Validation

Pavel Afonine

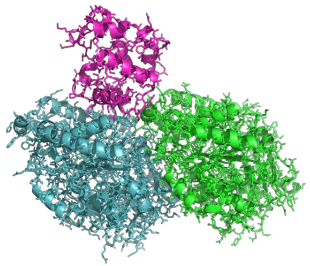
Phenix team

Lawrence Berkeley National Lab, California, USA

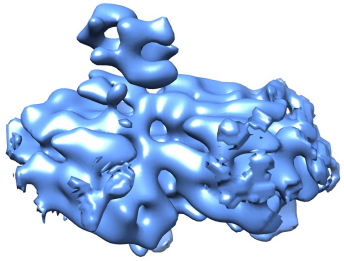
May 10th, 2023
MCCS, Madrid

Validation

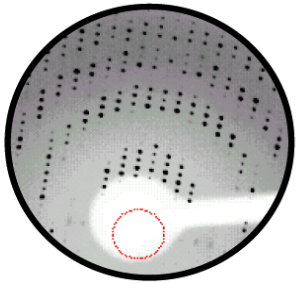
Model



Data



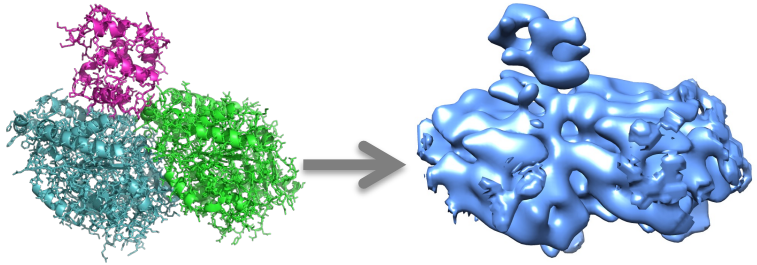
or



Cryo-EM

Diffraction

Model to data fit

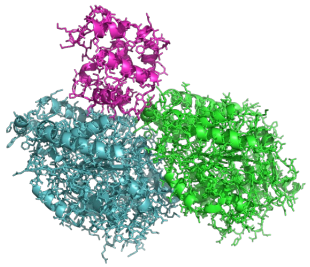


Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations

Validation tools: Crystallography *vs* Cryo-EM

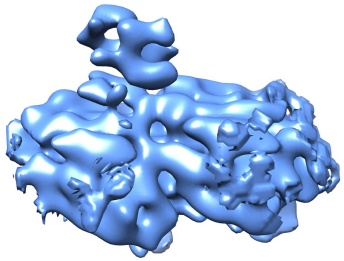
Exact same

Model

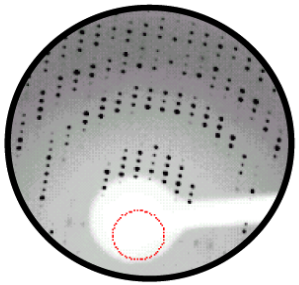


Different

Data



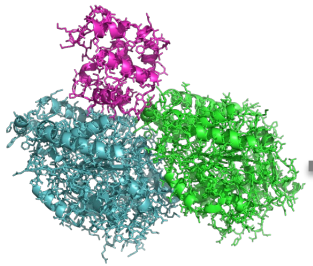
or

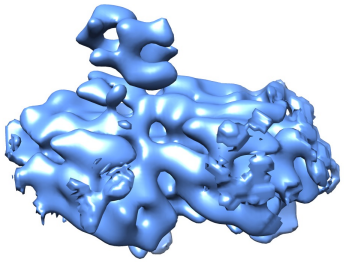


Cryo-EM

Diffraction

Model to data fit





Similar

Validation: *why to do?*

- Can help to:
  - save (a lot of) time
  - produce better models
  - Set correct expectations

- Subjectivity:
  - lot's of manual steps that depend on skills, pressure and ethics

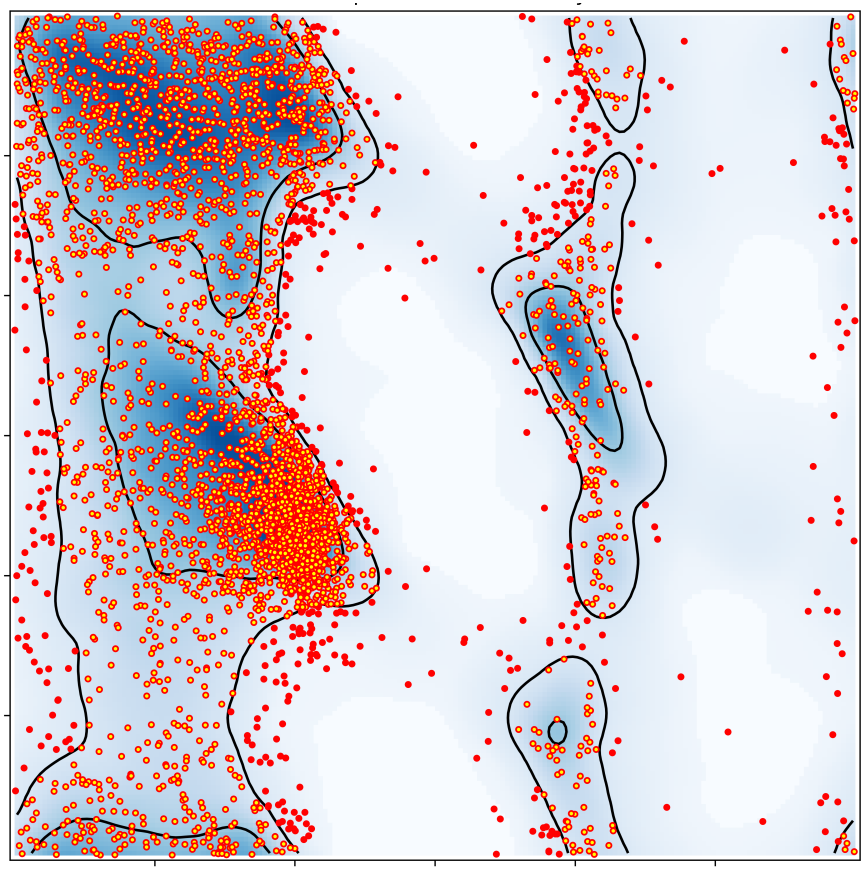
- Software isn't perfect

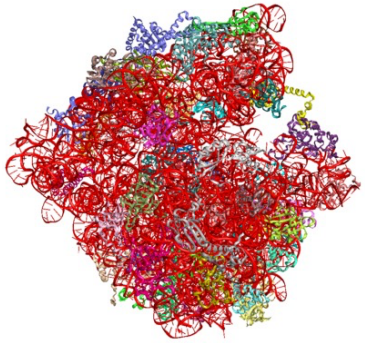
- Databases are not perfect

Lack of validation will be discovered (sooner or later)!

Validation: why to do?

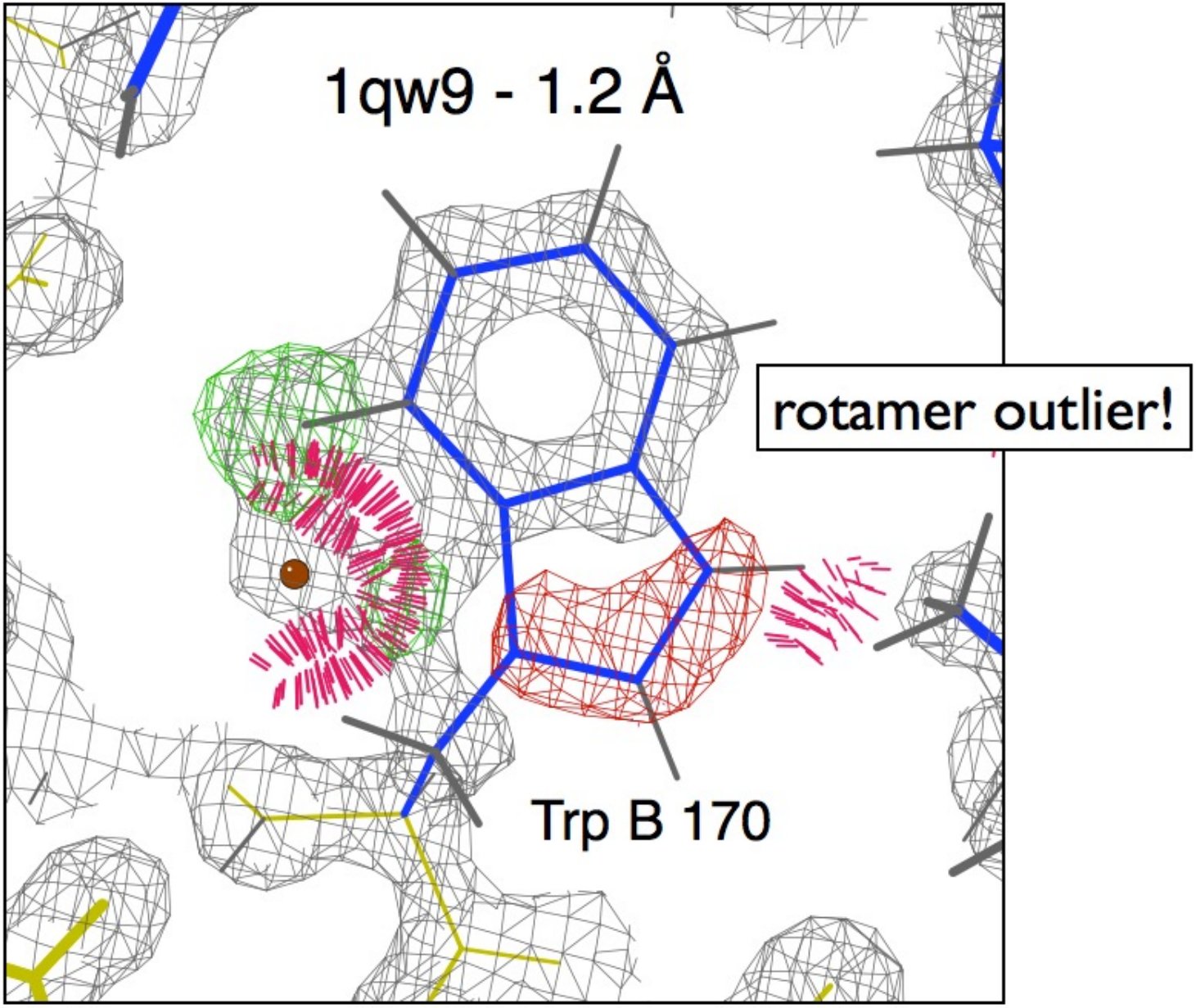
(2019) Nature 570: 400-404 | PDB: 6o9j 3.9Å



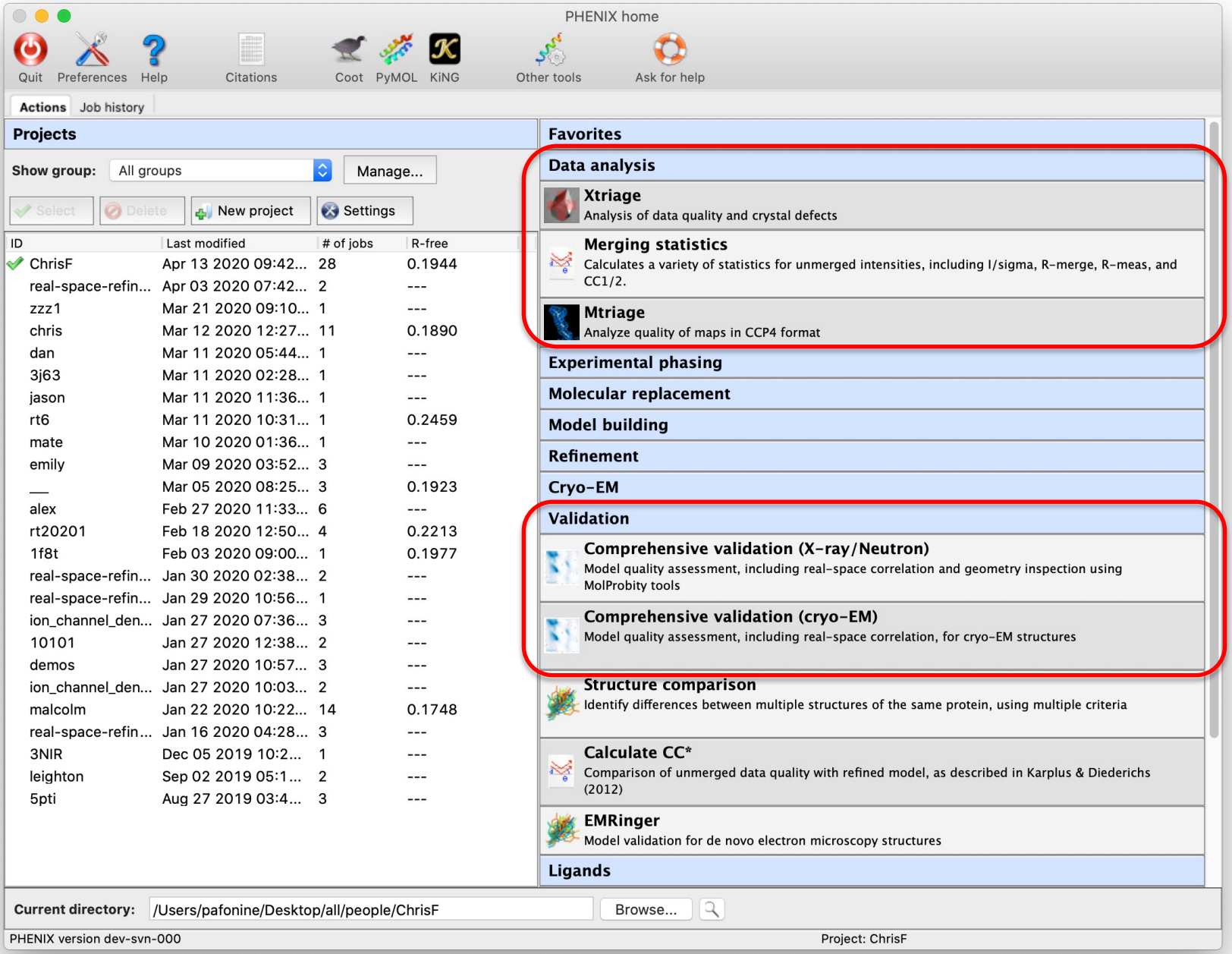


Metric	6o9j	Expected
Clashscore	70	Less than 10
Ramachandran favored, %	59	More than 98
Ramachandran outliers, %	15	0
Rotamer outliers, %	23	0
C_β deviations, %	0.5	0

Validation: why to do?



Validation tools in Phenix

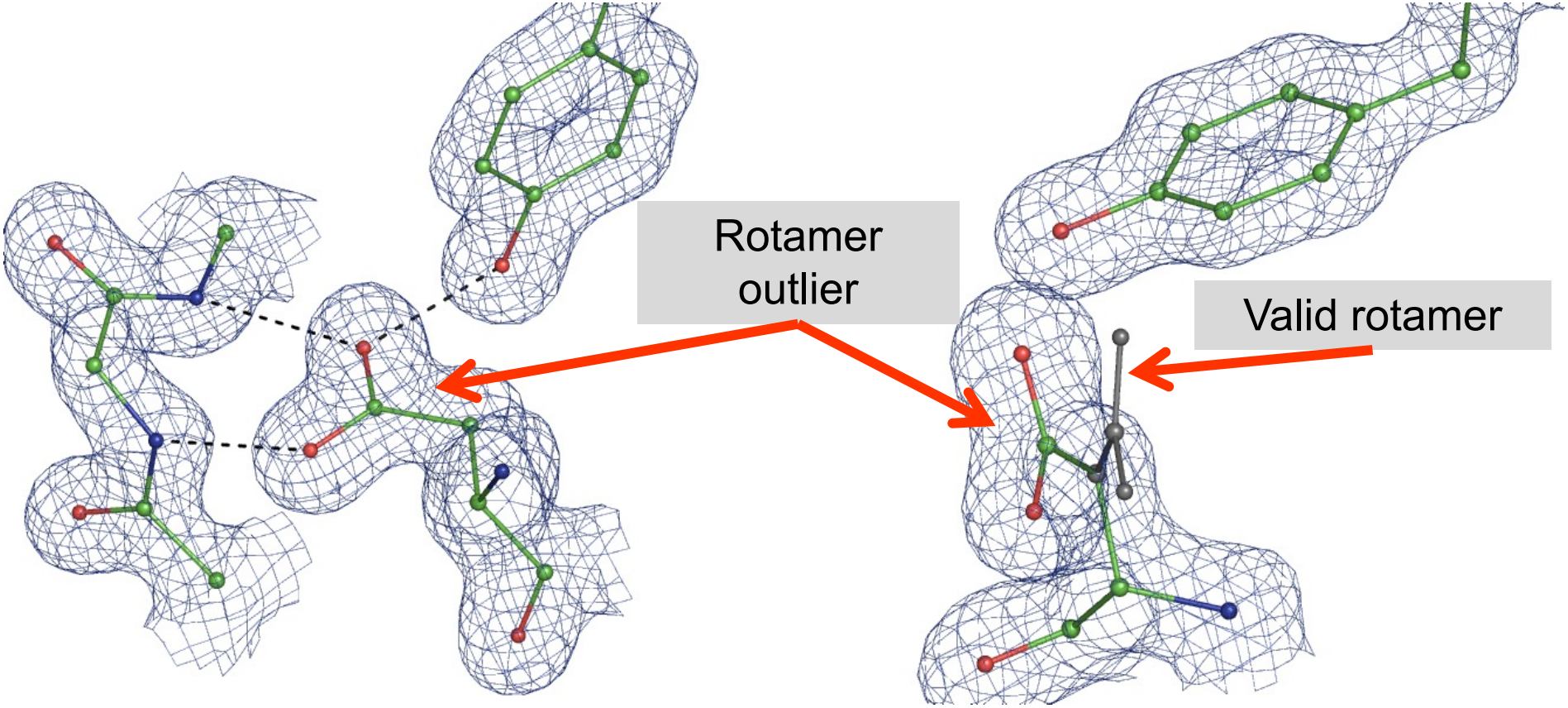


Model validation

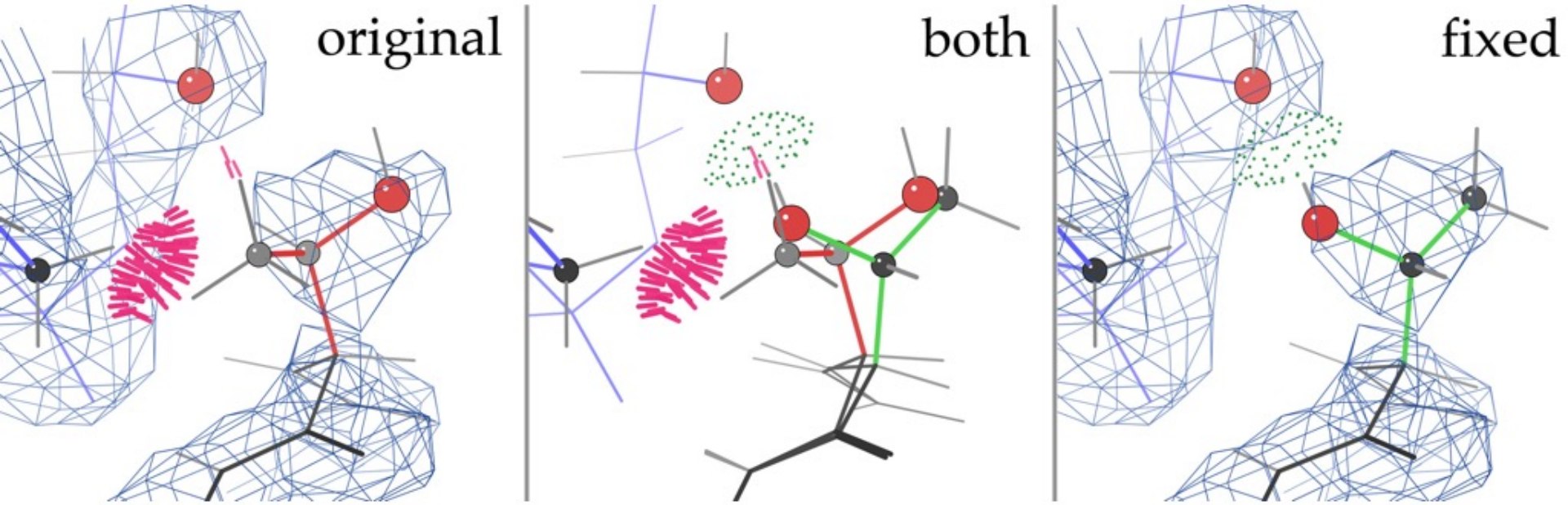
- **Table 1 items (not a complete list!)**
  - Content (macromolecule, ligands, NCS, …)
  - Bond/angle RMSDs / RMSZ
  - Molprobity:
    - Clashscore
    - Ramachandran plot (favorite, outliers)
    - Rotamer outliers
    - C-beta deviations
  - Incomplete residues
  - Solvent content
  - ADP (mean, Bonded $<B_i-B_j>$)
  - ……..

Model validation: amino-acid side-chain rotamers

- An outlier ≠ wrong
  - However, each outlier has to be explained



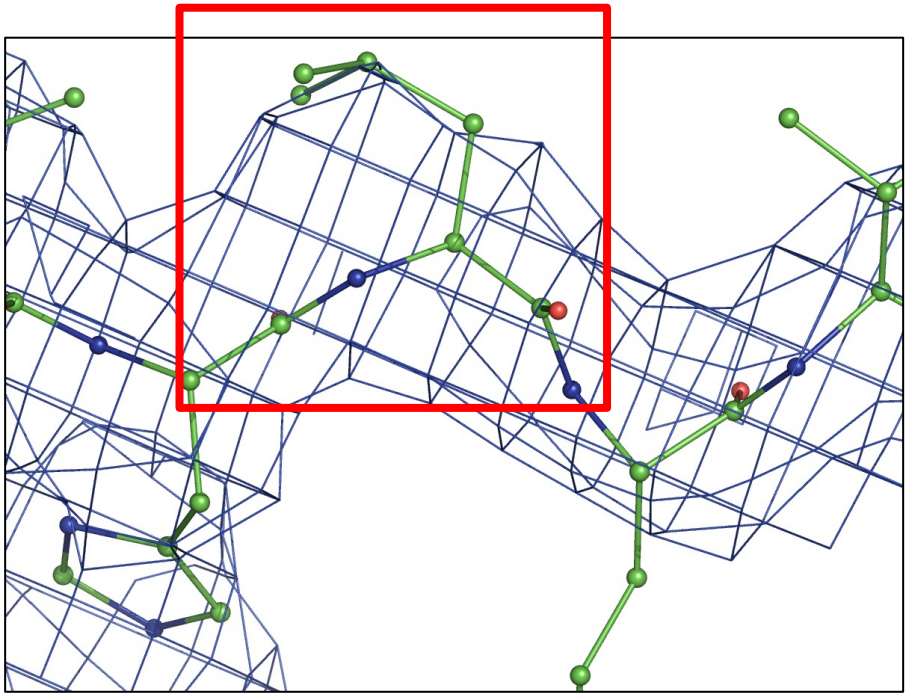
Model validation: amino-acid side-chain rotamers



Model validation: amino-acid side-chain rotamers

- Low-resolution maps

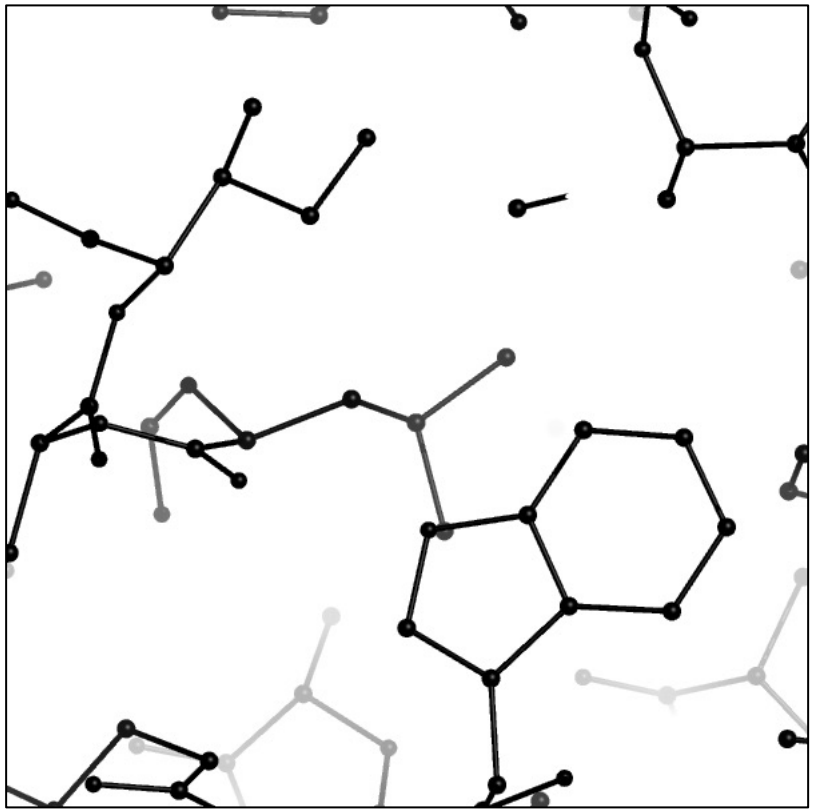
Side chain lacking density may be forced into main chain density and become a rotamer outlier



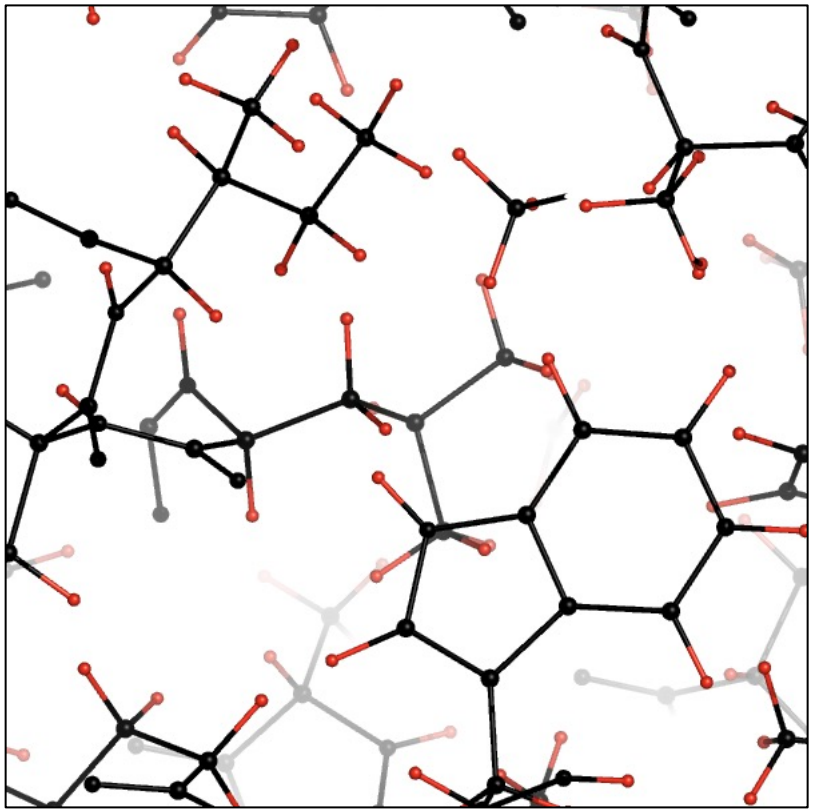
Phenix refinement (real- and reciprocal-space) use rotamer-specific restraints on torsion chi-angles

Model validation: clashes

- Half of the atoms in a protein molecule
- Make most interatomic contacts
  - Using H in refinement helps prevent or eliminate clashes



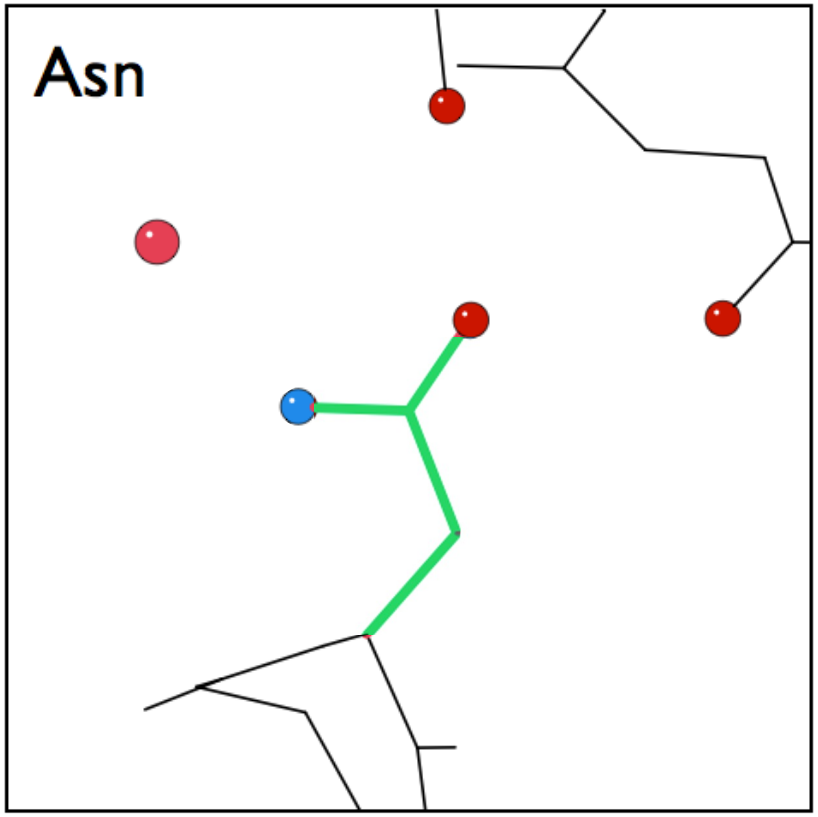
No H atoms

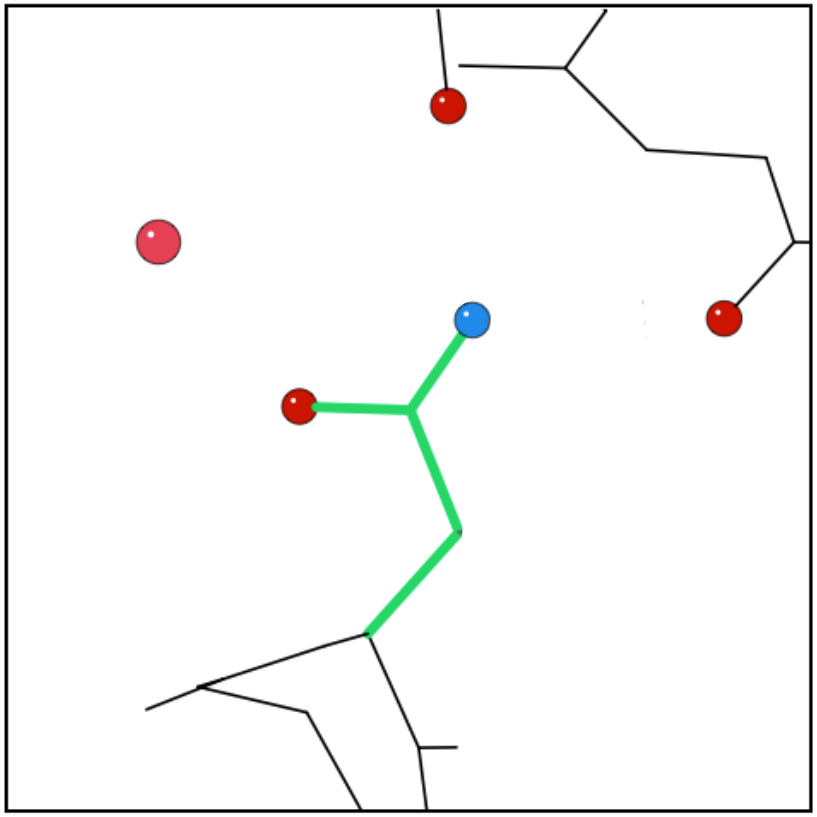


H atoms added

Model validation: clashes

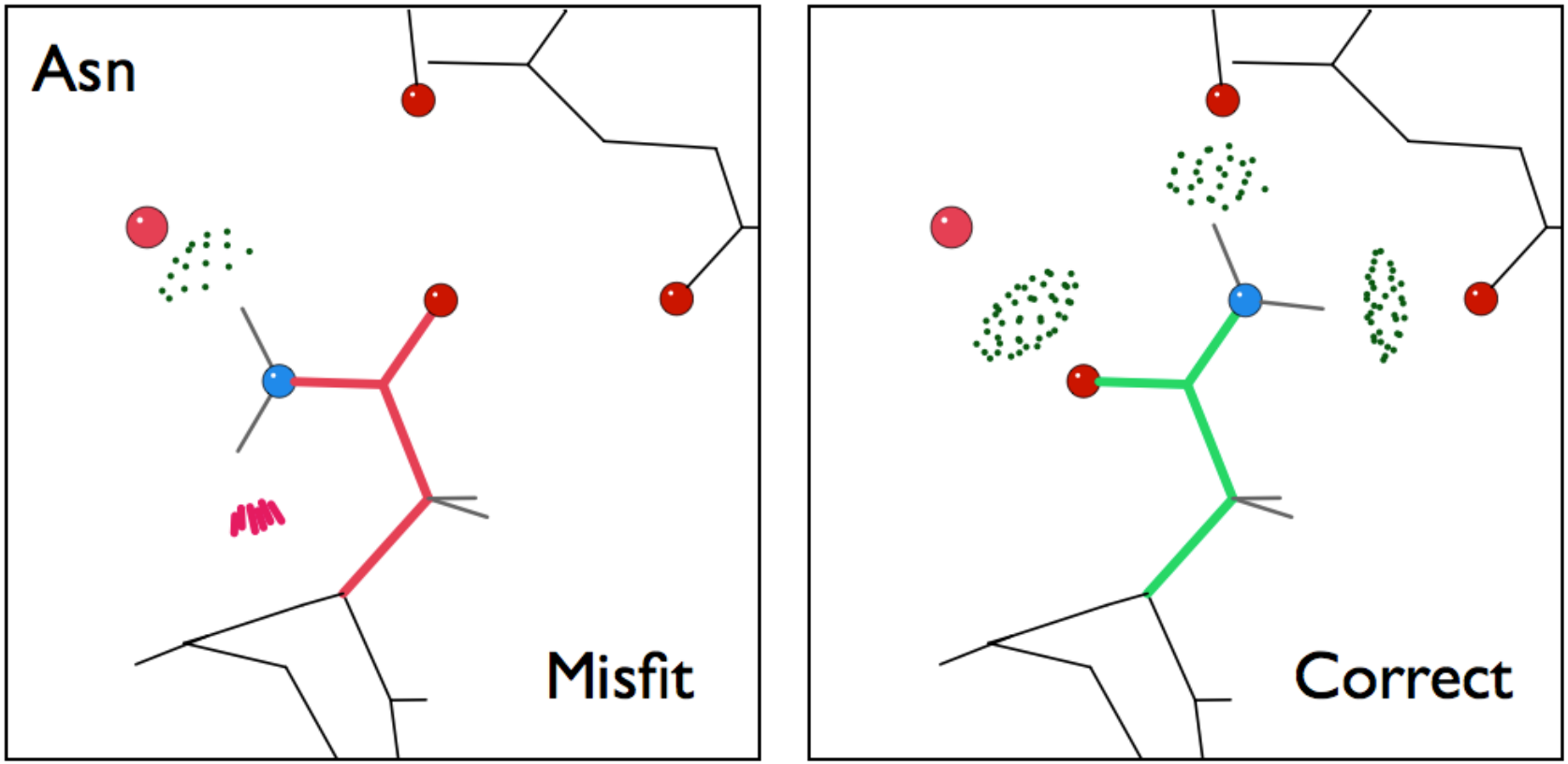
- N/Q/H flips (asparagine/glutamine/histidine)
  - Based on clash analysis
  - Requires H present





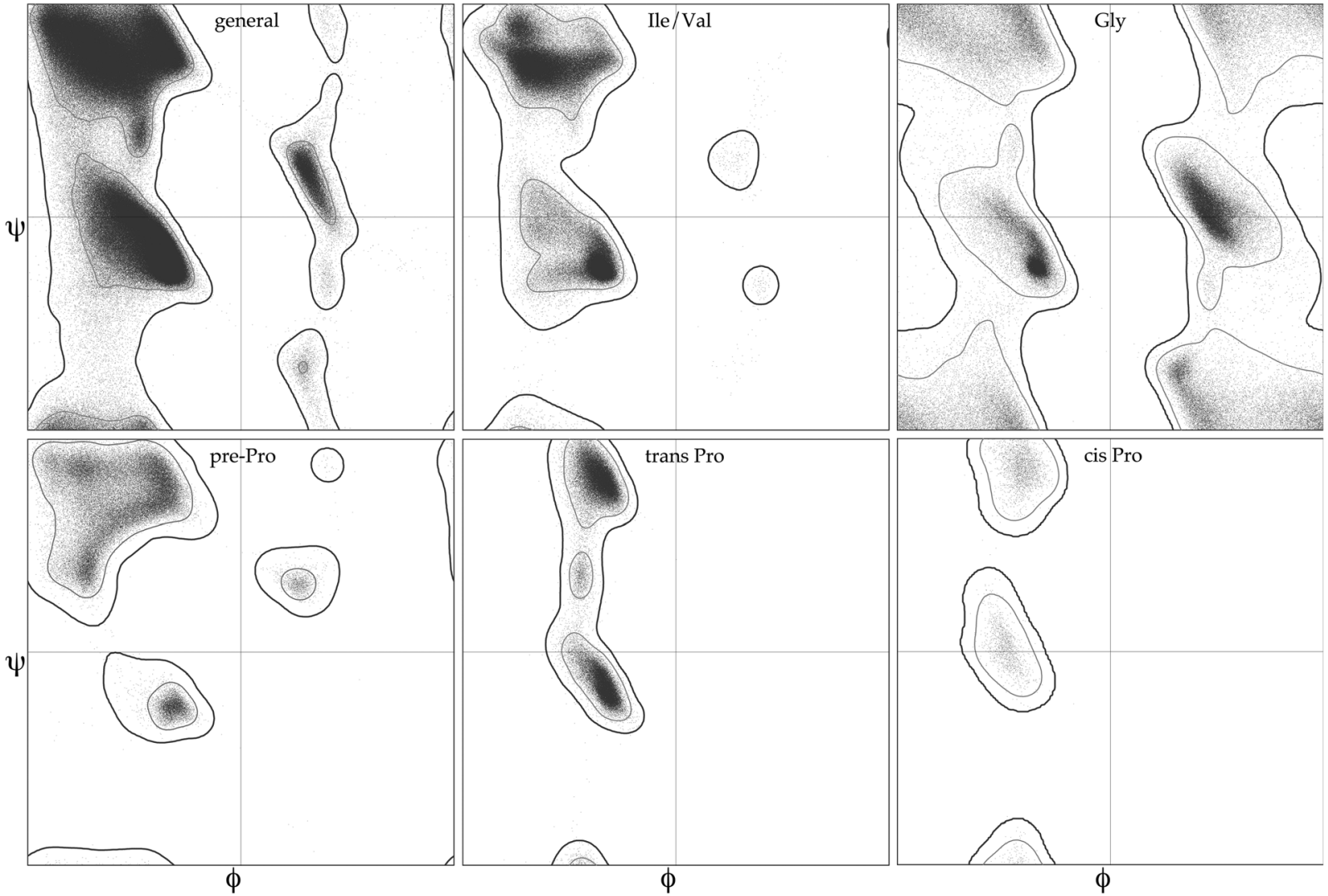
Model validation: clashes

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

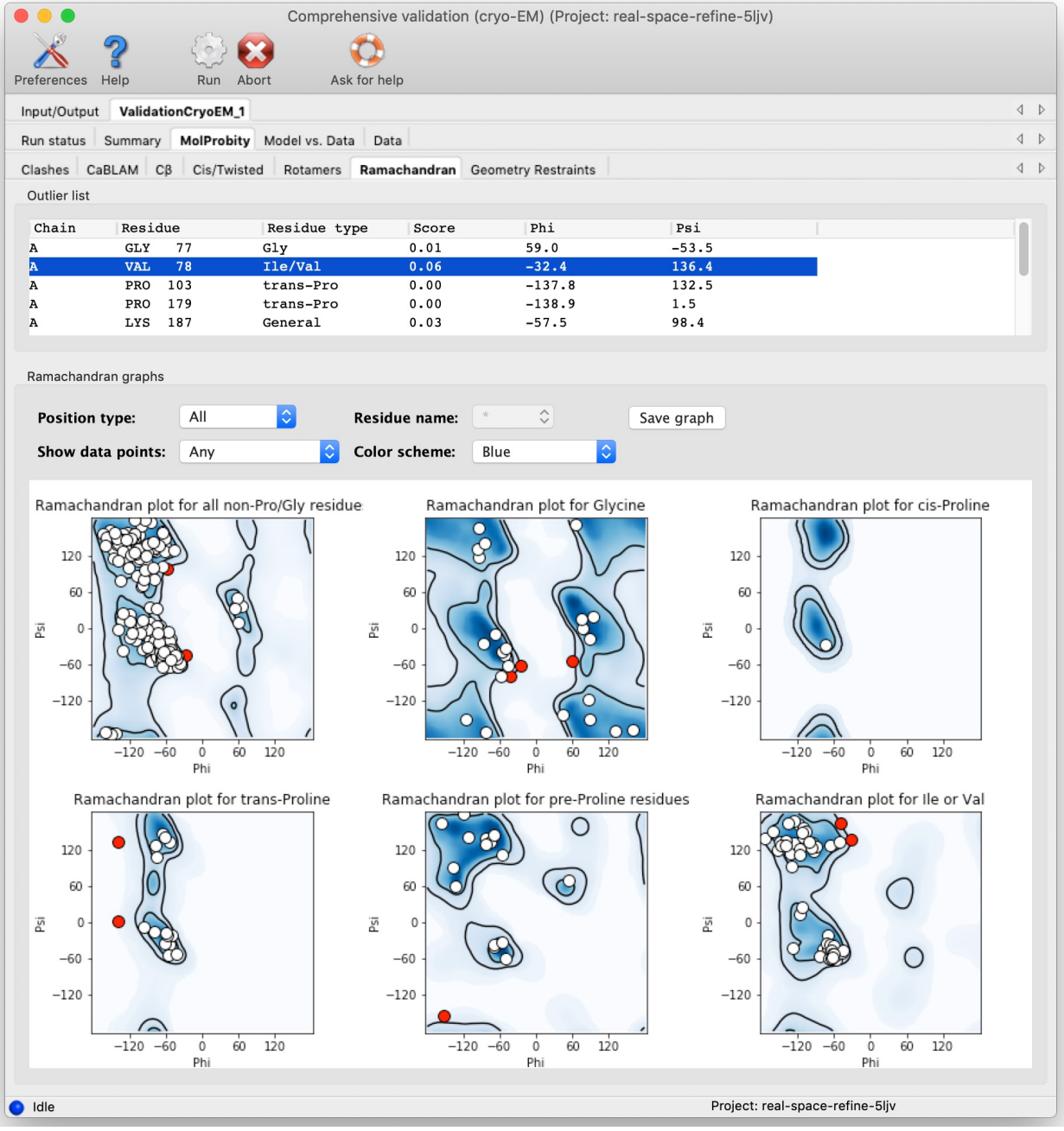


Model validation: Ramachandran plot

- Different plots for different classes of residues



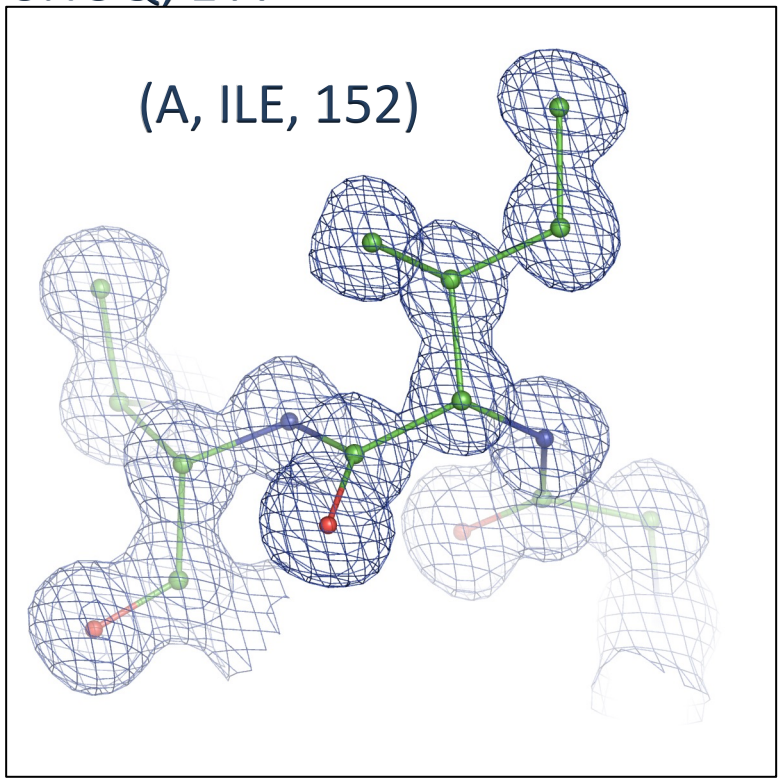
Model validation: Ramachandran plot

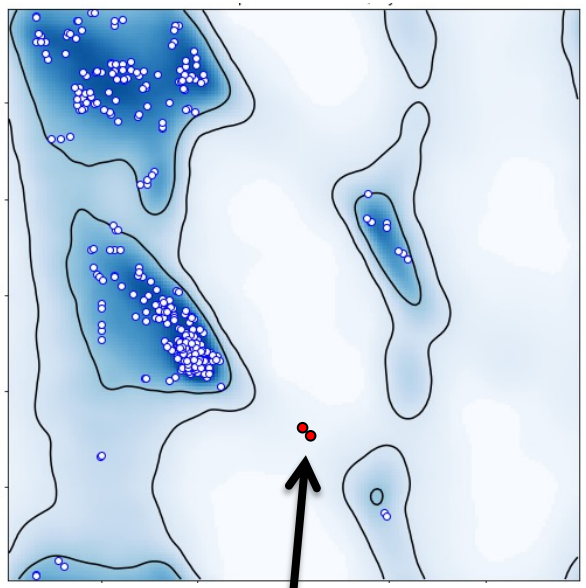


Model validation: Ramachandran plot

- A Ramachandran plot outlier ≠ wrong

3NOQ, 1 Å





Outliers:

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

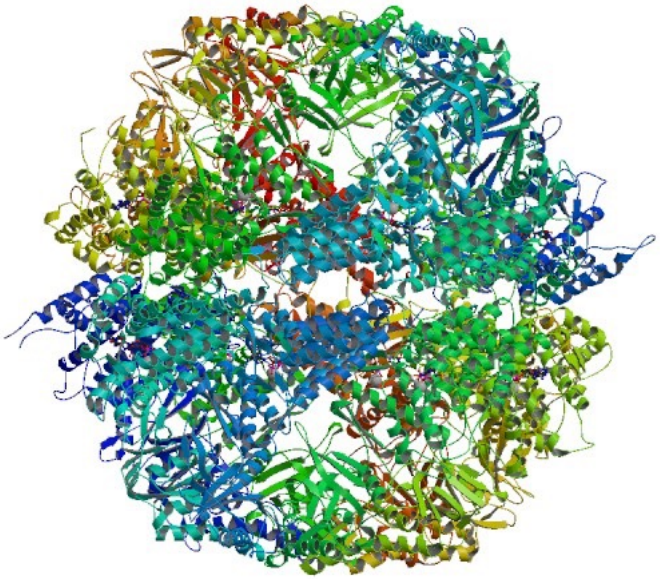
- All outliers need to be explained (supported by the data)

Model validation and refinement

- Validation metrics progressively become refinement goals
    - Ramachandran plot restraints
    - Cβ deviation restraints
    - Secondary structure restraints
    - Restraints on χ angles of amino-acid side-chain rotamers
- As result, validation becomes less capable of catching problems

Model validation

PNAS, 2019 116 (39) 19513-19522

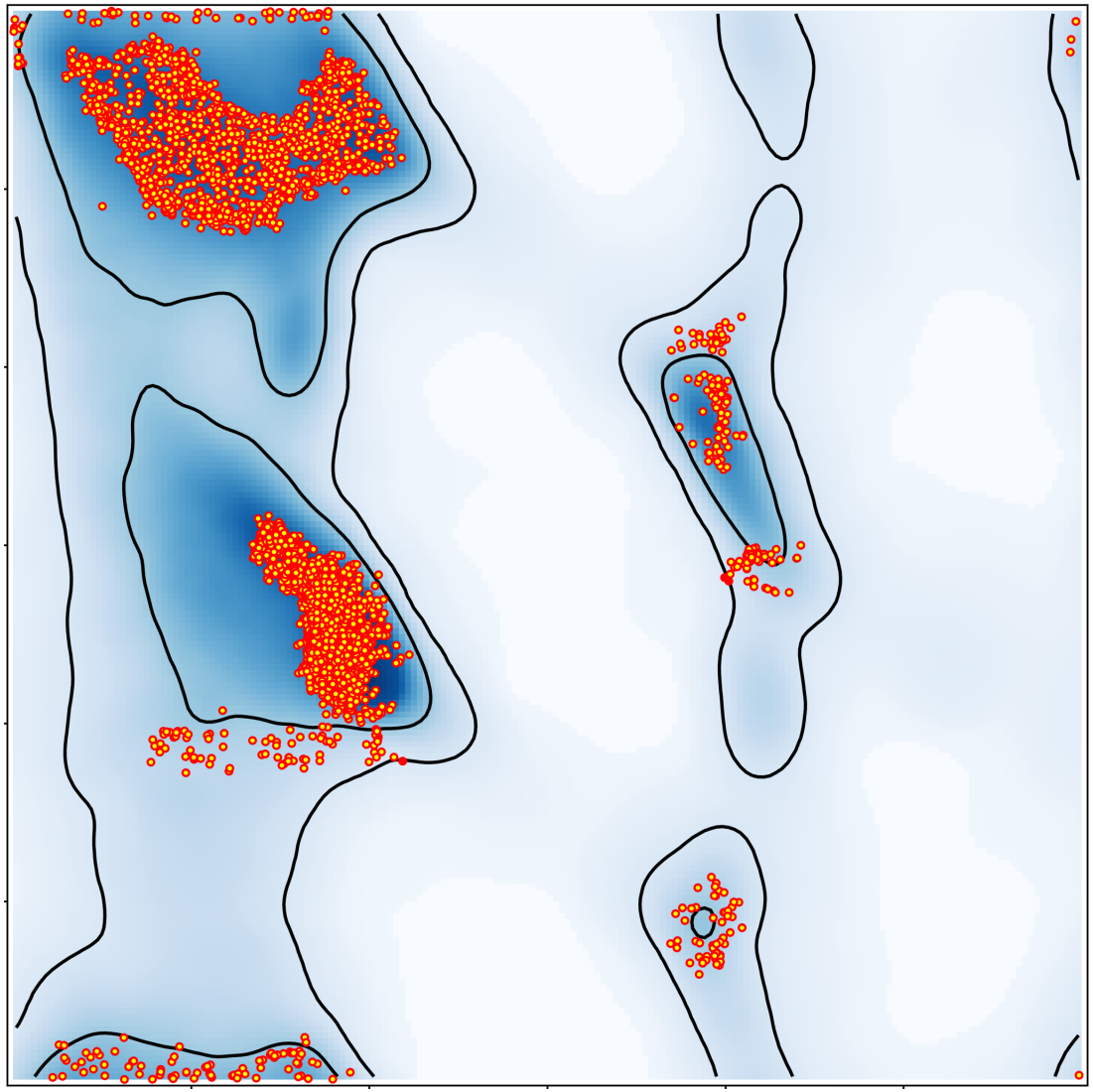


Metric / PDB code		6KS6
Clashscore		7.7
Rama. (%)	favored	96.4
	outliers	0.2
Rotamer outliers (%)		0
C_β deviations		0
RMSD	Bond (Å)	0.001
	Angle (°)	0.396
Resolution (Å)		3.0

Perfect statistics! All looks just great!

Model validation: *Ramachandran plot*

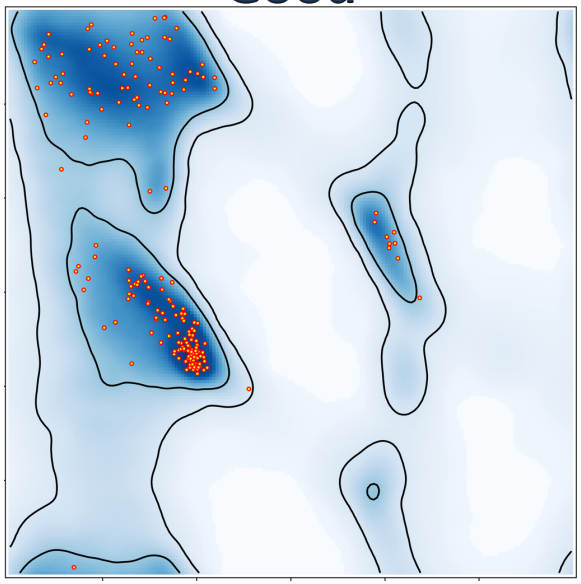
6KS6



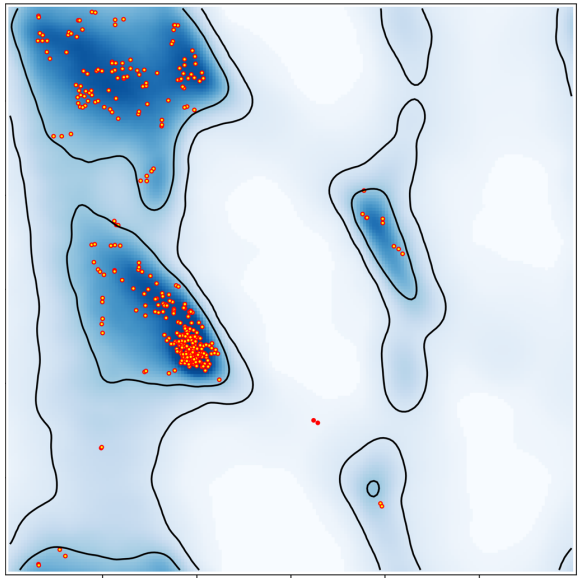
Odd Ramachandran plot. How we know this?

Model validation: *Ramachandran plots*

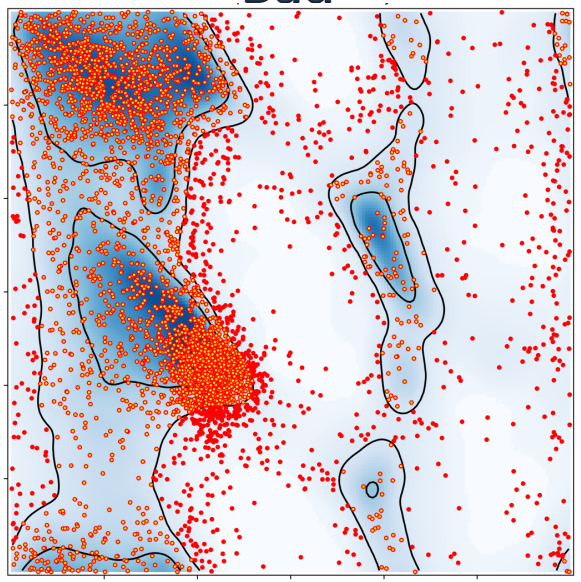
Good



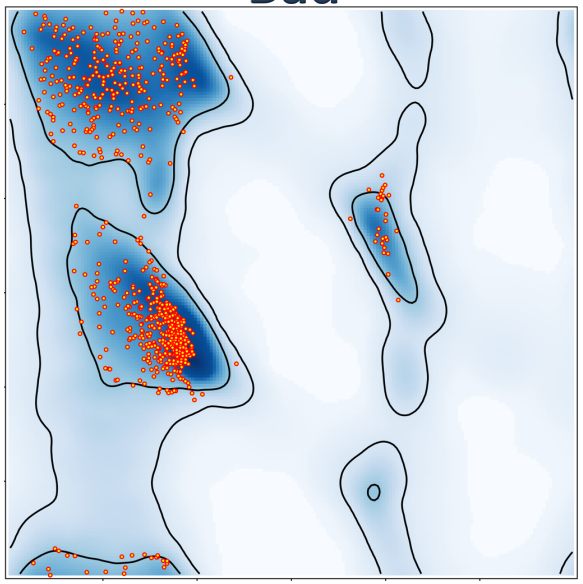
Good



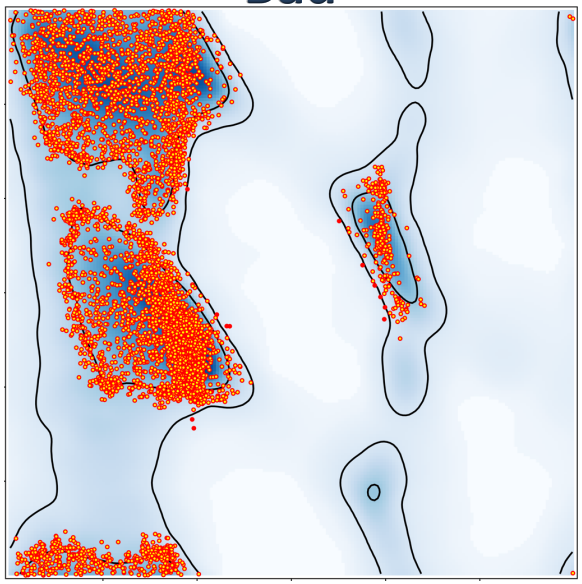
Bad



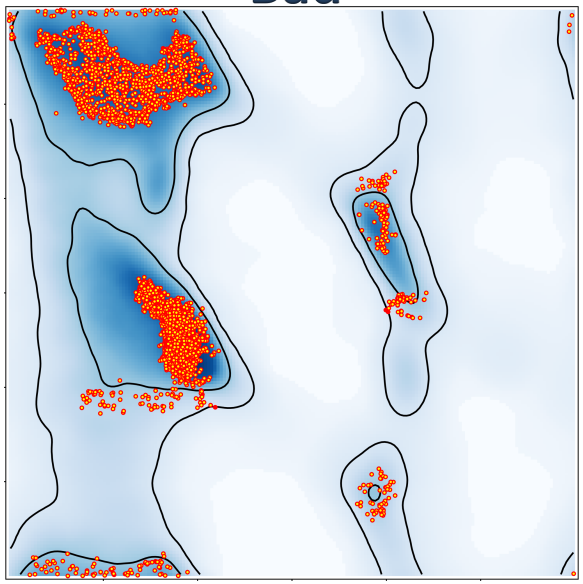
Bad



Bad



Bad



Model validation: *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

- Ramachandran Z-score is good at spotting odd plots
- Used in PDBREDO. Implemented in *Phenix* in all validation
- One number, simple criteria:
  - Poor: $|Z| > 3$ Suspicious: $2 < |Z| < 3$ Good: $|Z| < 2$

Structure

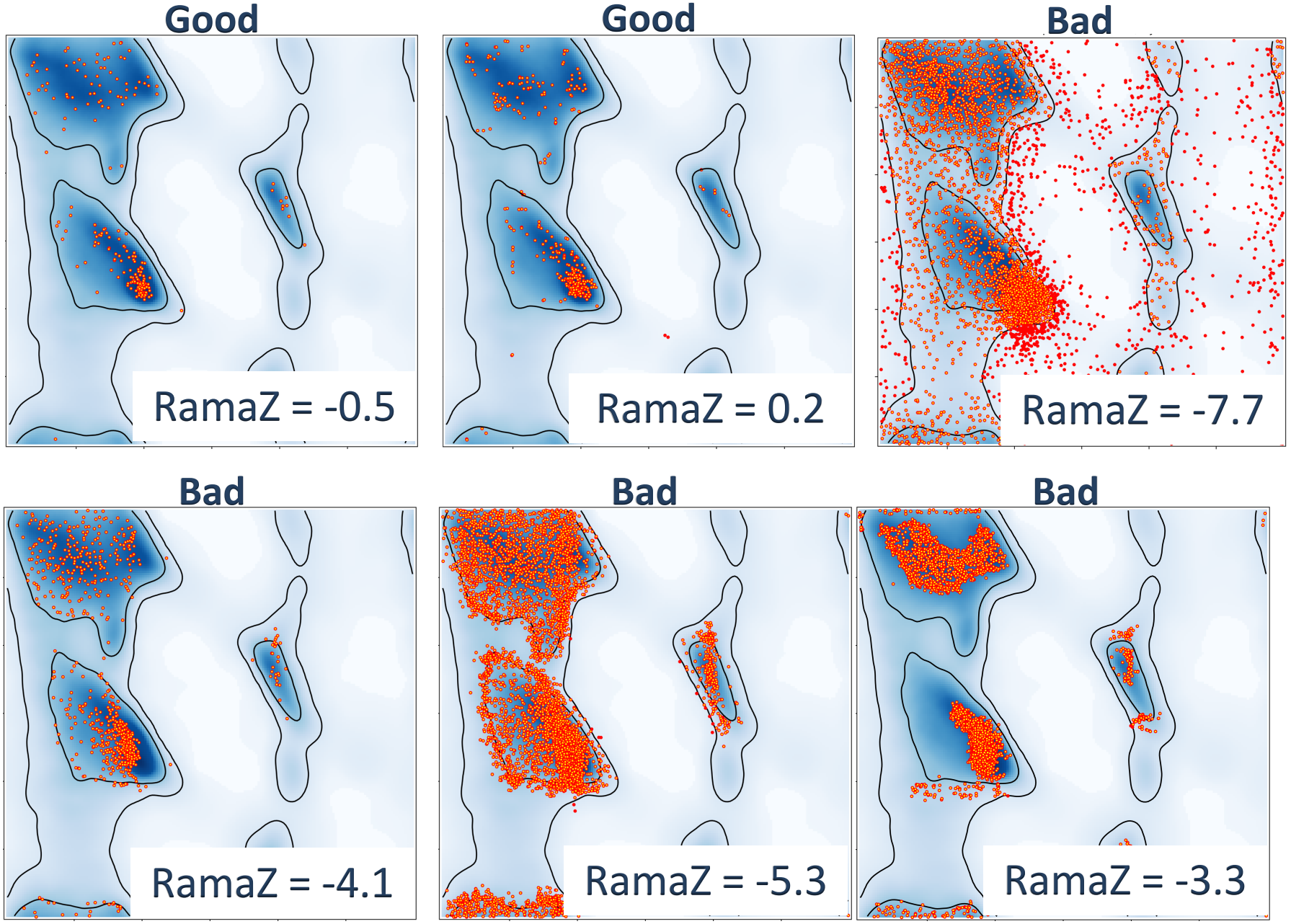
CellPress

Resource

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

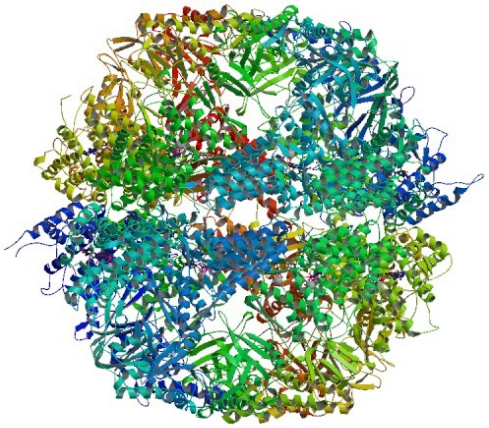
Oleg V. Sobolev,[1,5,*] Pavel V. Afonine,[1] Nigel W. Moriarty,[1] Maarten L. Hekkelman,[2,3] Robbie P. Joosten,[2,3,*] Anastassis Perrakis,[2,3] and Paul D. Adams[1,4]

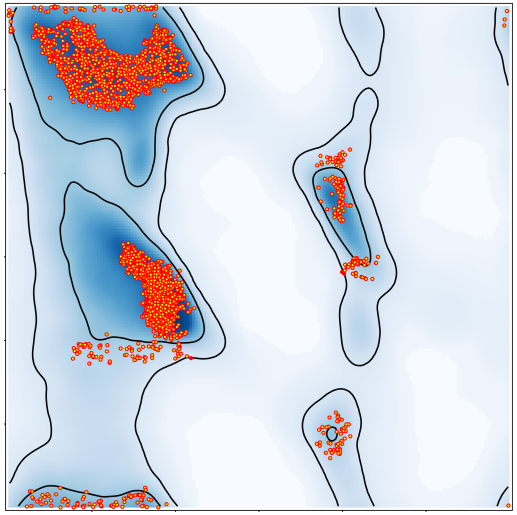
Model validation: *Ramachandran plot Z-score*



Model validation: *Ramachandran plot Z score*

PNAS, 2019 116 (39) 19513-19522





Metric / PDB code		6KS6
Clashscore		7.7
Rama. (%)	favored	96.4
	outliers	0.2
Rotamer outliers (%)		0
C_β deviations		0
RMSD	Bond (Å)	0.001
	Angle (°)	0.396
Resolution (Å)		3.0

RamaZ = -3.3

Poor: |Z| > 3 Suspicious: 2 < |Z| < 3 Good: |Z| < 2

Always think 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…

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

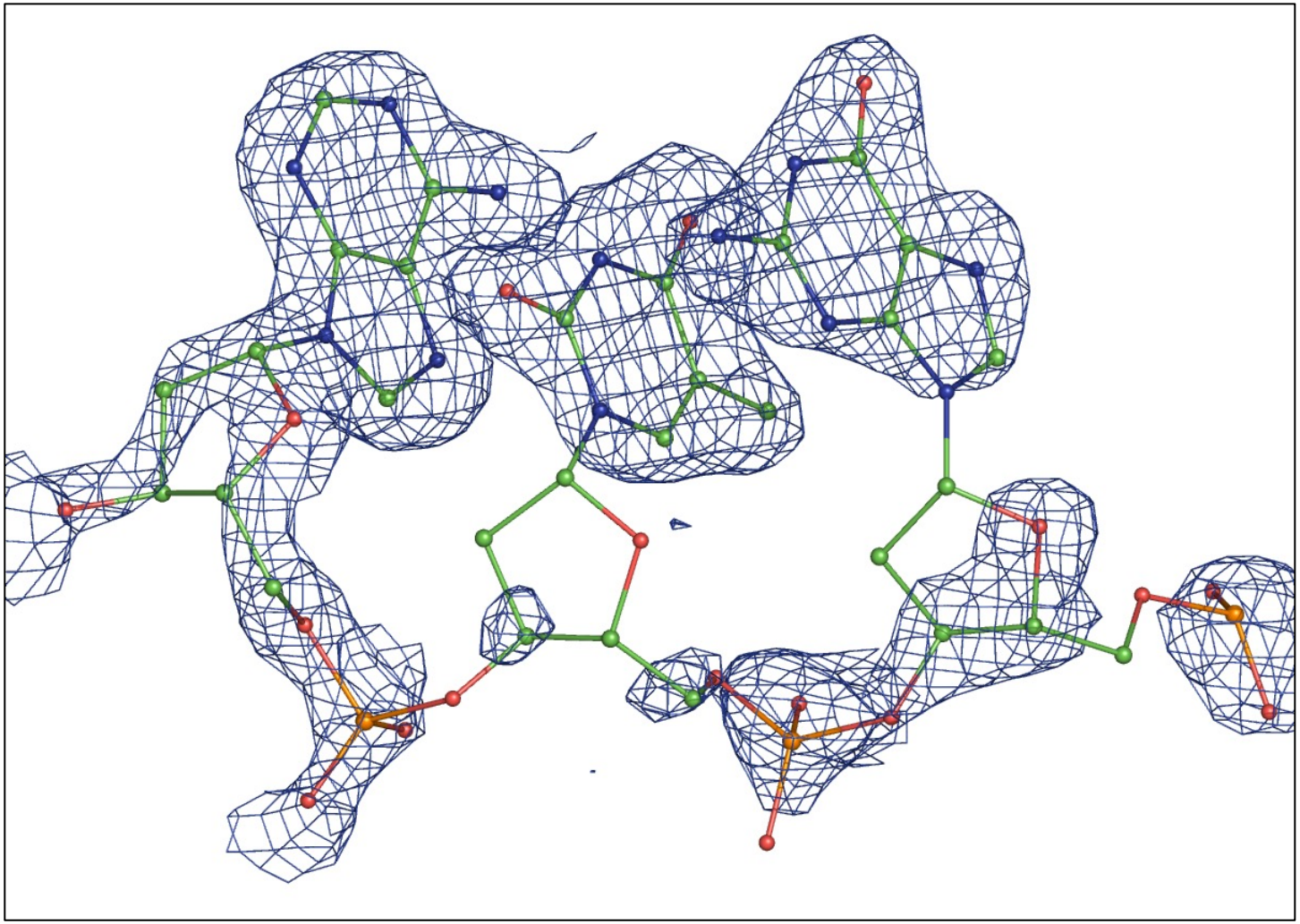
Bonds		Angles	
0.000 - 0.035:	2645	0.000 - 9.313:	4208
0.035 - 0.070:	19	9.313 - 18.626:	9
0.070 - 0.106:	13	18.626 - 27.939:	3
0.106 - 0.141:	5	27.939 - 37.252:	4
0.141 - 0.176:	3	37.252 - 46.565:	0
0.176 - 0.211:	0	46.565 - 55.878:	0
0.211 - 0.246:	0	55.878 - 65.191:	2
0.246 - 0.281:	0	65.191 - 74.504: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

Data completeness

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

2mFo-DFc , 1σ



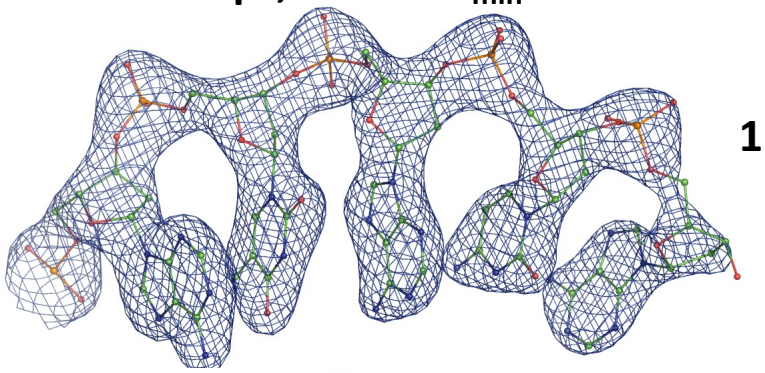
Local vs Global

Completeness by resolution:

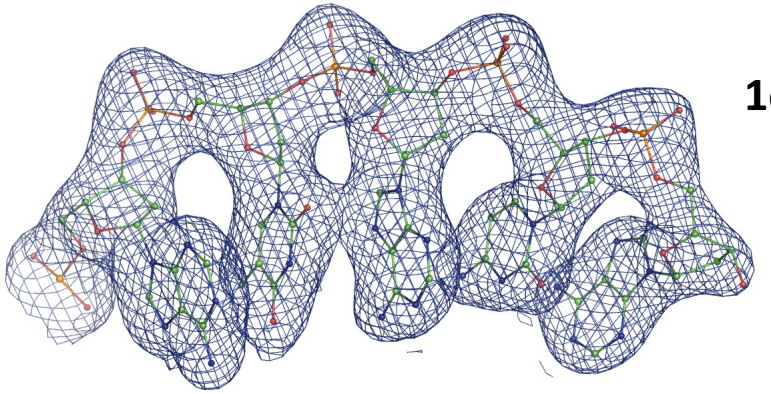
```
 19.9274 -  3.2441 0.78
  3.2441 -  2.5767 0.99
  2.5767 -  2.2515 1.00
  2.2515 -  2.0459 1.00
  2.0459 -  1.8993 0.99
```

Overall completeness in d_{min}-inf: 0.95

Fcalc maps, full set d_{min}-inf

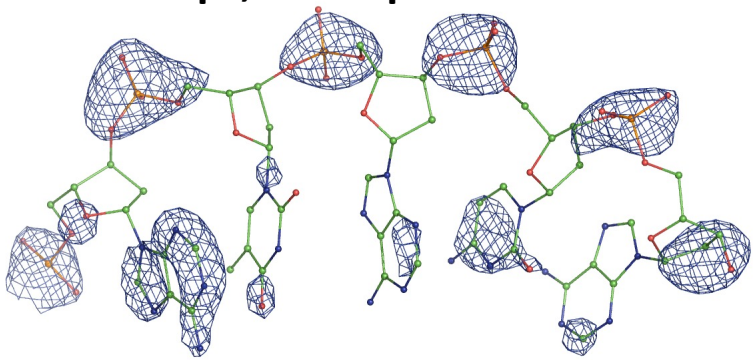


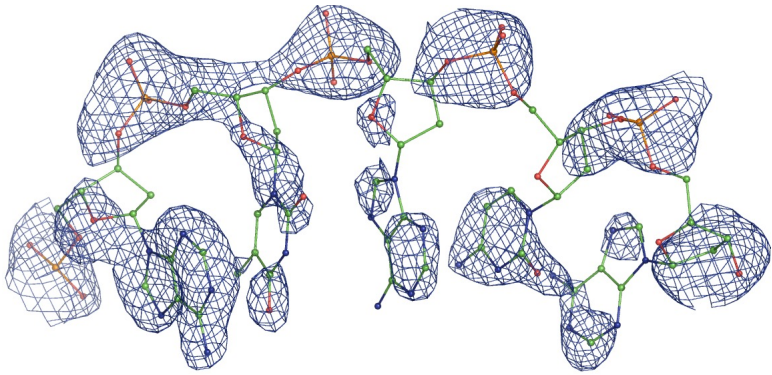
1.5σ map cutoff



1σ map cutoff

Fcalc maps, incomplete set



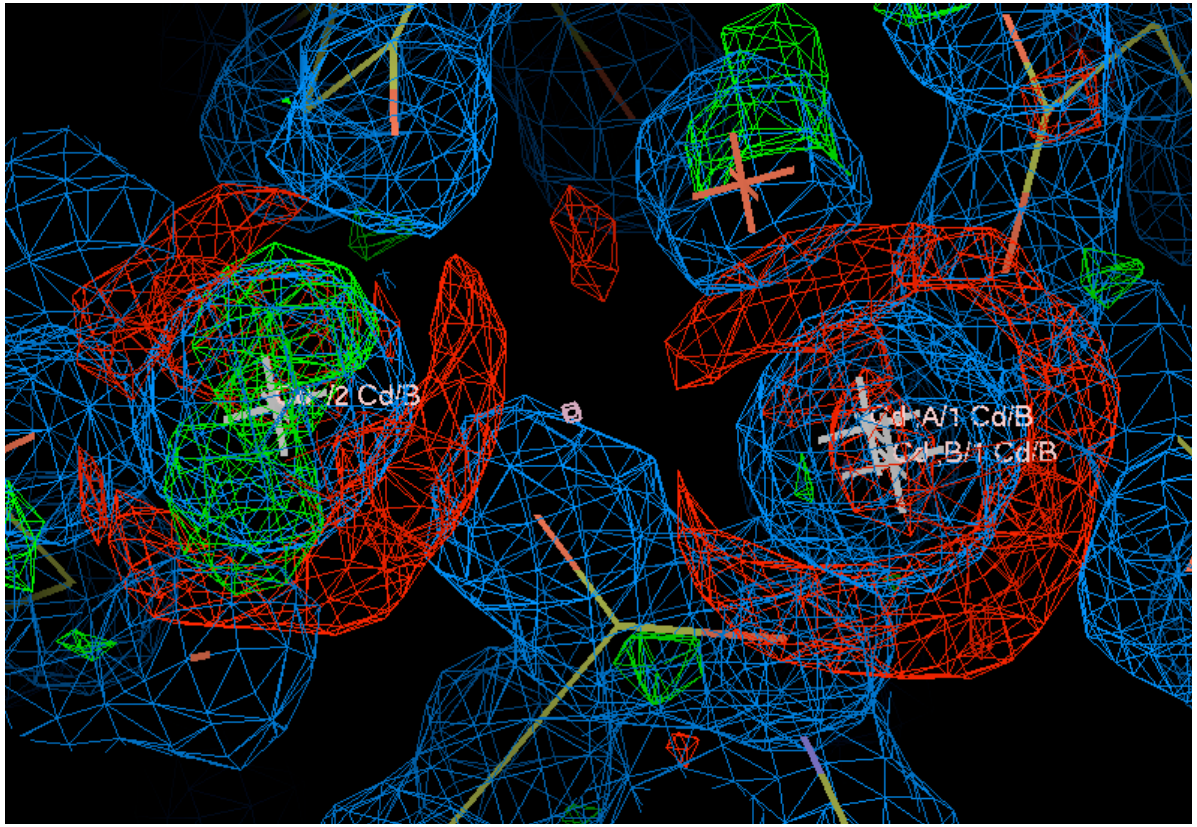


Data incompleteness distorts maps

Validation – Sequence register errors

```
MASTER  GFVDLTLHDQVSMEHPVKLLFGKCVEGMVEIVYTFLSSTLKSLE
Chain A GFVDLTRHDQVSMEHPGKLLFGK--EGMVEIVYTF-----KSLE
Chain B GFVDLTRHDQVSMEHPGKLLFGK--EGMVEIVYTFVSSTLKSLE
Chain C GFVDLTRHDQVSMEHPGKLLFGKKVEGMVEIVYTFVSSTLKSLE
Chain D GFVDLTRHDQVSMEHPGKLLFGKKVEGMVEIVYTFLSSTLKSLE
        ****** ********* ******  **********     ****
```

Map and model errors



- Errors in model parameters
- Fourier artifacts

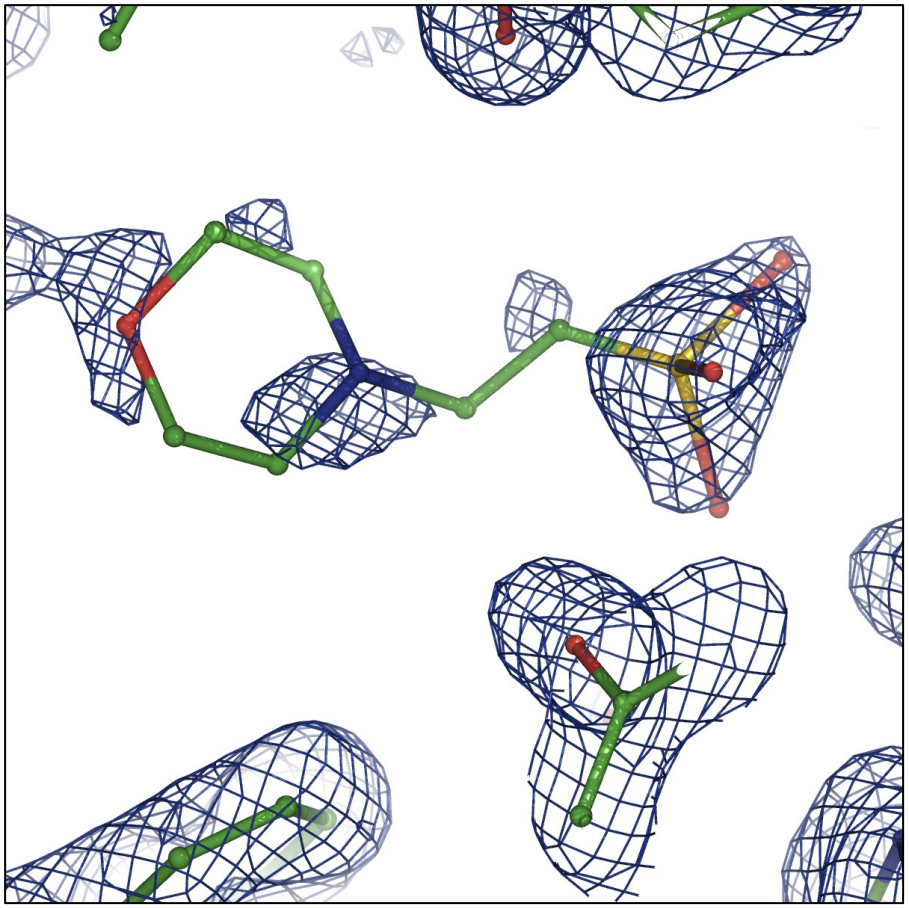
Things to try:

- Refine occupancy, anisotropic ADP, anomalous f’ & f’’, define charge.
- Do many refinement cycles (so refinement converges)
- Consider alternative ion

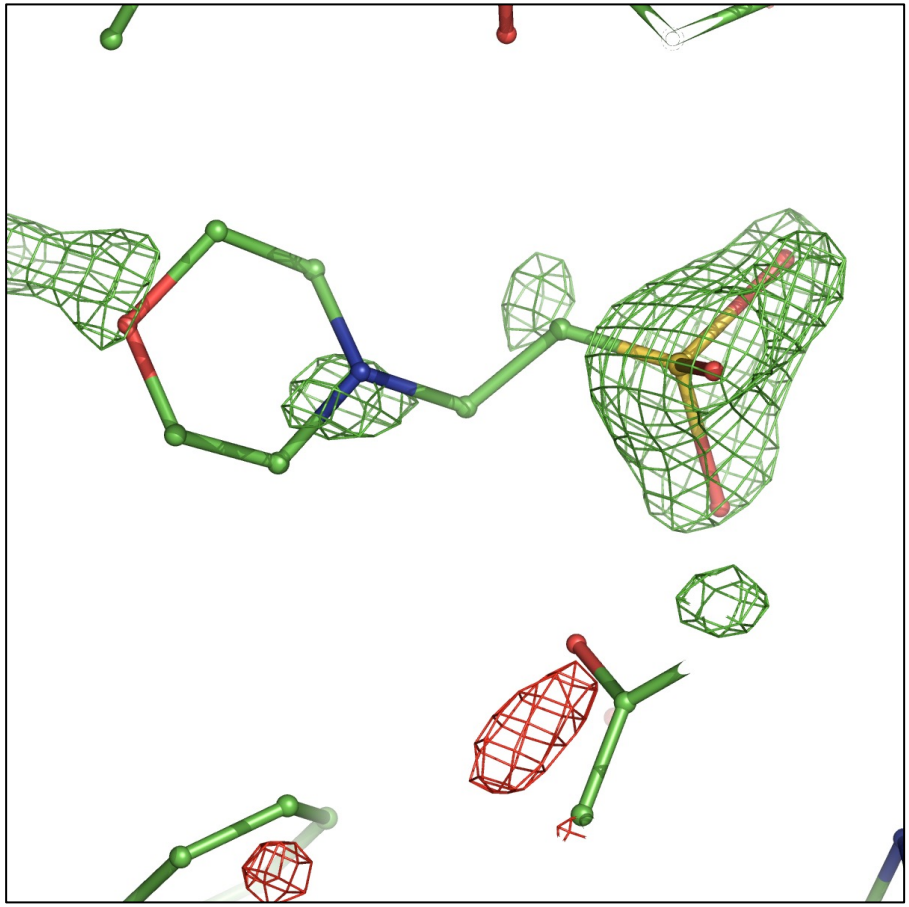
Ligands and Polder map

PDB code: 1ABA, Resolution: 1.45 Å

2mFo-DFc (1σ)



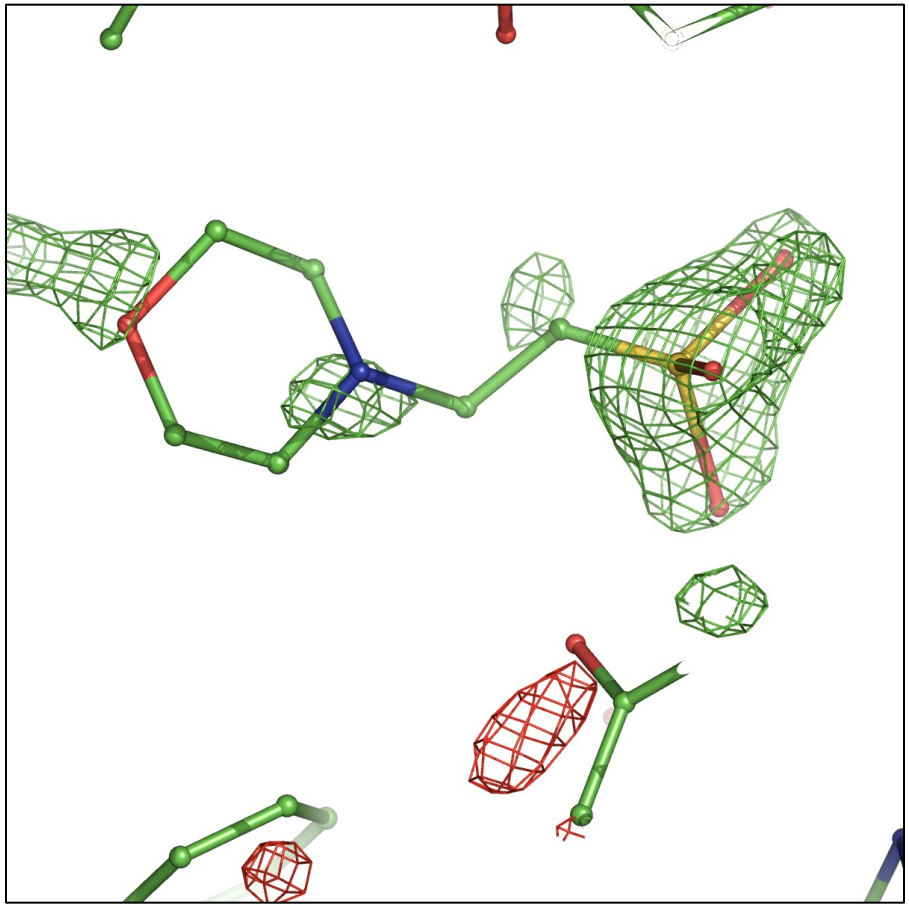
mFo-DFc (±3σ)



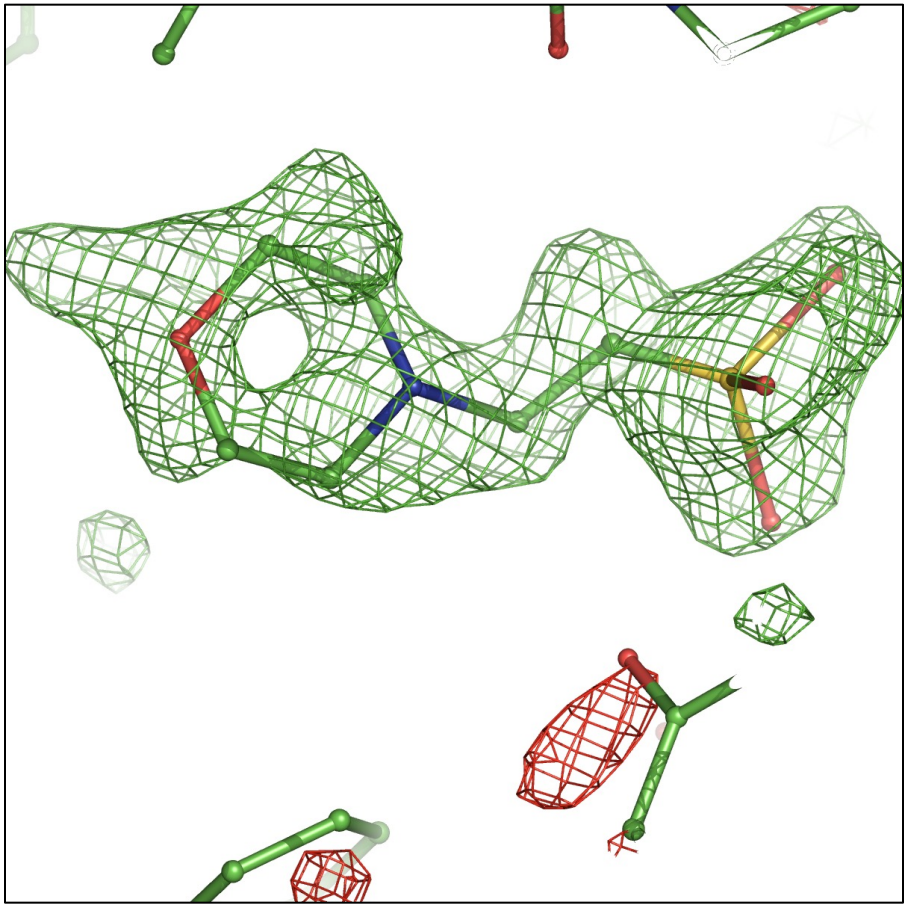
Ligands and Polder map

PDB code: 1ABA, Resolution: 1.45 Å

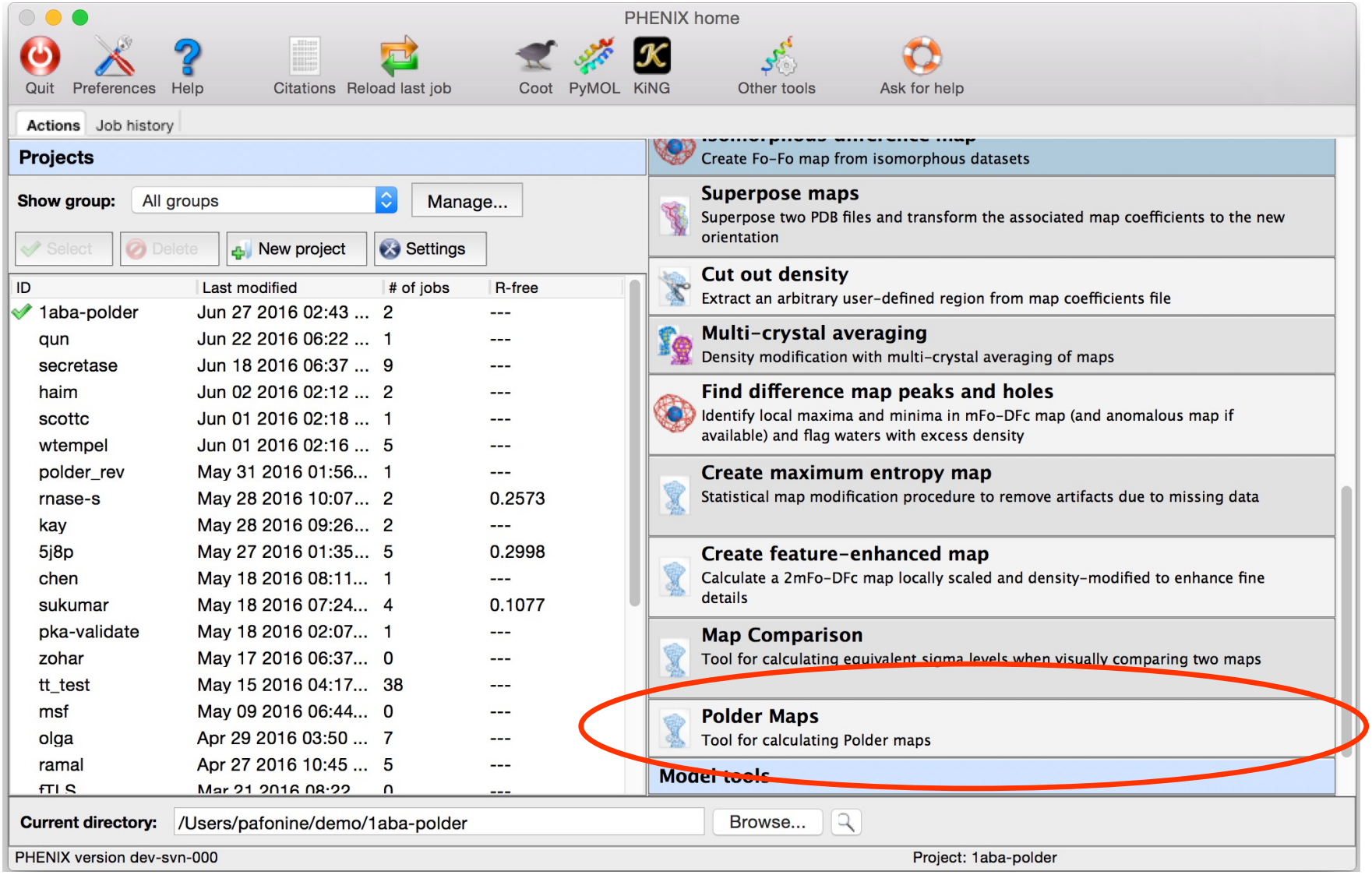
mFo-DFc (±3σ)



Polder mFo-DFc (±3σ)

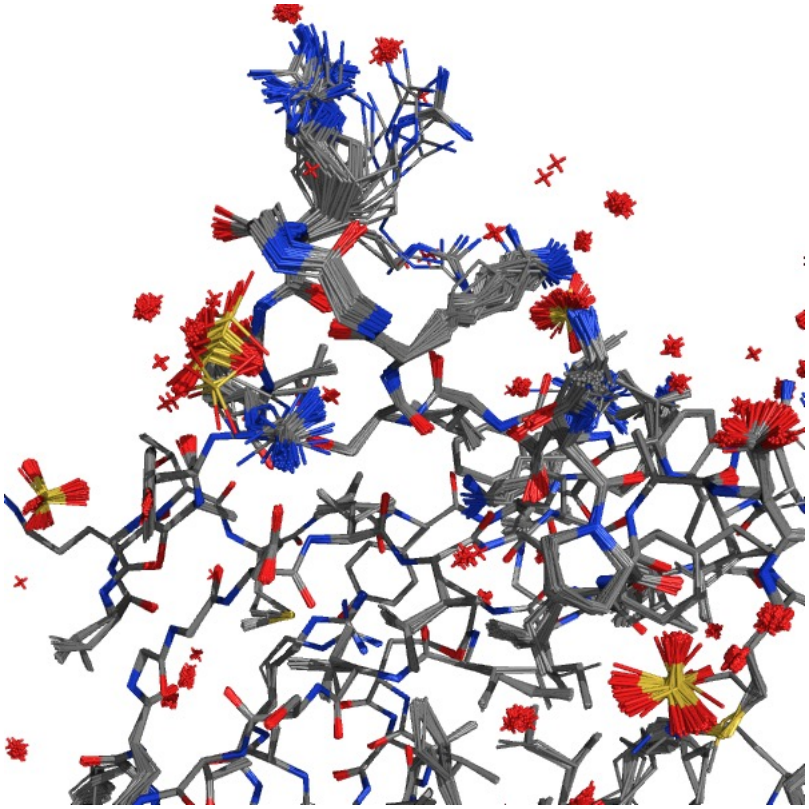


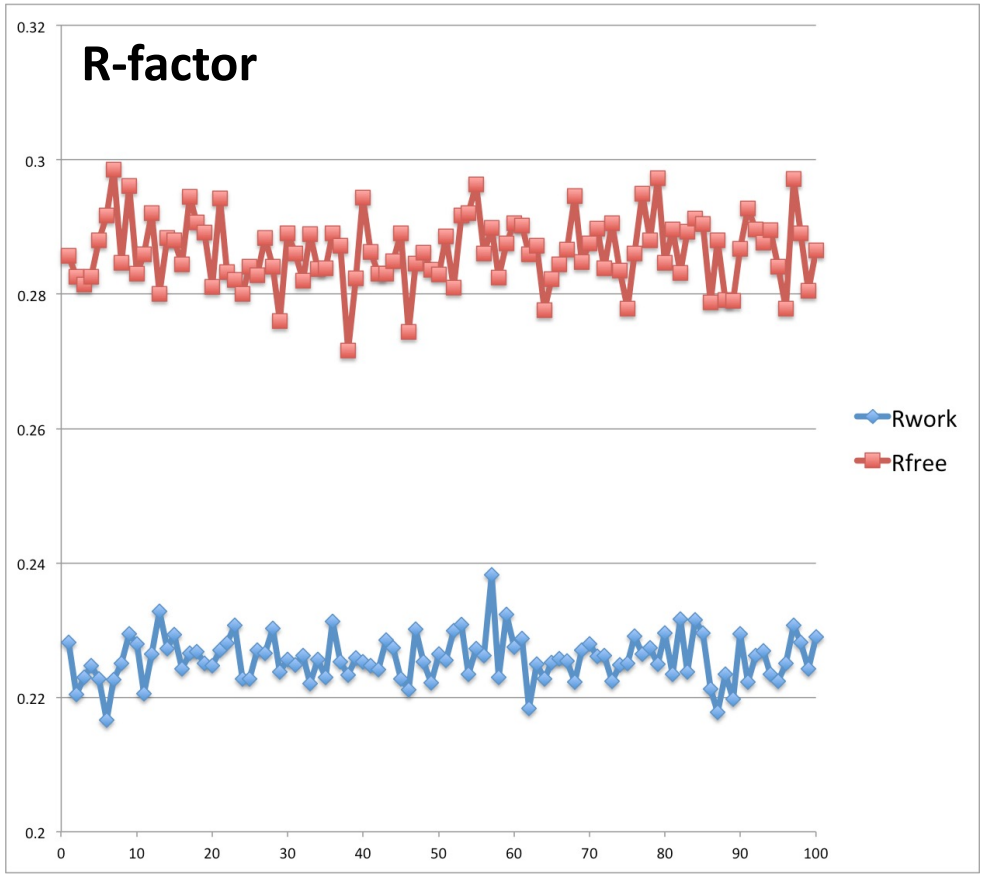
Ligands and Polder map



Estimating and using uncertainty

100 identical refinement runs each one starting with slightly perturbed model





Support

- Feedback, questions, help

  phenixbb@phenix-online.org
  bugs@phenix-online.org
  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

Subscribe to *Phenix* mailing list: www.phenix-online.org