

Model validation

Pavel Afonine

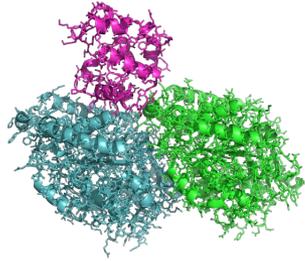
Lawrence Berkeley National Laboratory (LBNL)

September 26th, 2024

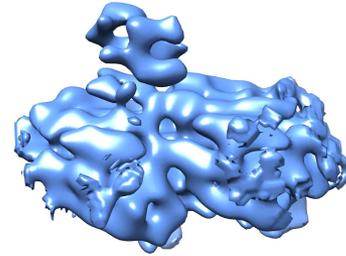
BNL

Validation

Model

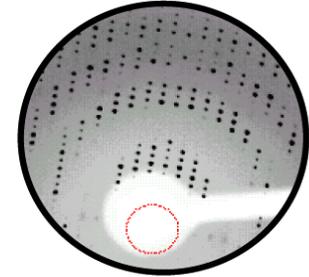


Data



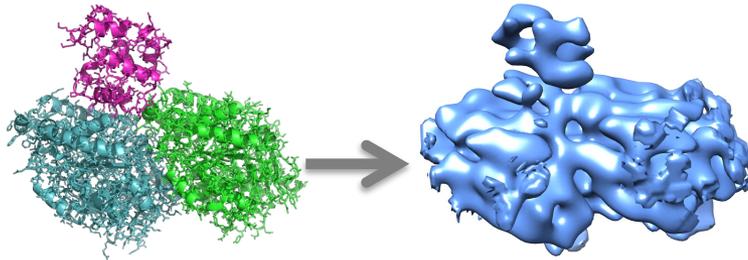
Cryo-EM

or



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

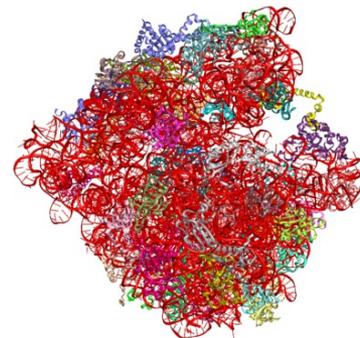
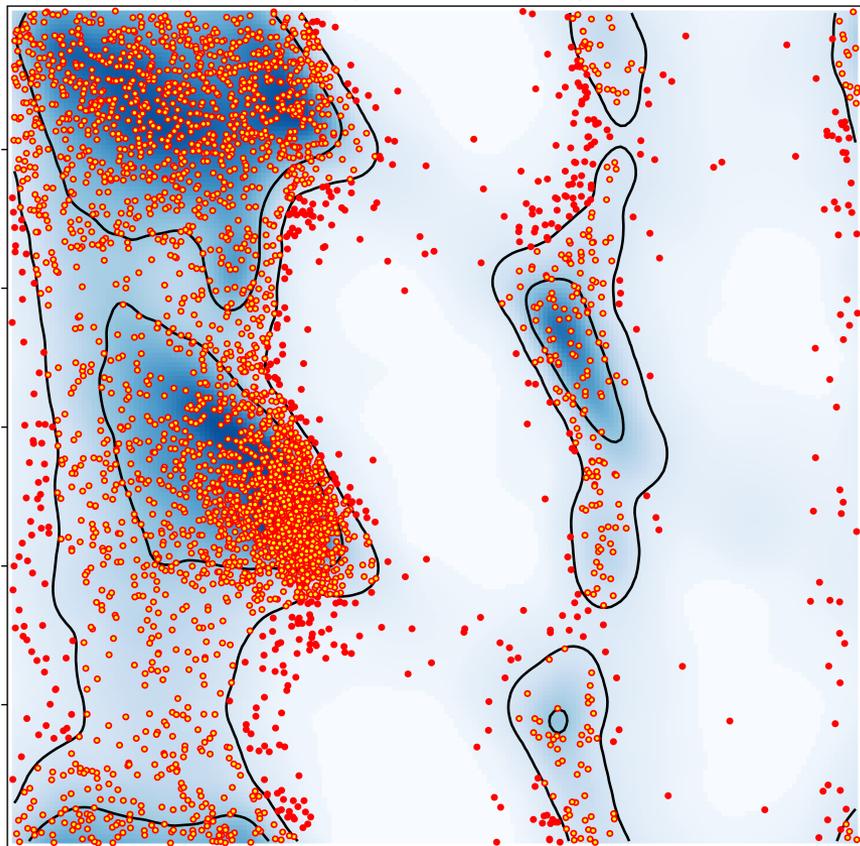
Nat. Struct. Biol. 7, 687–692 (2000); retracted 6 July 2009

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



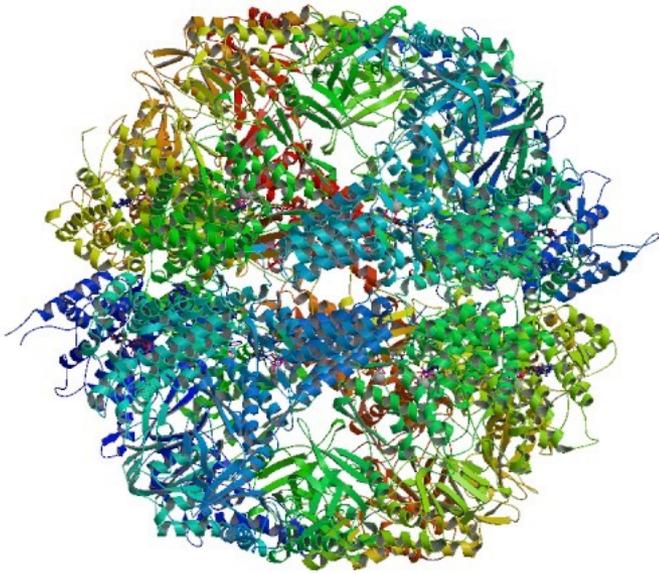
Metric	6o9j	Expected
Clashscore	70	Less than 10
Ramachandran favored, %	59	More than 98
Ramachandran outliers, %	15	0
Rotamer outliers, %	23	0
C _β deviations, %	0.5	0

Using validation tools as refinement goals

- In low-resolution refinement we use extra restraints to compensate for lack of data:
 - Ramachandran plot restraints
 - C β 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

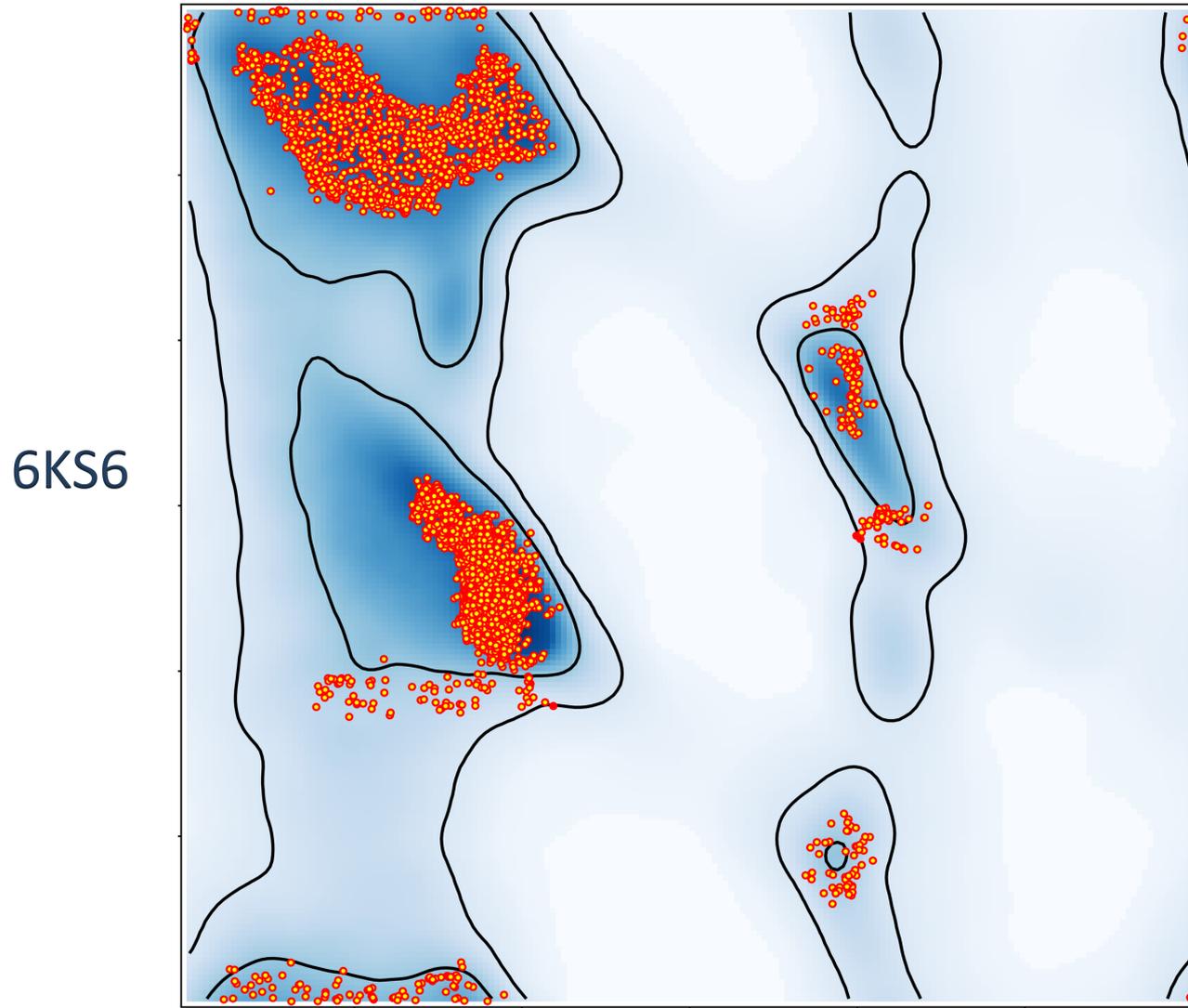
PNAS, 2019 116 (39) 19513-19522



Metric / PDB code		6KS6
Clashscore		7.7
Rama. (%)	avored	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*



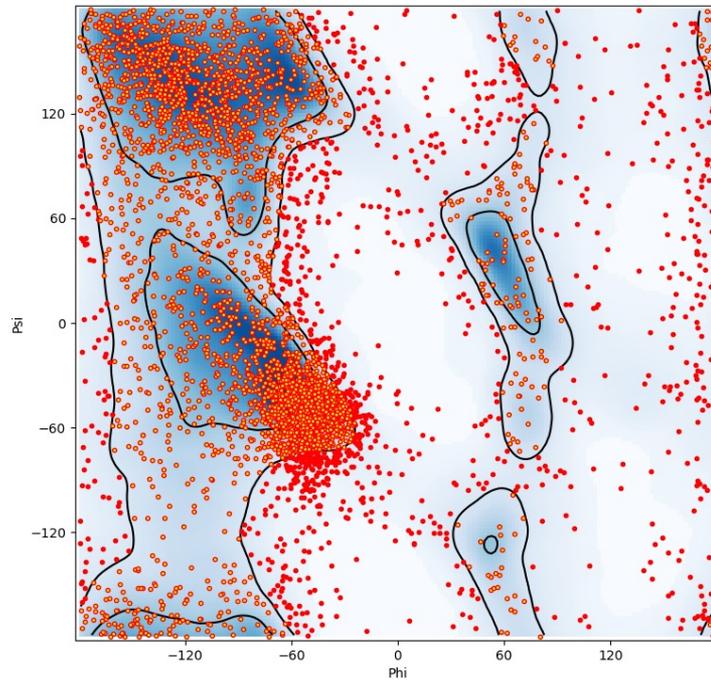
Odd Ramachandran plot. How we know this?

Ramachandran plot restraints

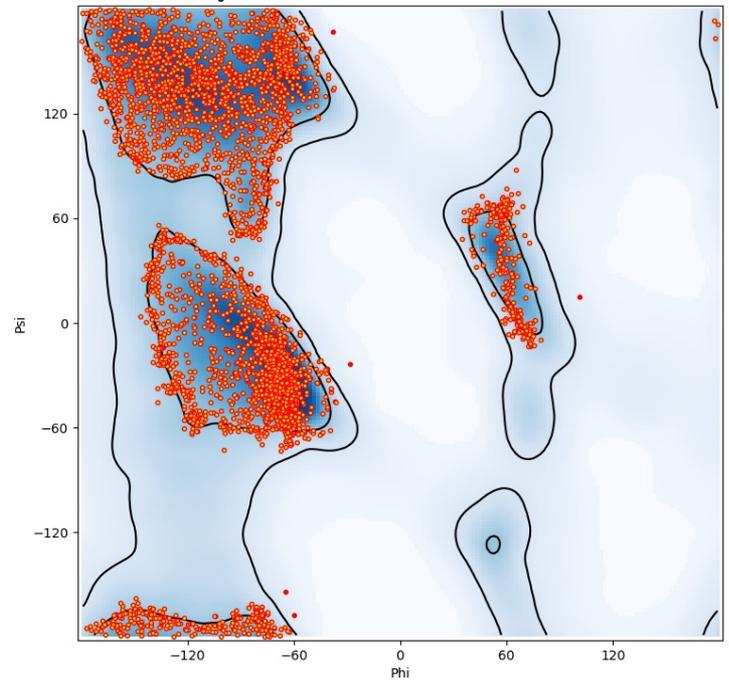
- Always use at low resolution
- Do not use to fix existing outliers

PDB code: 5a9z

Original



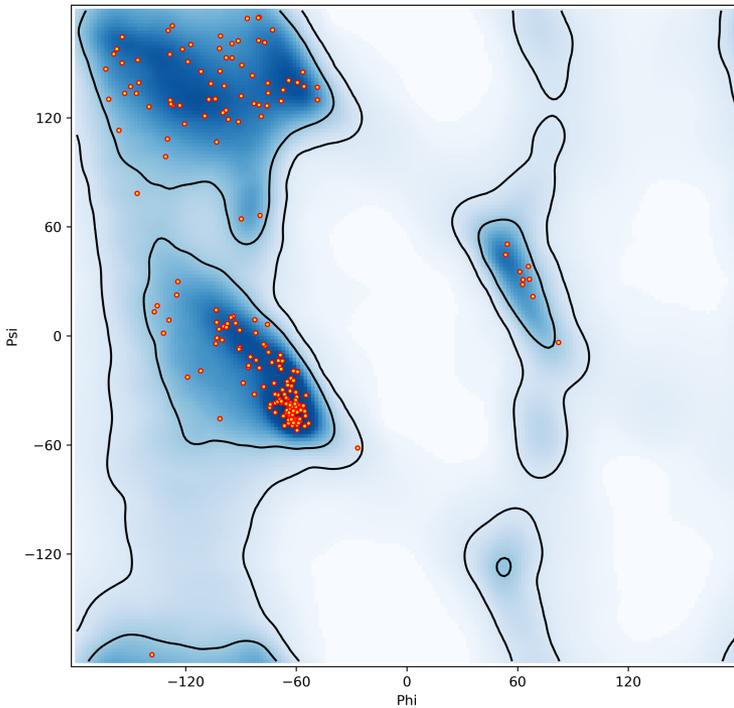
Refined with Ramachandran plot restraints



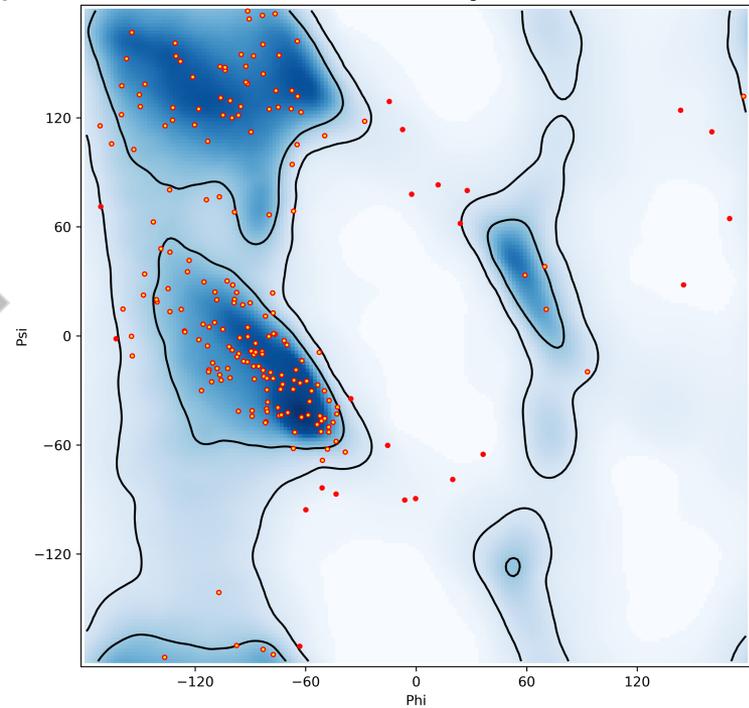
Ramachandran plot restraints

- Ramachandran plot restraints
 - Use to stop outliers from occurring

Before refinement



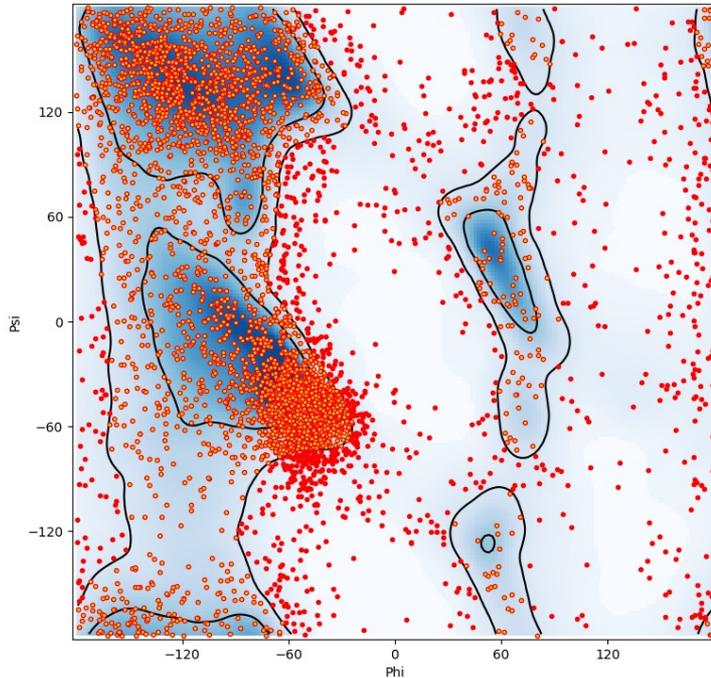
After refinement
(No Ramachandran plot restraints)



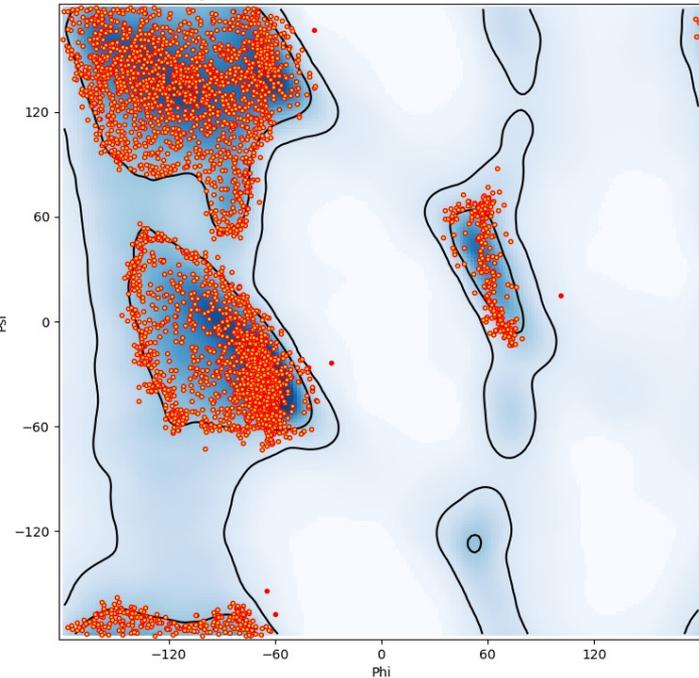
Ramachandran plot restraints

- What is wrong with this plot?

Original

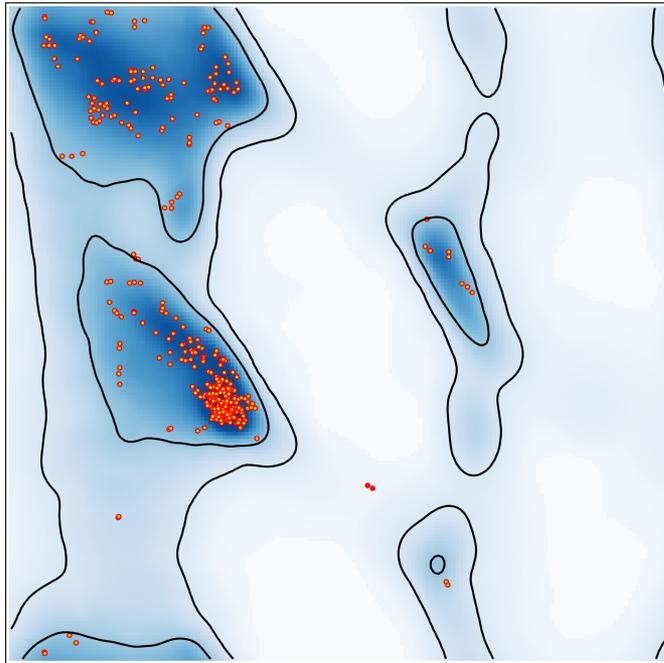


Refined with Ramachandran plot restraints



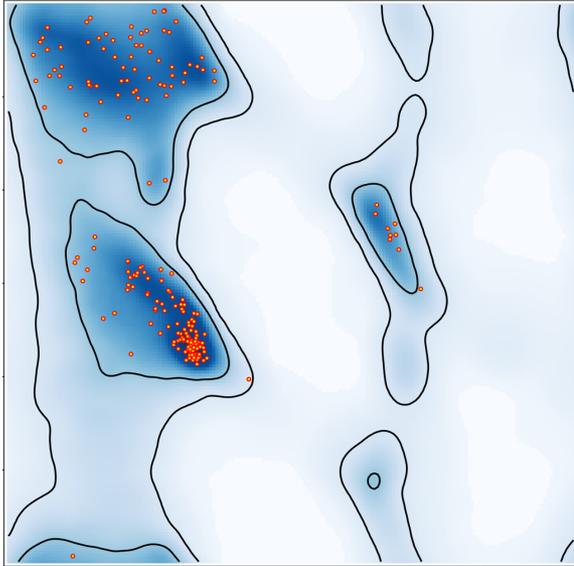
Ramachandran plot restraints

- It is very different from what we expect!

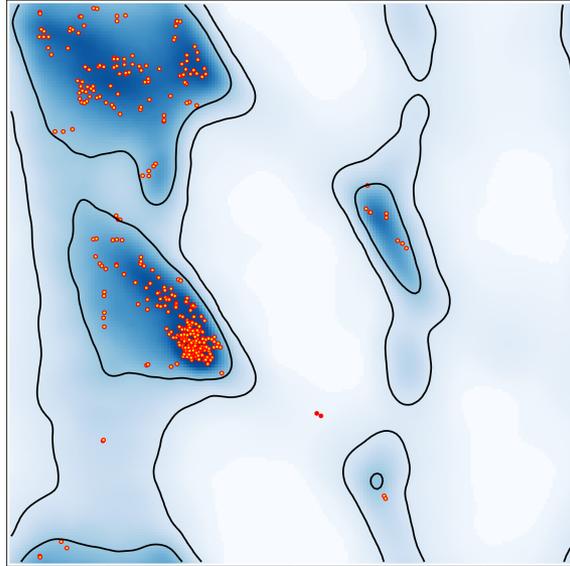


How you can tell good vs bad plot?

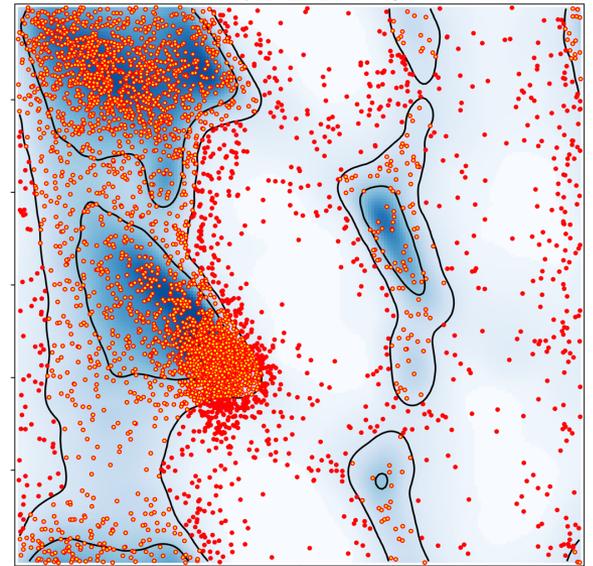
Good



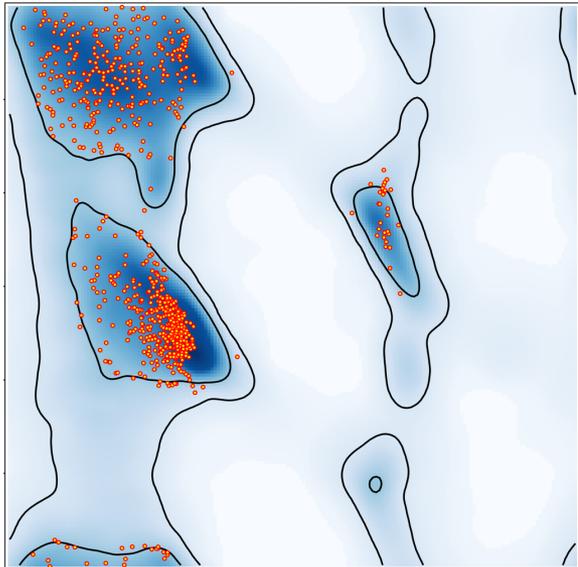
Good



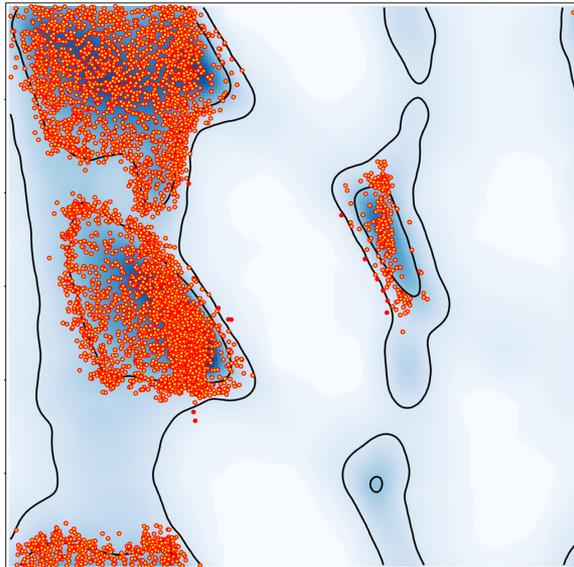
Bad



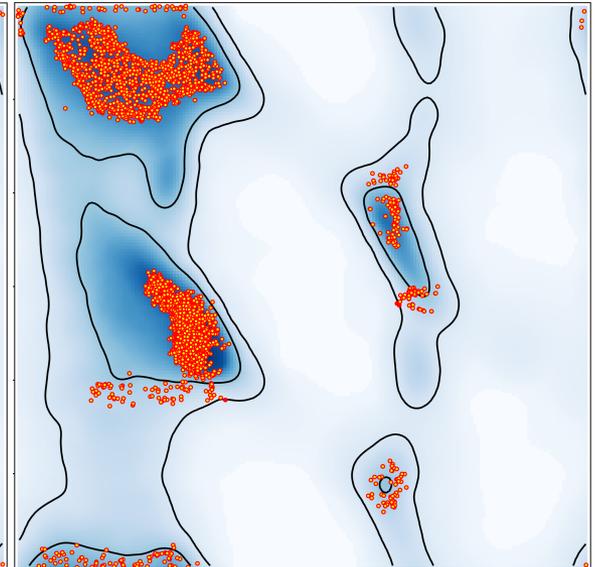
Bad



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

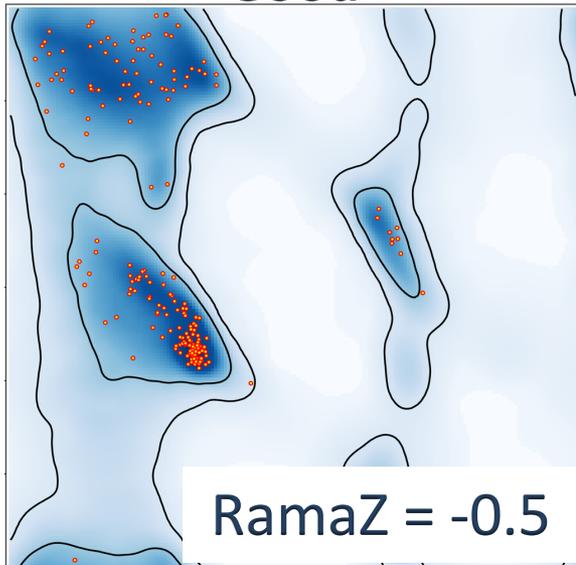
Resource

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

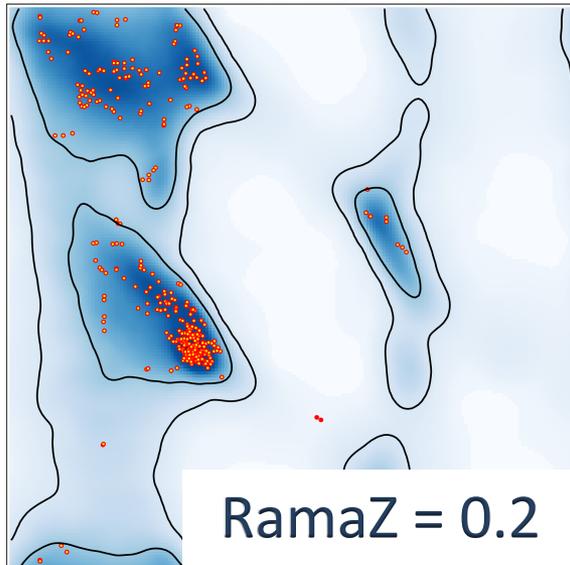
Oleg V. Sobolev,^{1,5,*} Pavel V. Afonine,¹ Nigel W. Moriarty,¹ Maarten L. Hekkelman,^{2,3} Robbie P. Joosten,^{2,3,*} Anastassis Perrakis,^{2,3} and Paul D. Adams^{1,4}

Model validation: *Ramachandran plot Z-score*

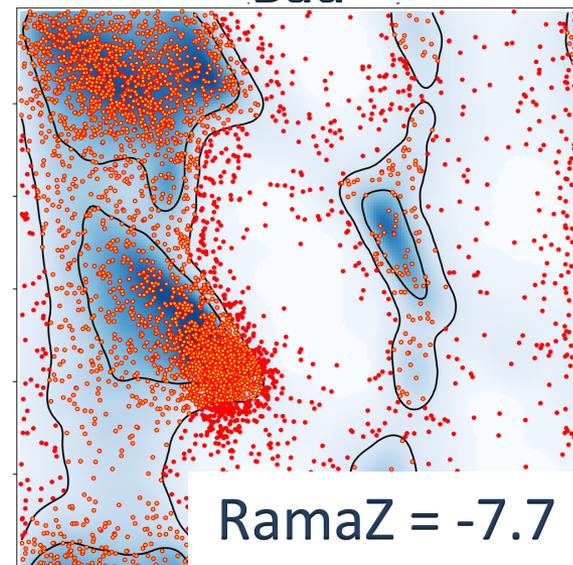
Good



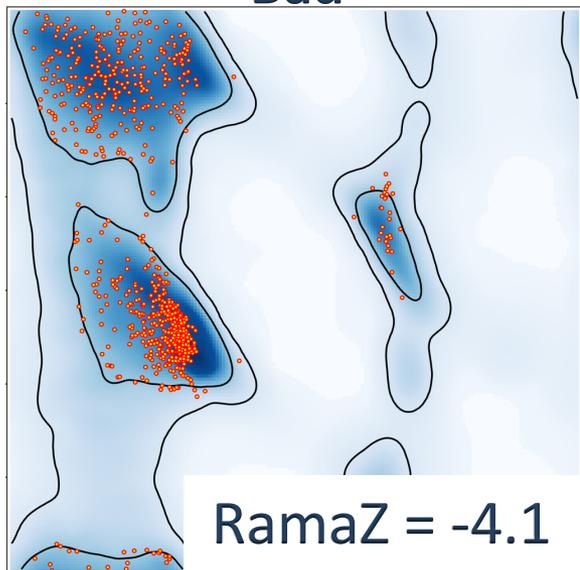
Good



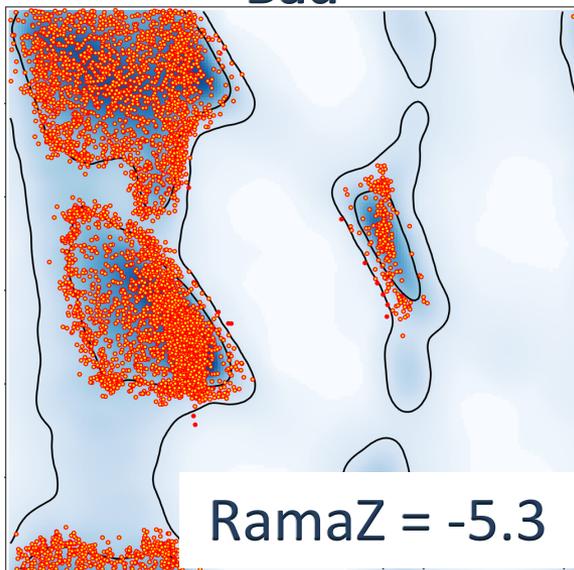
Bad



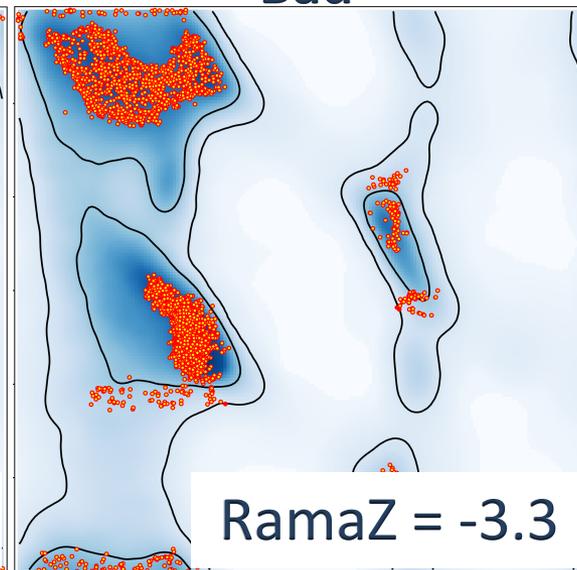
Bad



Bad

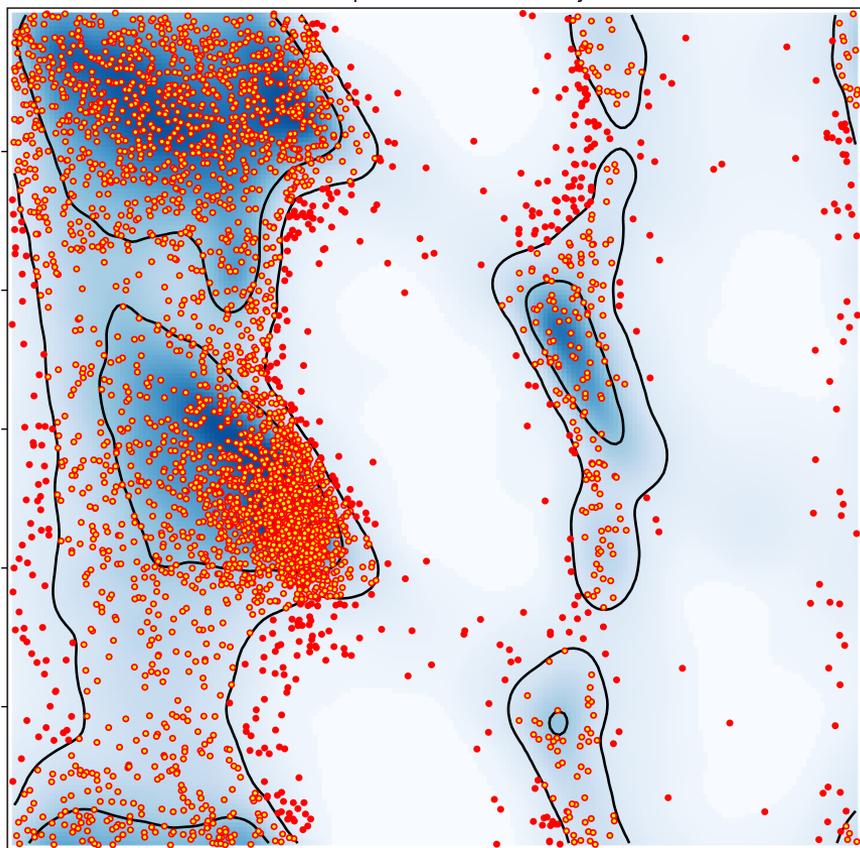


Bad



Validation: **why to do?**

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

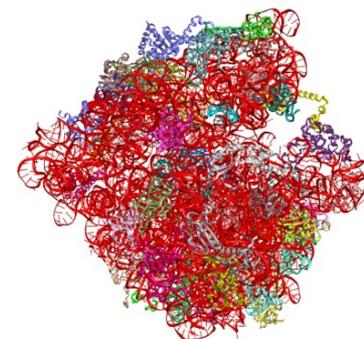


RamaZ = -3.3

Poor: $|Z| > 3$

Suspicious: $2 < |Z| < 3$

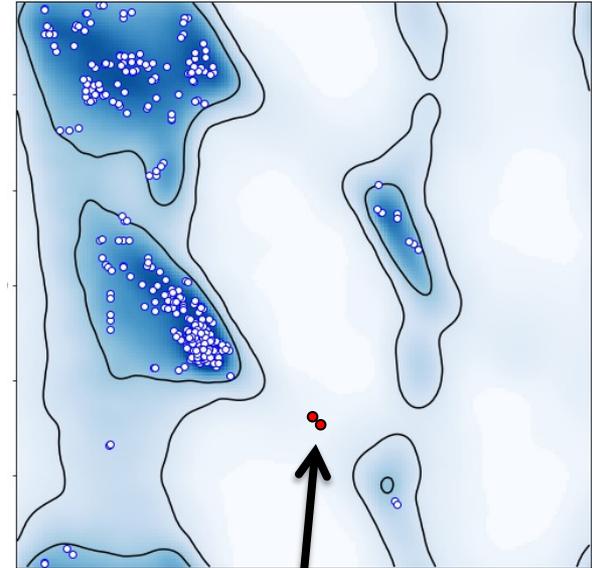
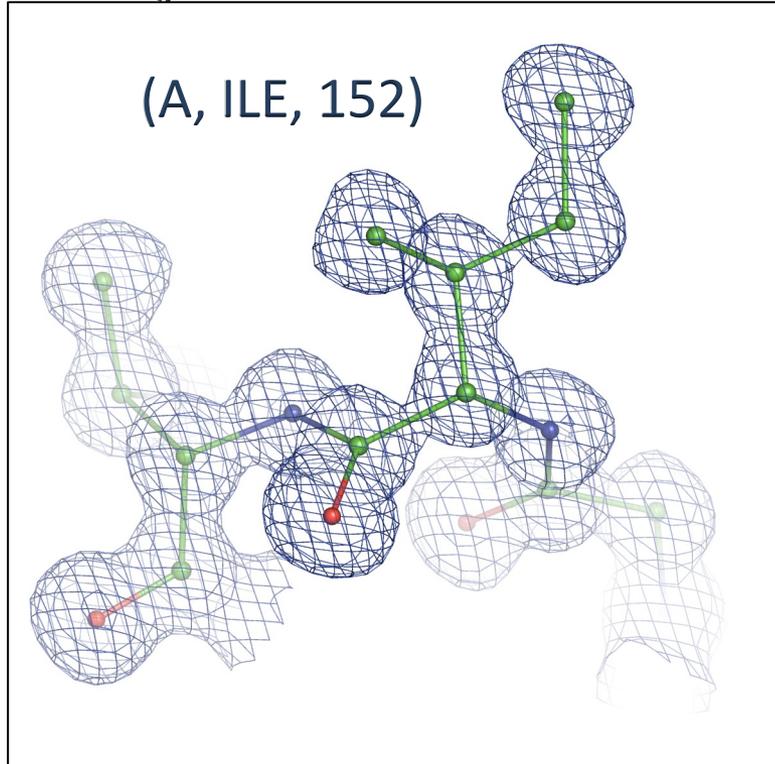
Good: $|Z| < 2$



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

An outlier \neq wrong

3NOQ, 1 Å



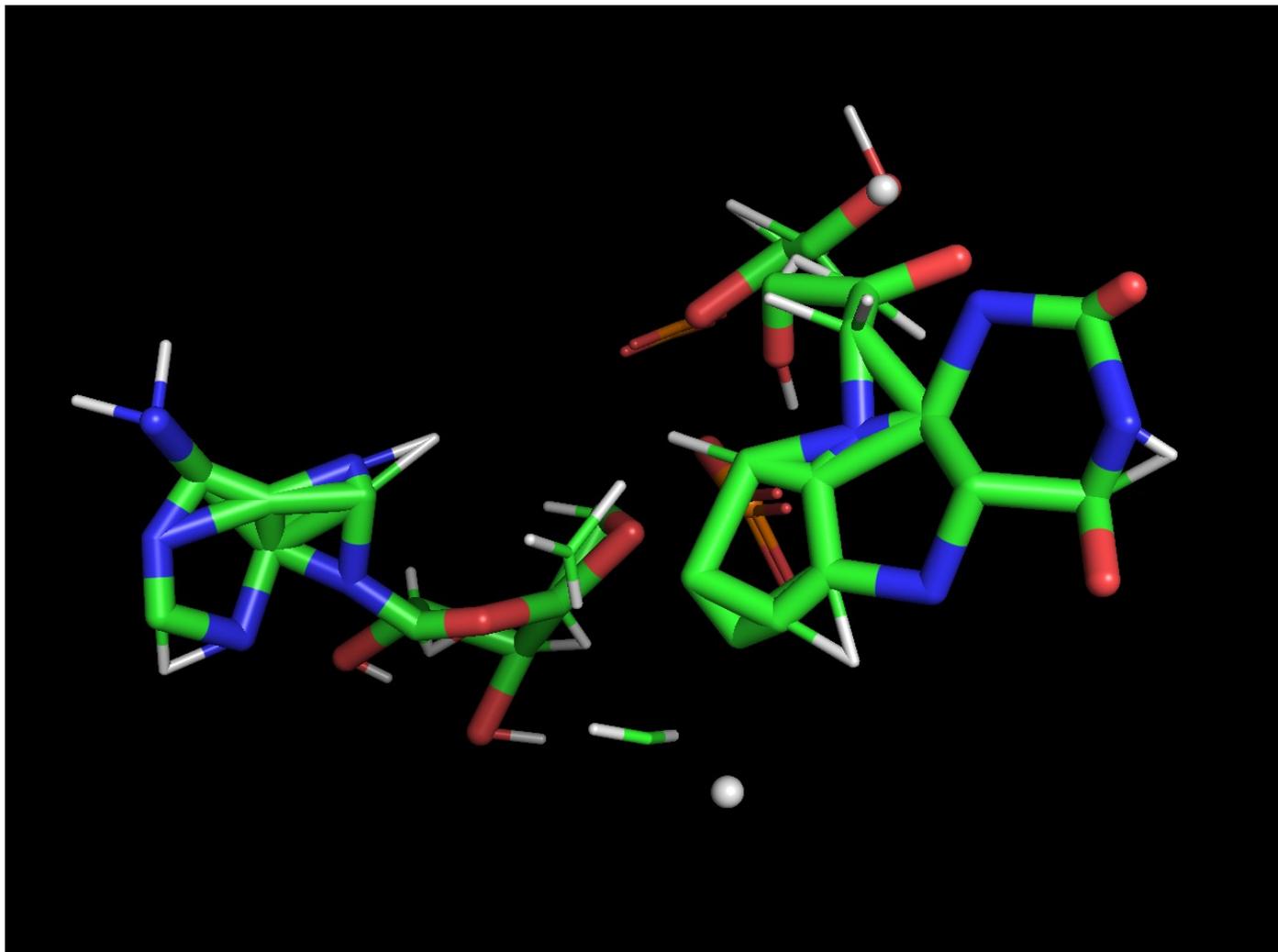
Outliers:

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

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

Local vs Global

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



Local vs Global

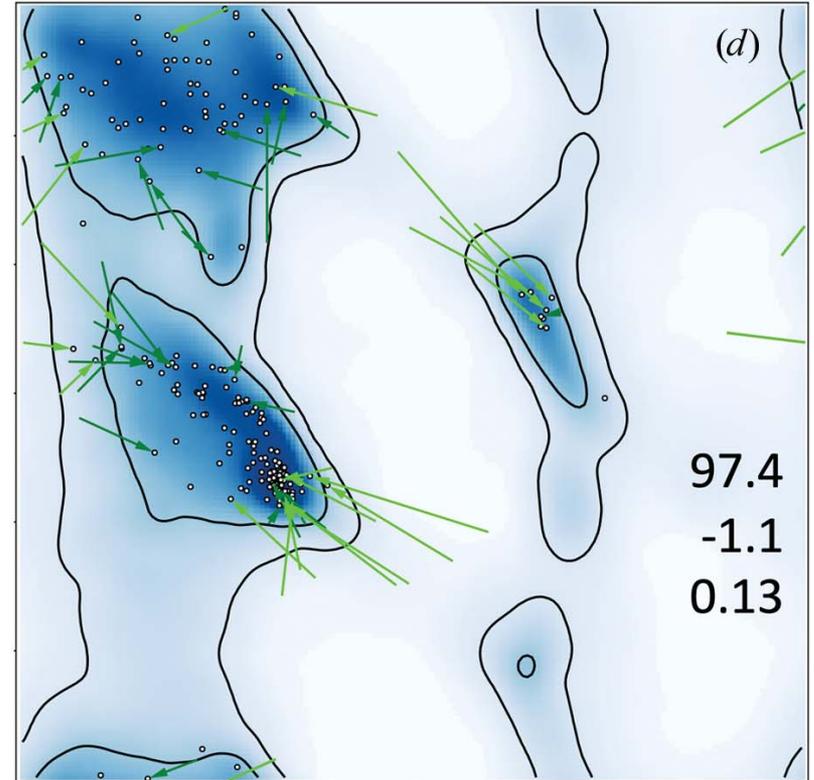
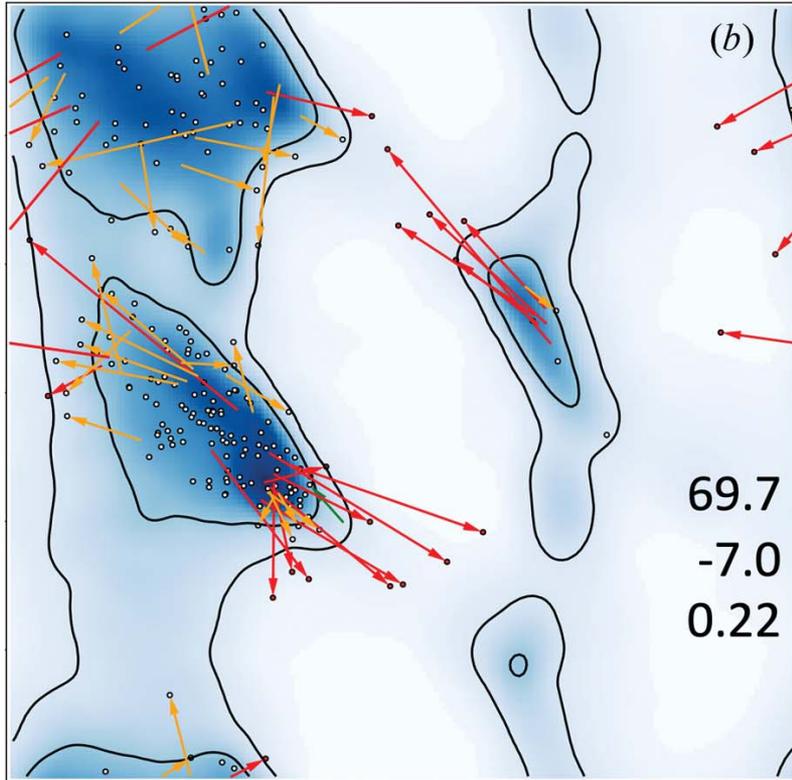
- RMSD from ideal: bonds = 0.01Å angles = 1.6°

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  GFVDLTLHDQVSMEHPVKLLFGKCVEGMVEIVYTFVLSSTLKSLE
Chain A  GFVDLTRHDQVSMEHPGKLLFGK--EGMVEIVYTF-----KSLE
Chain B  GFVDLTRHDQVSMEHPGKLLFGK--EGMVEIVYTFVVSSTLKSLE
Chain C  GFVDLTRHDQVSMEHPGKLLFGKKVEGMVEIVYTFVVSSTLKSLE
Chain D  GFVDLTRHDQVSMEHPGKLLFGKKVEGMVEIVYTFVLSSTLKSLE
***** ***** ***** *****
```

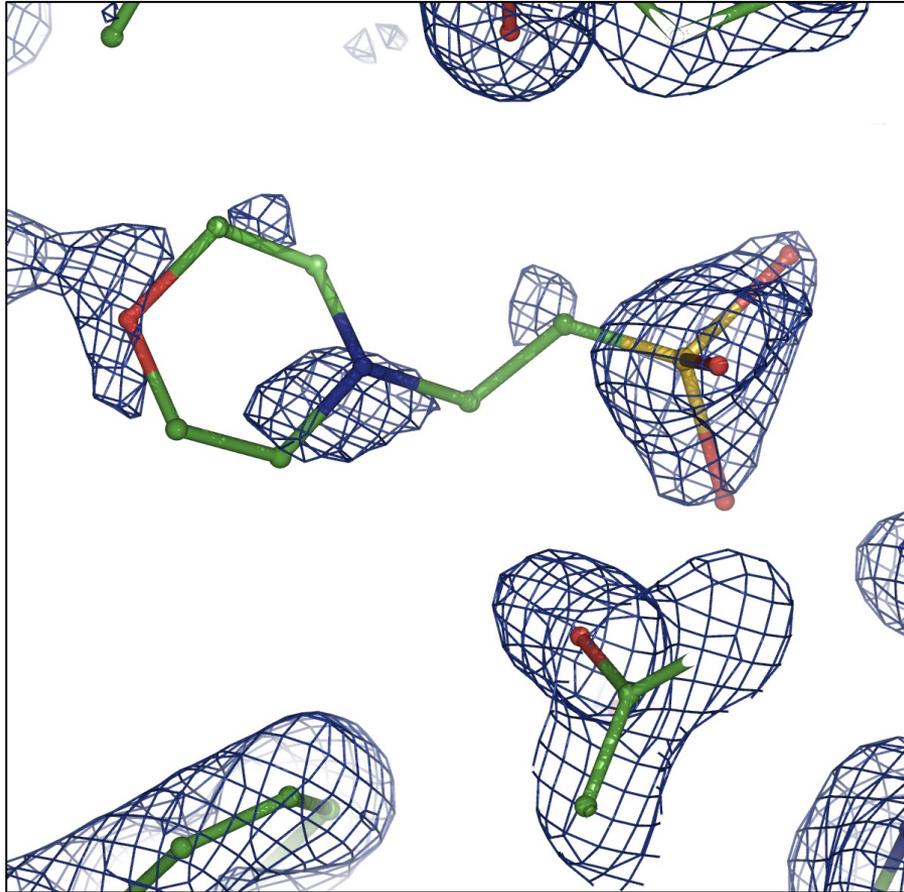
Comparama: phenix.comparama



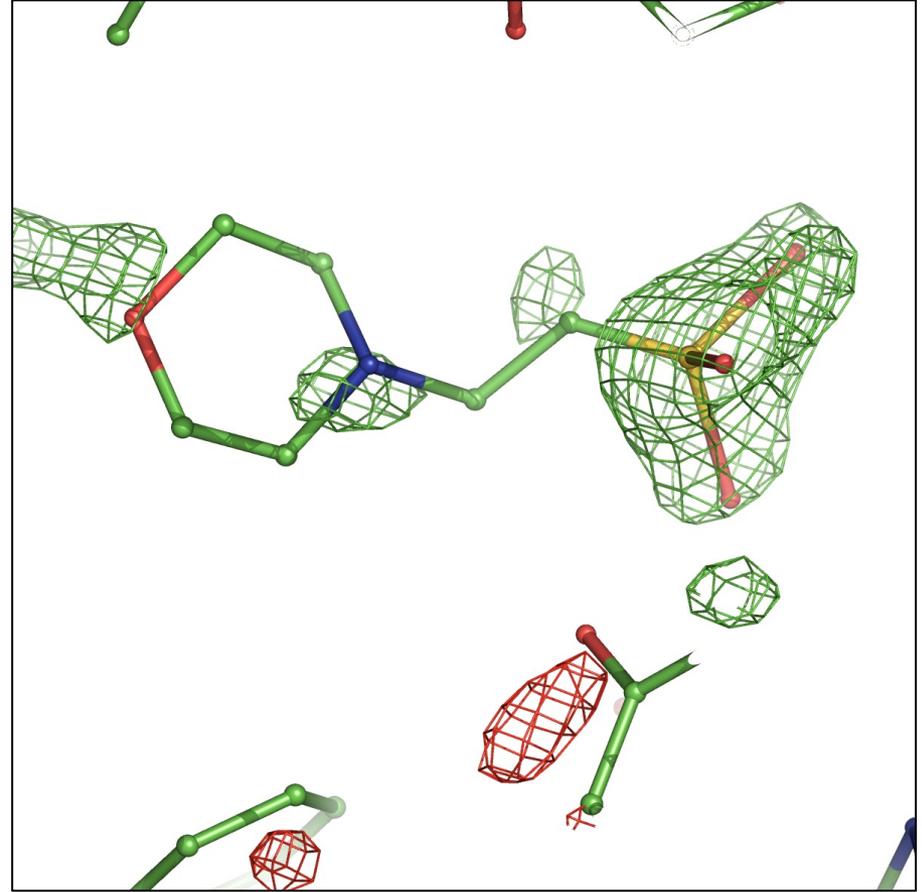
Ligands and Polder map

PDB code: 1ABA, Resolution: 1.45 Å

2mFo-DFc (1σ)



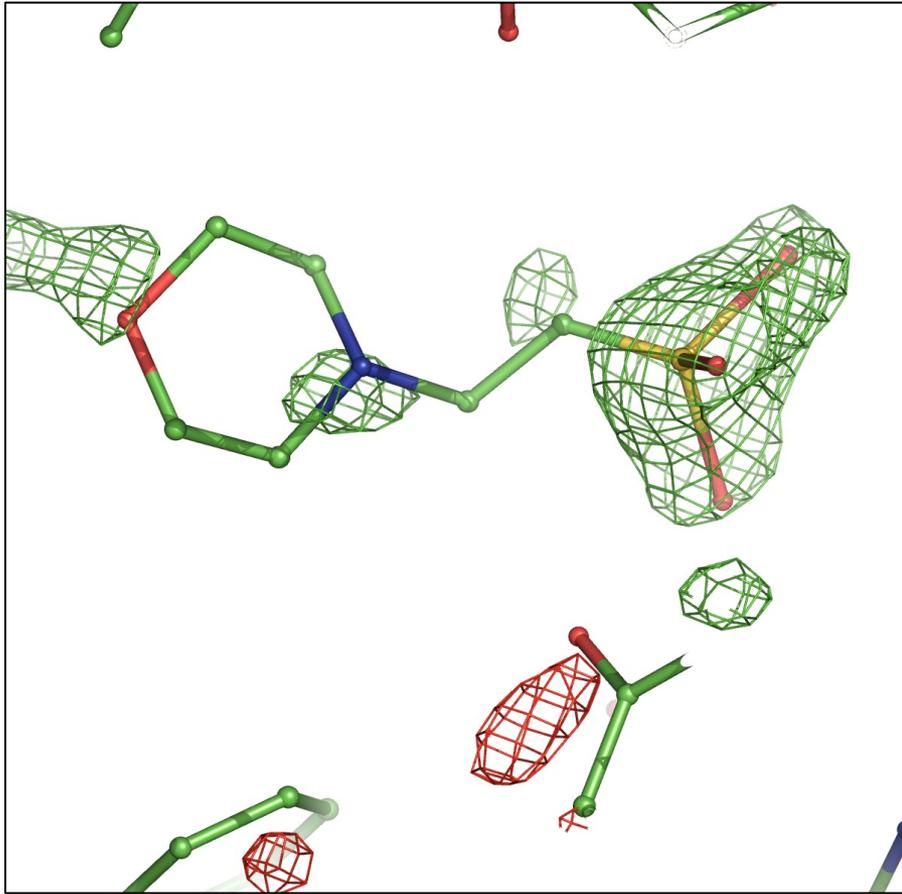
mFo-DFc ($\pm 3\sigma$)



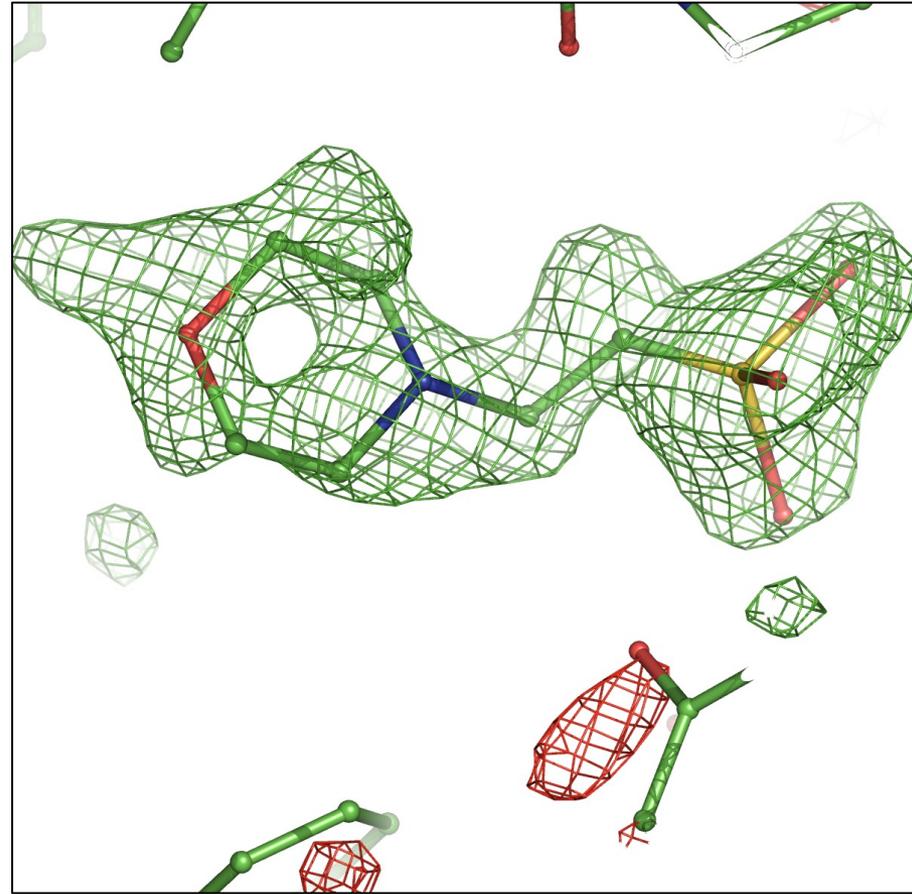
Ligands and Polder map

PDB code: 1ABA, Resolution: 1.45 Å

mFo-DFc ($\pm 3\sigma$)

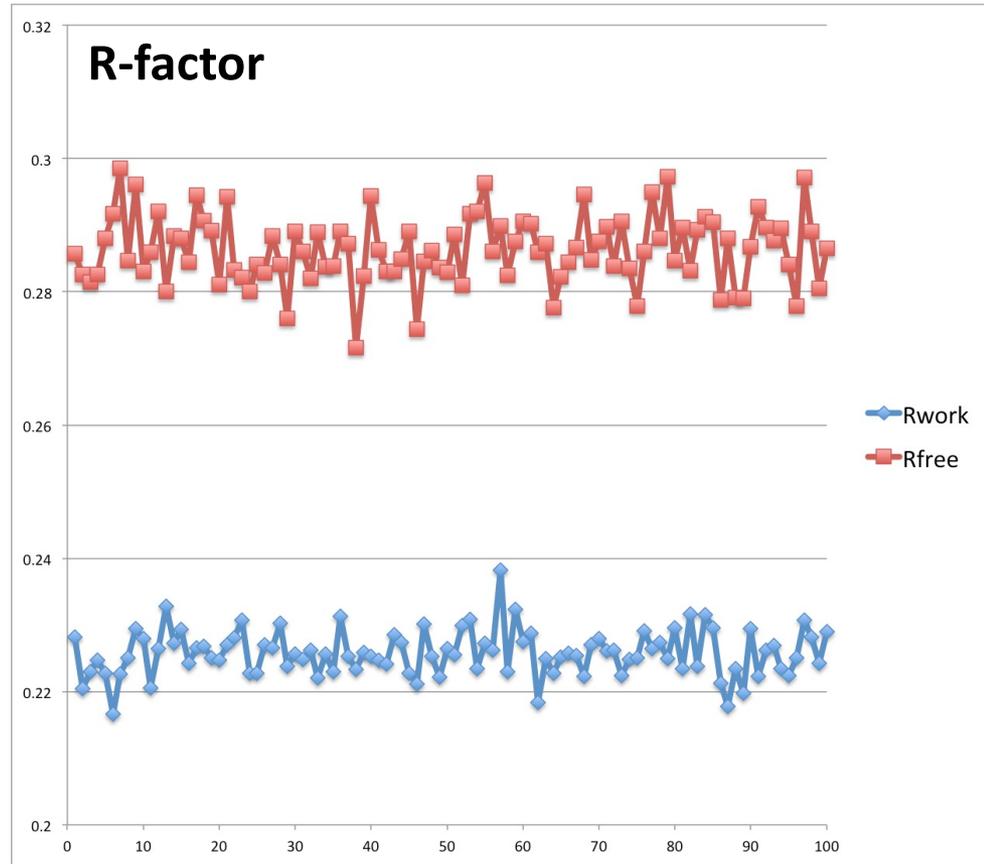
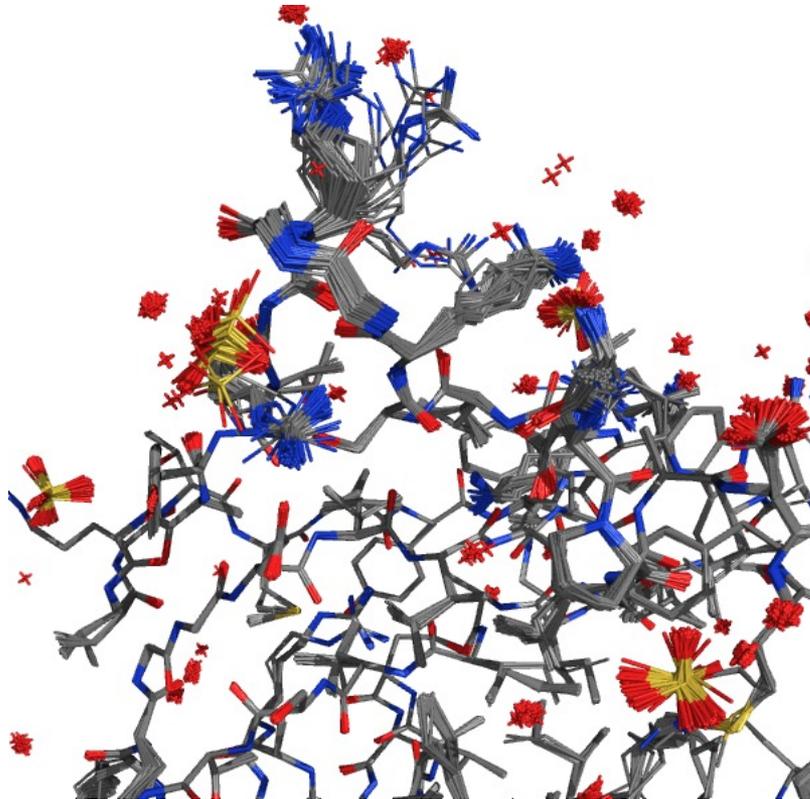


Polder mFo-DFc ($\pm 3\sigma$)



Estimating and using uncertainty

100 identical refinement runs each one starting with slightly perturbed model



Refinement run

PDB deposition

The screenshot displays the Phenix software interface. At the top, the title bar reads "Phenix home". Below it is a toolbar with icons for Quit, Preferences, Help, Citations, Reload last job, ChimeraX, Coot, PyMOL, KiNG, Tools, Help, and Server. The main window is divided into two panes. The left pane, titled "Projects", contains a table of project entries. The right pane shows a hierarchical menu of actions, with the "PDB Deposition" section highlighted in orange.

Projects

Show group: All groups [v] Manage...

Select Delete New project Import project Settings

ID	Last modified	# of jobs	R-free
✓ AF_POMGNT2_1	Jun 05 2024 11:46...	3	---
bugs	May 30 2024 02:38...	12	---
02_test_comma...	May 24 2024 01:20...	17	---
tests	May 22 2024 11:15...	67	0.2650
AF_bromodomai...	May 16 2024 10:37...	1	---
AF_7mjs_H_Pre...	Mar 19 2024 09:54...	1	---
groel_dock_refine	Mar 19 2024 09:28...	4	---
bugs_playground	Mar 07 2024 04:43...	13	---
fmodel	Feb 28 2024 02:44...	30	---
SEACOAST	Feb 13 2024 01:09...	7	---
AF_7mjs_H_Pre...	Jan 03 2024 10:19 ...	4	---
joint_XN	Nov 02 2023 03:49...	50	0.0989
AF_7mjs_H_Pre...	Apr 13 2023 02:18 ...	20	---
AF_7mjs_H_Pre...	Apr 13 2023 09:35 ...	0	---
AF_POMGNT2_0	Mar 31 2023 07:07...	3	---
AF_POMGNT2	Mar 30 2023 09:07...	6	---
7brm	Mar 17 2023 11:39...	25	---
7mjs_wcsbw	Mar 17 2023 09:31...	33	---
presentation	Mar 15 2023 02:00...	17	---
bughaton	Mar 06 2023 03:23...	8	---
-----	---	---	---

maps (create, manipulate, compare)

Enhanced maps (Polder, FEM, density-modified...)

Model building

Refinement

Ligands

Cryo-EM: Map analysis, symmetry, manipulation

Validation and map-based comparisons

Map improvement

Docking, model building and rebuilding

Refinement

Models: Superpose, search, compare, analyze symmetry

Modification, minimization and dynamics

PDB Deposition

- Prepare model for PDB deposition**
Finalize mmCIF files for deposition to the PDB
- Get PDB validation report**
Retrieve a validation report from the PDB
- Generate "Table 1" for journal**
Extraction of final model statistics for publication

Program search

Current directory: /Users/dcliebschner/Documents/AF_POMGNT2_1 Browse... 🔍

Phenix version 1.21.1-5286-000 Project: AF_POMGNT2_1

PDB deposition

mmCIF format is mandatory for deposition as of 2019



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