Crystallographic Structure Validation and Relevant Phenix Tools

Pavel Afonine

Phenix developer

Lawrence Berkeley National Labs
Berkeley, CA, USA

Comprehensive validation — why?

- Problems detected early can save a lot of time later
- Subjectivity in data interpretation
  - Human interpret the maps: experience, skills, pressure
- Human program the software
  - Programs may contain bugs
- Subjectivity in model building and refinement
  - Model parameterization, target weights, starting points
  - Luck of data = multiple possibilities for interpretation
- Post-refinement pre-deposition manipulations
  - Hand editing files: removing waters, hydrogens, ANISOU
- Misusing quality metrics
  - Choose single water or decide about twinning using R-factor
- Fraud or honest mistakes

Comprehensive validation — quality filters

- Crystallographer (you)
  - Software you use
    - Your boss
      - Reviewers (of your paper)
        - PDB deposition (software and people)
          - Community (those who eventually may come across of your structure and use it)

- Unnoticed (intentionally or not) problems
  - Likely to be discovered anyway, sooner or later
  - There may be negative consequences

Comprehensive validation — mistakes & consequences

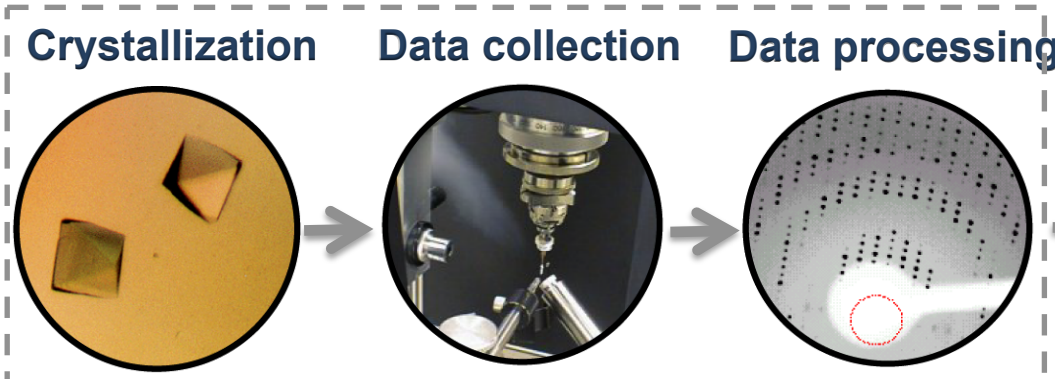
Retraction: Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution

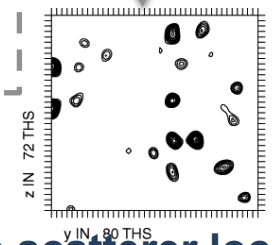
Michael A Hanson & Raymond C Stevens
Nat. Struct. Biol. 7, 687–692 (2000); retracted 6 July 2009

In this paper, we described both the three-dimensional crystal structure of a botulinum toxin catalytic domain separated from the holotoxin (BoNT/B-LC, PDB 1F82) and a structure of the toxin catalytic domain in complex with a peptide (Sb2–BoNT/B-LC, PDB 1F83). The complex was later refined and deposited in the Protein Data Bank (PDB 3G94). The apo structure (PDB 1F82) remains valid. However, because of the lack of clear and continuous electron density for the peptide in the complex structure, the paper is being retracted. We apologize for any confusion this may have caused.

- **H.M. Krishna Murthy – Protein Fabrication scandal**
  - **12 falsified structures and 10 related papers**
  - **1BEF, 1CMW, 1DF9, 2QID, 1G40, 1G44, 1L6L, 2OU1, 1RID, 1Y8E, 2A01, and 2HR0**
  - **Murthy's falsified data ended up affecting 449 papers**

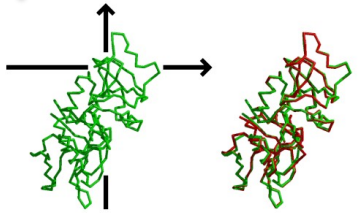
Comprehensive validation — what and when?



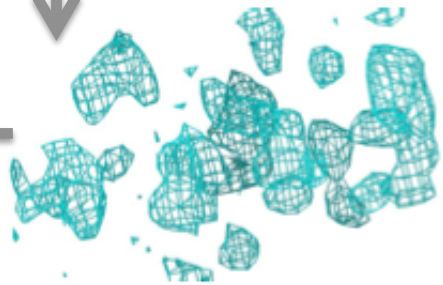


Anomalous scatterer location

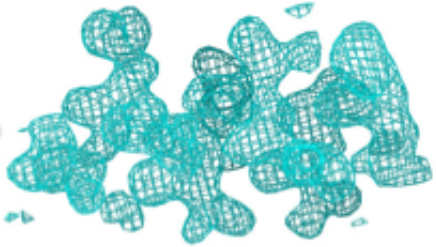
Molecular replacement



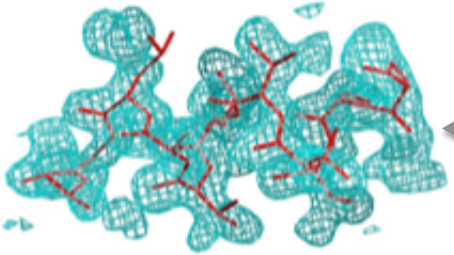
Phase determination



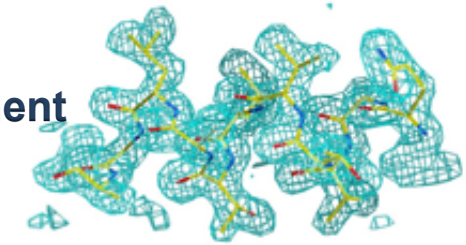
Phase improvement



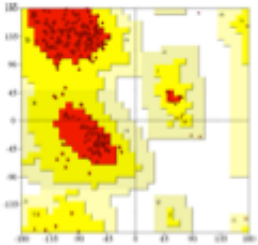
Map interpretation



Model refinement



Model validation



Comprehensive validation — what and when?

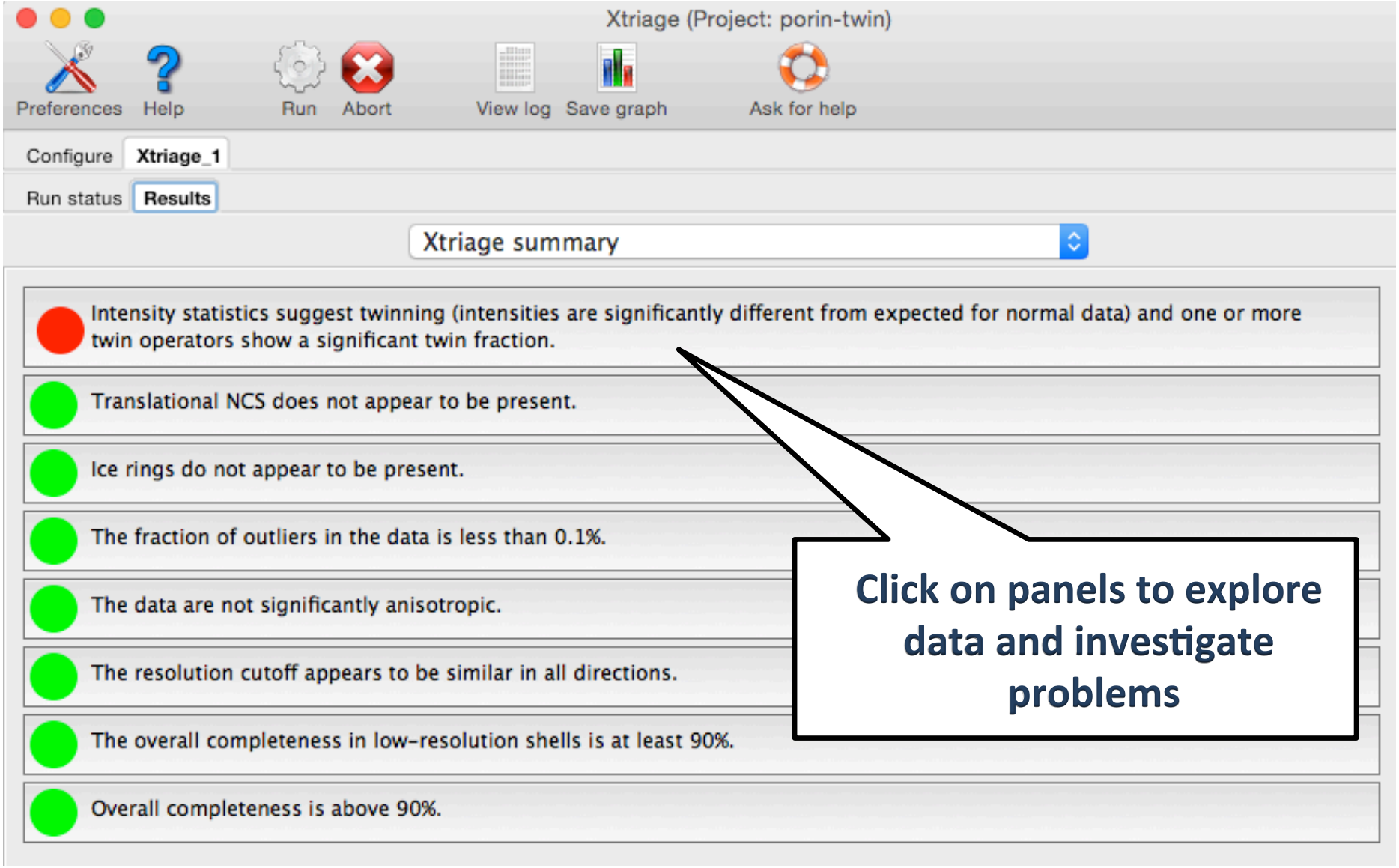
- **What:**
  - **Data**
    - **Global: overall R/R_{FREE}, completeness**
    - **Local: R, completeness, <Fobs> per resolution bin**
  - **Model**
    - **Global: bond/angle rmsd, rotamer, Ramachandran outliers**
    - **Local: local (per residue) geometry distortions**
  - **Model to Data Fit**
    - **Global: overall RSCC**
    - **Local: RSCC per atom or per residue, map values**

- **When:**
  - **Throughout the process of structure determination**

Comprehensive validation — data (check list)

- Matthews coefficient probabilities
- Completeness by resolution
- Wilson plot sanity
- Detection of translational NCS (tNCS)
- Analysis of systematic absences and combination of tNCS with current space group
- Anomalous signal from measurability analysis
- Symmetry and twinning analyses
- Alternative point-group symmetry (can be detected on the basis of an R-value analyses)

Comprehensive validation — Xtriage to check data

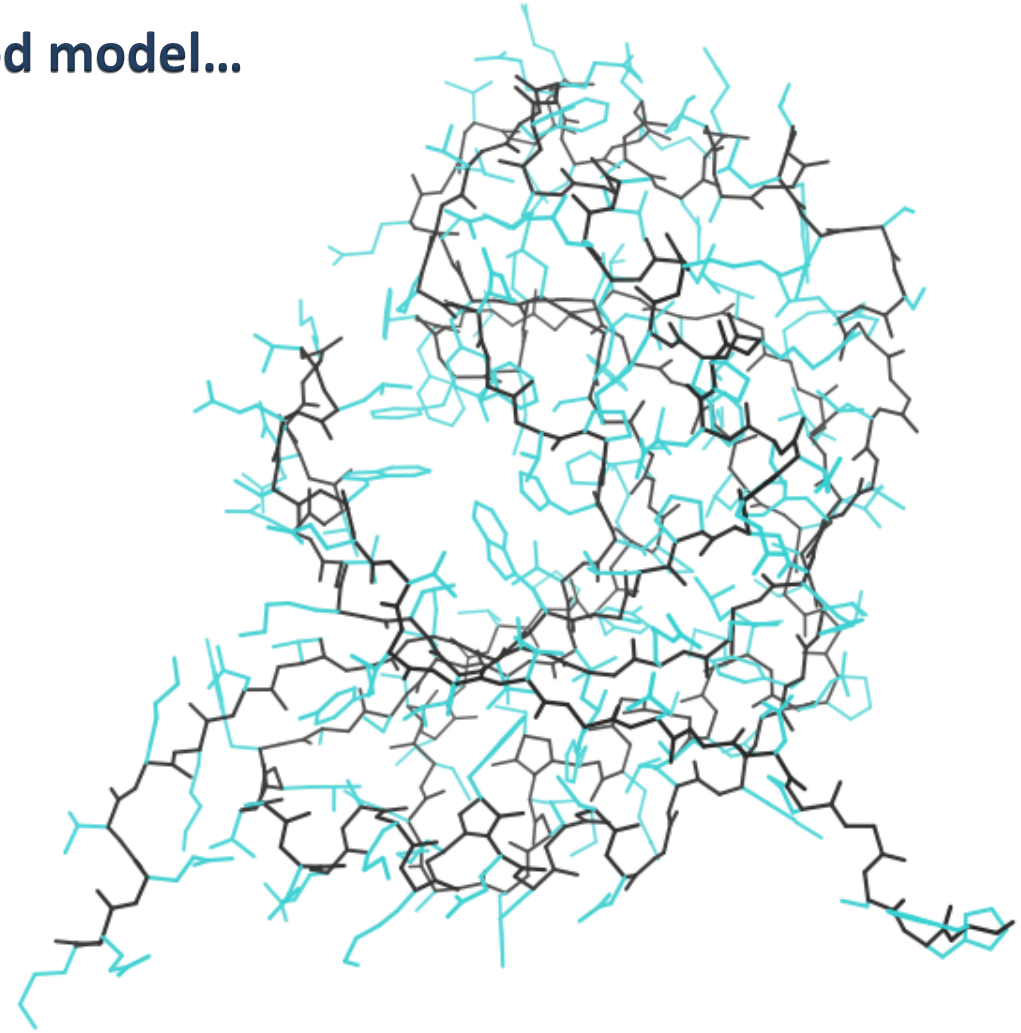


Comprehensive validation — Model

- A good model should be sound in all senses
  - Physically
    - packing, contacts, crystal solvent content
  - Chemically
    - correct local geometry, bonds, angles, etc
  - Crystallographically
    - R-factors, B-factors, map fit, bulk-solvent, symmetry
  - Statistically
    - No under-modeling (under-refinement), no over-fitting (over-modeling)
    - Model global quality figures are expected to be in agreement with corresponding values found in similar structures

Comprehensive validation — Model (steric clashes)

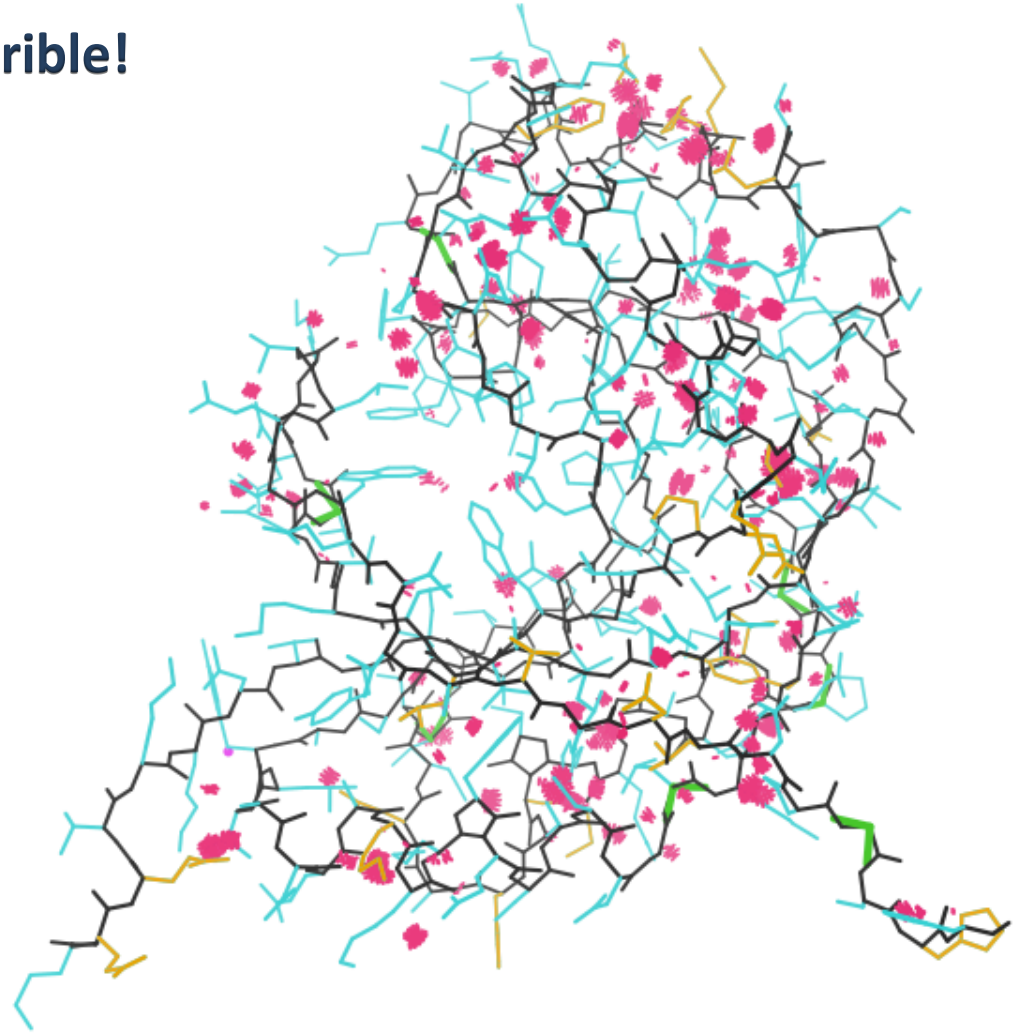
Looks like a good model…



Cyclic Nucleotide Phosphodiesterase (2.4 Å)

Comprehensive validation — Model (steric clashes)

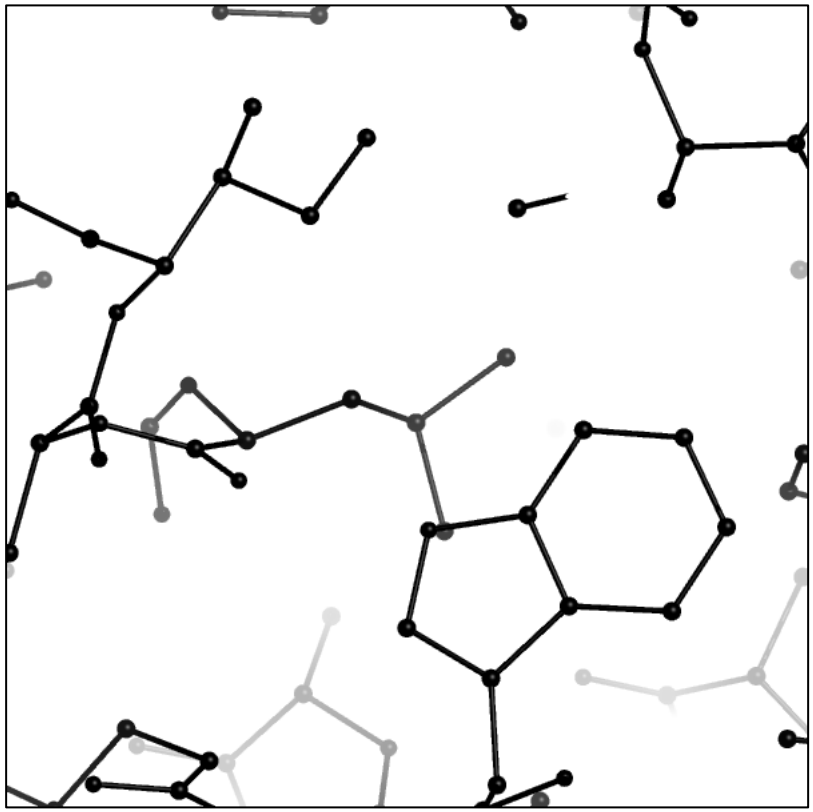
… in fact it is terrible!



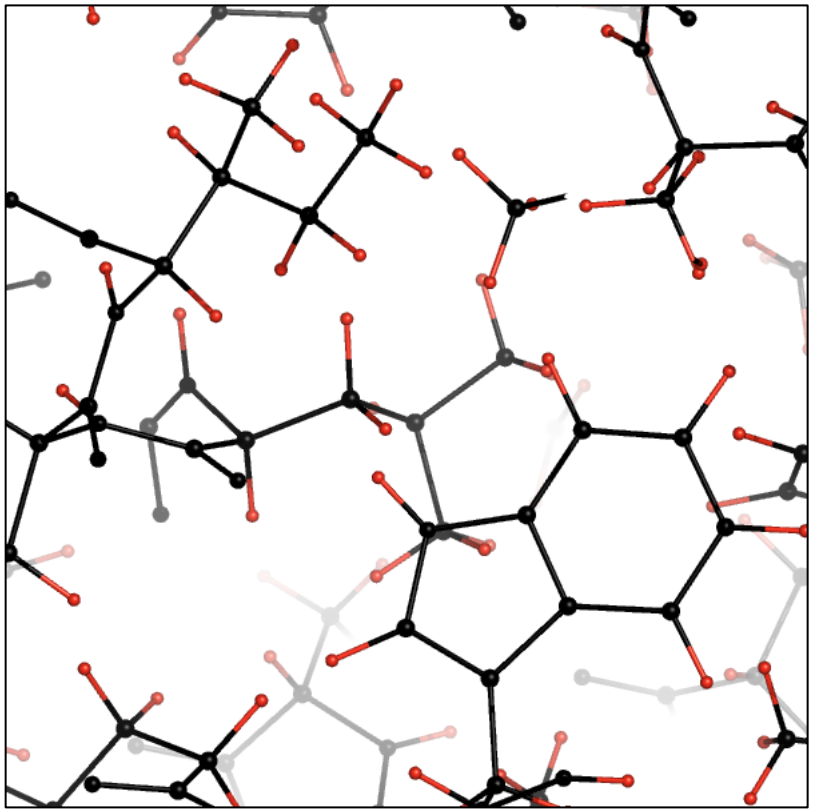
Cyclic Nucleotide Phosphodiesterase (2.4 Å)

Comprehensive validation — Fixing model clashes

- Refinement with H usually helps, because:
  - Half of the atoms in a protein molecule
  - Make most interatomic contacts
- Use weight optimization in refinement
- Resolve severe clashes manually



No H atoms



H atoms added

Comprehensive validation — Model (Rotamers)

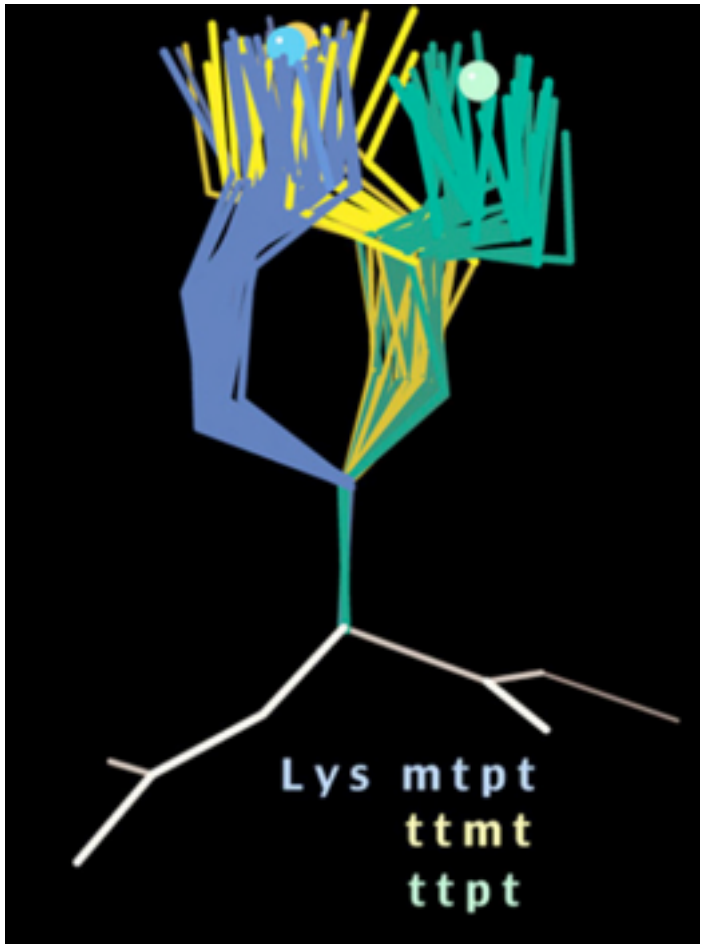
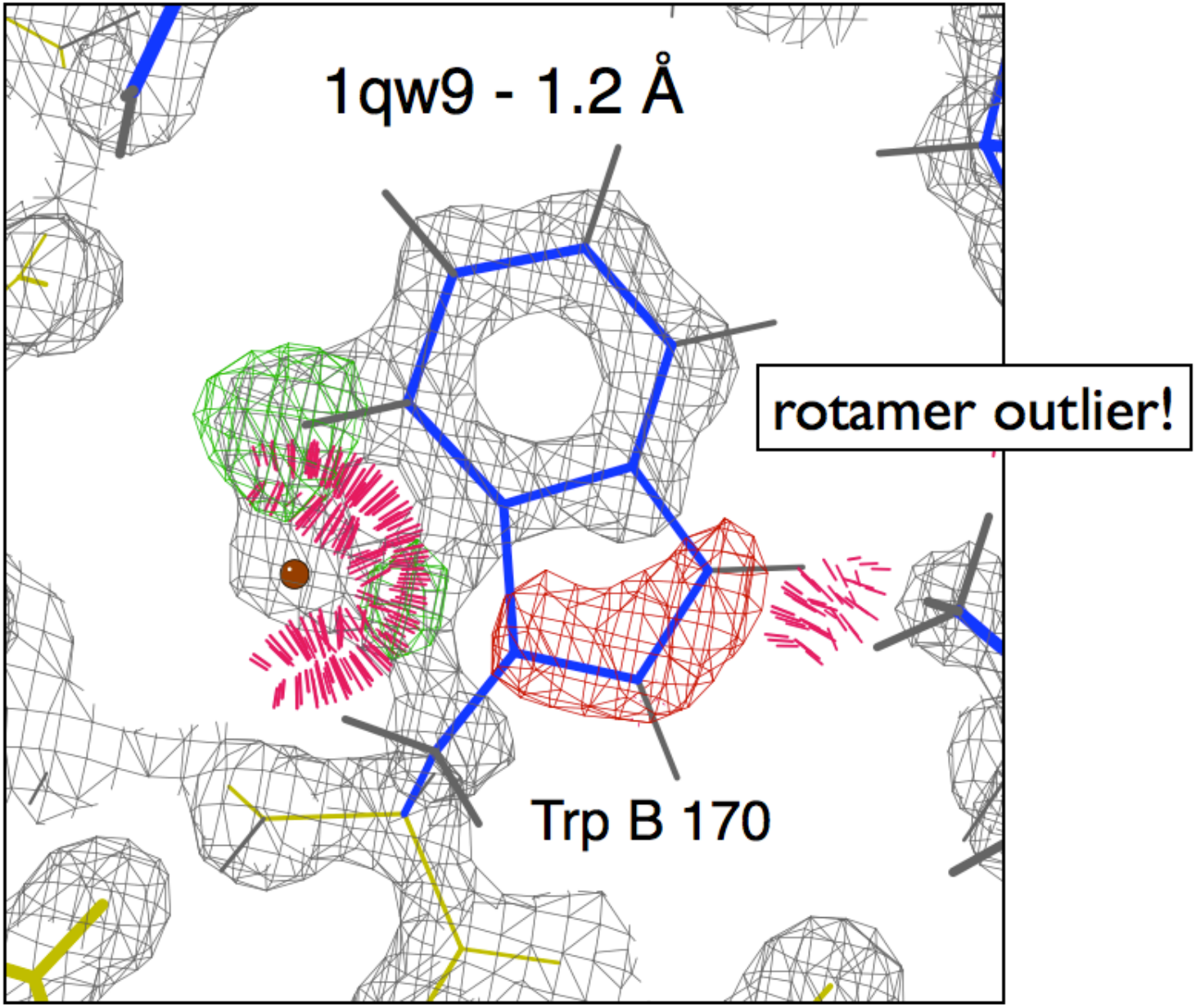


Image from Jane and David Richardson, Duke University

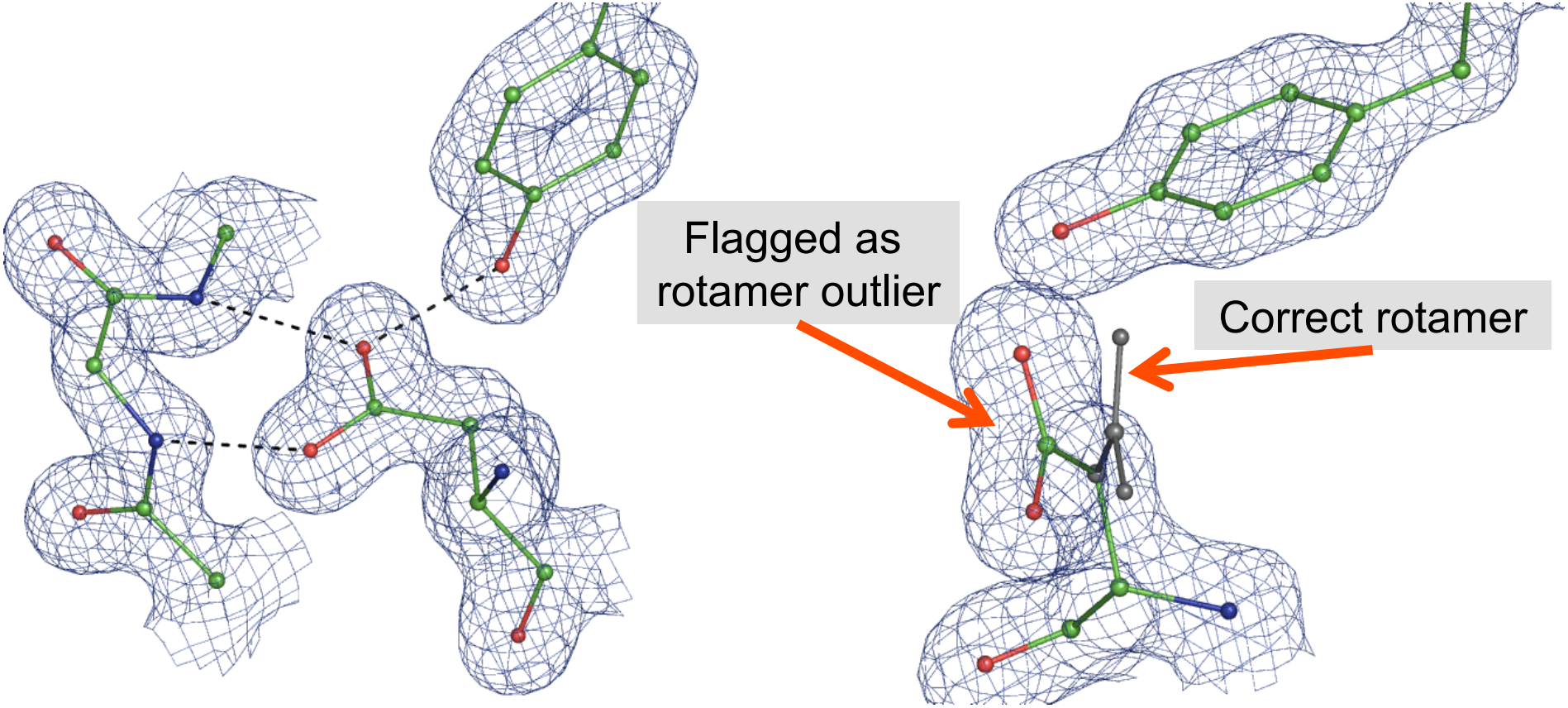
- Steric clashes between atoms within amino acid side chains lead to preferred conformations, called rotamers
- Different rotamers are generated by rotation of side chain torsion angle

Comprehensive validation — Model (rotamer outliers)

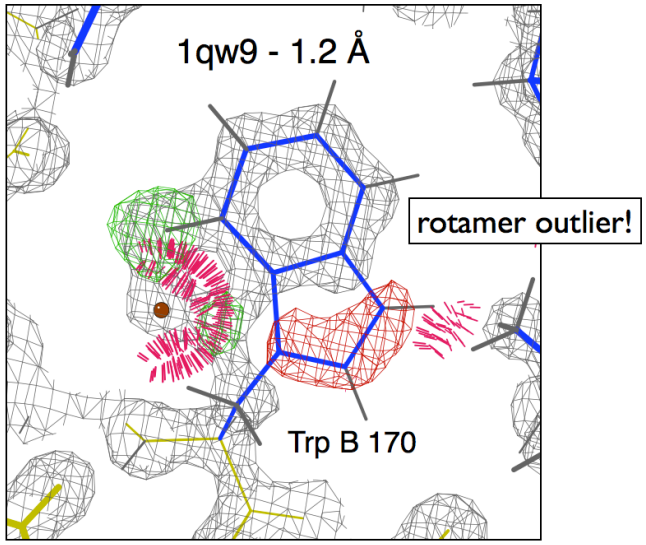


Comprehensive validation — valid rotamer outliers

- **Not everything flagged as outlier is actually bad**
  - **However, each outliers has to be explained**



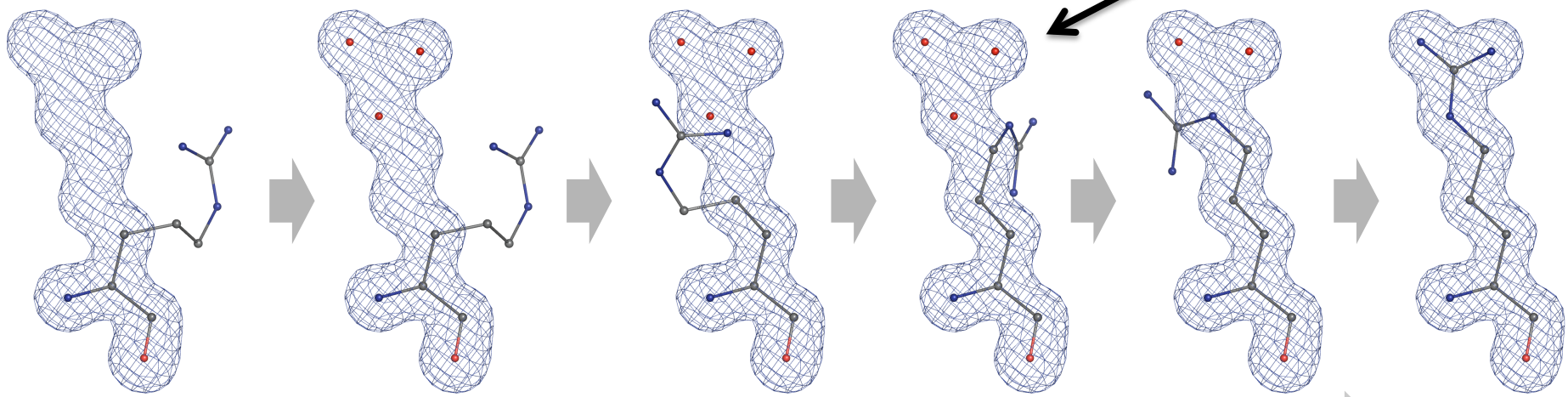
Comprehensive validation — Model (rotamer outliers)



- In proteins 99% of the side chains obey known rotameric conformations
- Often errors are fixed manually but can now be fixed automatically
- A systematic search through rotamer space is combined with a fit-to-density score in Phenix

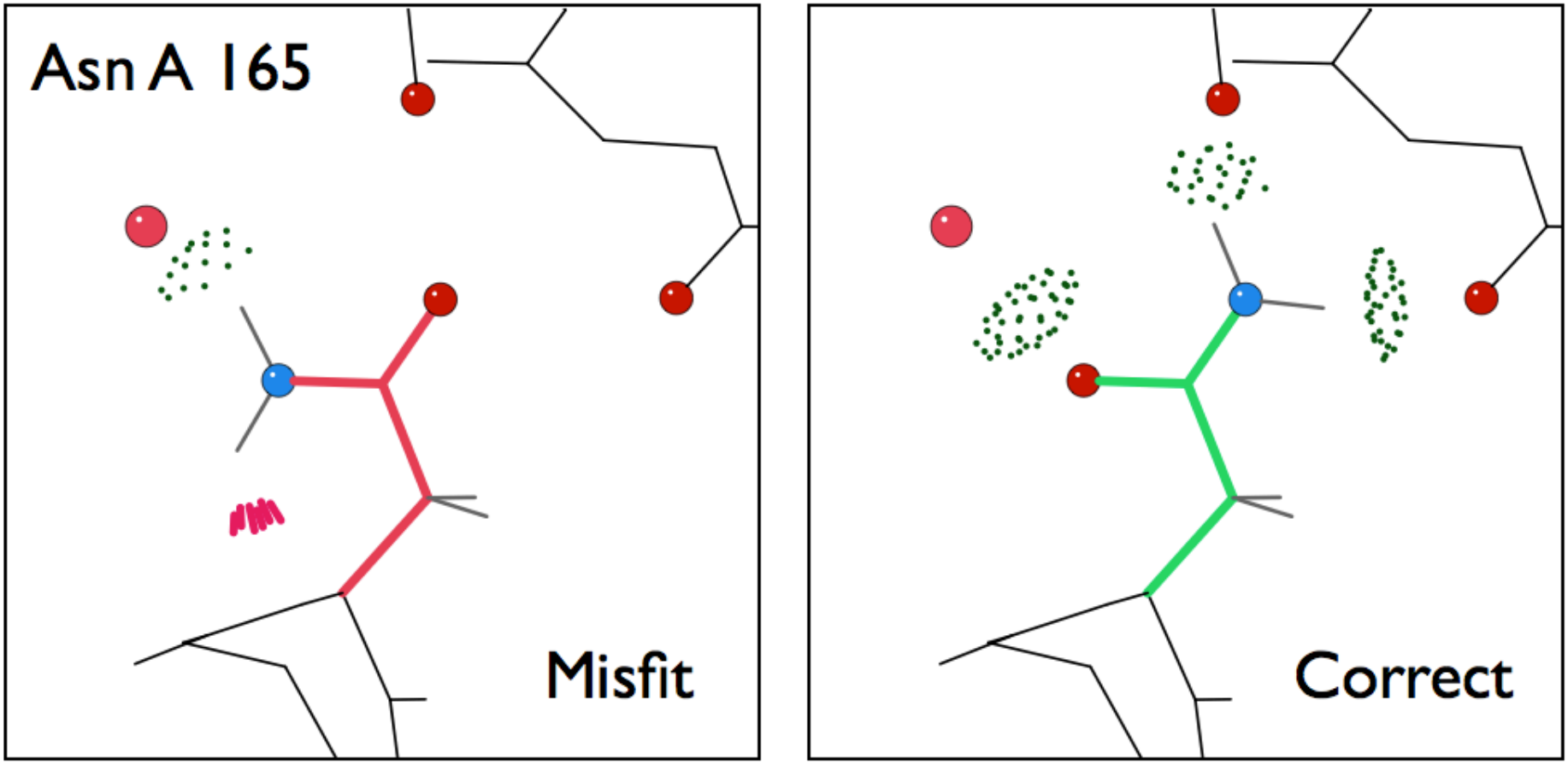
Find best map fitting rotamer…

… remove water if necessary



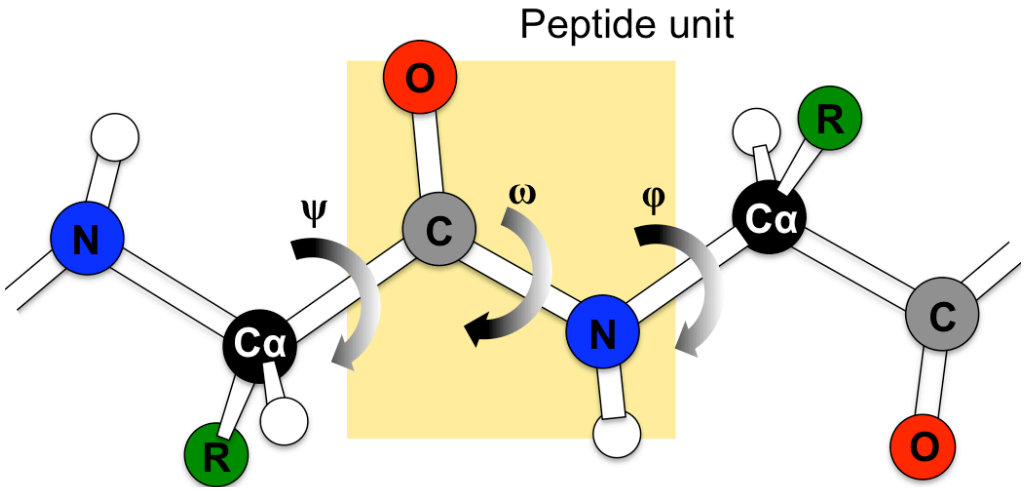
START ~ 1 second for ARG, 1/1000 second for SER RESULT

N/Q/H flips — H atoms help to resolve ambiguity



- *Phenix.refine* does this correction automatically
- If automation fails, fix manually during validation

Comprehensive validation — Model (Ramachandran plot)

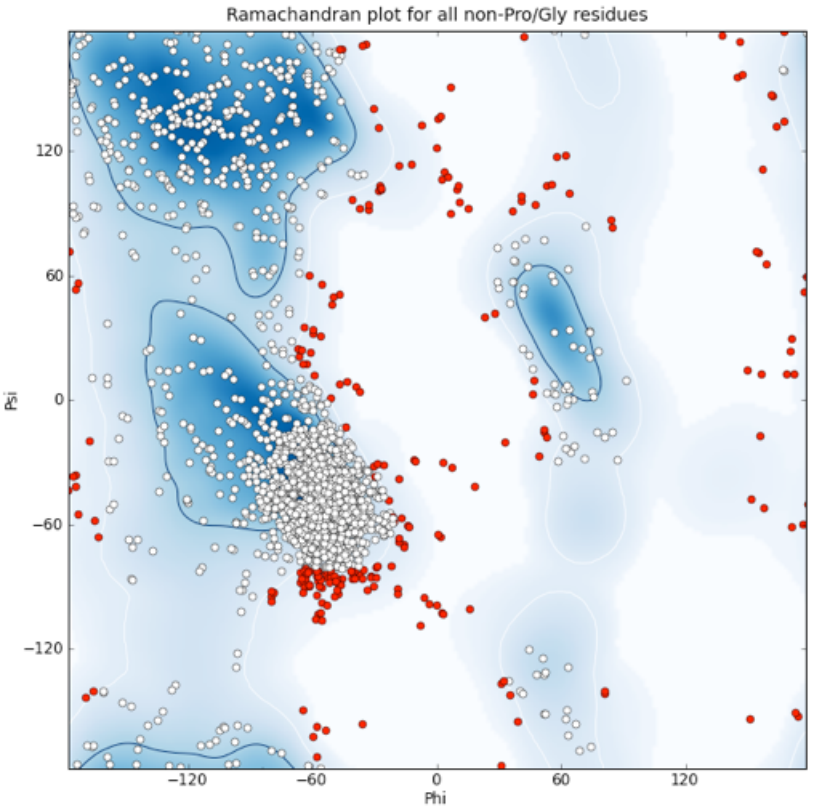


G. N. Ramachandran

The handedness of amino acids, and the steric clashes that occur, given the side chain attachment to the mainchain, results in limits on the distribution of mainchain torsion angles.

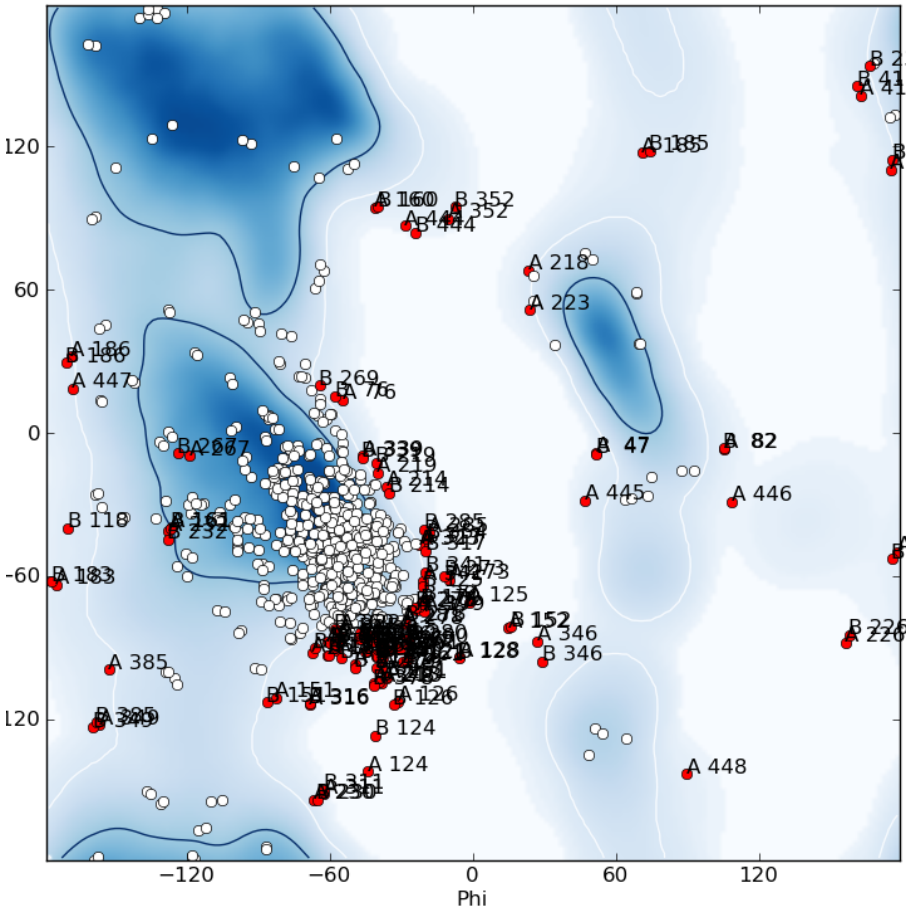
Comprehensive validation — Model (Ramachandran plot)

- A protein structure should conform to prior expectations
- Most (98%+) residues should have a mainchain conformation consistent with the Ramachandran distribution
- A small percentage (0.2%) of residue may show Ramachandran outliers (they are not necessarily errors!)
- Outliers can be seen in strained regions of the structure (e.g. in the active site)
- Any outliers need to be confirmed by detailed analysis

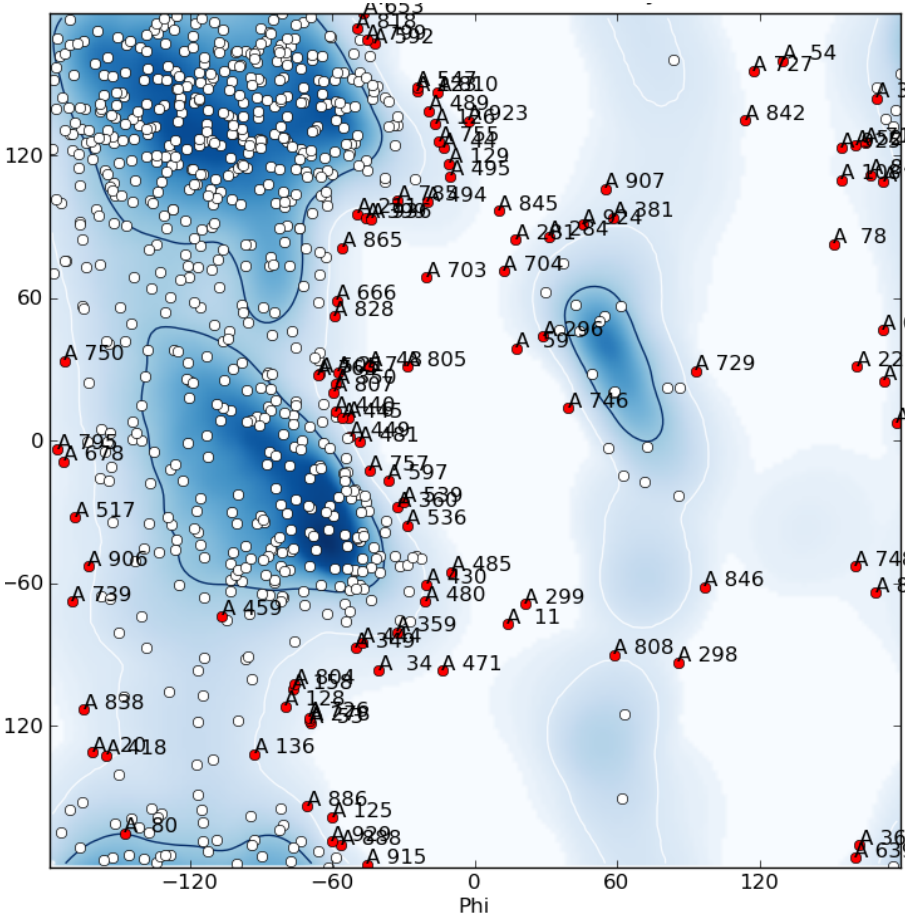


Comprehensive validation — Model (Ramachandran plot)

PDB code: 3MKU
17% outliers, 52% favored

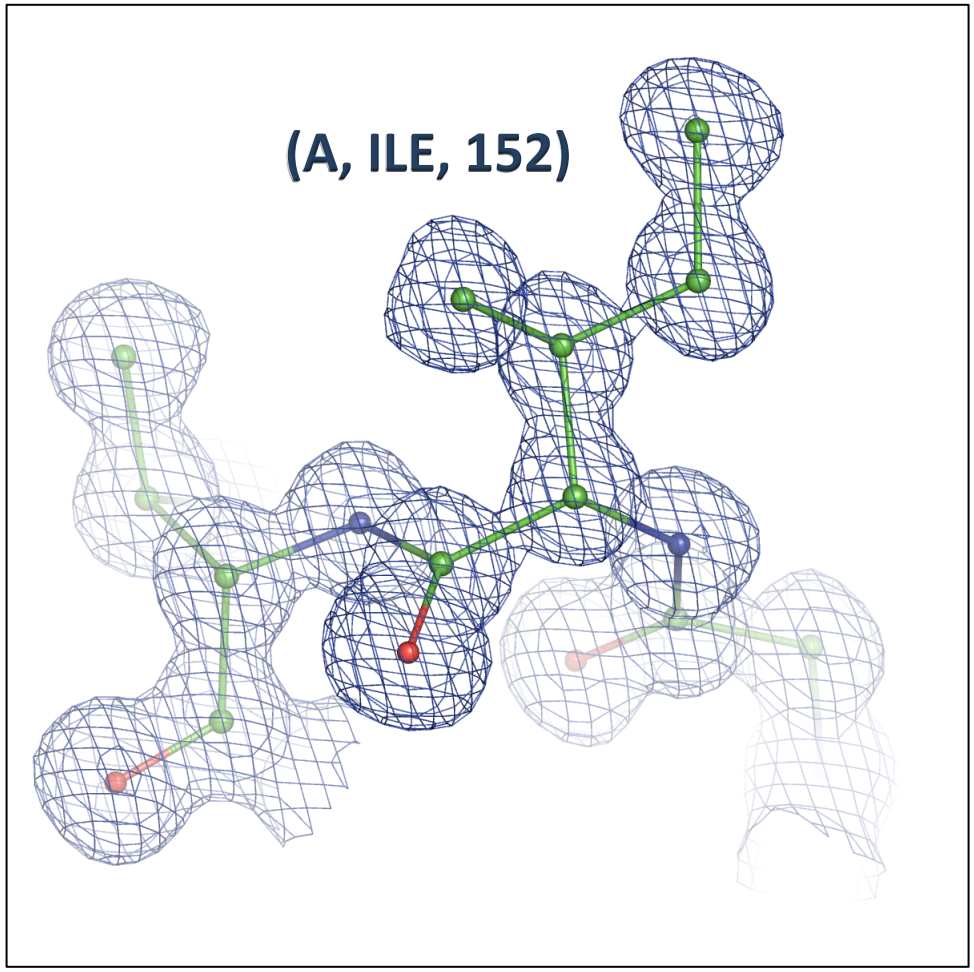


PDB code: 2INY,
15% outliers, 59% favored

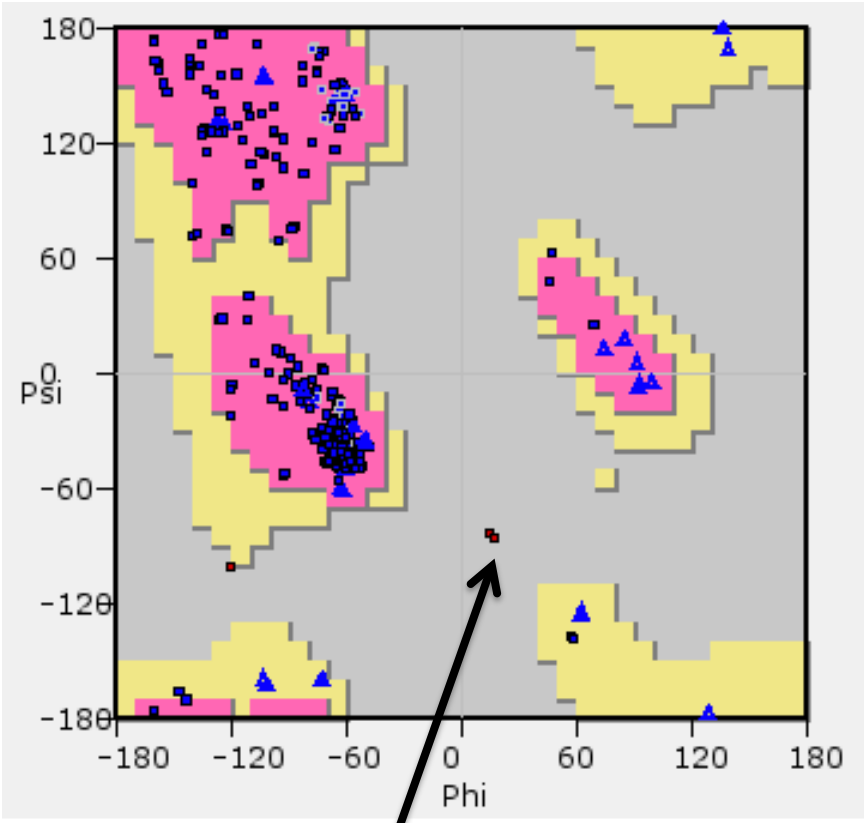


Comprehensive validation — Model (Ramachandran plot)

PDB code 3NOQ, 1 Å



Ramachandran plot in Coot

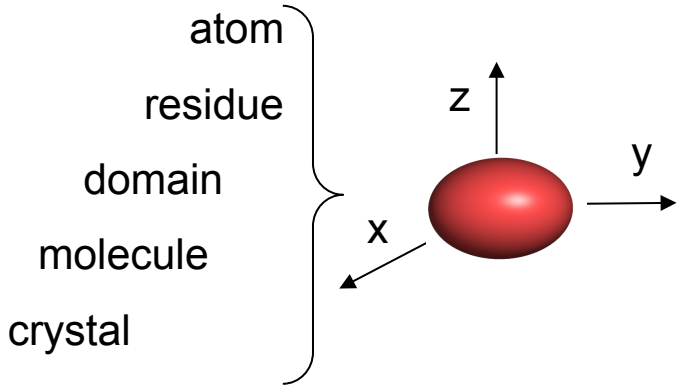


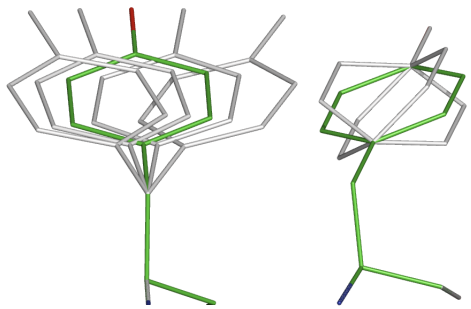
Outliers:

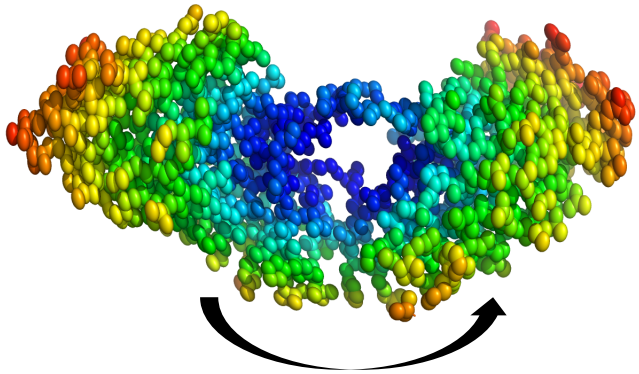
(A, ILE, 152), (B, ILE, 154)

Valid Ramachandran plot outliers: justified by the density map

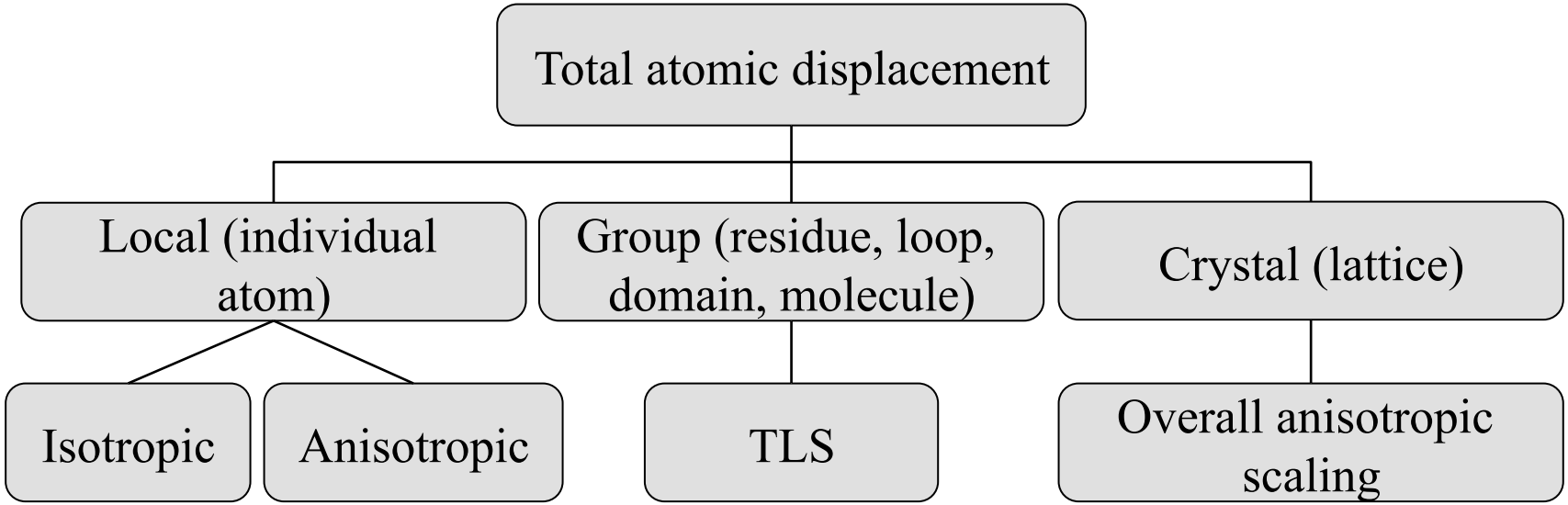
Atomic Displacement Parameters (ADP, B-factors)







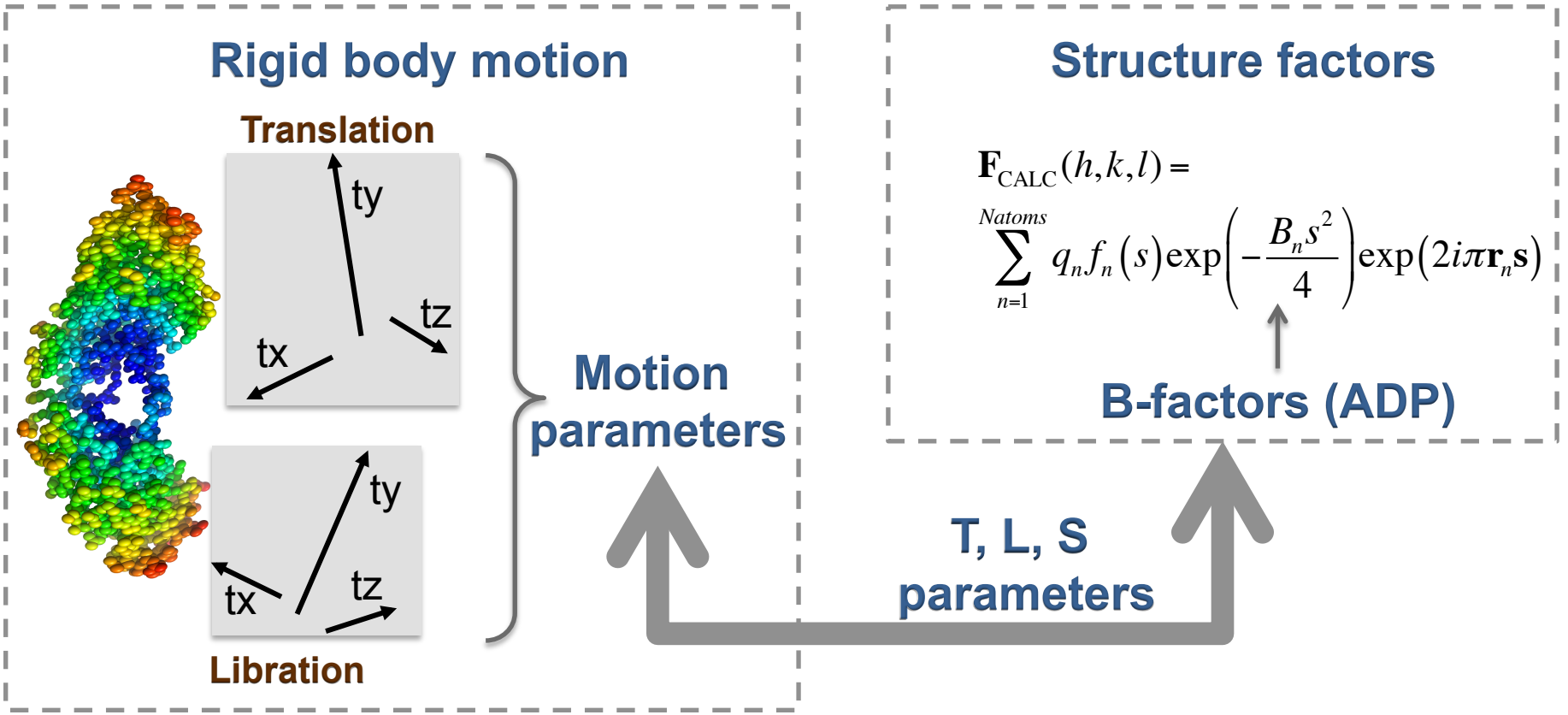
- Parameterization for refinement



$$U_{TOTAL} = U_{CRYST} + U_{GROUP} + U_{LOCAL}$$

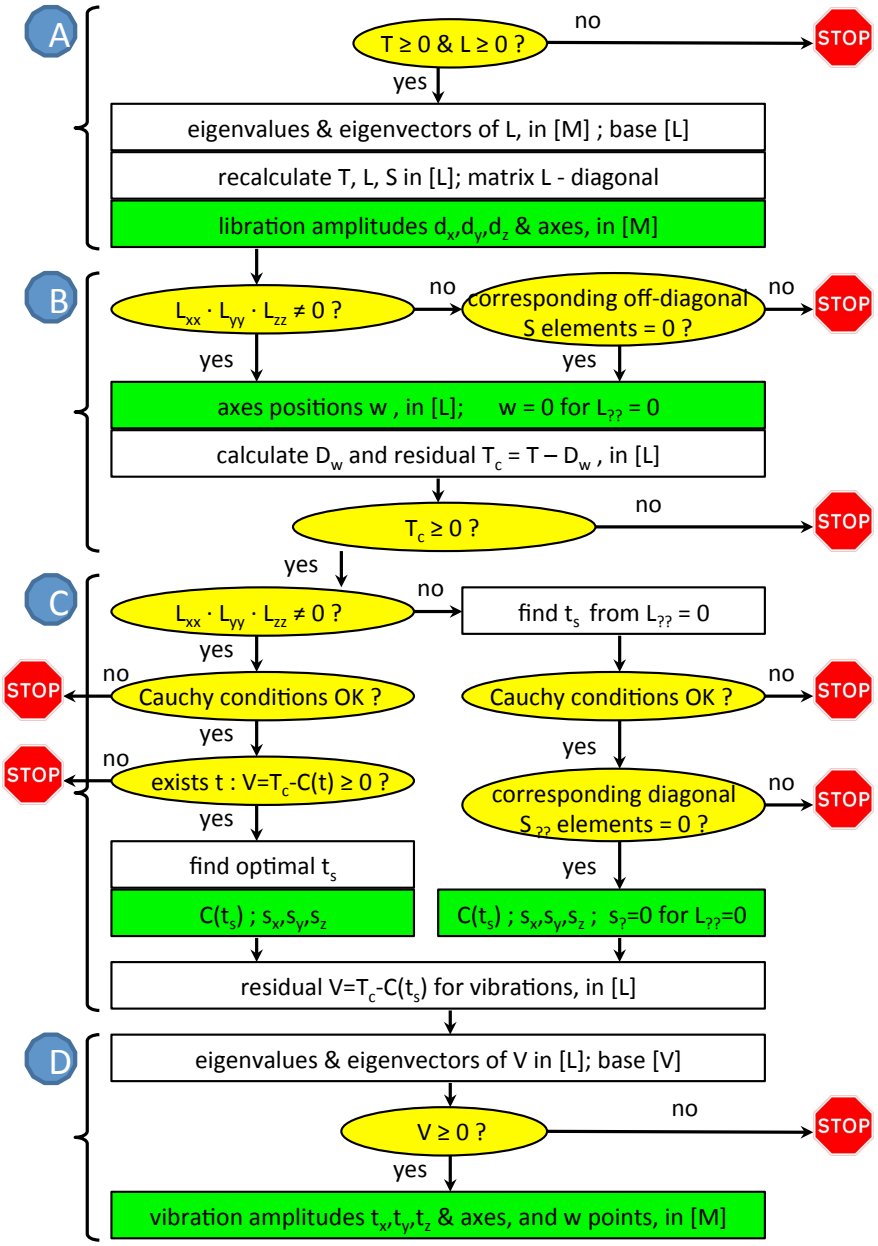
TLS parameters and rigid body motion

- TLS parameters is a way to pack descriptors of rigid-body motion into a form suitable to calculate structure factors

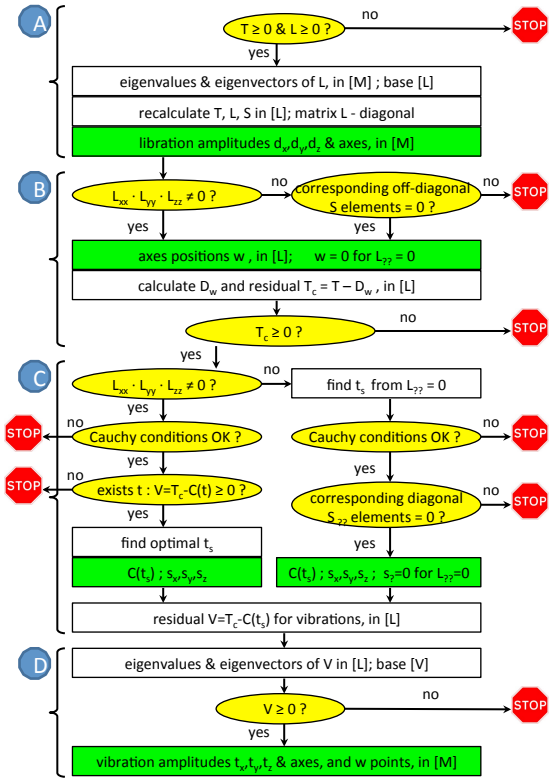


- TLS do not directly provide parameters of rigid body motion
  - Descriptions of motions need to be extracted from TLS matrices

TLS parameters as descriptors of rigid body motion



TLS parameters as descriptors of rigid body motion



- To describe rigid-body motion TLS matrices must satisfy several conditions
- Broken condition means TLS cannot describe rigid-body motion and therefore do not make physical sense
- PDB survey:
  - Total models: 105,000
  - Contain TLS: 25,284
  - Files with bad TLS: 21,540 (85%)

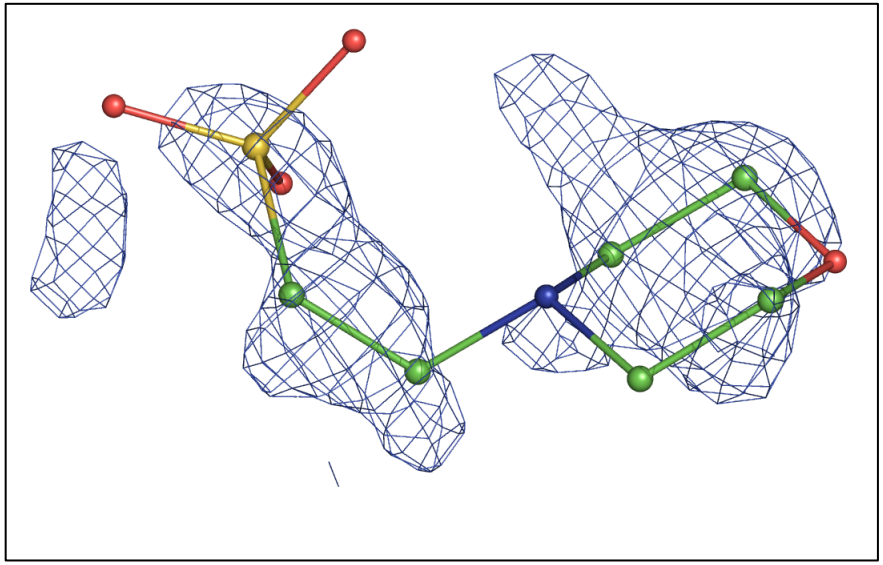
From deep TLS validation to ensembles of atomic models (…)
A. Urzhumtsev, P. Afonine, A. Benschoten, J. Fraser & P. Adams
Acta Cryst. D, 2015 (accepted)

Available in Phenix (work in progress)

Comprehensive validation — Ligands

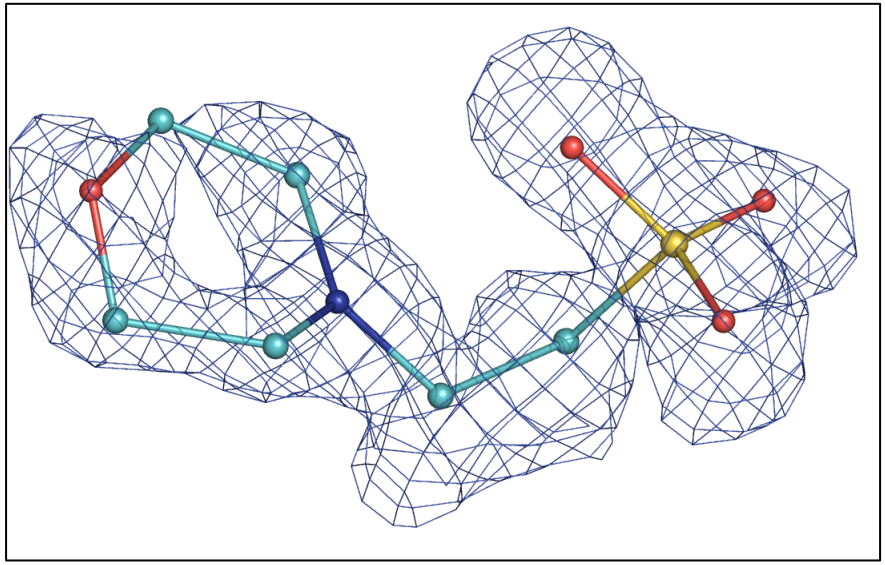
- PDB code: 1SE6, resolution 1.6Å
  - **Techniques, tools and best practices for ligand electron-density analysis and results from their application to deposited crystal structures.**
    Pozharski, Weichenberger and Rupp, Acta Cryst D69, 150-167 (2013)

Wrong



2mFo-DFc

Correct



FEM

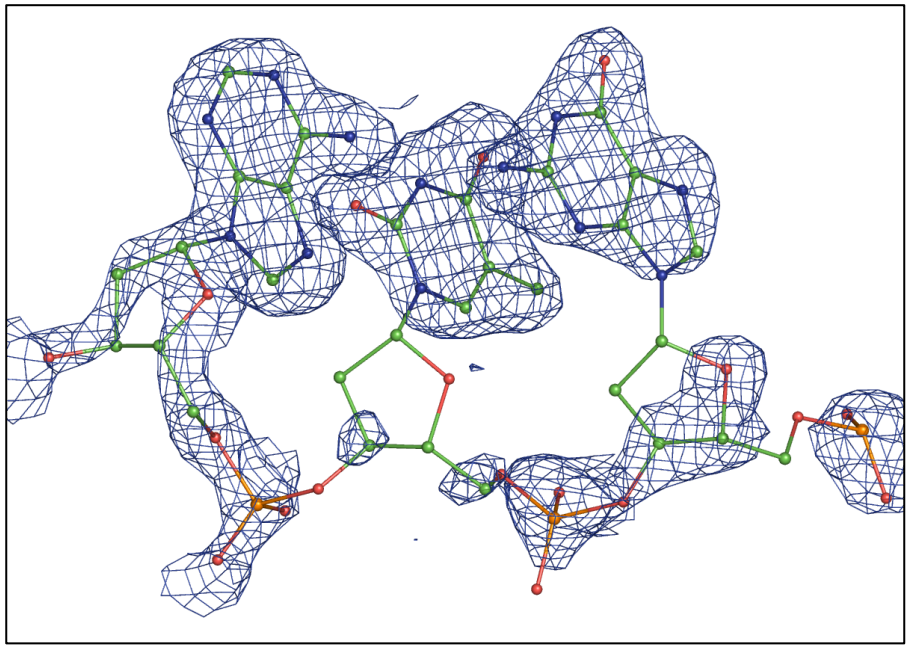
Comprehensive validation — Ligands and maps

- phenix.composite_omit_map
  - With or without Simulated Annealing refinement
- Simple OMIT residual map
  - Remove ligand, then compute mFo-DFc map
  - There are caveats
- Feature Enhanced Map (FEM)
  - Compute FEM with and without ligand

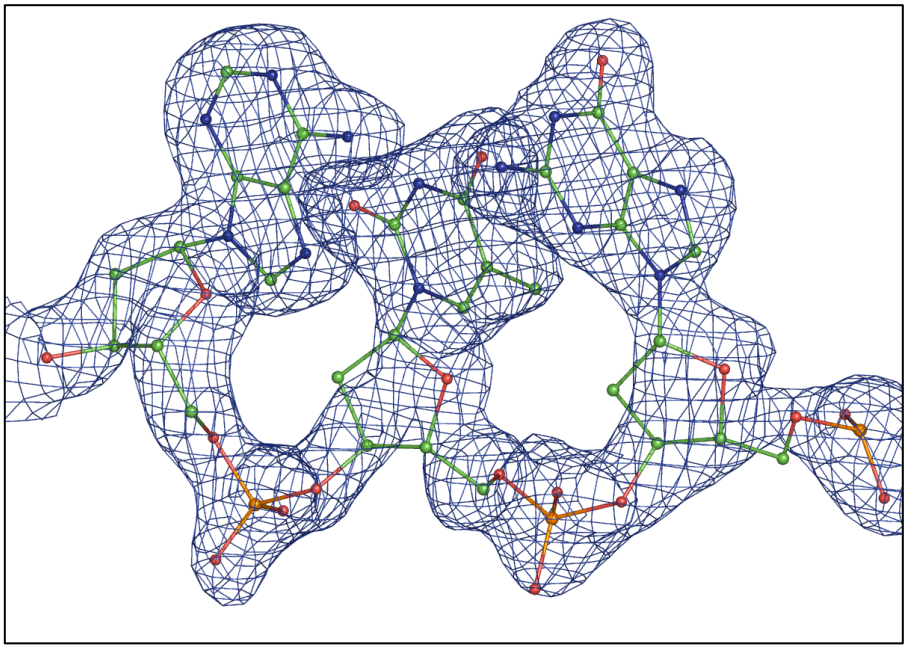
Comprehensive validation — Ligands and maps

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

2mFo-DFc , 1σ



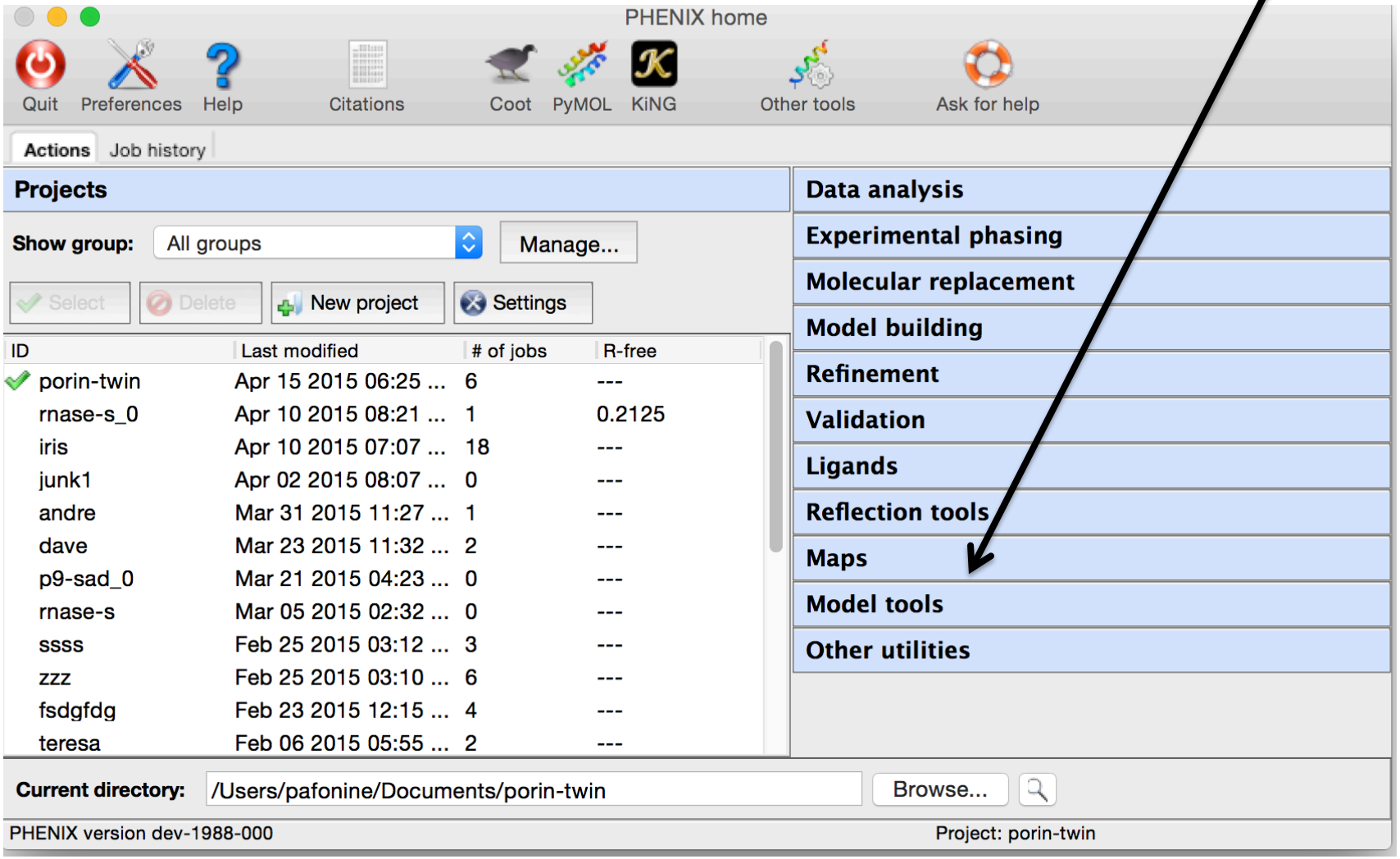
FEM, equivalent 1σ



FEM

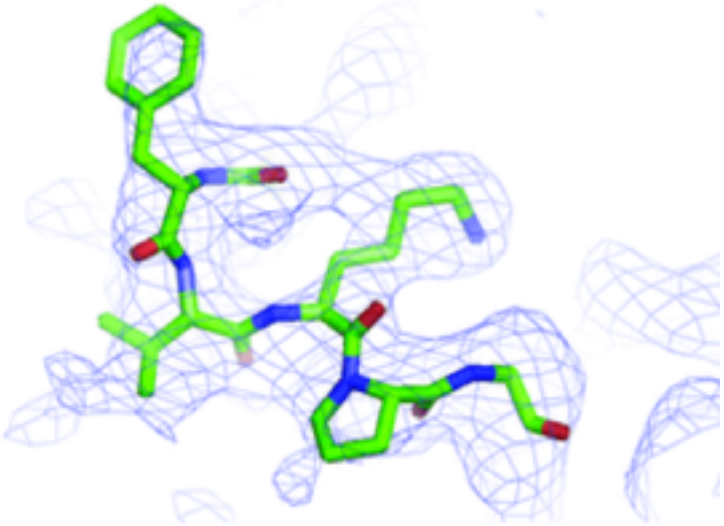
Comprehensive validation — Ligands and maps

Phenix tools to compute maps

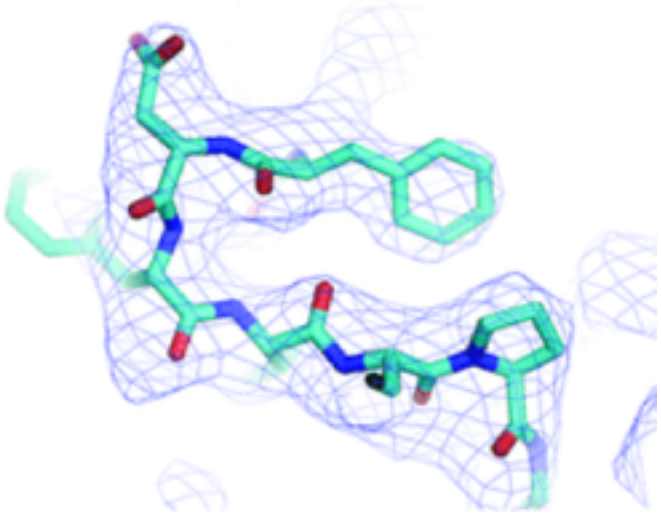


Comprehensive validation — Sequence register errors

1ZEN



2mFo-DFc model biased map and incorrect model



Corrected model in iterative build OMIT map

Phenix iterative build OMIT map

Terwilliger et al., Acta Cryst D64, 515-524, 2008

Comprehensive validation — Sequence register errors

- Check actual sequence against derived from PDB file:
  - Extract sequence from PDB file:

```
phenix.print_sequence model.pdb > model.seq
```

  - Align actual sequence with model.seq

- Example of a problem:

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

Local errors obscured by global numbers

- **Resolution 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 errors obscured by global numbers

- **Resolution 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**

Are you done with refinement?

- **Say you are refining a structure at 1.0 Å resolution and R-factors are:**
  R_{WORK} = 18% and R_{FREE} is 23%.

  Question: Are these values good? Is refinement completed?

Are you done with refinement?

- **Say you are refining a structure at 1.0 Å resolution and R-factors are: R_{WORK} = 18% and R_{FREE} is 23%.**

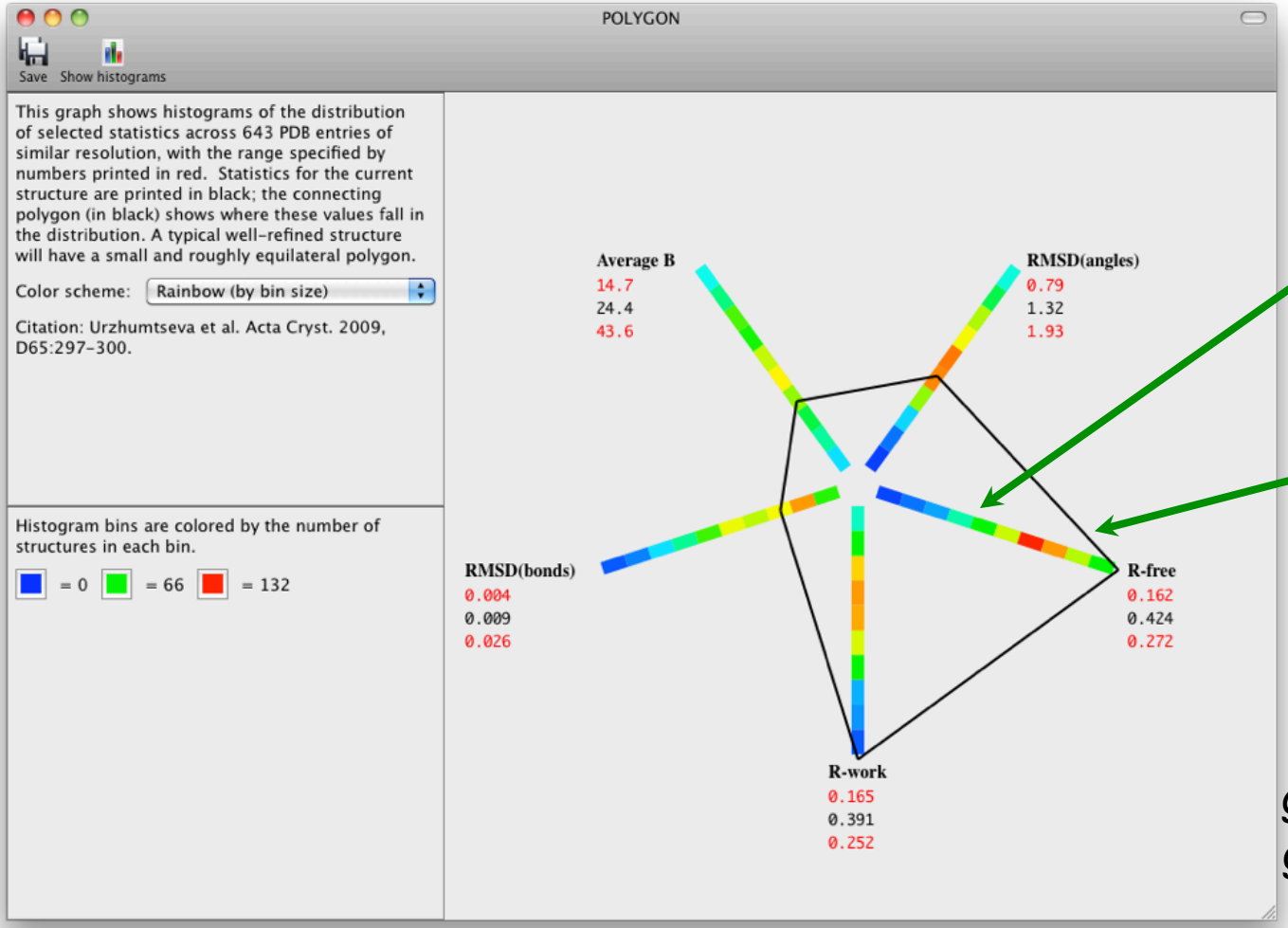
  Question: Are these values good? Is refinement completed?

- **PDB statistics: compare your structure with similar structures in the database**

R_{WORK} at 0.9-1.1Å			R_{FREE} at 0.9-1.1Å			R_{FREE}-R_{WORK} at 0.9-1.1Å		
0.10 -	0.12:	68	0.11 -	0.13:	16	0.00 -	0.01:	8
0.12 -	**0.14:**	**94**	0.13 -	0.15:	56	0.01 -	0.01:	22
0.14 -	0.16:	73	**0.15 -**	**0.17:**	**97**	0.01 -	0.02:	56
0.16 -	**0.18:**	**17 <<<**	0.17 -	0.18:	69	**0.02 -**	**0.03:**	**62**
0.18 -	0.20:	12	0.18 -	0.20:	14	0.03 -	0.03:	58
0.20 -	0.21:	3	0.20 -	0.22:	12	0.03 -	0.04:	29
0.21 -	0.23:	5	**0.22 -**	**0.24:**	**3 <<<**	0.04 -	0.04:	14
0.23 -	0.25:	0	0.24 -	0.26:	4	**0.04 -**	**0.05:**	**10 <<<**
0.25 -	0.27:	0	0.26 -	0.28:	1	0.05 -	0.06:	6
0.27 -	0.29:	2	0.28 -	0.30:	2	0.06 -	0.06:	9

- **Answer: the R-factors are not good, the structure needs some more work**

Are you done with refinement? POLYGON will tell!



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

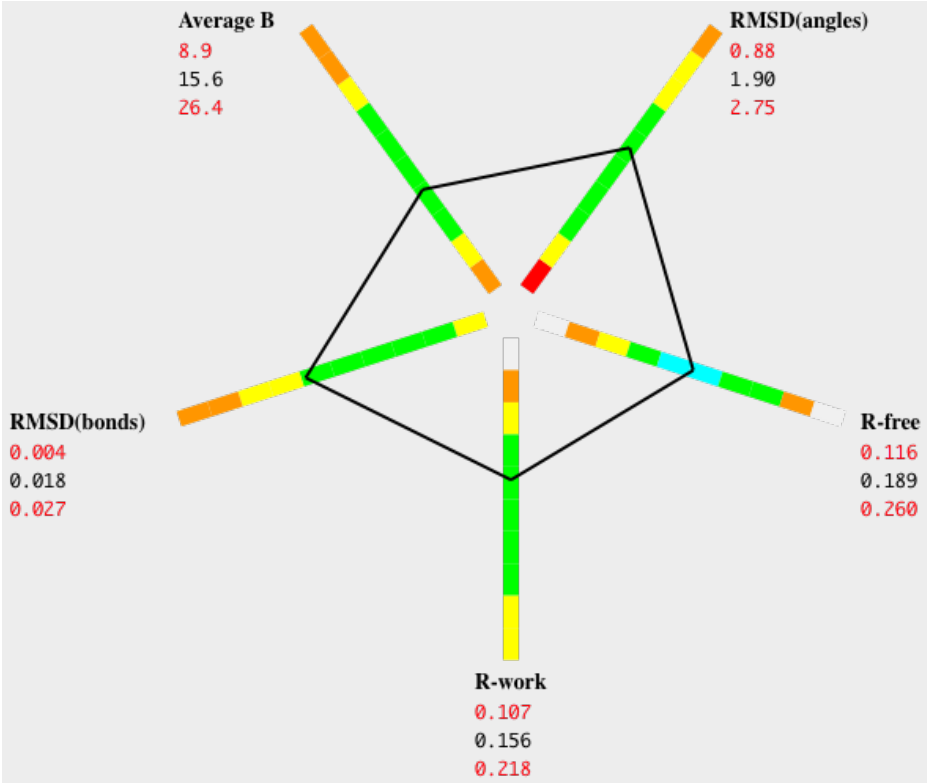
The black polygon shows where the statistics for the user's structure fall in each histogram

The structure used to generate this figure has good geometry relative to the PDB, but very poor R-factors.

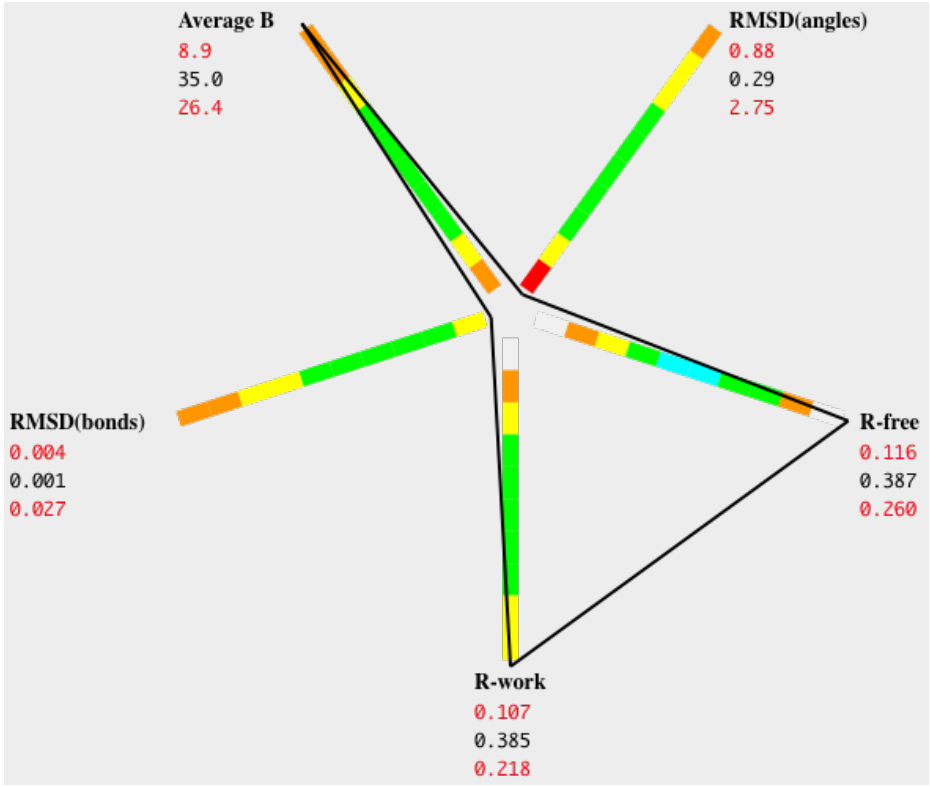
Crystallographic model quality at a glance.
L.Urzhumtseva, P.V.Afonine, P.D.Adams & A.Urzhumtsev. *Acta Cryst.* **D**65, 297-300 (2009)

Are you done with refinement? POLYGON examples

Likely overall good model



Clearly there are problems



Are you done with refinement? POLYGON examples

PDB: 1eic

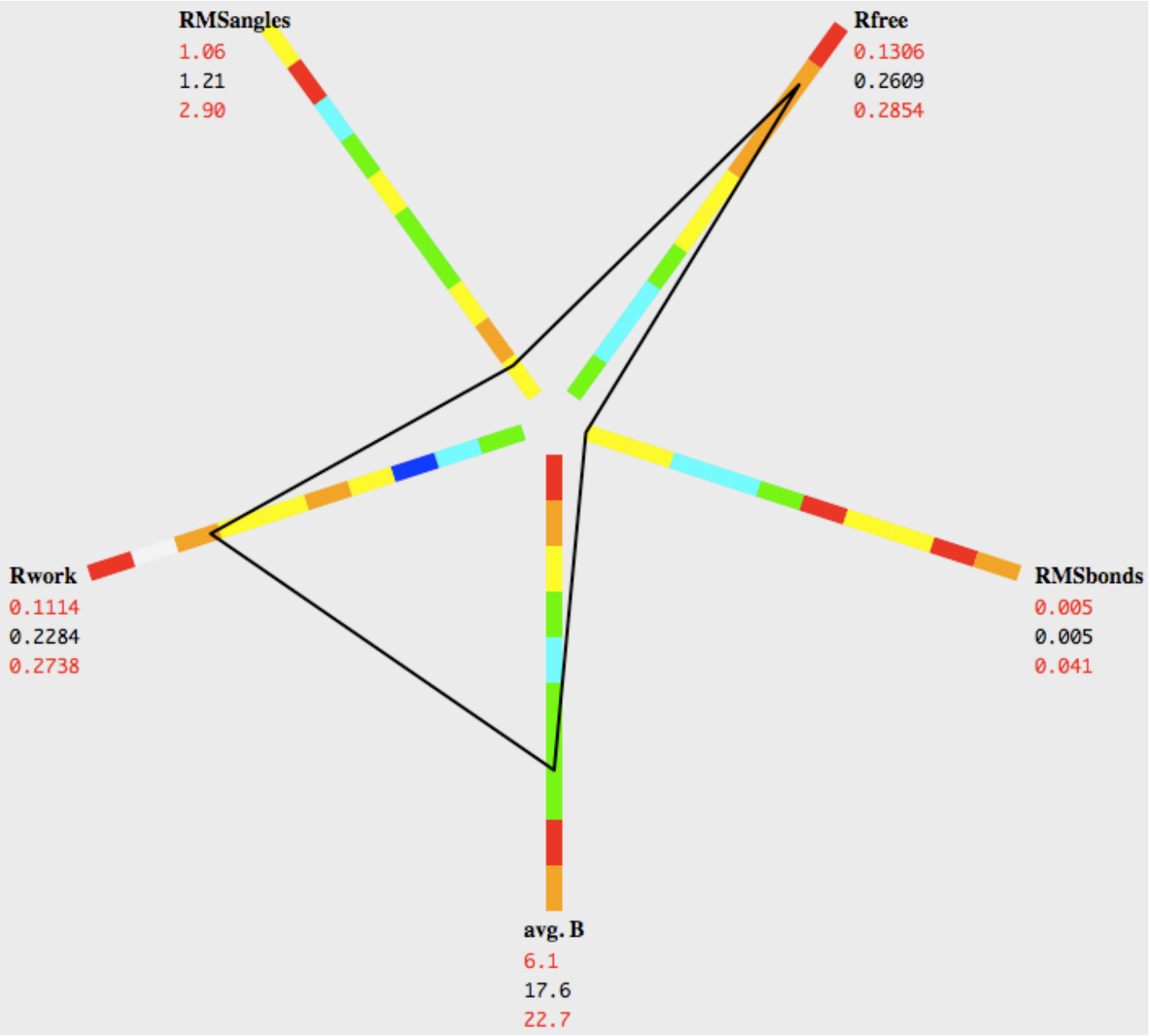
Resolution: 1.4Å

Deposition year: 2000

PUBLISHED:

Rwork = 20%

Rfree = 25%

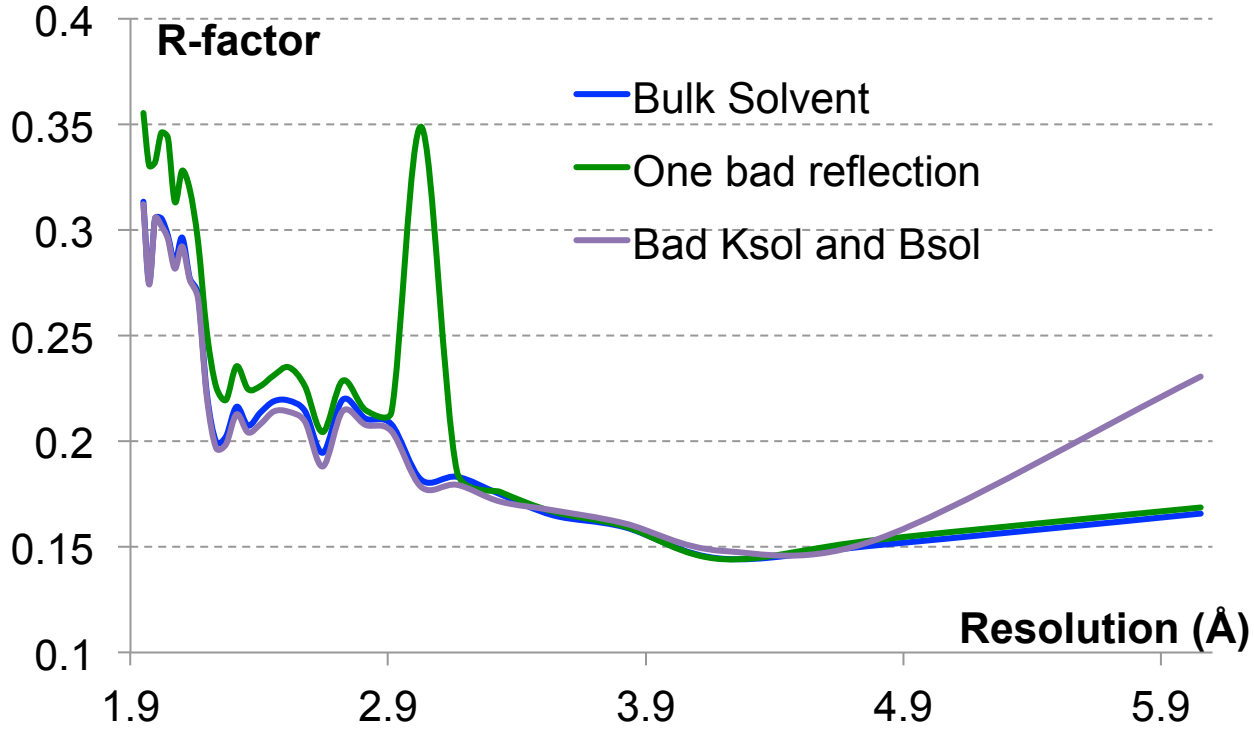


Are you done with refinement? POLYGON examples

- Structure from PDB: 1eic (resolution = 1.4Å; deposition year: 2000)
  - PUBLISHED: Rwork = 20% Rfree = 25%
- Problems:
  - No 'riding' H atoms;
  - All atoms are isotropic;
  - Suboptimal weights, refinement is not converged, incomplete solvent
- Fixing the model with PHENIX
  - Add and refine H as riding model
  - Update ordered solvent
  - Refine atoms as anisotropic (except H and water)
  - Optimize X-ray/Restraints weights

FINAL MODEL: Rwork = 14% Rfree = 17%

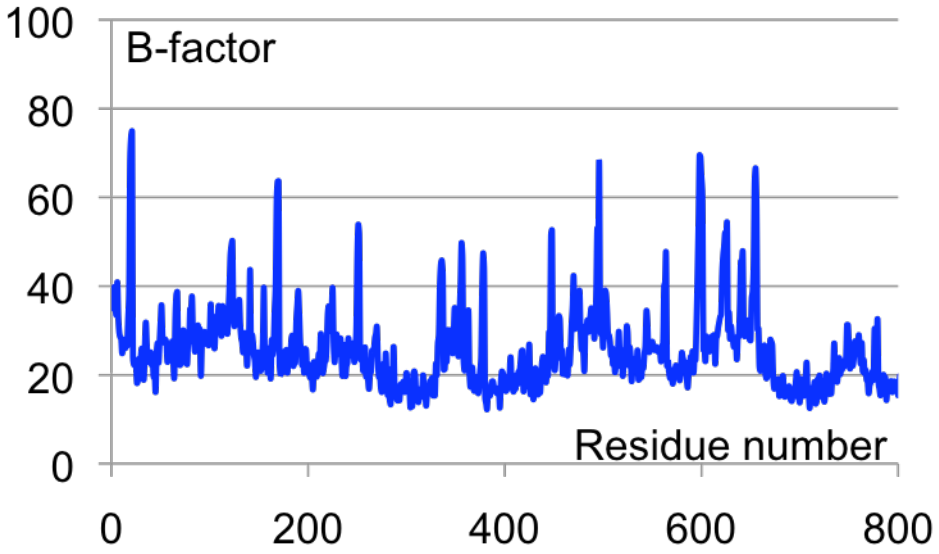
Statistics per resolution

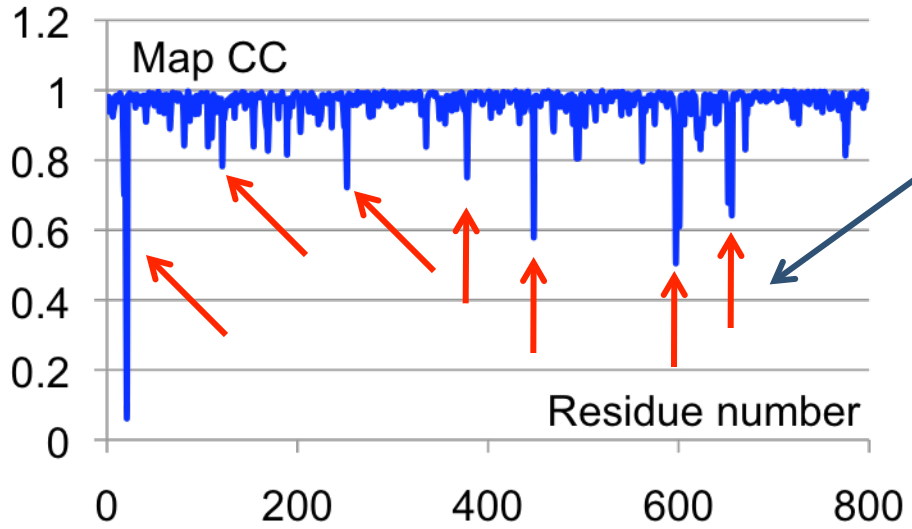


- **R-factor in resolution bins helps to identify:**
  - **Problem with bulk-solvent modeling**
  - **Problems at high resolution**
  - **Artifacts (green line):**

```
INDE      3     5   -42 IOBS= 99999.999 SIGIOBS=      0.000
```

Local model quality — B-factor and map CC per residue

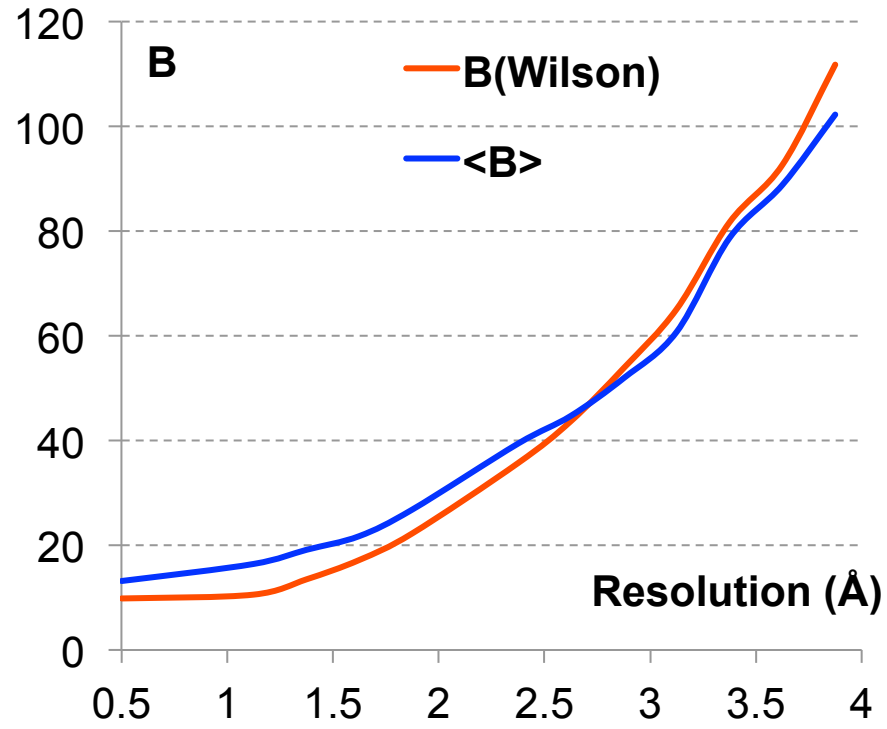




Indicates problem places

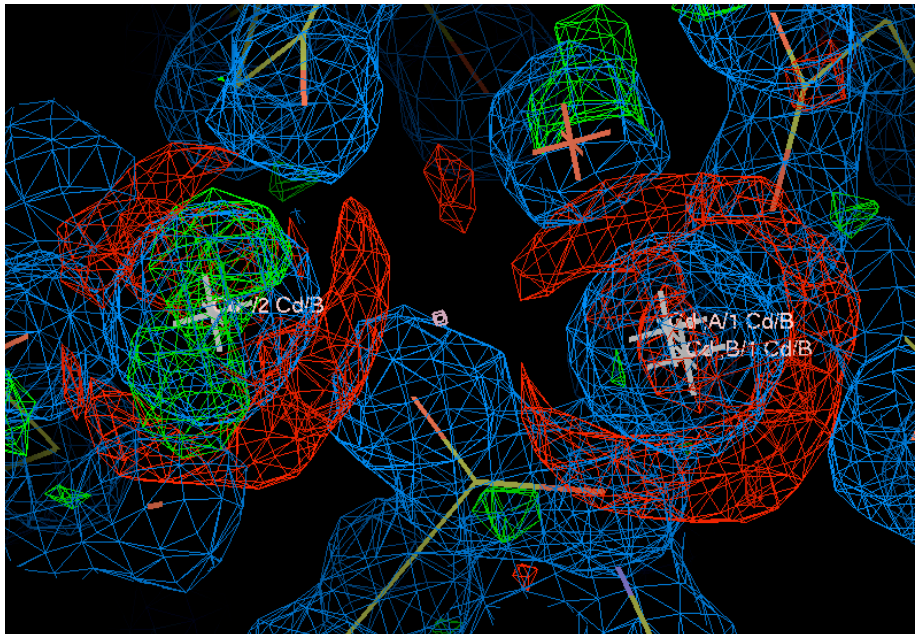
Mean and Wilson B-factor

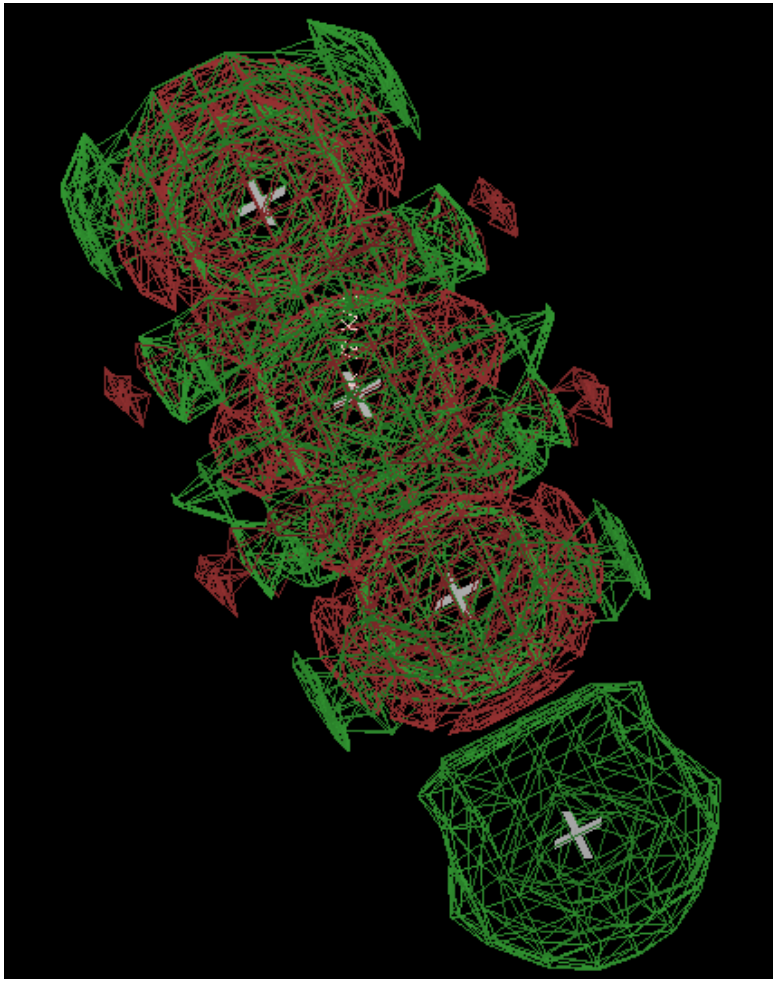
Resolution	B(Wilson)	<B>	Models
0.00-1.00	9.77	13.11	94
1.00-1.25	10.58	16.44	401
1.25-1.50	13.50	19.14	1050
1.50-1.75	17.20	21.76	3600
1.75-2.00	22.27	26.82	5510
2.25-2.50	35.70	39.42	3385
2.50-2.75	43.71	44.73	2844
2.75-3.00	53.86	51.94	1628
3.00-3.25	65.11	60.76	780
3.25-3.50	81.69	78.70	165
3.50-3.75	92.67	88.84	100
3.75-4.00	111.83	102.29	30



- Mean B and Wilson B are usually similar
  - Wilson B is dominated by strongly diffracting (lower B) atoms that contribute more to high-res reflections
    - Wilson B represents the lower end of the range of B-factors
      - Discrepancy between Wilson B and mean B is not important

Fourier truncation artifacts

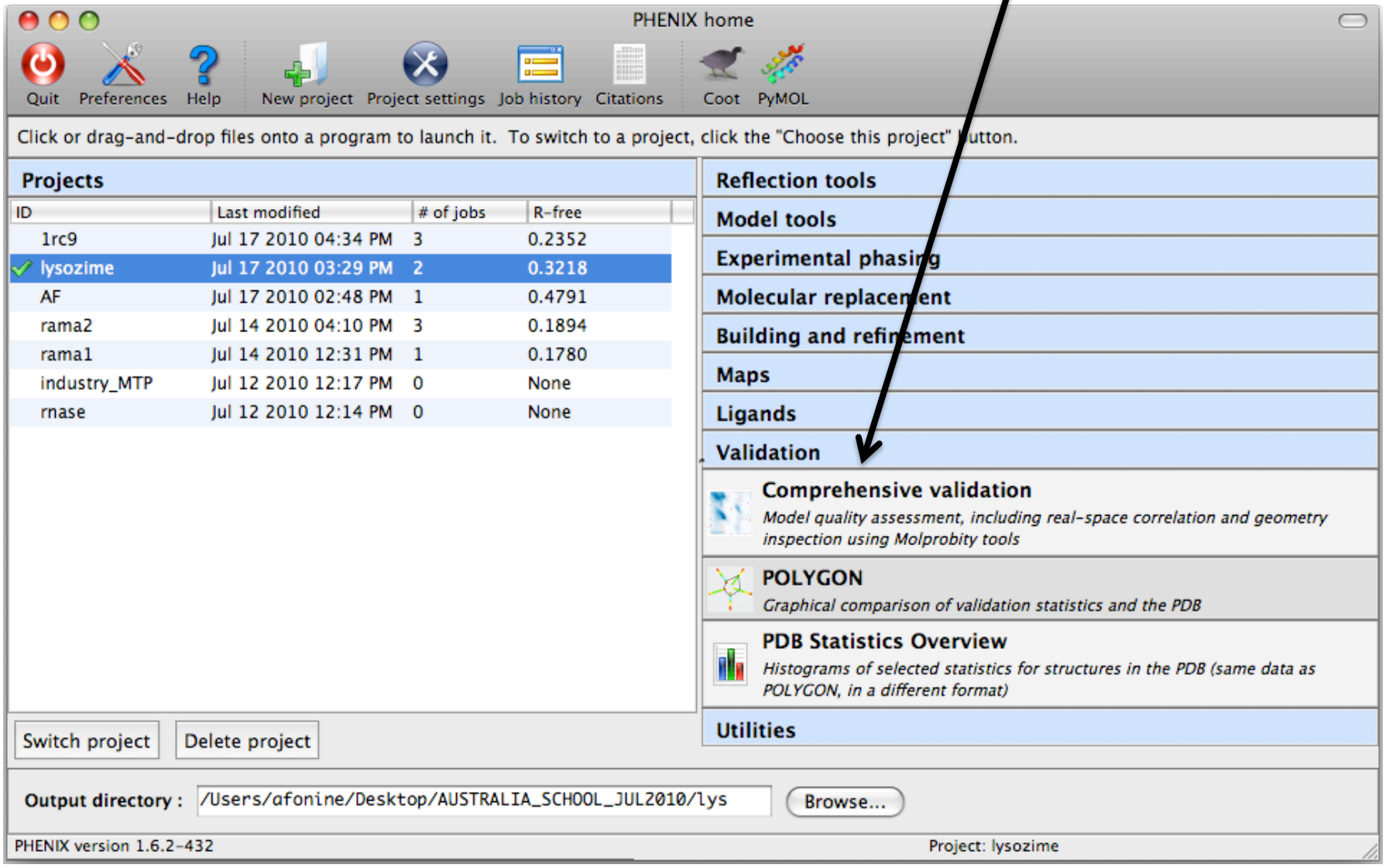




- **Don’t attempt to build water into noise !**

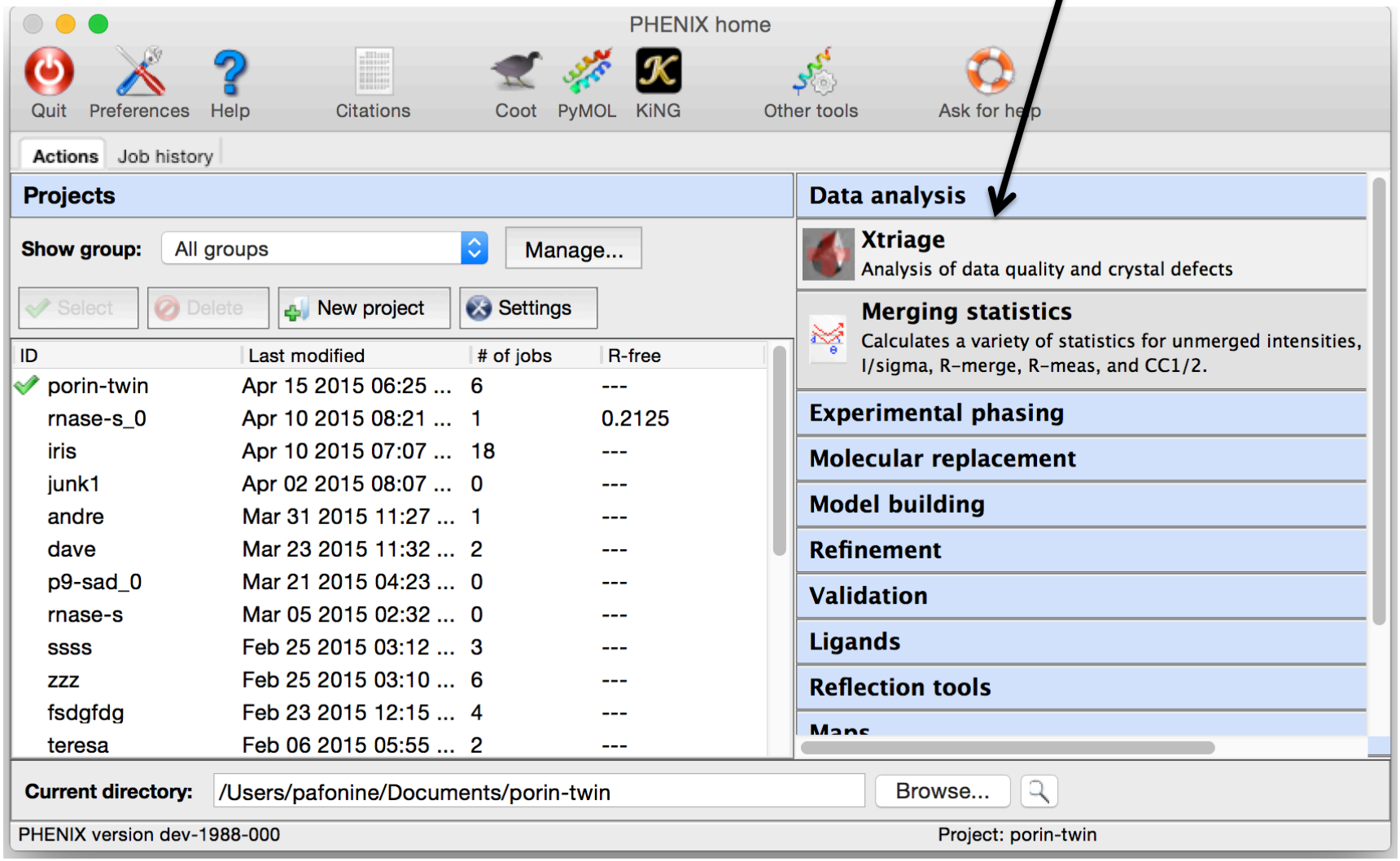
Phenix tools for validation

Validation tab



Phenix tools for validation

Data analysis - Xtriage

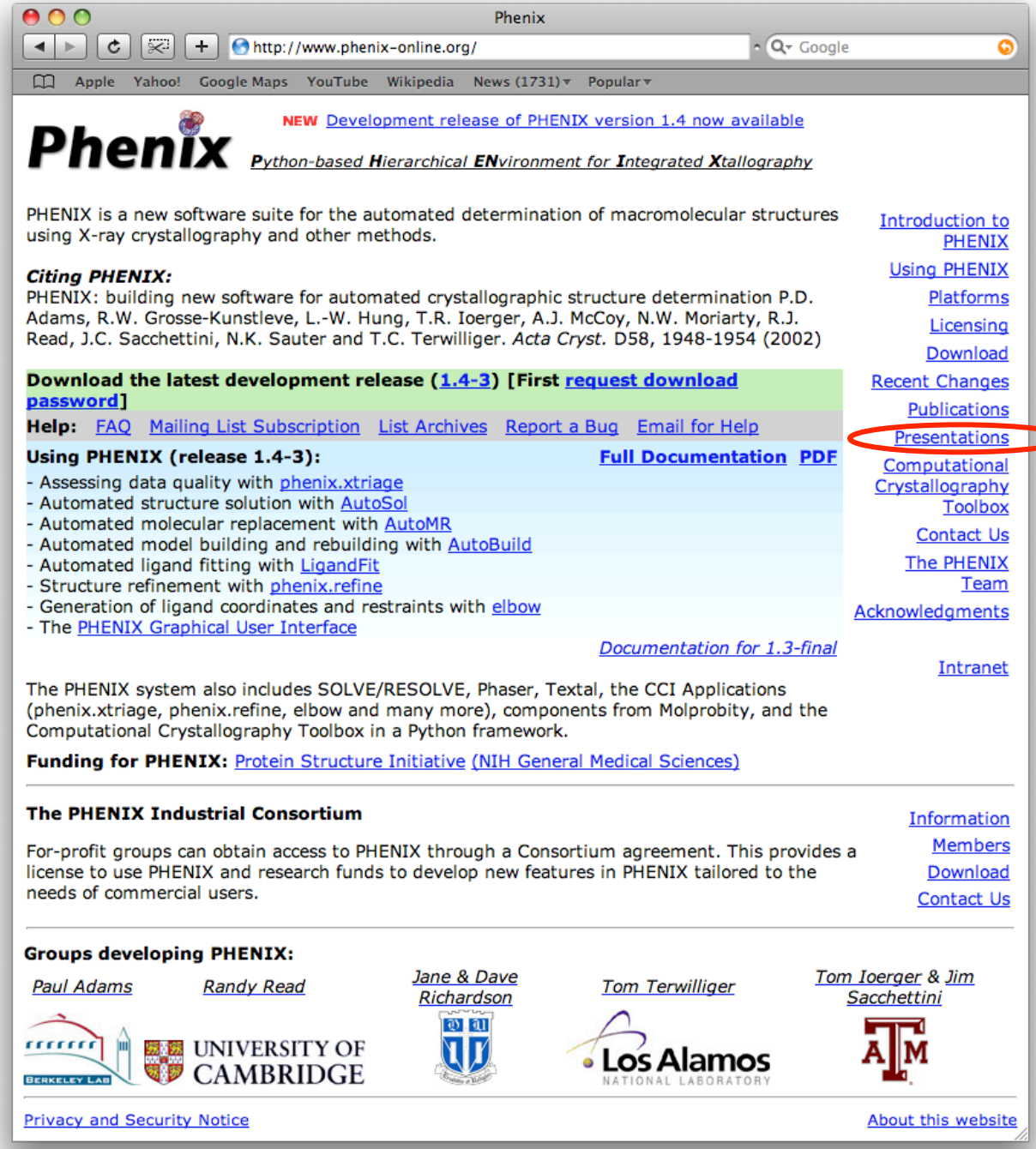


Acknowledgments

- **Lawrence Berkeley Laboratory**
  - **Paul Adams**, Nat Echols, Jeff Headd, Ralf Grosse-Kunstleve, Nigel Moriarty, Nicholas Sauter, Peter Zwart
- **Los Alamos National Laboratory**
  - Tom Terwilliger, Li-Wei Hung
- **Cambridge University**
  - Randy Read, Airlie McCoy, Laurent Storoni, Gabor Bunkoczi, Robert Oeffner
- **Duke University**
  - Jane Richardson & David Richardson, Ian Davis, Vincent Chen, Jeff Headd, Chris Williams, Bryan Arendall, Laura Murray
- **Collaborations**
  - David Baker and Frank DiMaio
  - Andy Karplus and Dale Tronrud
  - Rhiju Das and Fang Chou
  - Paul Langan and Marat Mustyakimov
- **Others**
  - Garib Murshudov & Alexi Vagin
  - Kevin Cowtan, Paul Emsley, Bernhard Lohkamp
  - Alexandre Urzhumtsev & Vladimir Lunin
  - David Abrahams
  - PHENIX Testers & Users: James Fraser, Herb Klei, Warren Delano, William Scott, Joel Bard, Bob Nolte, Frank von Delft, Scott Classen, Ben Eisenbraun, Phil Evans, Felix Frolow, Christine Gee, Miguel Ortiz-Lombardia, Blaine Mooers, Daniil Prigozhin, Miles Pufall, Edward Snell, Eugene Valkov, Erik Vogan, Andre White, and many more
- **Funding:**
- NIH/NIGMS:
  - *P01GM063210, P50GM062412, P01GM064692, R01GM071939*
- Lawrence Berkeley Laboratory
- PHENIX Industrial Consortium

BERKELEY LAB
Lawrence Berkeley National Laboratory

Phenix



This presentation (PDF file) and much more

www.phenix-online.org