# *Model validation*

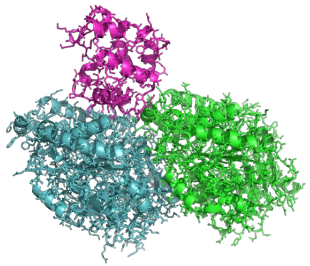

## Pavel Afonine

*Lawrence Berkeley National Laboratory (LBNL)*

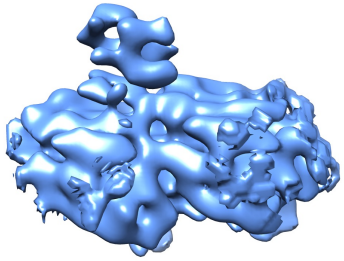

**September 26th, 2024**
**BNL**

# Validation

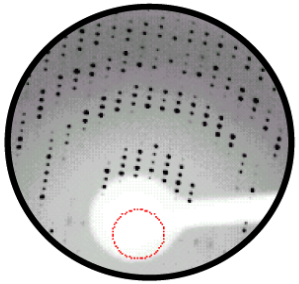
## Model



## Data



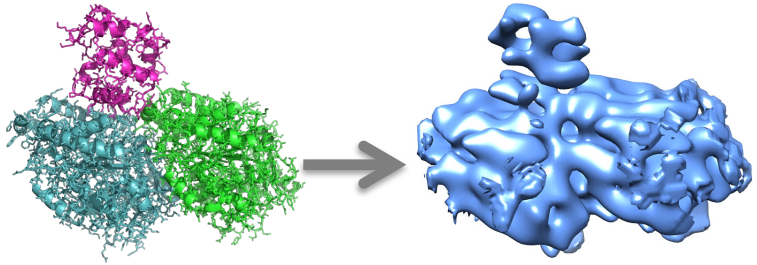
or



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: why to do?

- **Problems detected early can save a lot of time later**
- **Subjectivity**
  - Manual map interpretation: experience, skills, pressure
  - Model parameterization, target weights, starting points
  - Lack of data = multiple possibilities for interpretation
- **Human program the software**
  - Programs may contain bugs
- **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**

# Validation: why to do?

- Helps to
  - save time
  - produce better models
  - set correct expectations
- Minimize fraud or honest mistakes

# Validation: why to do?

- **Quality filters:**
  - You
  - Software you use
  - Your boss
  - Reviewers (of your paper)
  - PDB deposition (software and people)
  - Community

- **Unnoticed (intentionally or not) problems**
  - Likely discovered anyway, sooner or later

# Validation: why to do?

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

Michael A Hanson & Raymond C Stevens

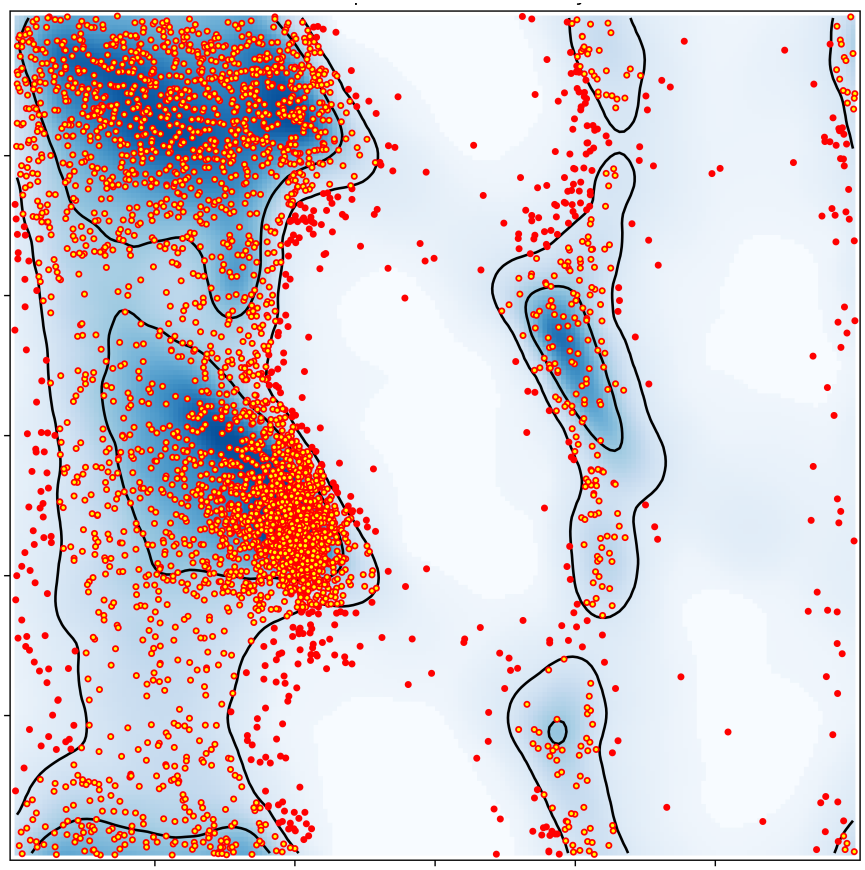
*Nat. Struct. Biol.* 7, 687–692 (2000); retracted 6 July 2009

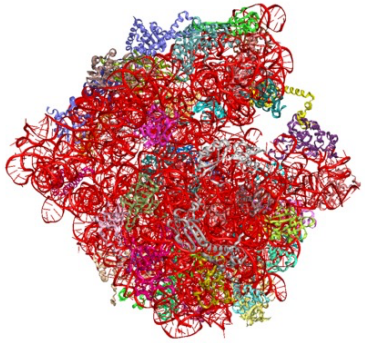
In this paper, we described both the three-dimensional crystal structure of a botulinum toxin catalytic domain separated from the holotoxin (BoNT/B-LC, PDB 1F82) and a structure of the toxin catalytic domain in complex with a peptide (Sb2–BoNT/B-LC, PDB 1F83). The complex was later refined and deposited 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 (University of Alabama) – 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 at that time

# Validation: why to do?

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





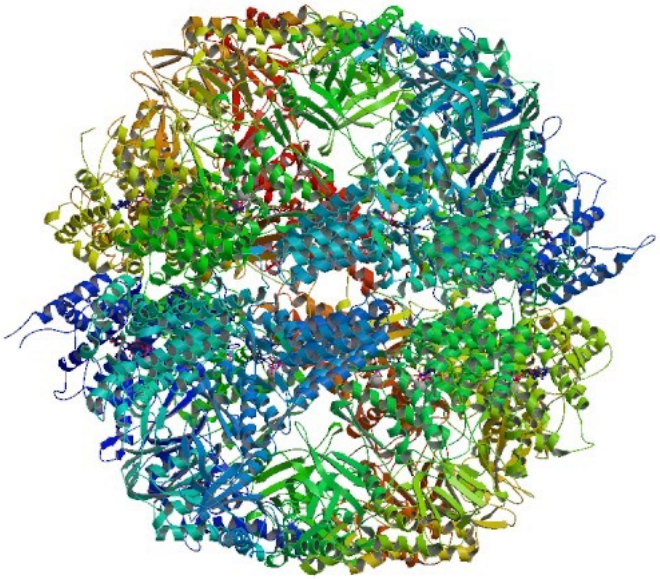
| Metric | 6o9j | Expected |
|---|---|---|
| Clashscore | 70 | Less than 10 |
| Ramachandran favored, % | 59 | More than 98 |
| Ramachandran outliers, % | 15 | 0 |
| Rotamer outliers, % | 23 | 0 |
| $C_\beta$ deviations, % | 0.5 | 0 |

# Using validation tools as refinement goals

- In low-resolution refinement we use extra restraints to compensate for lack of data:
    - Ramachandran plot restraints
    - Cβ deviation restraints
    - Secondary structure restraints
    - Restraints on χ angles of amino-acid side-chain rotamers
- These are standard validation tools… using them as restraints compromises their validation power

# Model validation

## PNAS, 2019 116 (39) 19513-19522

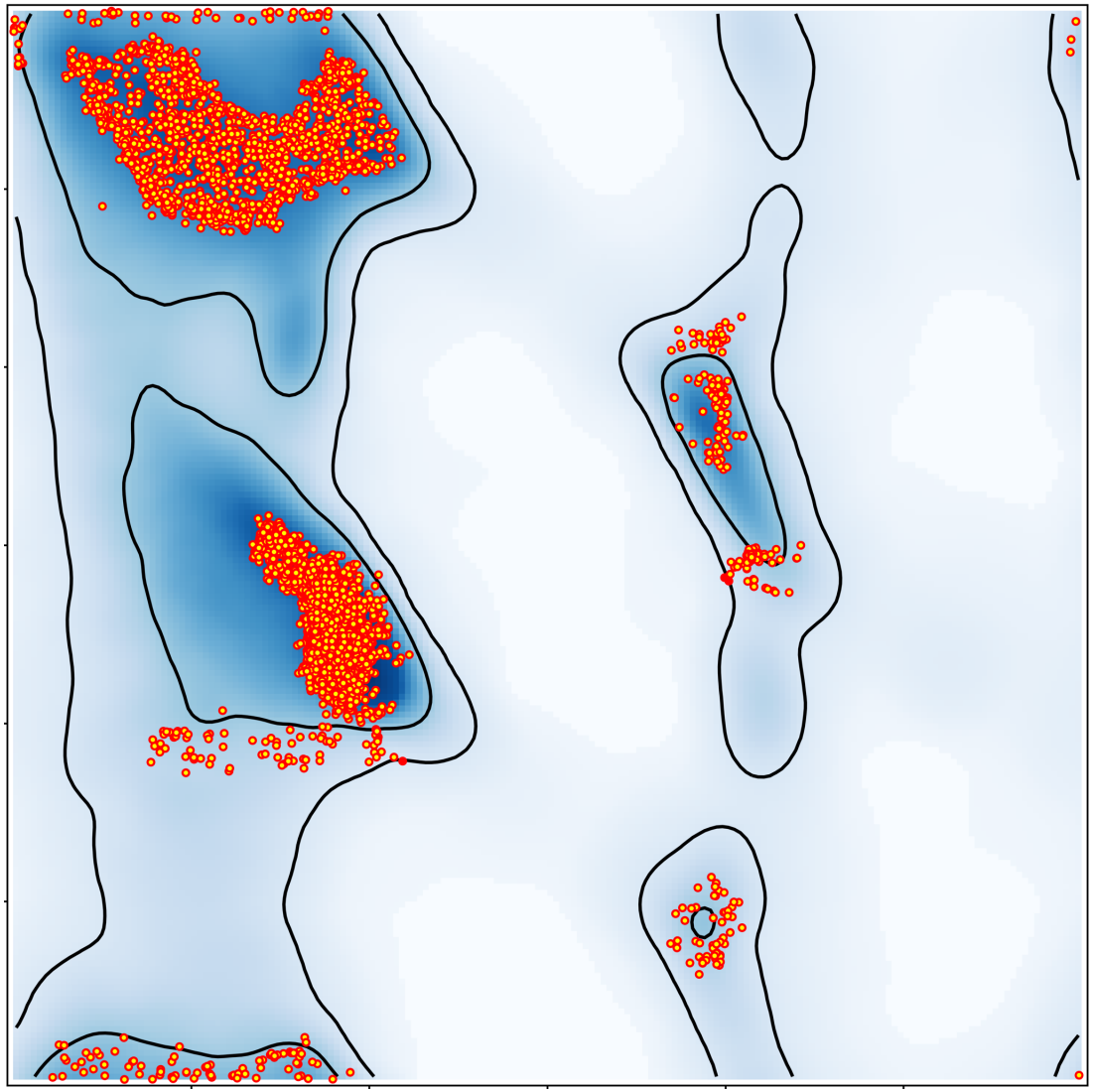


| Metric / PDB code | | 6KS6 |
|---|---|---|
| Clashscore | | 7.7 |
| Rama. (%) | favored | 96.4 |
| | outliers | 0.2 |
| Rotamer outliers (%) | | 0 |
| $C_\beta$ deviations | | 0 |
| RMSD | Bond (Å) | 0.001 |
| | Angle (°) | 0.396 |
| Resolution (Å) | | 3.0 |

Perfect statistics! All looks just great!

# Model validation: *Ramachandran plot*

6KS6



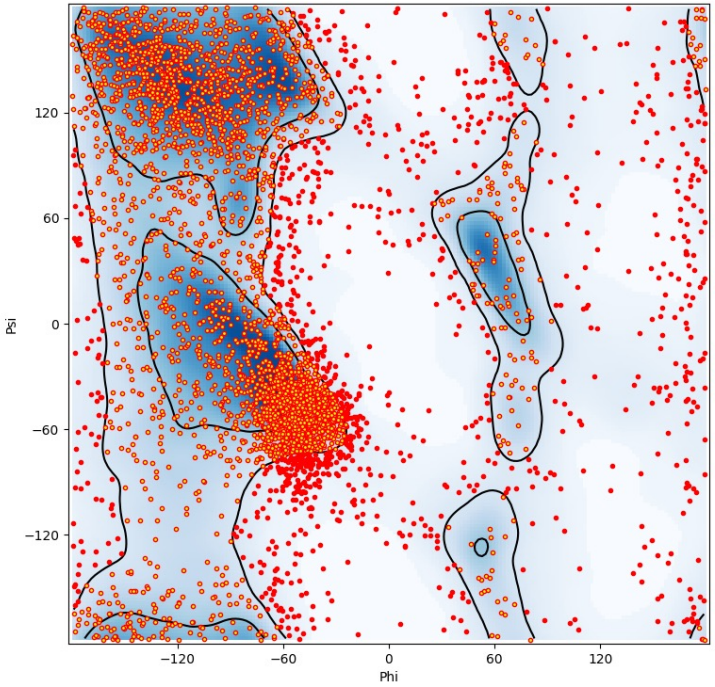
Odd Ramachandran plot. How we know this?

# Ramachandran plot restraints

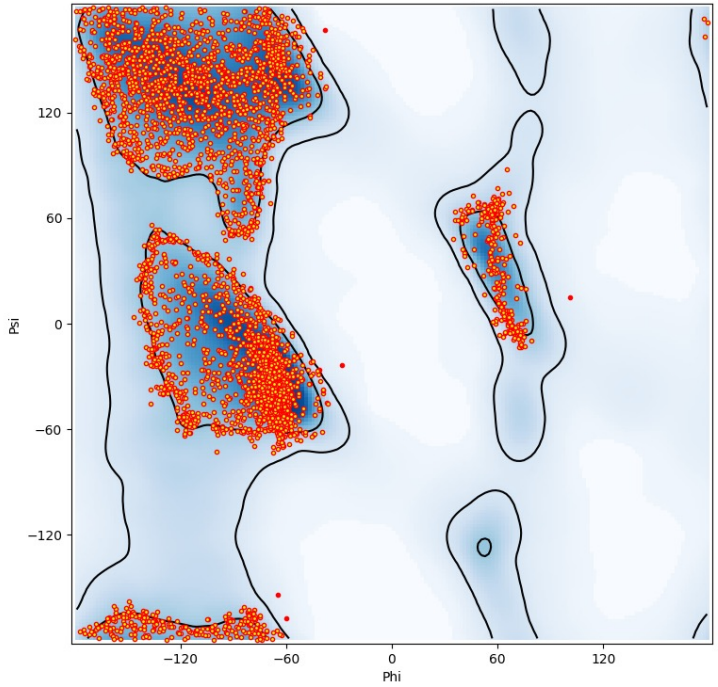
- Always use at low resolution
- Do not use to fix existing outliers

**PDB code: 5a9z**

Original



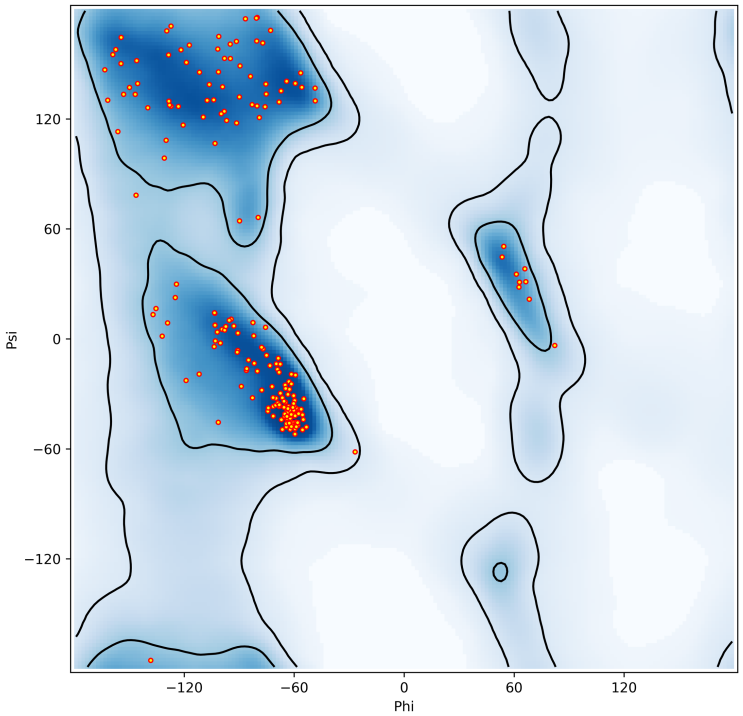

Refined with Ramachandran plot restraints



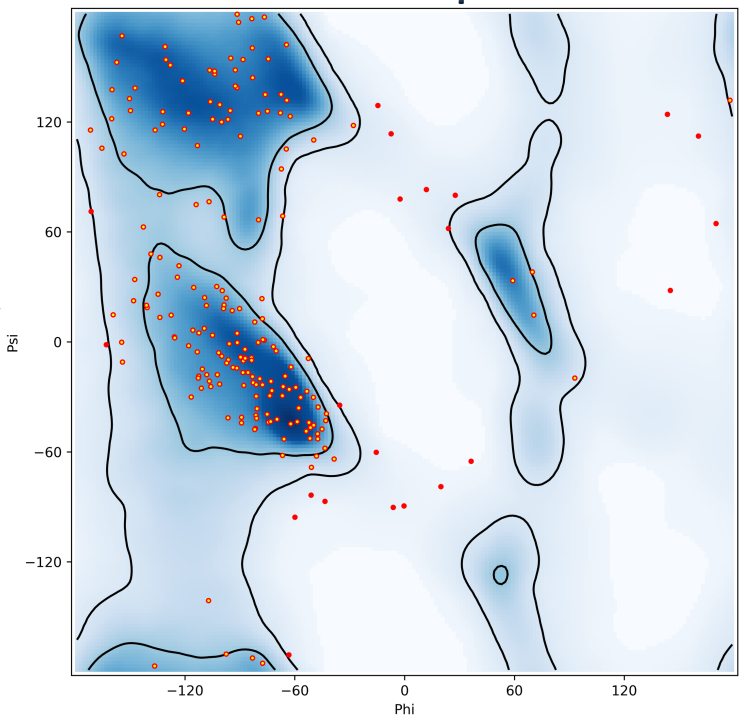
# Ramachandran plot restraints

- Ramachandran plot restraints
  - Use to stop outliers from occurring

Before refinement



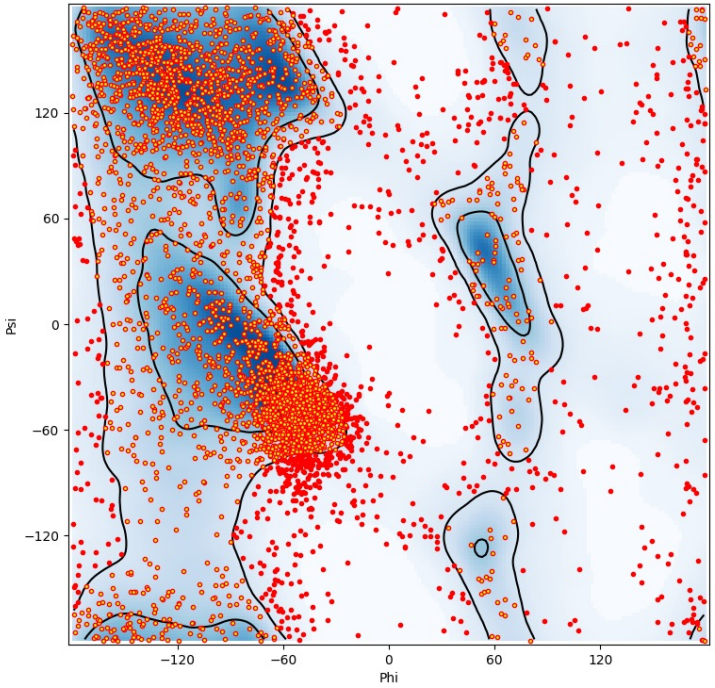

After refinement
(No Ramachandran plot restraints)



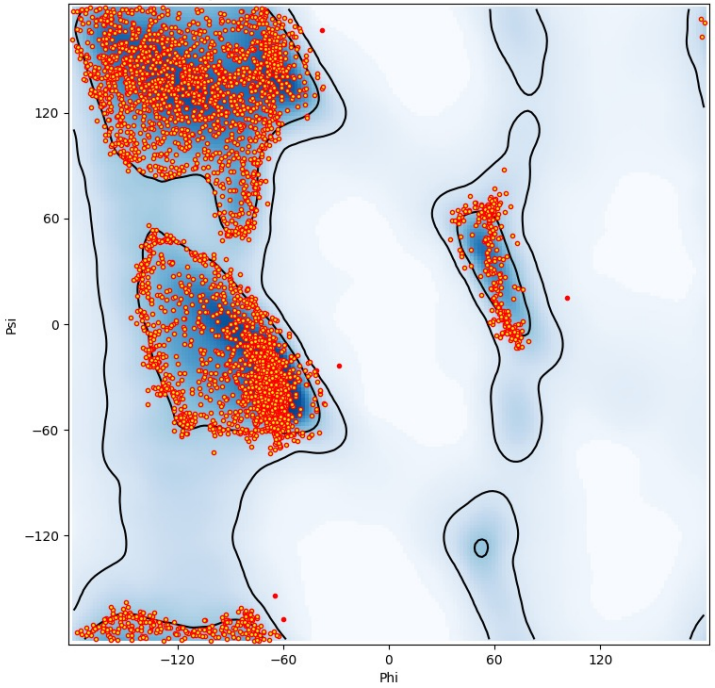
# Ramachandran plot restraints

- What is wrong with this plot?

Original

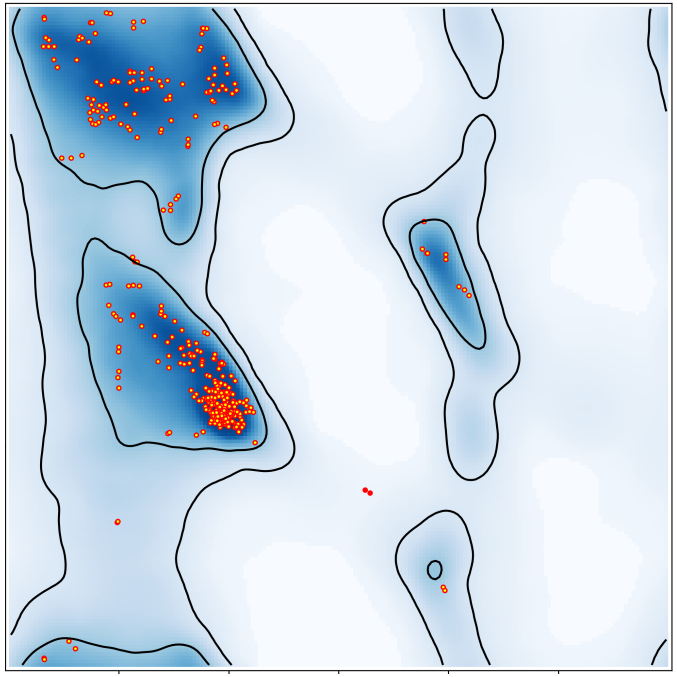



Refined with Ramachandran plot restraints



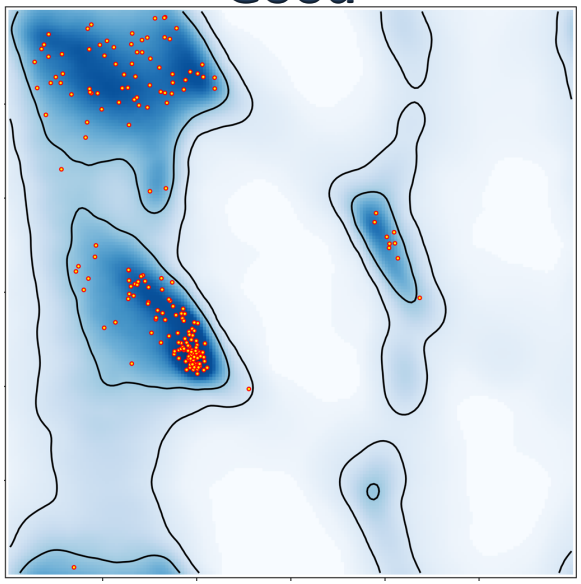
# Ramachandran plot restraints

- It is very different from what we expect!

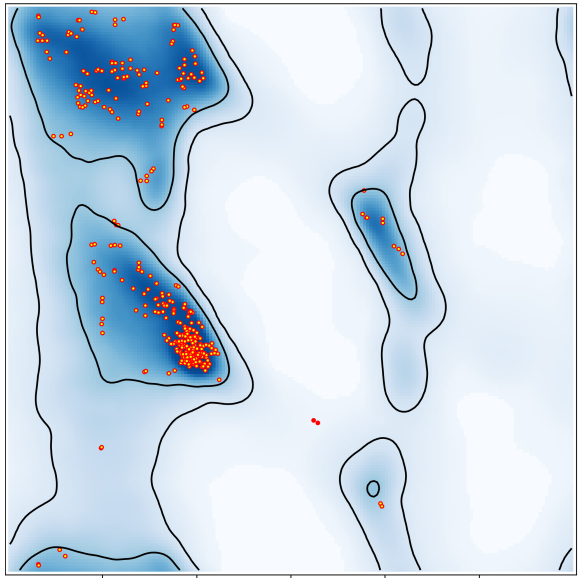


# How you can tell good vs bad plot?

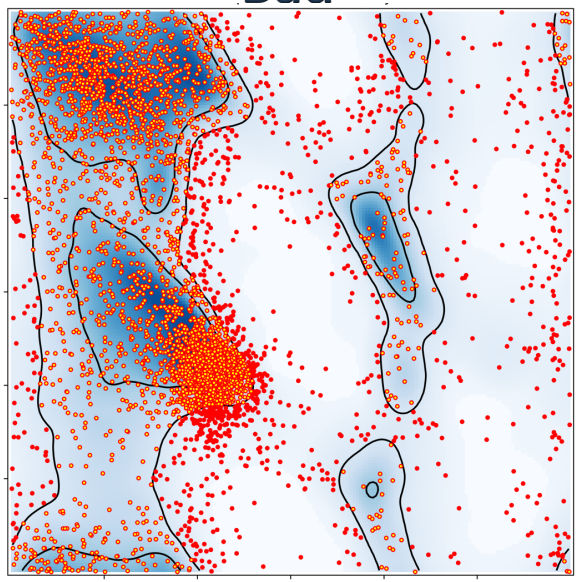
Good



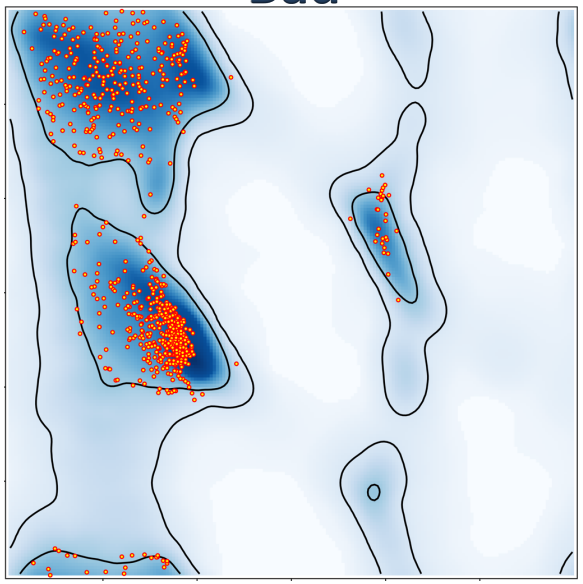
Good



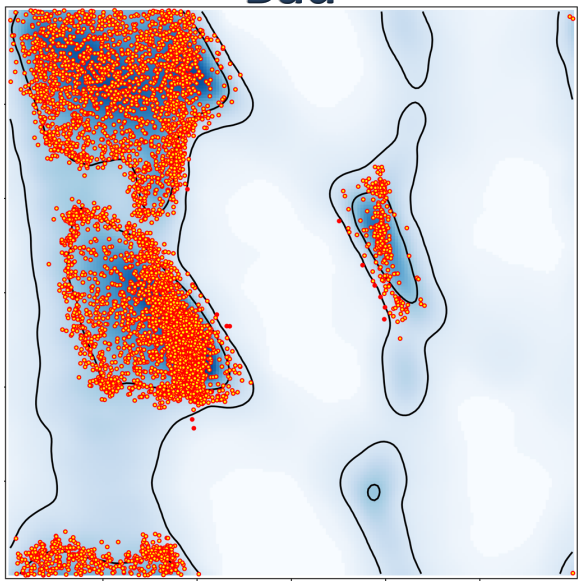
Bad



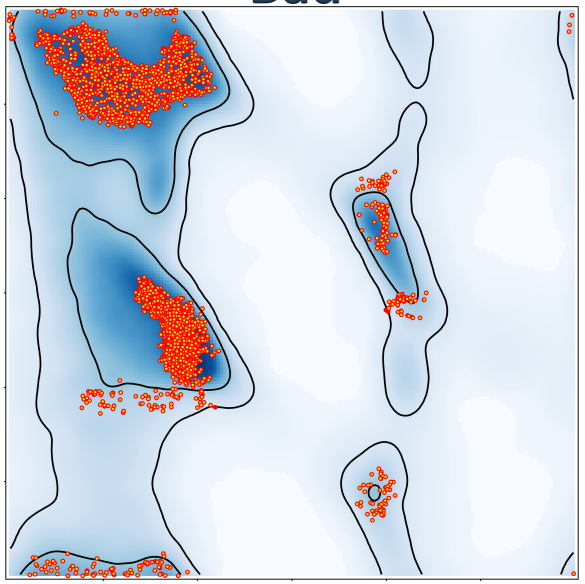
Bad



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

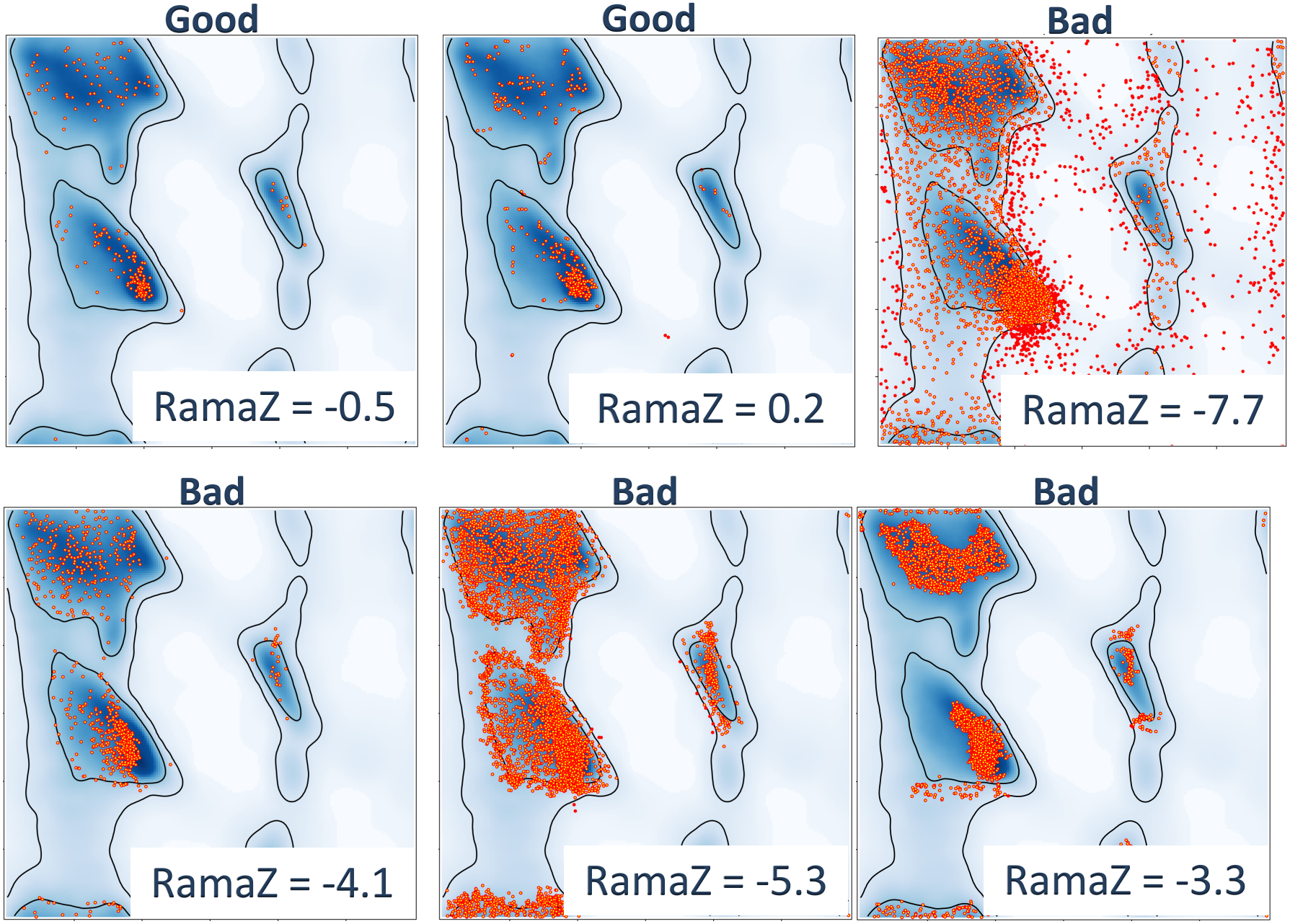
CellPress

Resource

**A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry**

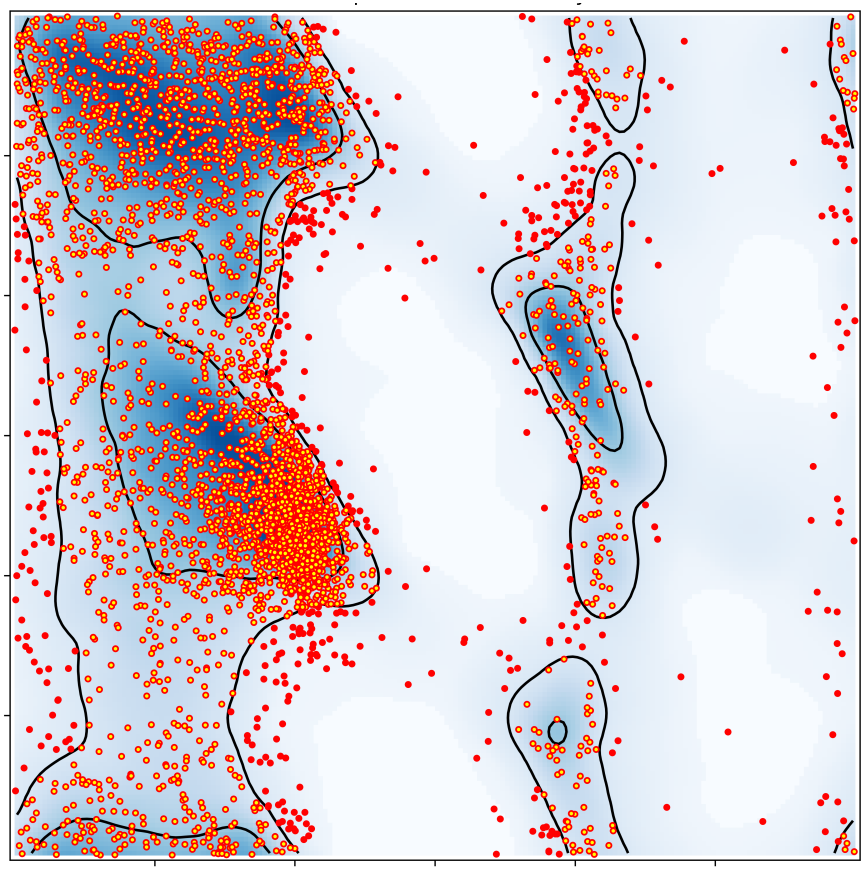
Oleg V. Sobolev,[1,5,*] Pavel V. Afonine,[1] Nigel W. Moriarty,[1] Maarten L. Hekkelman,[2,3] Robbie P. Joosten,[2,3,*] Anastassis Perrakis,[2,3] and Paul D. Adams[1,4]

# Model validation: *Ramachandran plot Z-score*

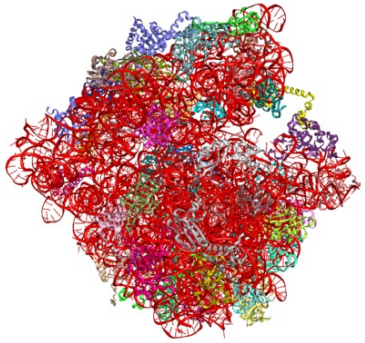


# Validation: why to do?

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



**RamaZ = -3.3**

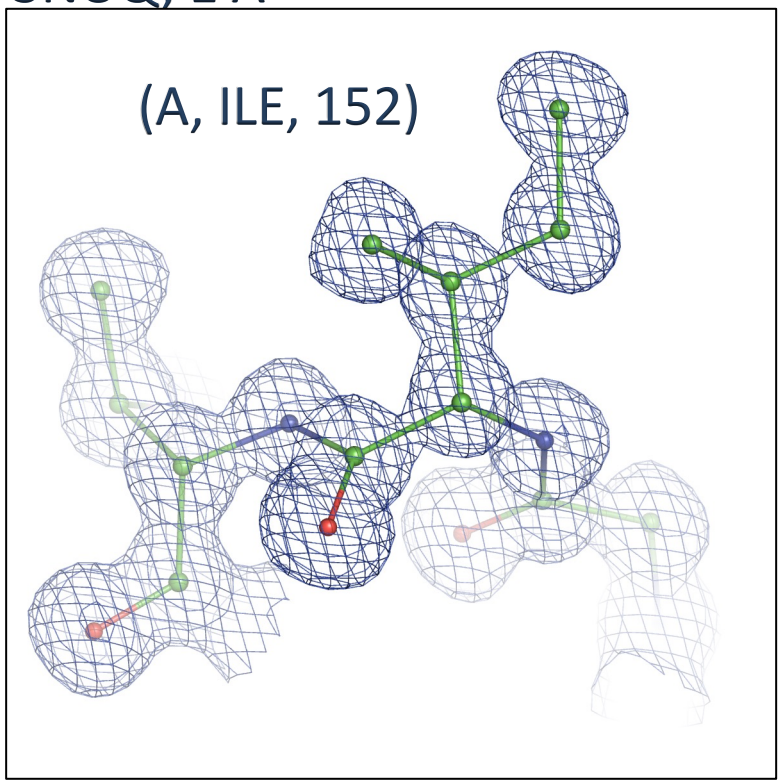


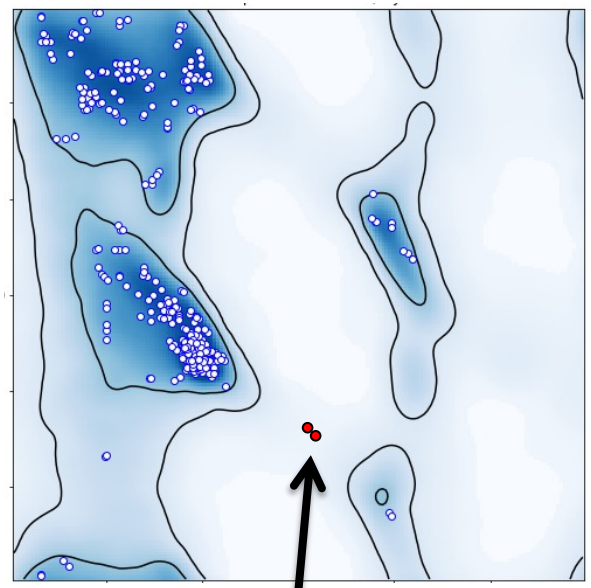
| Metric | 6o9j | Expected |
|---|---|---|
| Clashscore | 70 | Less than 10 |
| Ramachandran favored, % | 59 | More than 98 |
| Ramachandran outliers, % | 15 | 0 |
| Rotamer outliers, % | 23 | 0 |
| $C_\beta$ deviations, % | 0.5 | 0 |

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

# An outlier ≠ wrong

3NOQ, 1 Å



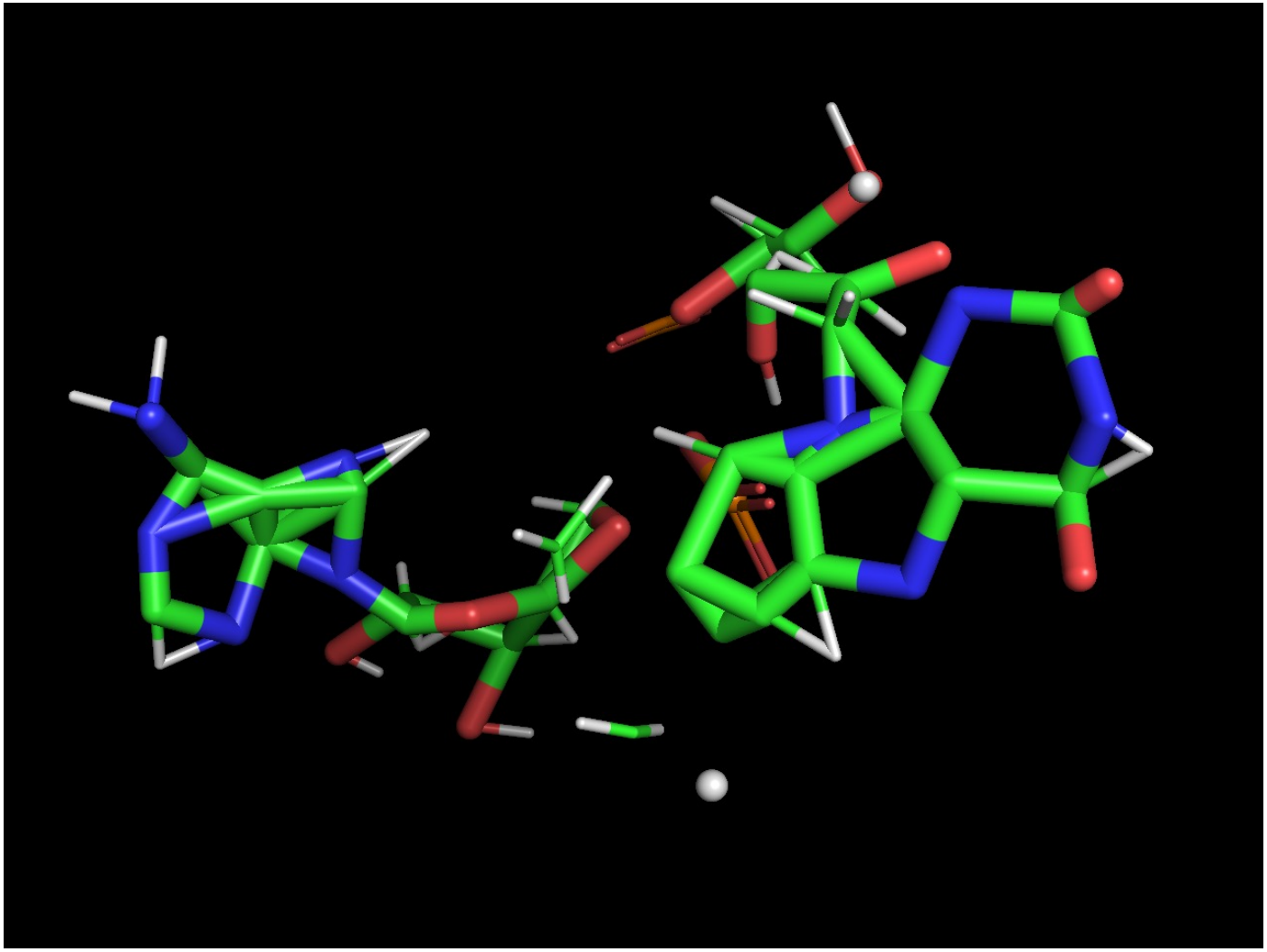
Outliers:

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

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

# Local *vs* Global

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



# Local *vs* Global

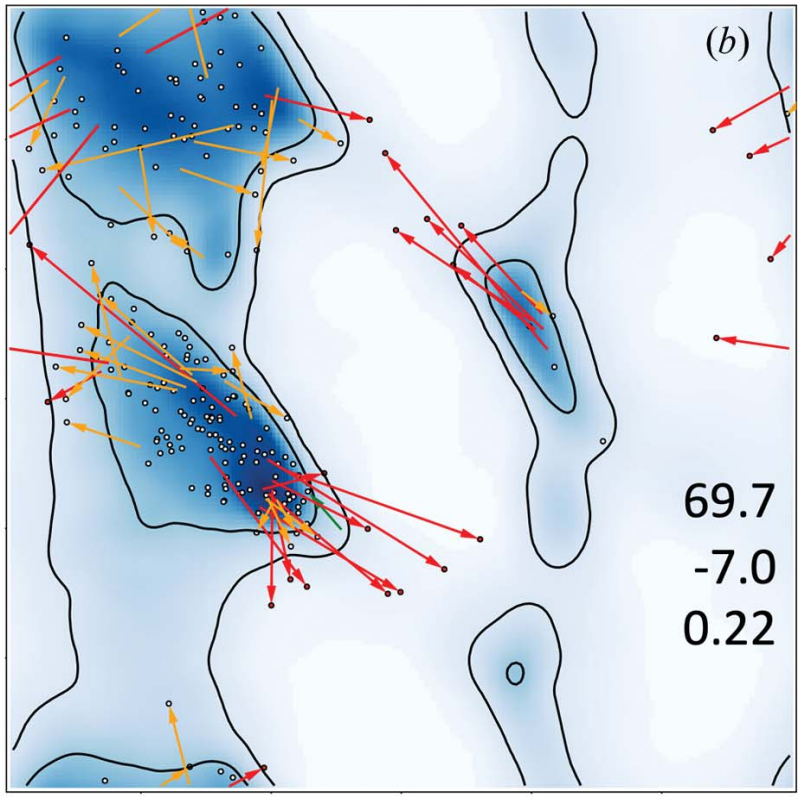
- RMSD from ideal: bonds = 0.01Å  angles = 1.6°

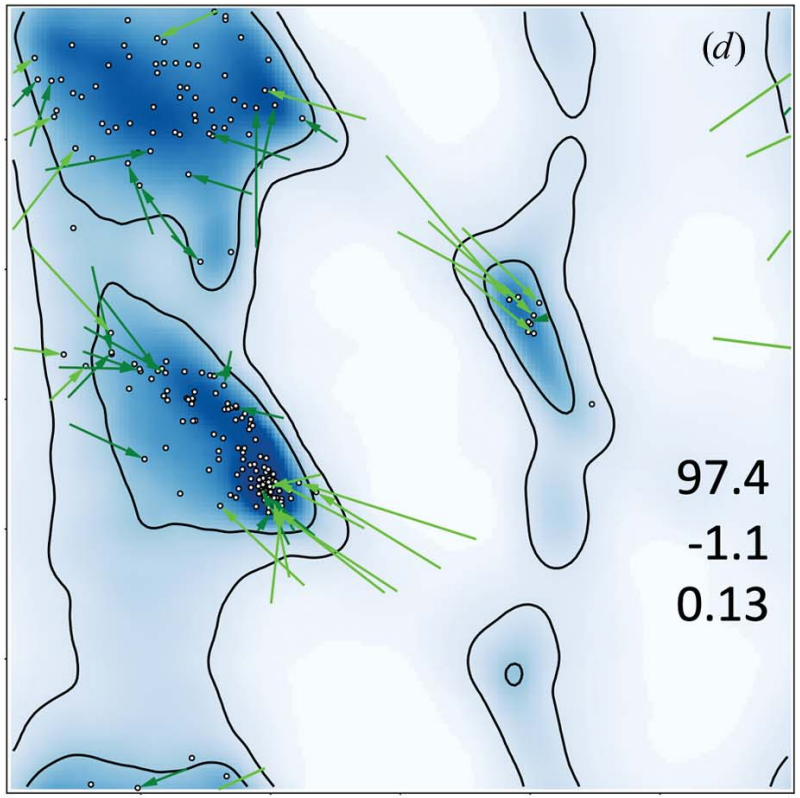
| Histogram of deviations from ideal values | | | |
|---|---|---|---|
| Bonds | | Angles | |
| 0.000 - 0.035: | 2645 | 0.000 - 9.313: | 4208 |
| 0.035 - 0.070: | 19 | 9.313 - 18.626: | 9 |
| 0.070 - 0.106: | 13 | 18.626 - 27.939: | 3 |
| 0.106 - 0.141: | 5 | 27.939 - 37.252: | 4 |
| 0.141 - 0.176: | 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 |

# Validation – Sequence register errors

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

# Comparama: phenix.comparama

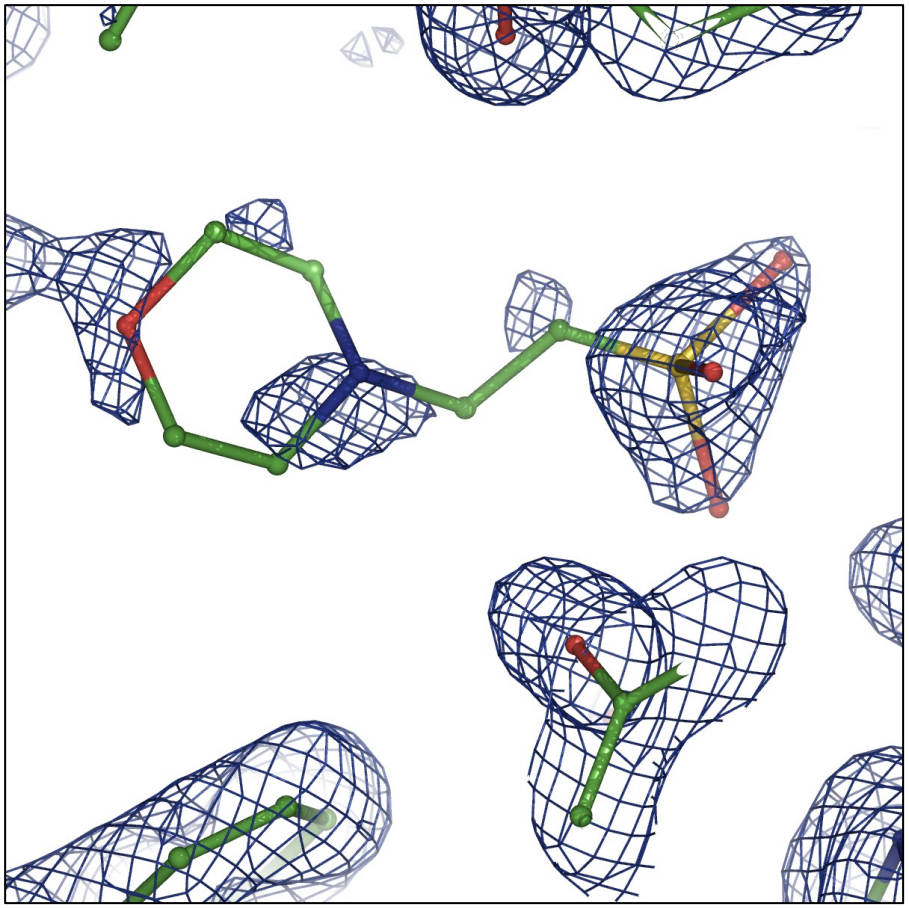


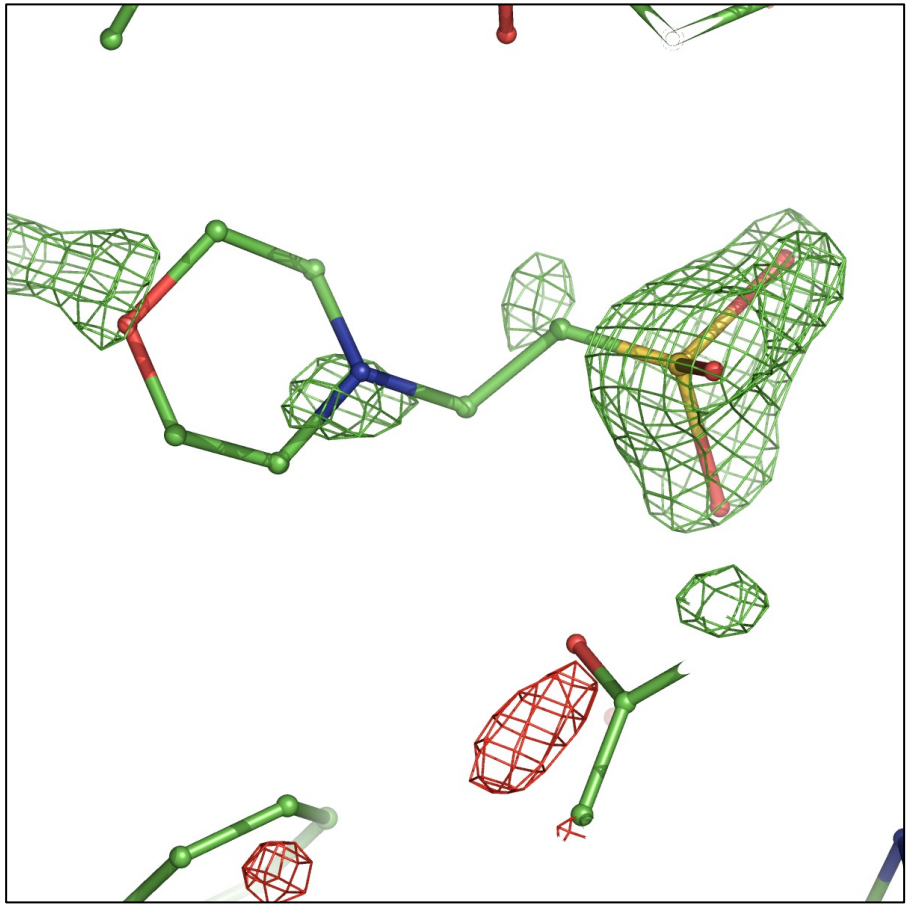
# Ligands and Polder map

## PDB code: 1ABA, Resolution: 1.45 Å

### 2mFo-DFc (1σ)



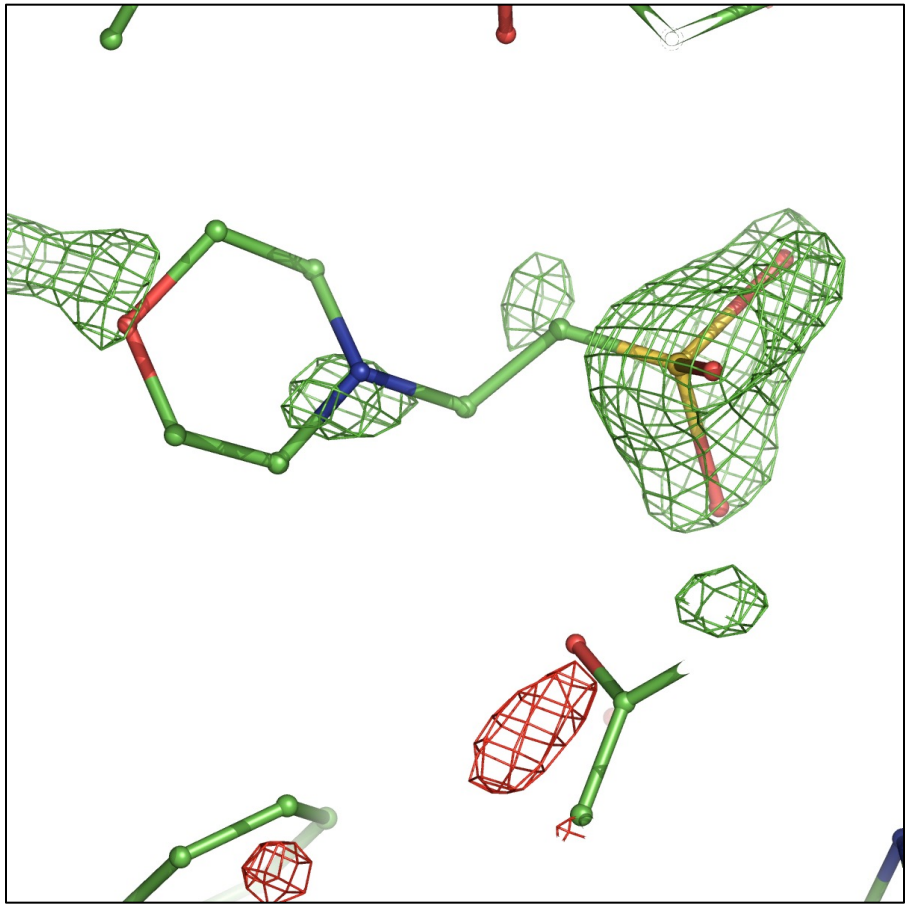
### mFo-DFc (±3σ)



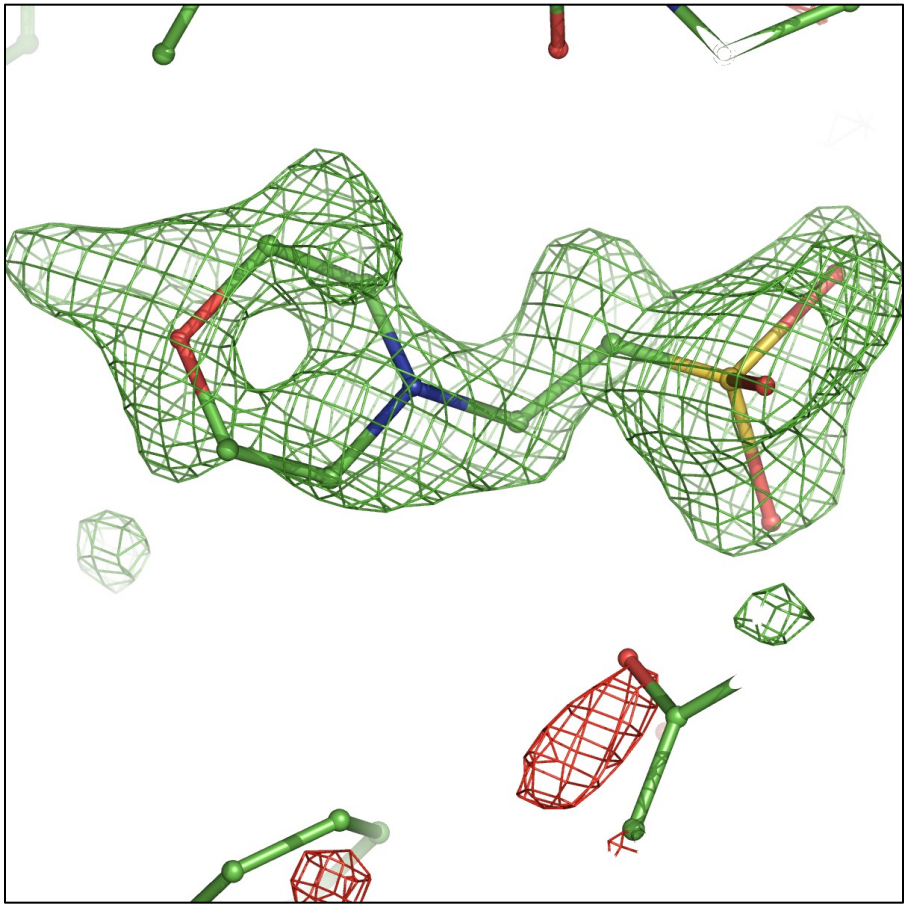
# Ligands and Polder map

## PDB code: 1ABA, Resolution: 1.45 Å

### mFo-DFc (±3σ)

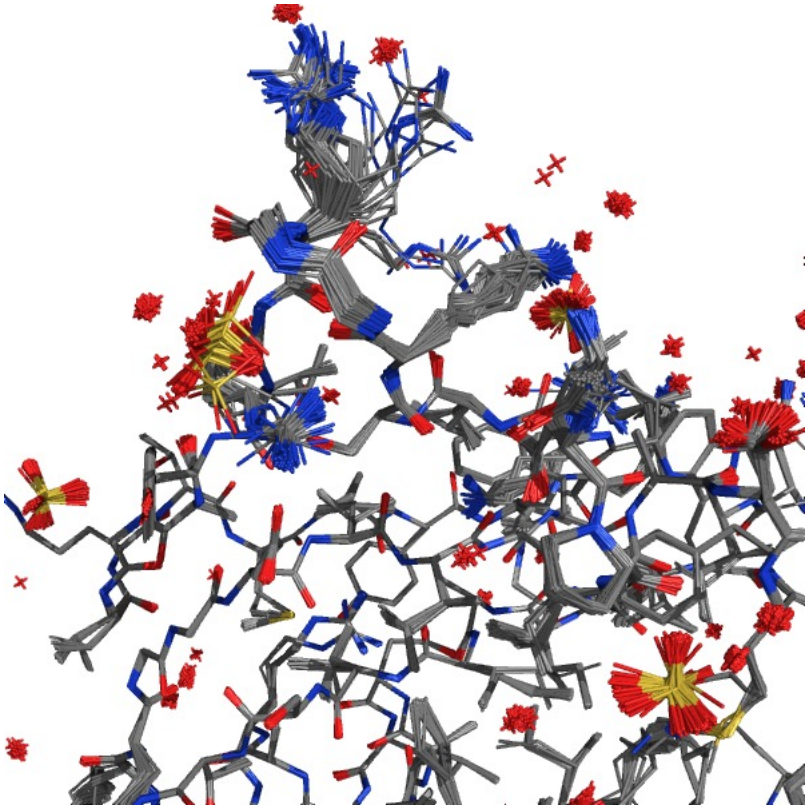


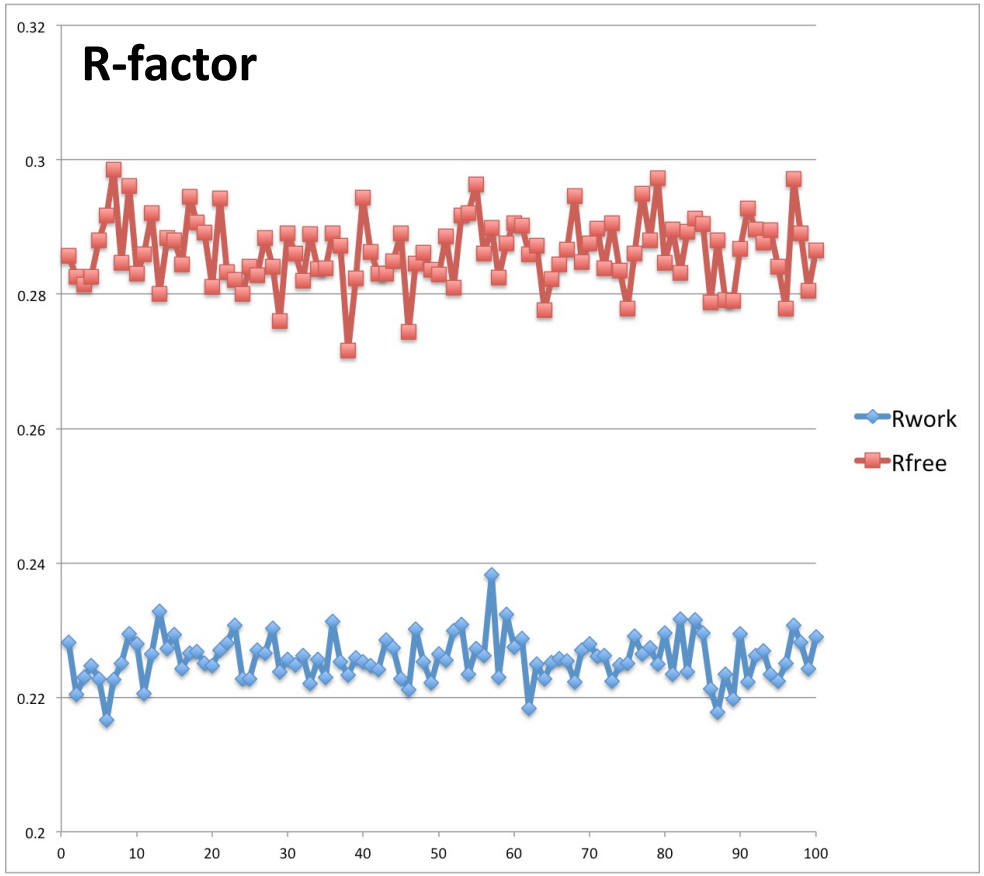
### Polder mFo-DFc (±3σ)



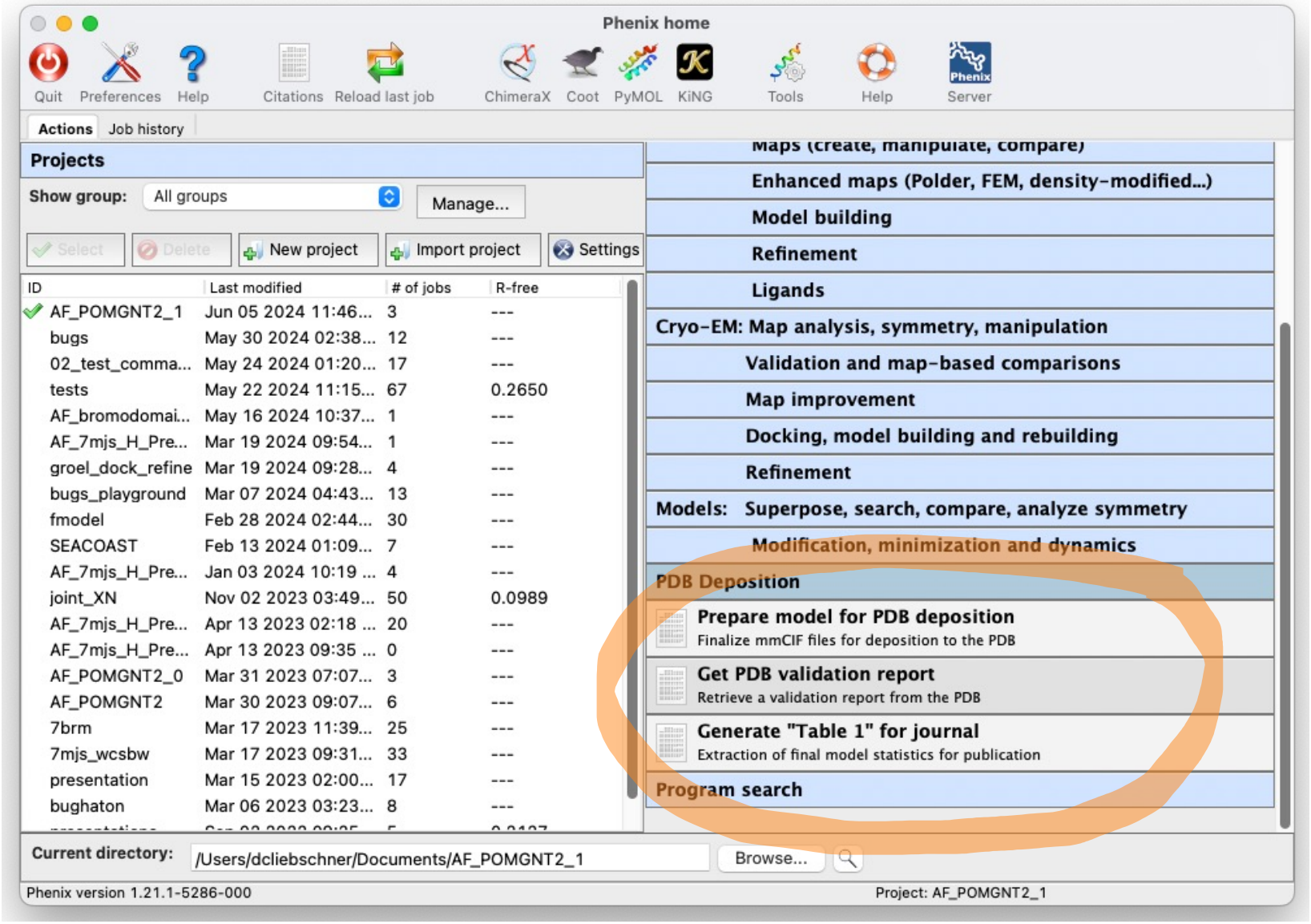
# Estimating and using uncertainty

100 identical refinement runs each one starting with slightly perturbed model





# PDB deposition



# PDB deposition

mmCIF format is mandatory for deposition as of 2019

Acta Cryst D STRUCTURAL BIOLOGY

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