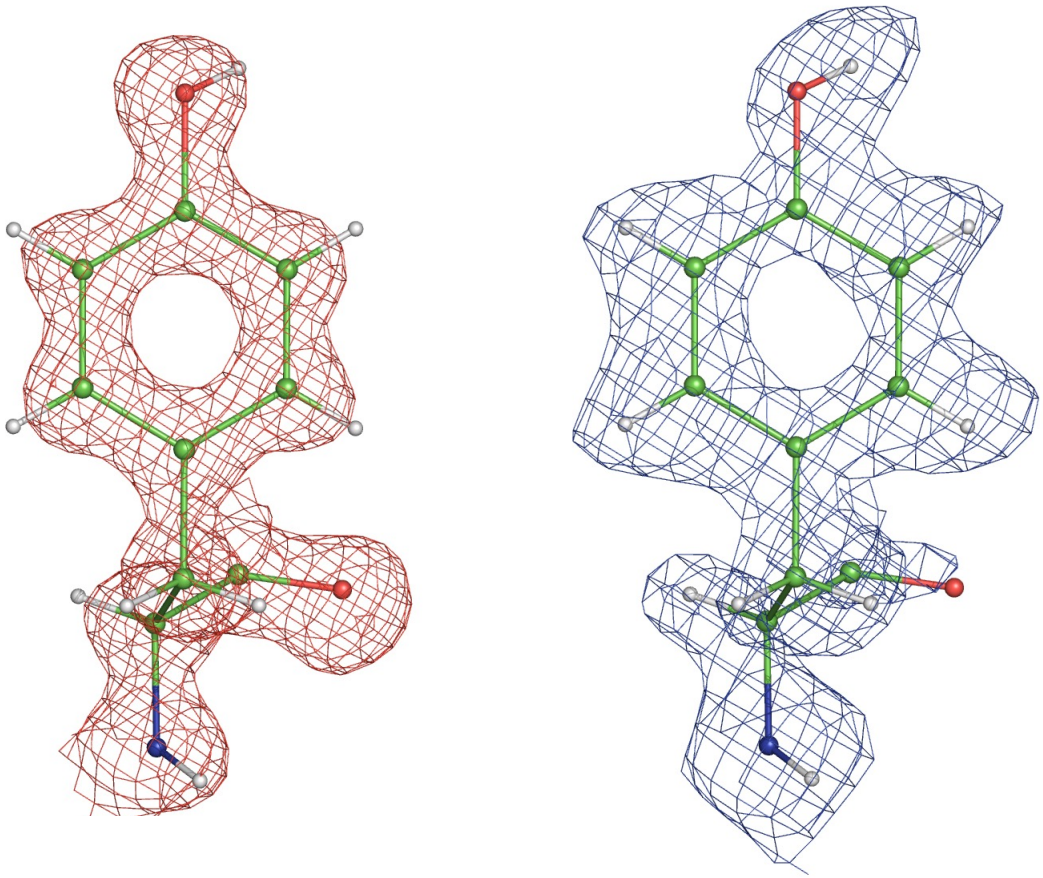




Hydrogens are best revealed by neutrons!

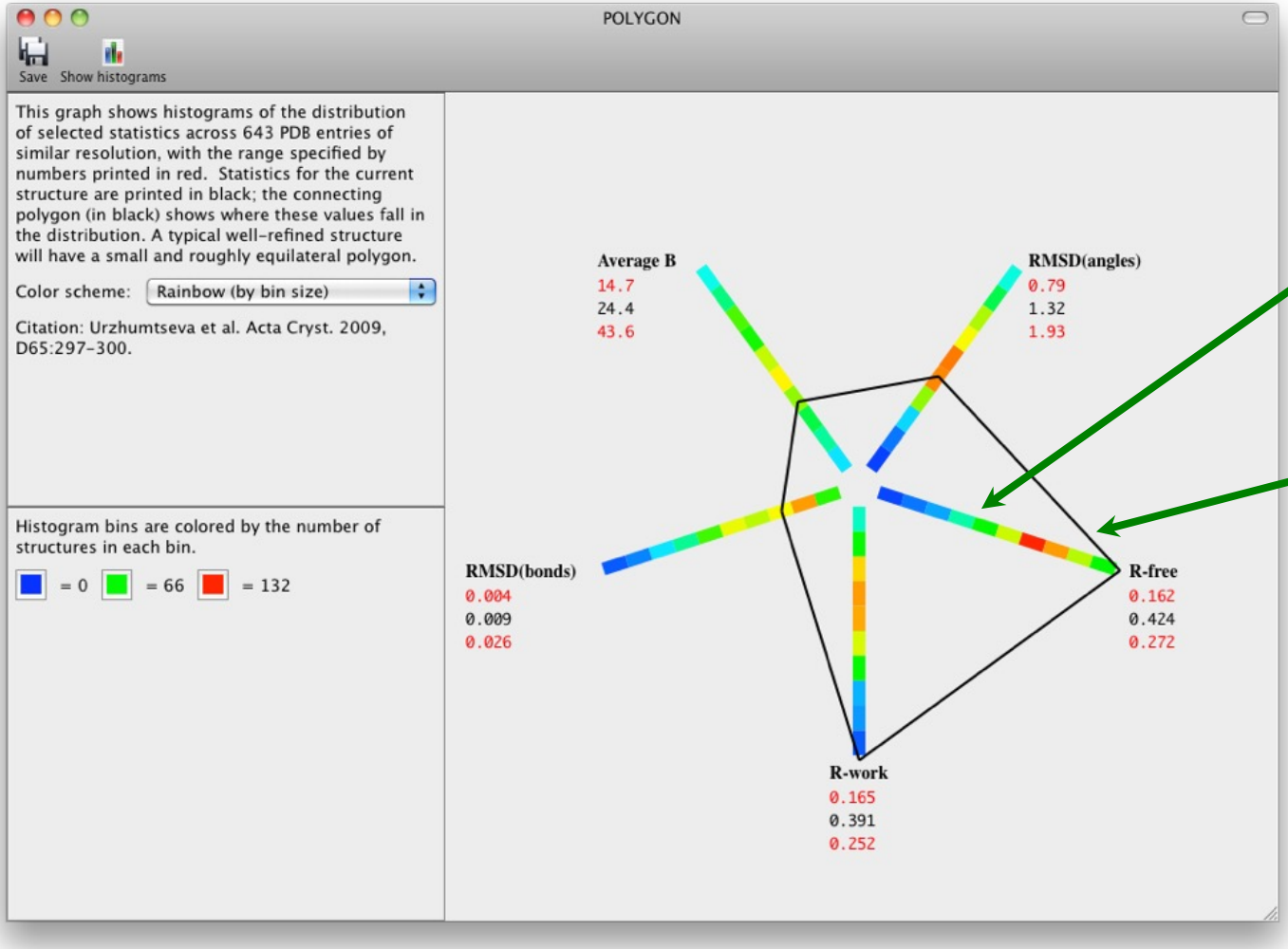
Nuclear density maps show H (D) at typical macromolecular resolutions (~2Å)

X-ray (1.1 Å) **Neutron (1.7 Å)**



2mFo-DFc maps at 1.5σ (Rubredoxin, PDB code: 3KKY)

Know when to stop



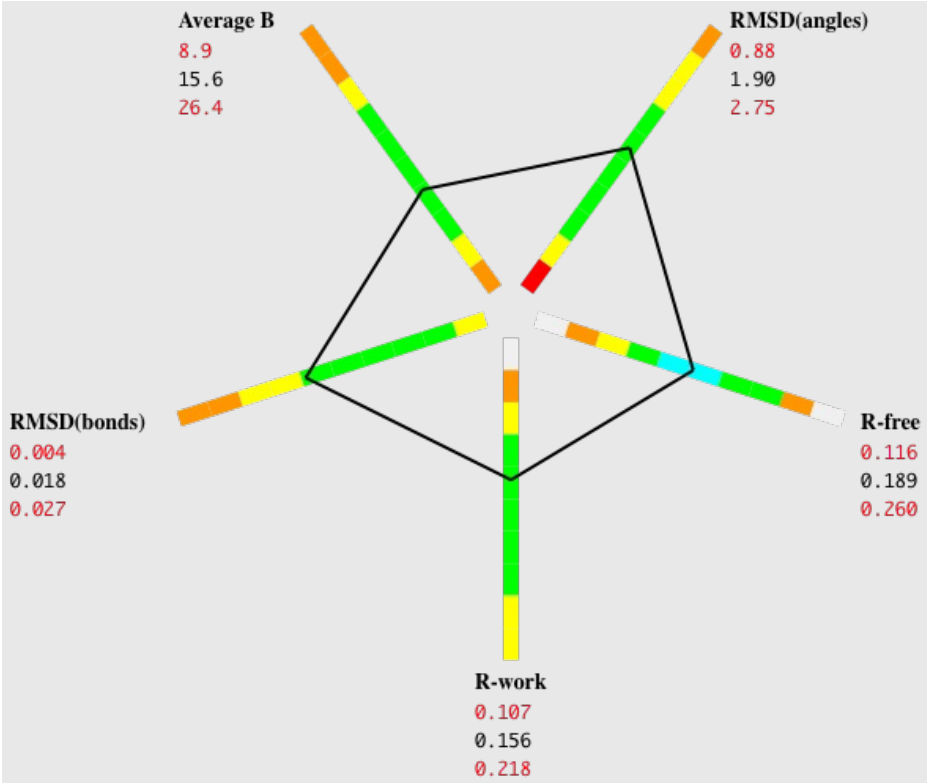
Colored bars are histograms showing distribution of values for structures at similar resolution

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

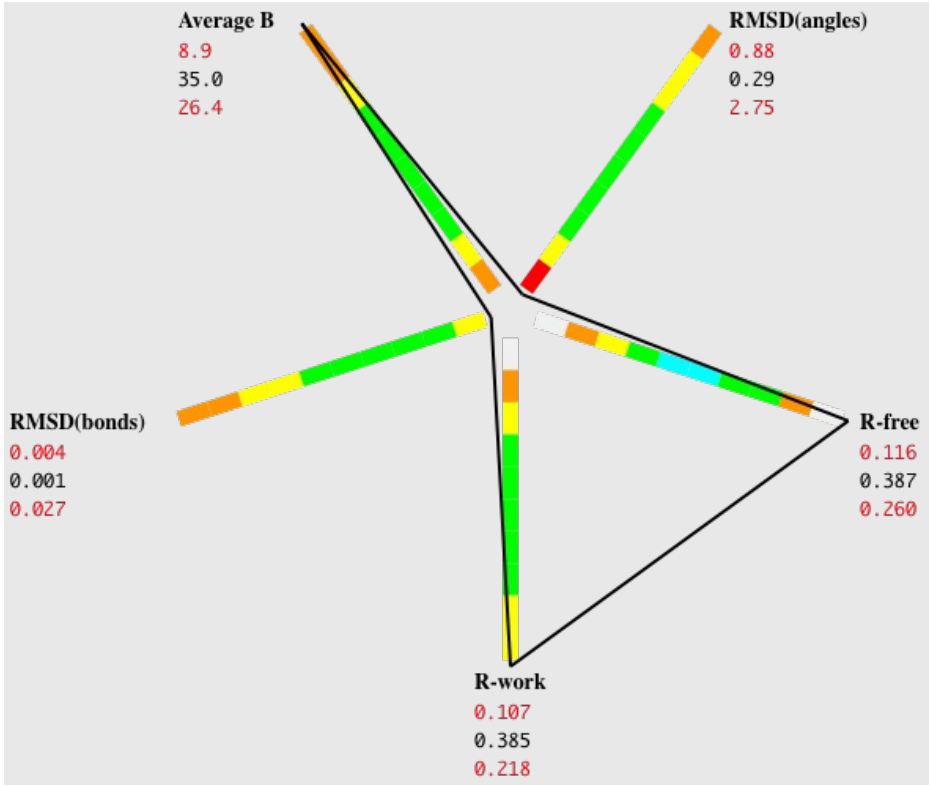
Crystallographic model quality at a glance.
L.Urzhumtseva, P.V.Afonine, P.D.Adams & A.Urzhumtsev. *Acta Cryst*. **D**65, 297-300 (2009)

Know when to stop

Likely overall good model

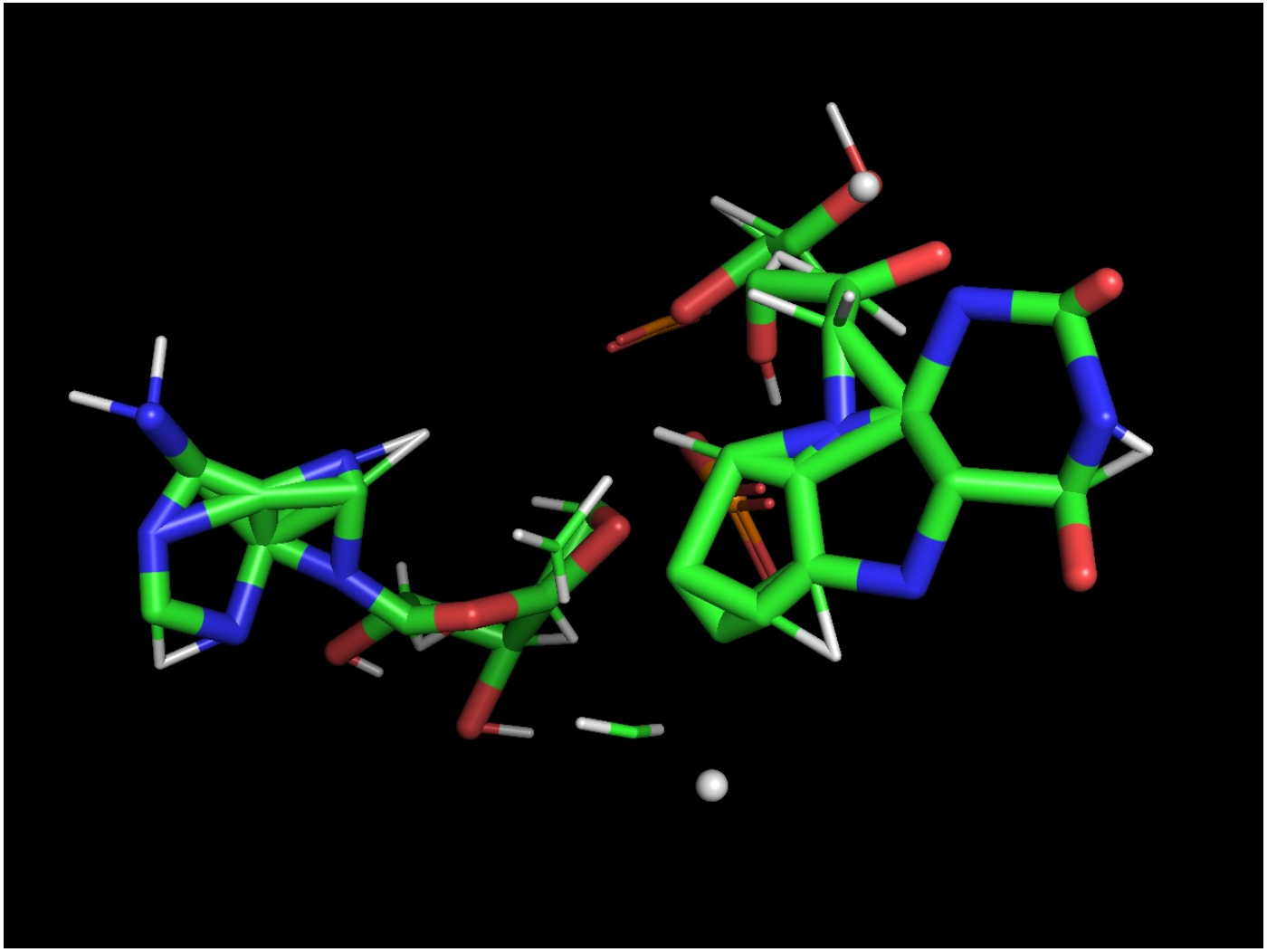


Clearly there are problems

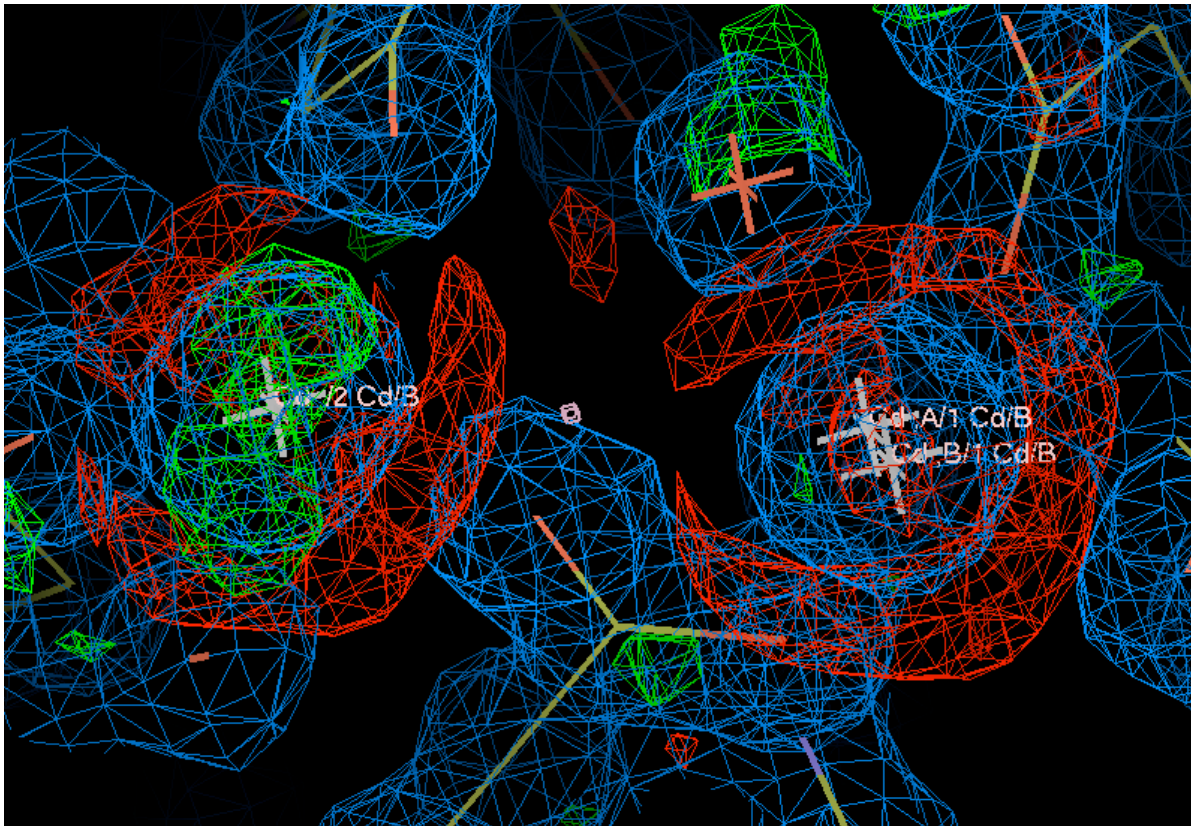


Local *vs* Global

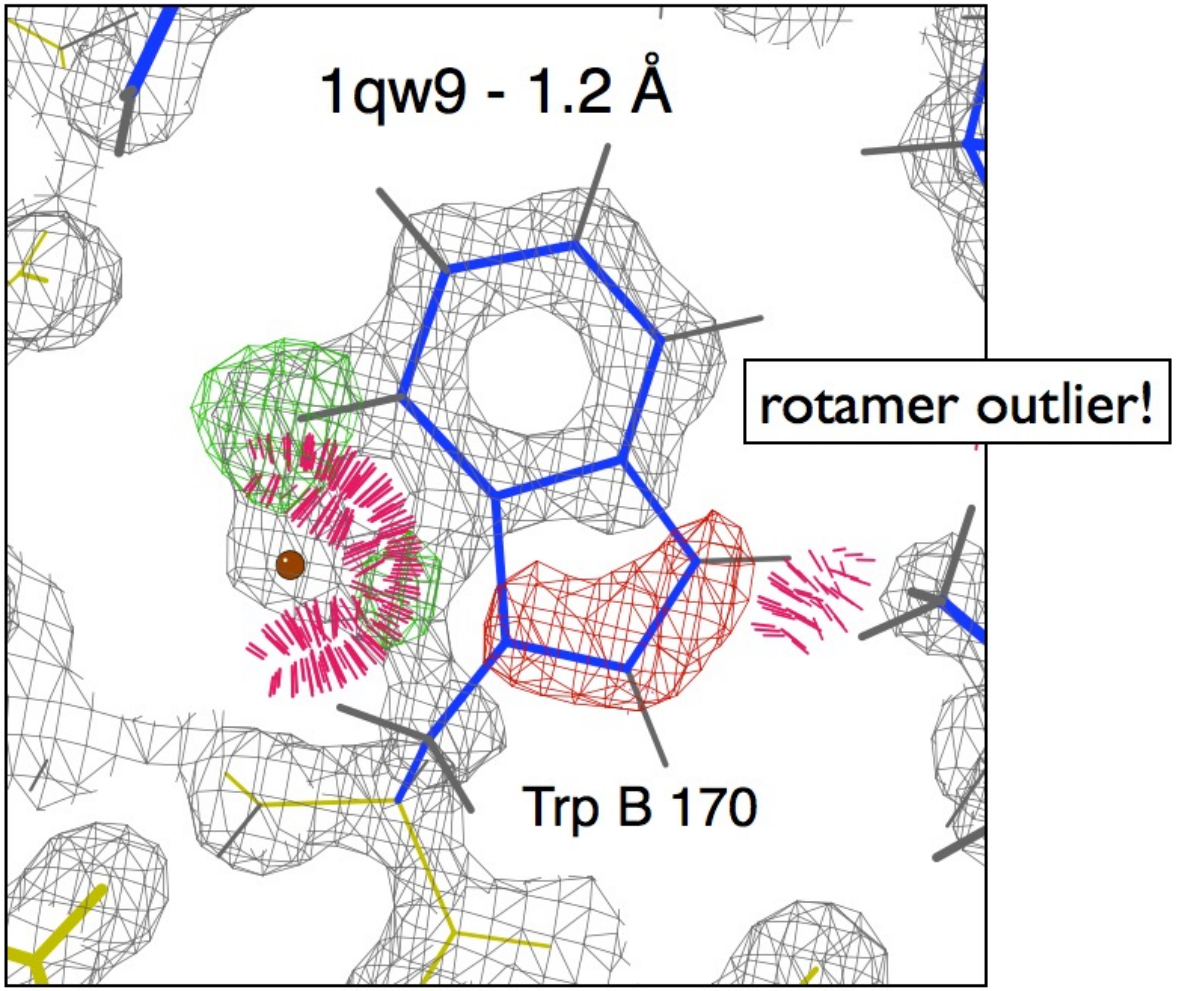
- R_{WORK}/R_{FREE} , bond/angle RMSDs etc do not report on local errors



Map and model errors



Not all modeling errors can be fixed by refinement



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

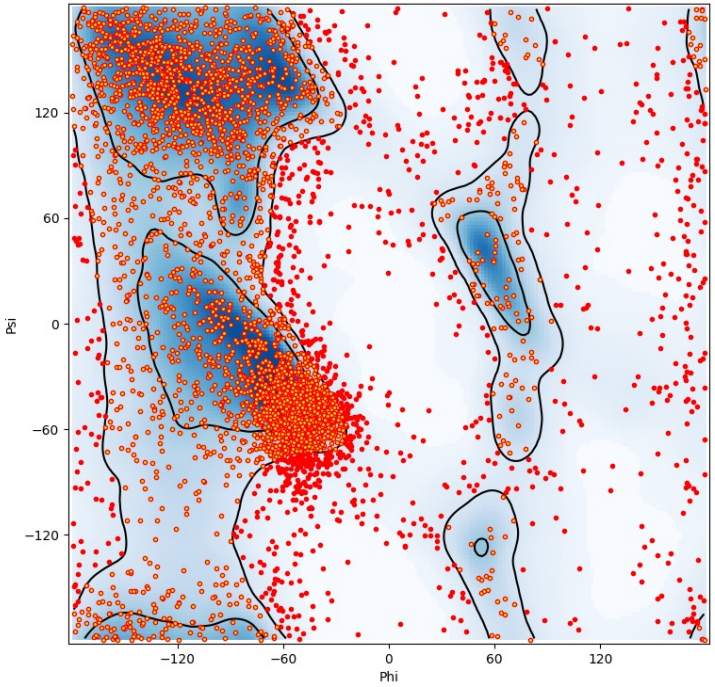
- Simulated annealing (SA)
- Model morphing
  - Only use if model has gross errors (correction requires large movements)
  - Do not use if model is relatively good and only needs small corrections

Ramachandran plot restraints

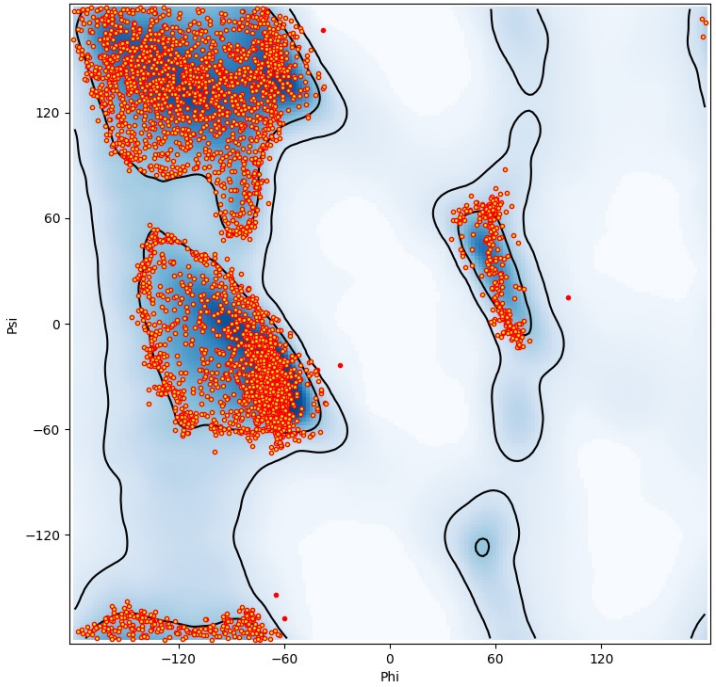
- Always use at low resolution
- Do not use to fix existing outliers

PDB code: 5a9z

Original



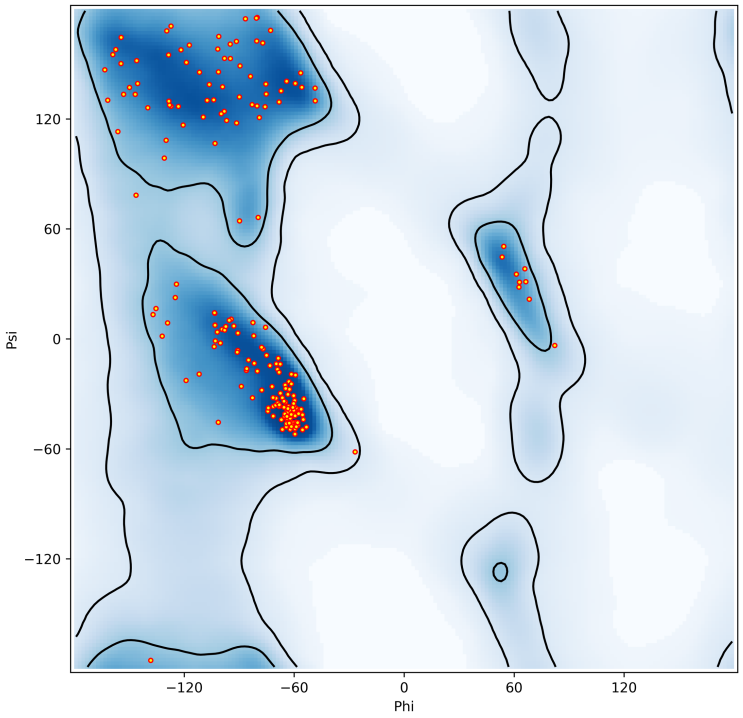
Refined with Ramachandran plot restraints



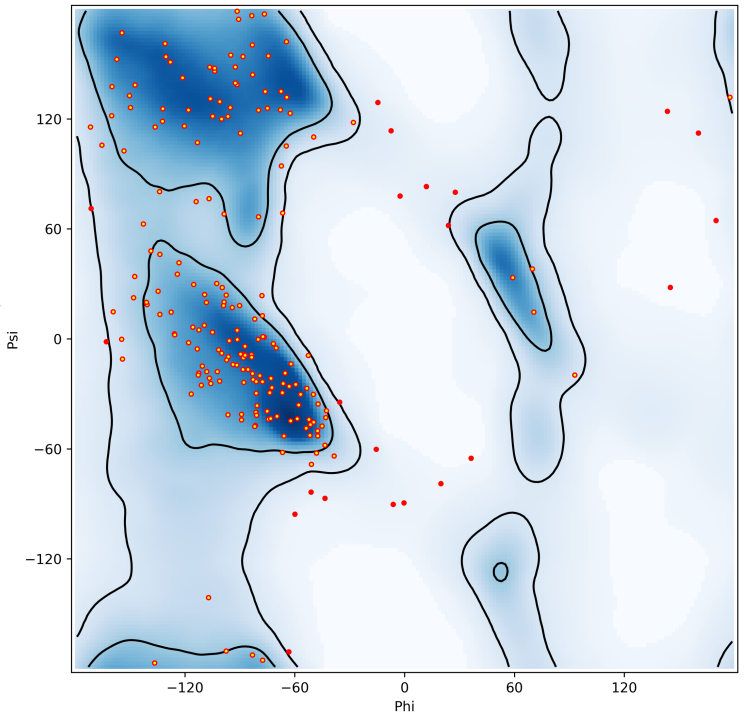
Ramachandran plot restraints

- Ramachandran plot restraints
  - Use to stop outliers from occurring

Before refinement



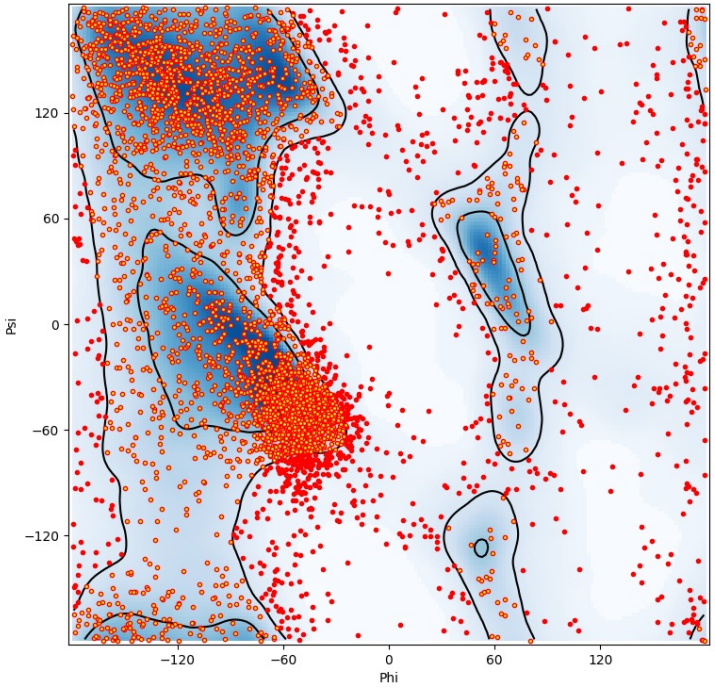
After refinement (No Ramachandran plot restraints)



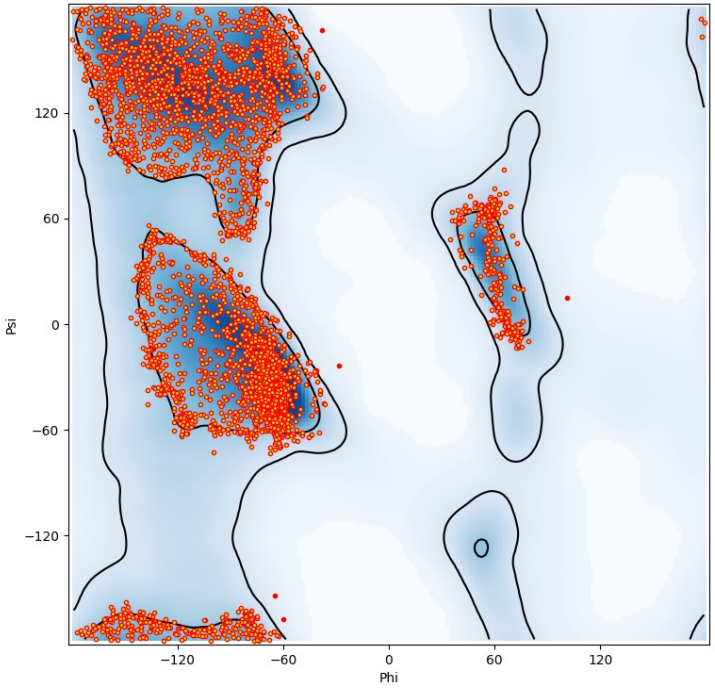
Ramachandran plot restraints

- What is wrong with this plot?

Original

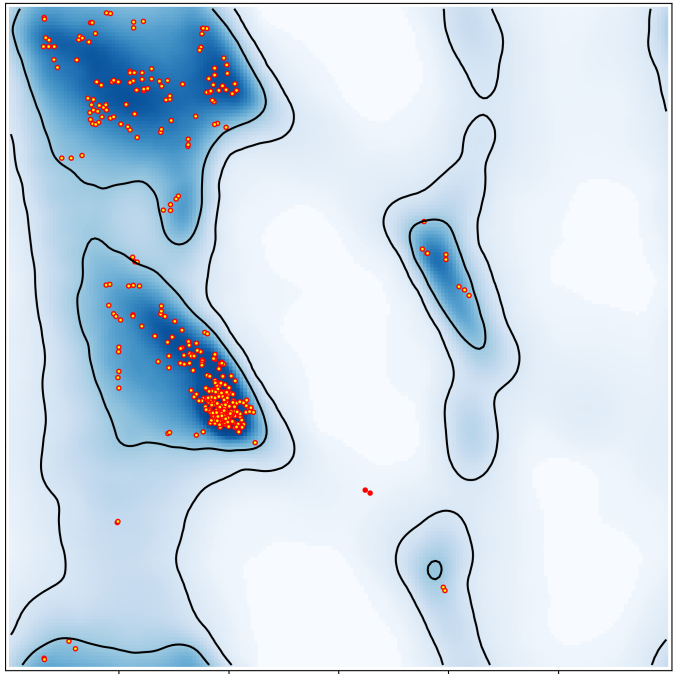


Refined with Ramachandran plot restraints



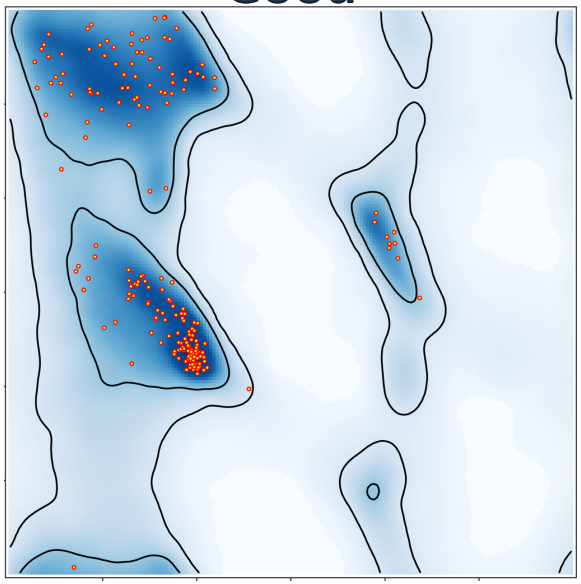
Ramachandran plot restraints

- It is very different from what we expect!

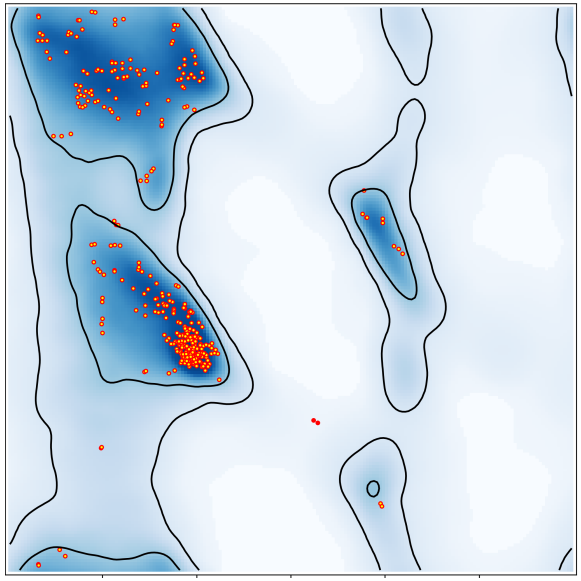


How you can tell good vs bad plot?

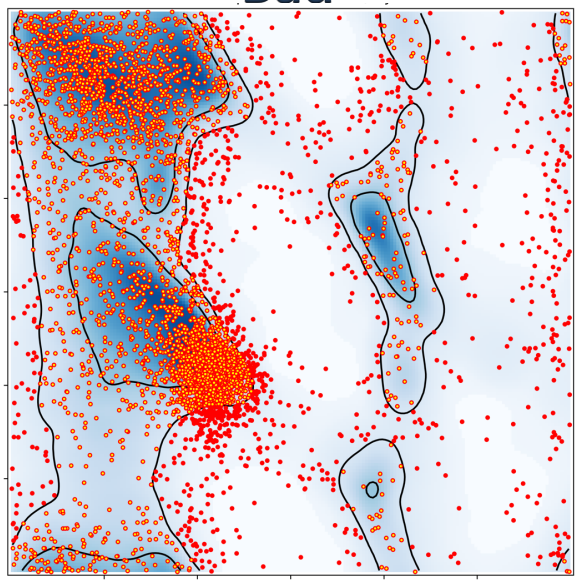
Good



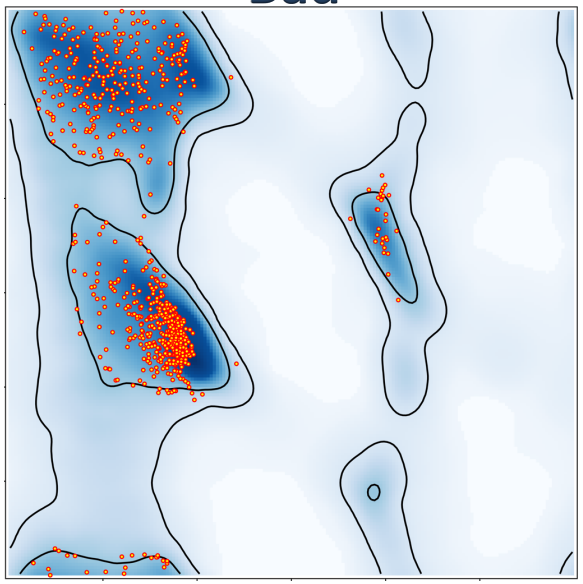
Good



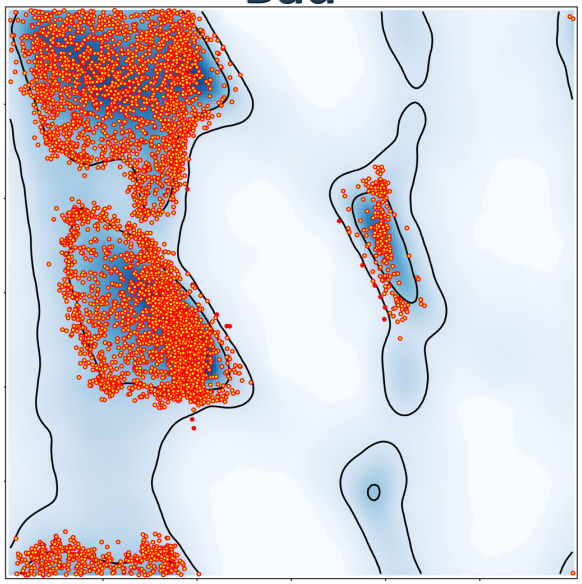
Bad



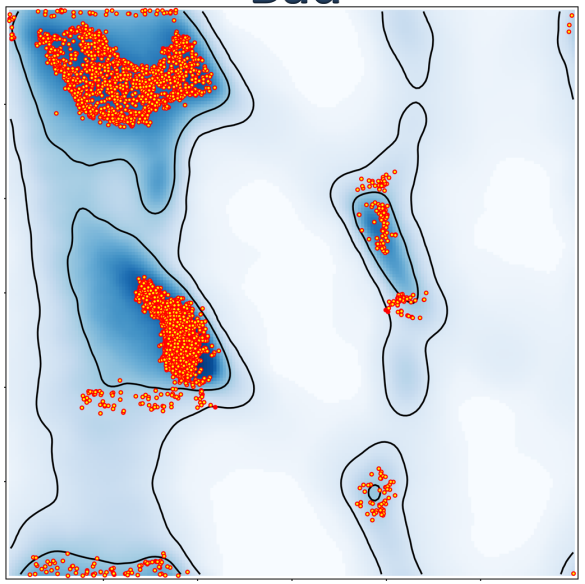
Bad



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

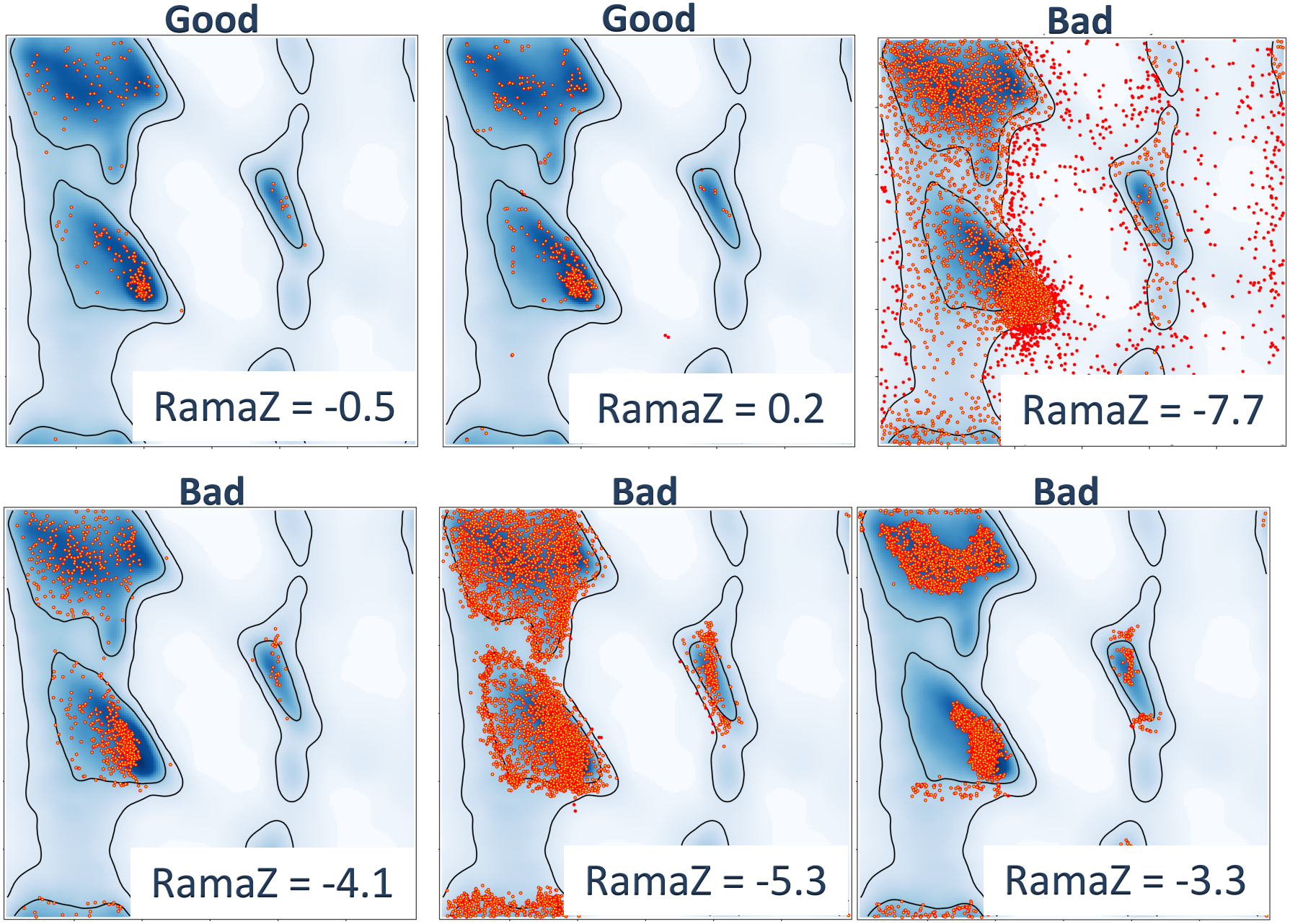
CellPress

Resource

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

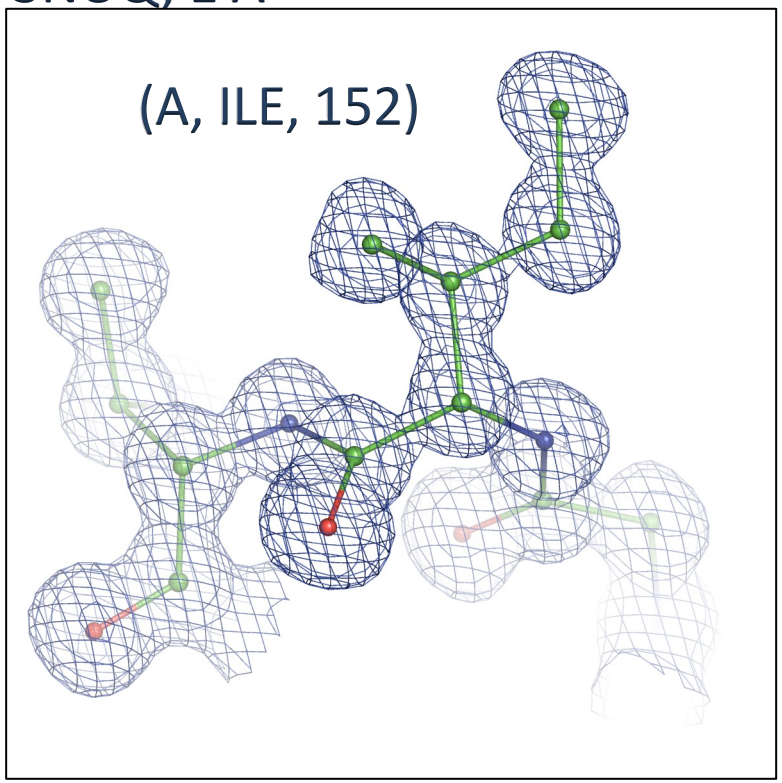
Oleg V. Sobolev,[1,5,*] Pavel V. Afonine,[1] Nigel W. Moriarty,[1] Maarten L. Hekkelman,[2,3] Robbie P. Joosten,[2,3,*] Anastassis Perrakis,[2,3] and Paul D. Adams[1,4]

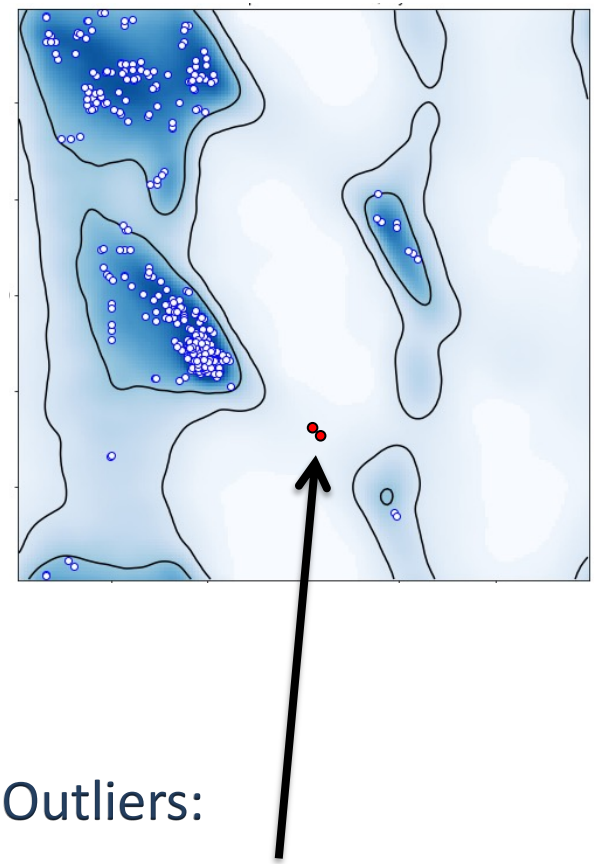
Model validation: *Ramachandran plot Z-score*



An outlier ≠ wrong

3NOQ, 1 Å





Outliers:

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

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

Phenix.refine inputs and outputs

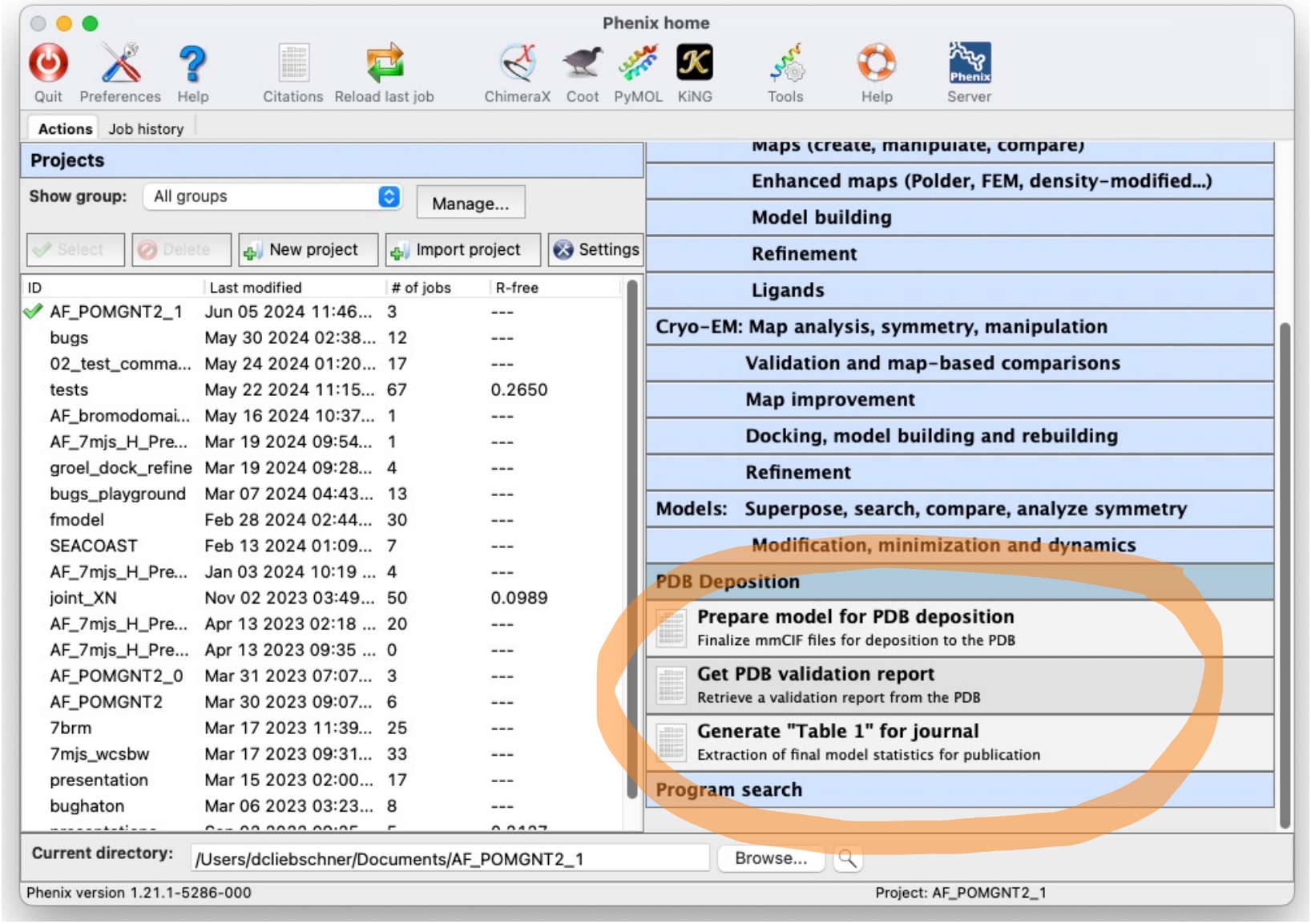
- Phenix.refine outputs
  - Atomic model (PDB, mmCIF)
  - .log file
  - .eff file – summary of all input parameters
  - MTZ file with copy of input data and 2Fo-Fc and Fo-Fc maps
  - .geo file – all restraints used

Phenix.refine inputs and outputs

- MTZ from `phenix.refine` contains

1. Verbatim copy of input data considered for use
2. Data that was actually used in refinement
3. Total model structure factors $\mathbf{F}_{model}$
4. Fourier maps
   - $2mF_{obs}-DF_{model}$ 'filled'
   - $2mF_{obs}-DF_{model}$
   - $mF_{obs}-DF_{model}$
   - Anomalous difference map (if anomalous data)

PDB deposition



PDB deposition

mmCIF format is mandatory for deposition as of 2019

ISSN 2059-7983

Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB)

Paul D. Adams,[a,b] Pavel V. Afonine,[a] Kumaran Baskaran,[c] Helen M. Berman,[d] John Berrisford,[e] Gerard Bricogne,[f] David G. Brown,[g] Stephen K. Burley,[d,h,i]* Minyu Chen,[j] Zukang Feng,[d] Claus Flensburg,[f] Aleksandras Gutmanas,[e] Jeffrey C. Hoch,[k]* Yasuyo Ikegawa,[j] Yumiko Kengaku,[j] Eugene Krissinel,[l] Genji Kurisu,[j]* Yuhe Liang,[d] Dorothee Liebschner,[a] Lora Mak,[e] John L. Markley,[c]* Nigel W. Moriarty,[a] Garib N. Murshudov,[m] Martin Noble,[n] Ezra Peisach,[d] Irina Persikova,[d] Billy K. Poon,[a] Oleg V. Sobolev,[a] Eldon L. Ulrich,[c] Sameer Velankar,[e]* Clemens Vonrhein,[f] John Westbrook,[d] Marcin Wojdyr,[f,l] Masashi Yokochi[j] and Jasmine Y. Young[d]

Received 21 February 2019
Accepted 3 April 2019

Edited by R. J. Read, University of Cambridge, England

PDB deposition: mmCIF facts

- Contains a lot more information than PDB
- Not intended to be human editable
  - You can read it but it is (much) harder than PDB
- Phenix tools generally produce output in mmCIF format
- Avoid editing by hand
  - Easy to make hard-to-recover mistakes

PDB deposition: CIF file confusion

- CIF is a file format

- CIF file can contain:
  - Ligand information
  - Atomic model
  - Reflection data
  - Any mixture of three above

PDB deposition: dos and don’ts

- Do not change the content of files from refinement for any reason:
  - Add/remove atoms (hydrogens, water)
  - Edit labels, header information

- Run Comprehensive validation (Phenix GUI) to address all outstanding issues before deposition

- Don’t panic if validation statistics reported by Phenix does not match PDB validation report
  - If that happens and presents a problem – start conversation with PDB stuff and involve Phenix developers

- Once all is deposited and up on the web – check everything: mistakes at PDB end happen

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