

Validation

Pavel Afonine

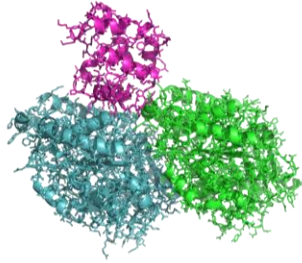
Phenix team

Lawrence Berkeley National Lab, California, USA

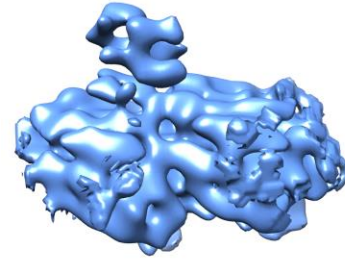
**May, 2025
MCCS, Madrid**

Validation

Model

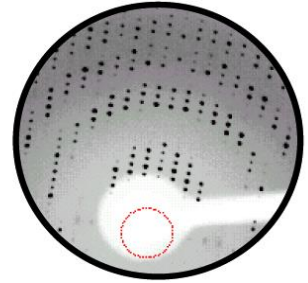


Data



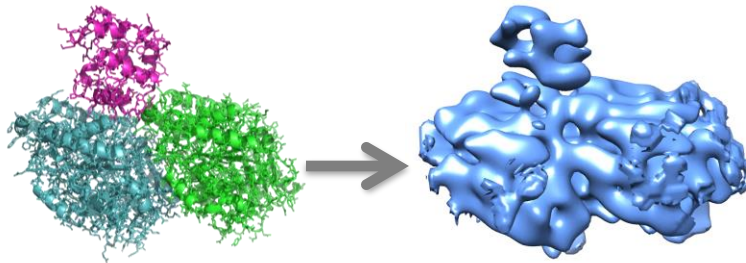
Cryo-EM

or



Diffraction

Model to data fit

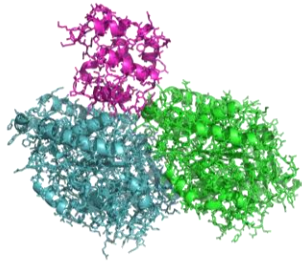


Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations

Validation tools: Crystallography vs Cryo-EM

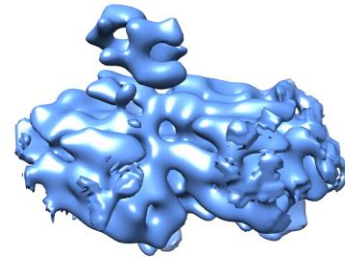
Exact same

Model



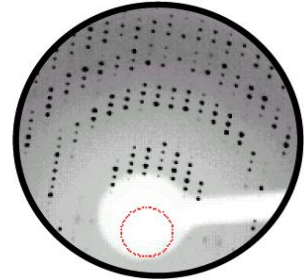
Different

Data



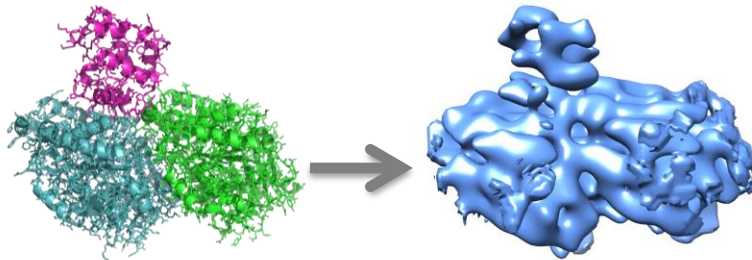
Cryo-EM

or



Diffraction

Model to data fit



Similar

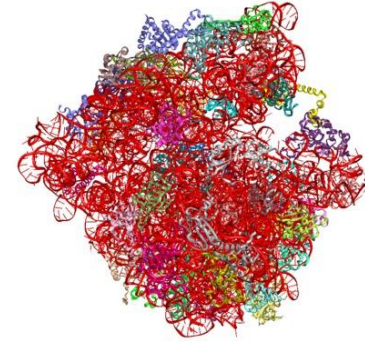
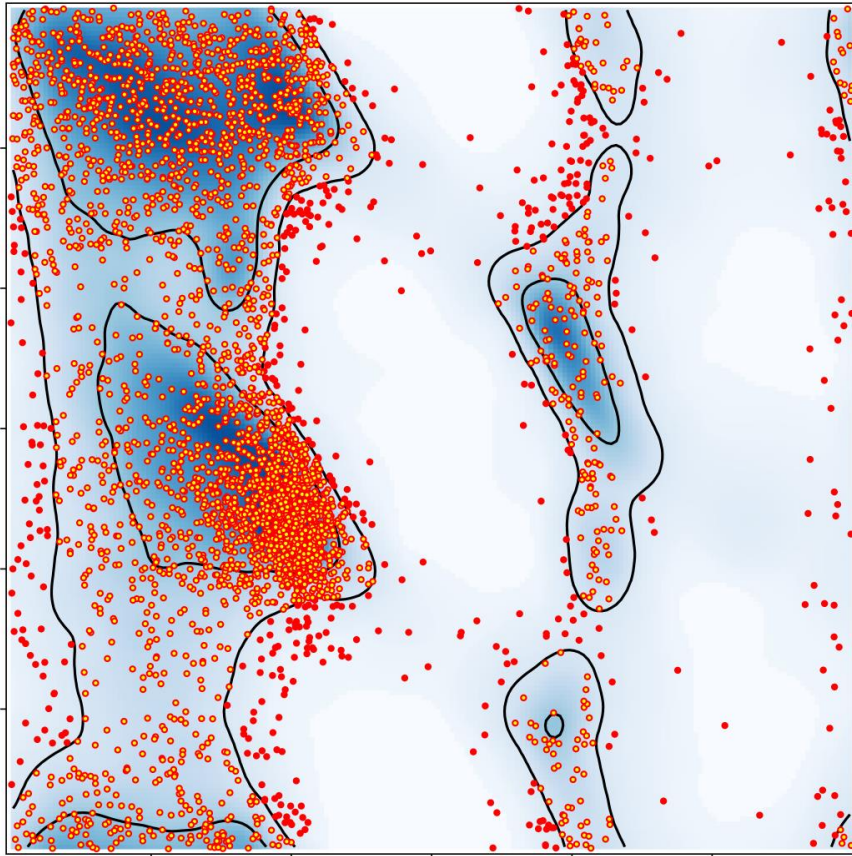
Validation: *why to do?*

- 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)!

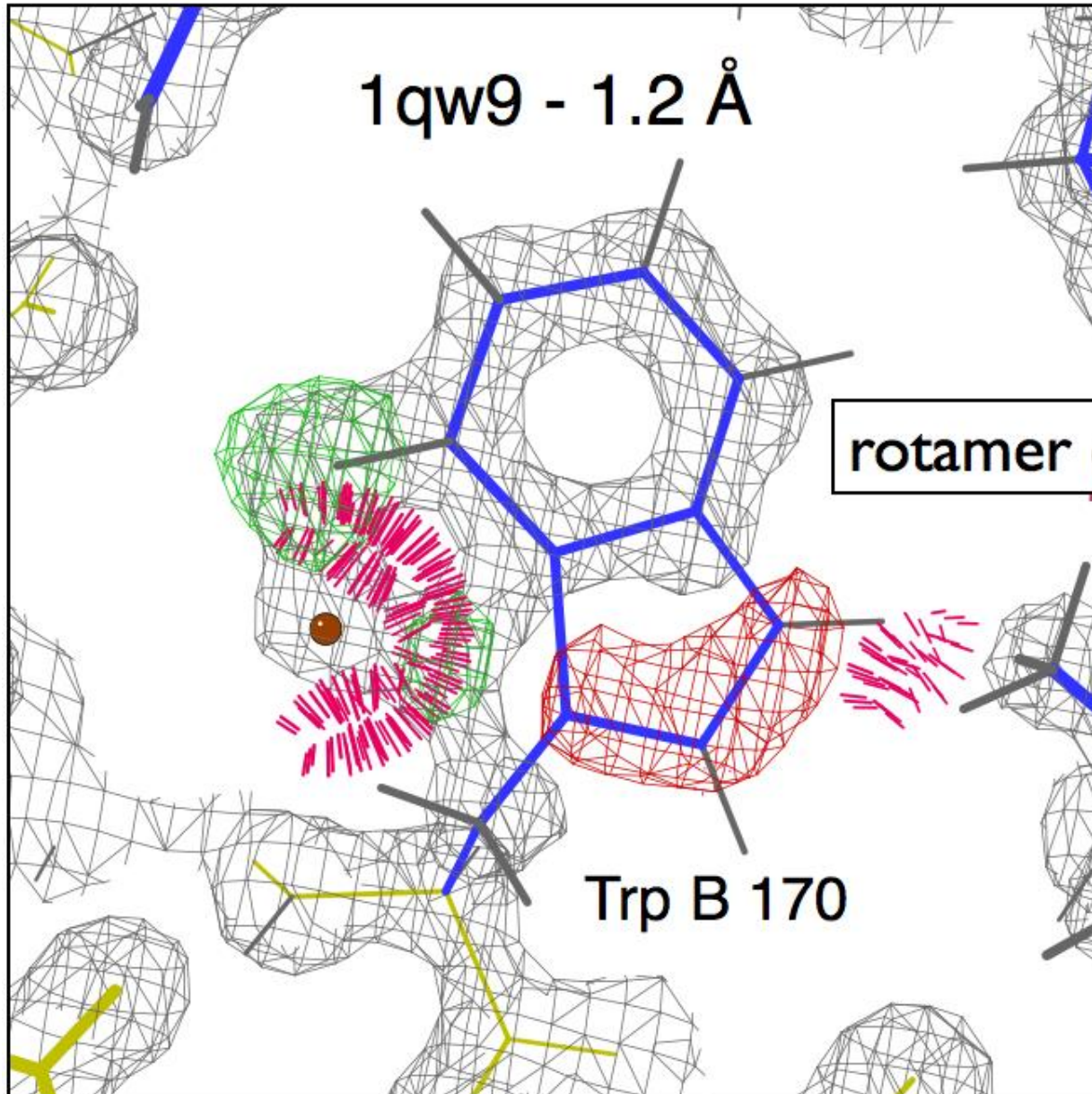
Validation: **why to do?**

(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





PHENIX home

 Quit  Preferences  Help  Citations  Coot  PyMOL  KING  Other tools  Ask for help

Actions Job history

Projects


Show group: All groups  Manage...


 Select  Delete  New project  Settings


ID	Last modified	# of jobs	R-free
✓ ChrisF	Apr 13 2020 09:42...	28	0.1944
real-space-refin...	Apr 03 2020 07:42...	2	---
zzz1	Mar 21 2020 09:10...	1	---
chris	Mar 12 2020 12:27...	11	0.1890
dan	Mar 11 2020 05:44...	1	---
3j63	Mar 11 2020 02:28...	1	---
jason	Mar 11 2020 11:36...	1	---
rt6	Mar 11 2020 10:31...	1	0.2459
mate	Mar 10 2020 01:36...	1	---
emily	Mar 09 2020 03:52...	3	---
—	Mar 05 2020 08:25...	3	0.1923
alex	Feb 27 2020 11:33...	6	---
rt20201	Feb 18 2020 12:50...	4	0.2213
1f8t	Feb 03 2020 09:00...	1	0.1977
real-space-refin...	Jan 30 2020 02:38...	2	---
real-space-refin...	Jan 29 2020 10:56...	1	---
ion_channel_den...	Jan 27 2020 07:36...	3	---
10101	Jan 27 2020 12:38...	2	---
demos	Jan 27 2020 10:57...	3	---
ion_channel_den...	Jan 27 2020 10:03...	2	---
malcolm	Jan 22 2020 10:22...	14	0.1748
real-space-refin...	Jan 16 2020 04:28...	3	---
3NIR	Dec 05 2019 10:2...	1	---
leighton	Sep 02 2019 05:1...	2	---
5pti	Aug 27 2019 03:4...	3	---

Favorites

Data analysis

 **Xtriage**
Analysis of data quality and crystal defects

 **Merging statistics**
Calculates a variety of statistics for unmerged intensities, including I/sigma, R-merge, R-meas, and CC1/2.

 **Mtriage**
Analyze quality of maps in CCP4 format

Experimental phasing


Molecular replacement


Model building


Refinement


Cryo-EM


Validation

 **Comprehensive validation (X-ray/Neutron)**
Model quality assessment, including real-space correlation and geometry inspection using MolProbity tools



 **Comprehensive validation (cryo-EM)**
Model quality assessment, including real-space correlation, for cryo-EM structures

 **Structure comparison**
Identify differences between multiple structures of the same protein, using multiple criteria

 **Calculate CC***
Comparison of unmerged data quality with refined model, as described in Karplus & Diederichs (2012)

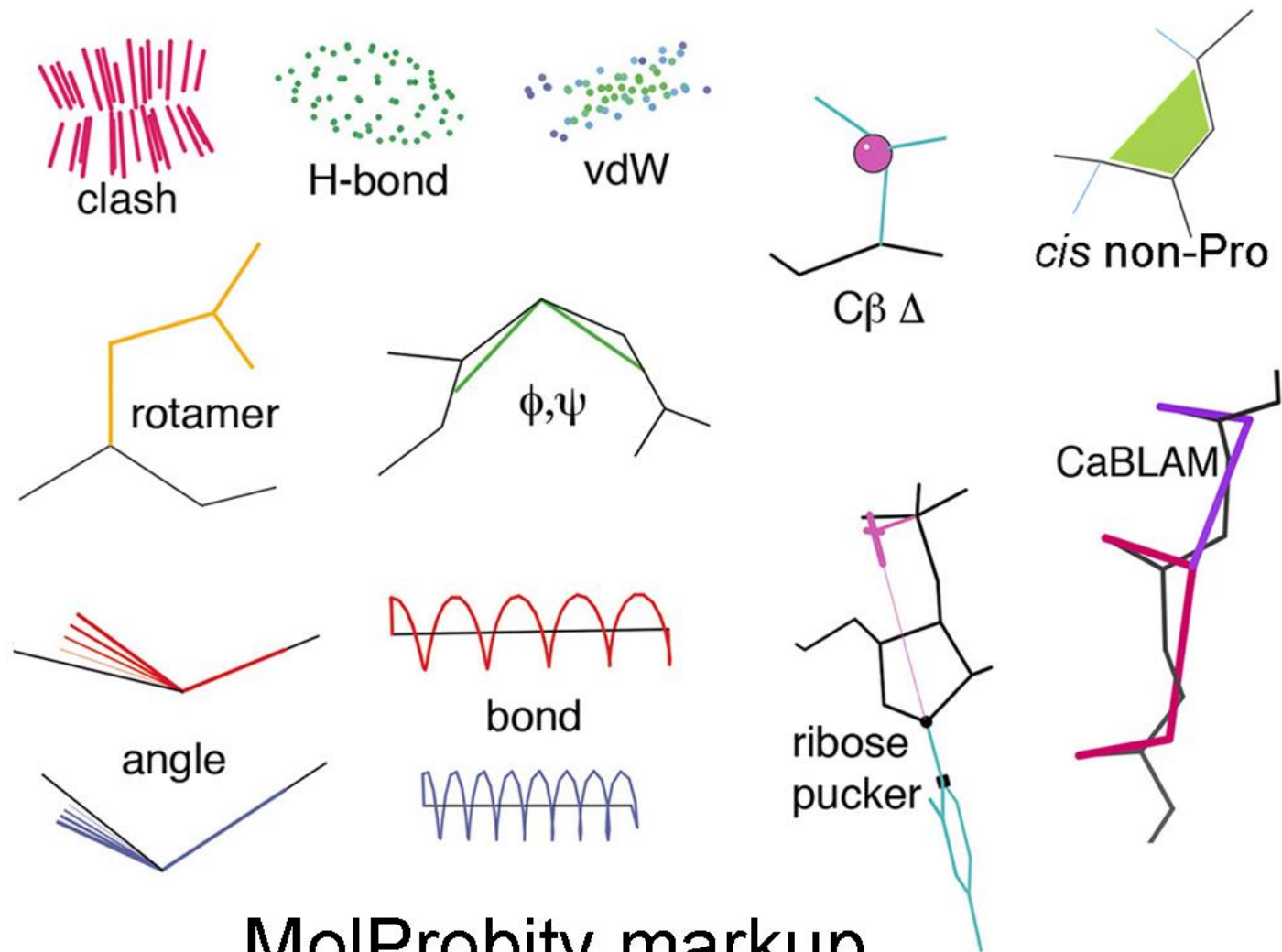
 **EMRinger**
Model validation for de novo electron microscopy structures

Ligands

Current directory: /Users/pafonine/Desktop/all/people/ChrisF  Browse... 

PHENIX version dev-svn-000 Project: ChrisF

Model validation



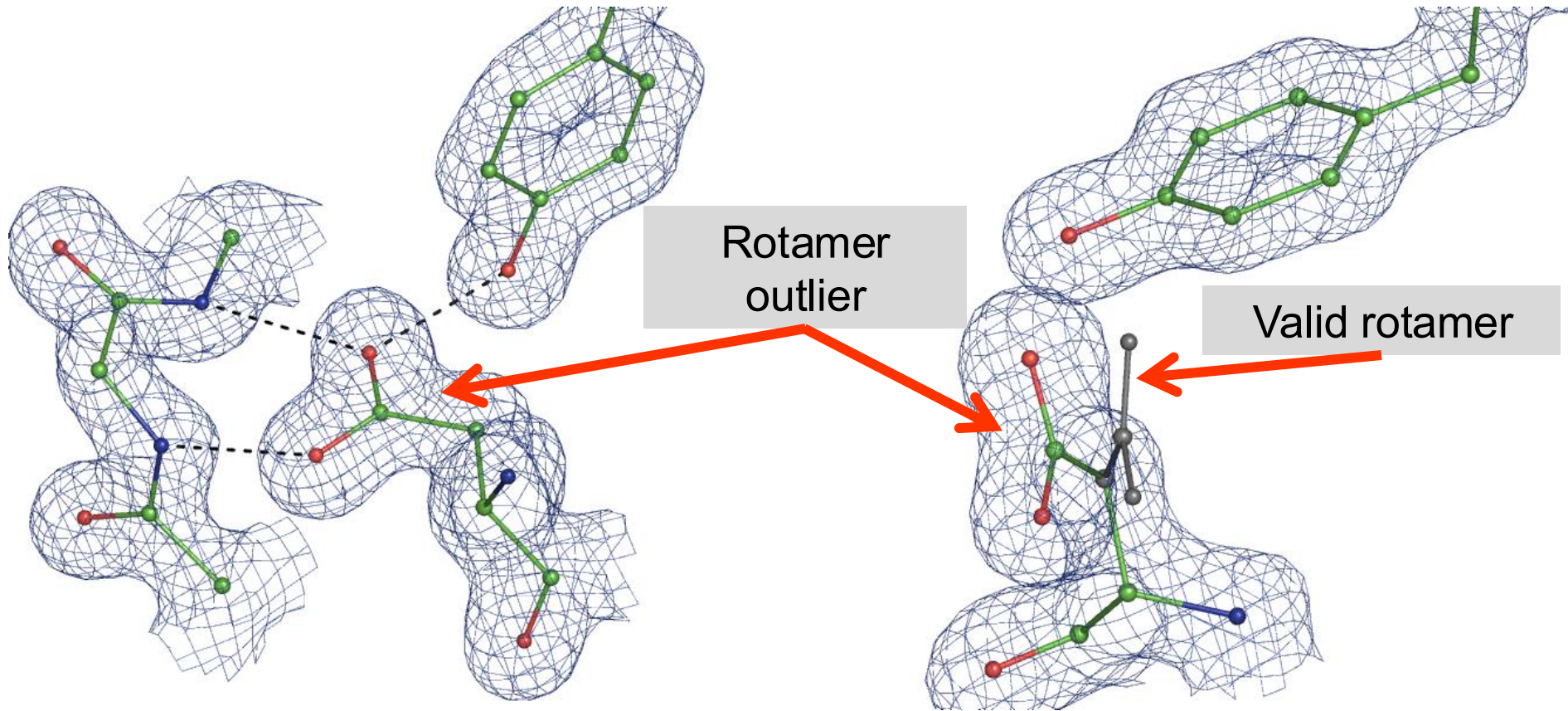
MolProbity markup

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 $\langle B_i - B_j \rangle$)
 -

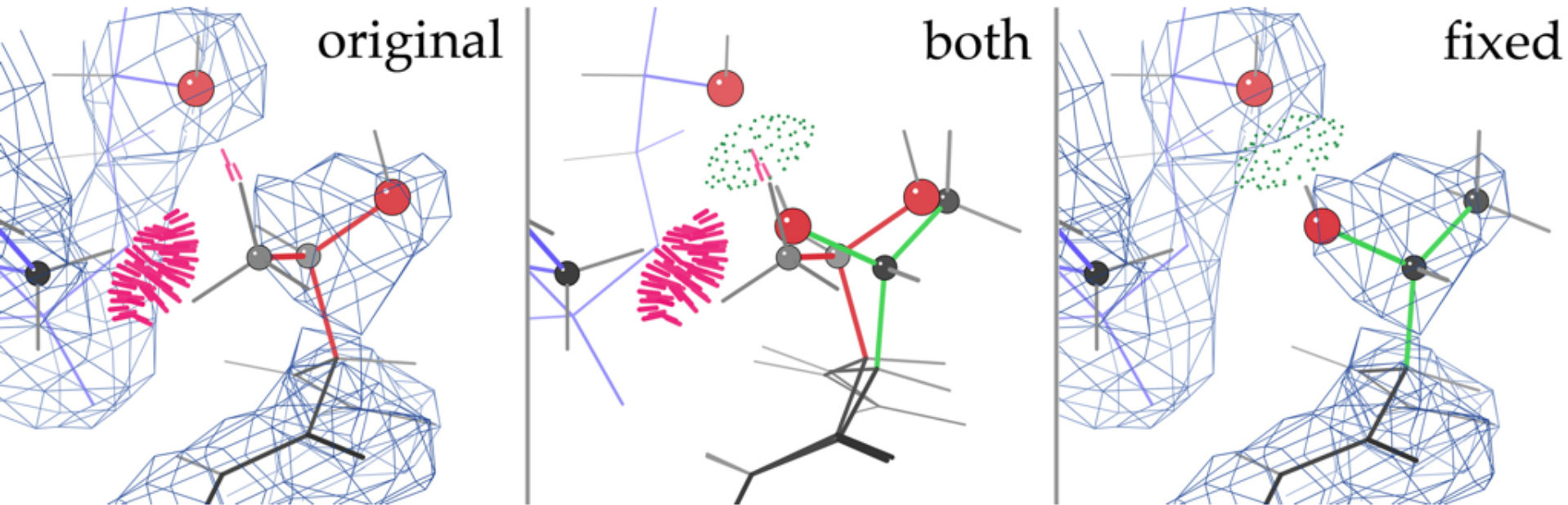
Model validation: amino-acid side-chain rotamers

- An outlier \neq wrong
 - However, each outlier has to be explained



Model validation: amino-acid side-chain rotamers

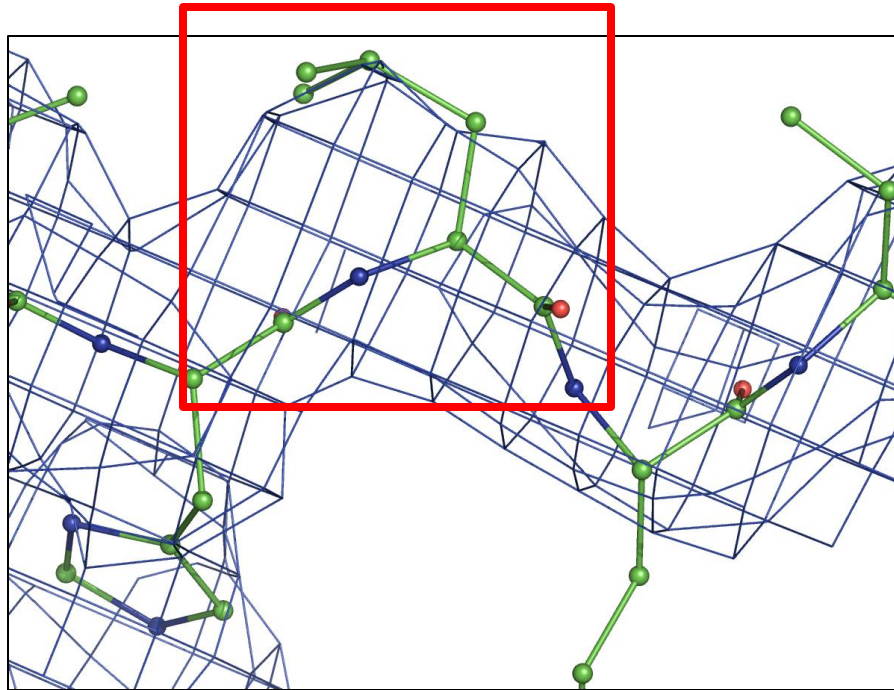
Thr O 3 from 1YHQ



Model validation: amino-acid side-chain rotamers

- Low-resolution maps

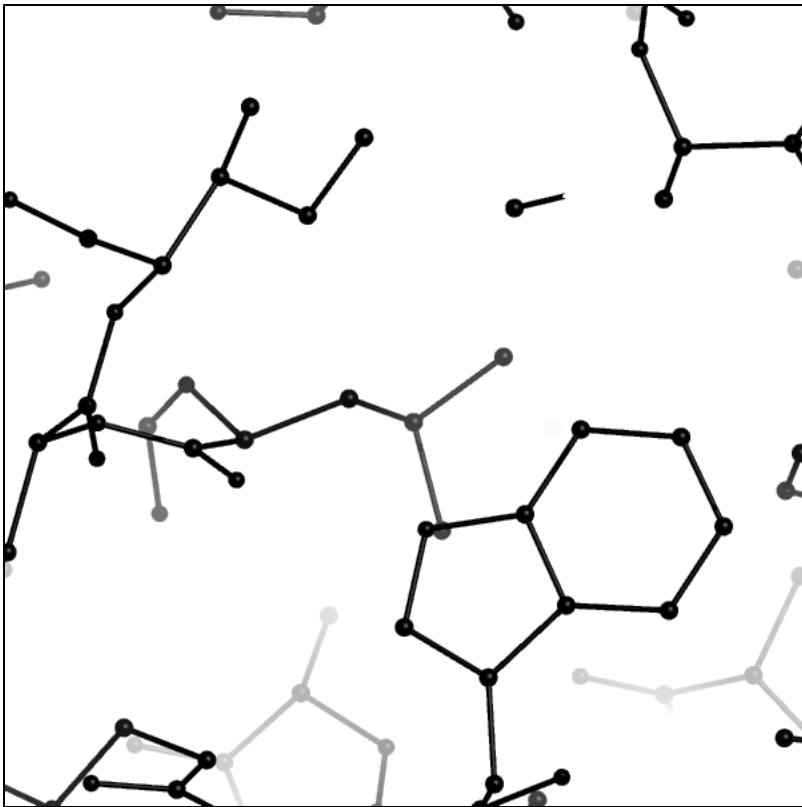
Side chain lacking density may be forced into main chain density and become a rotamer outlier



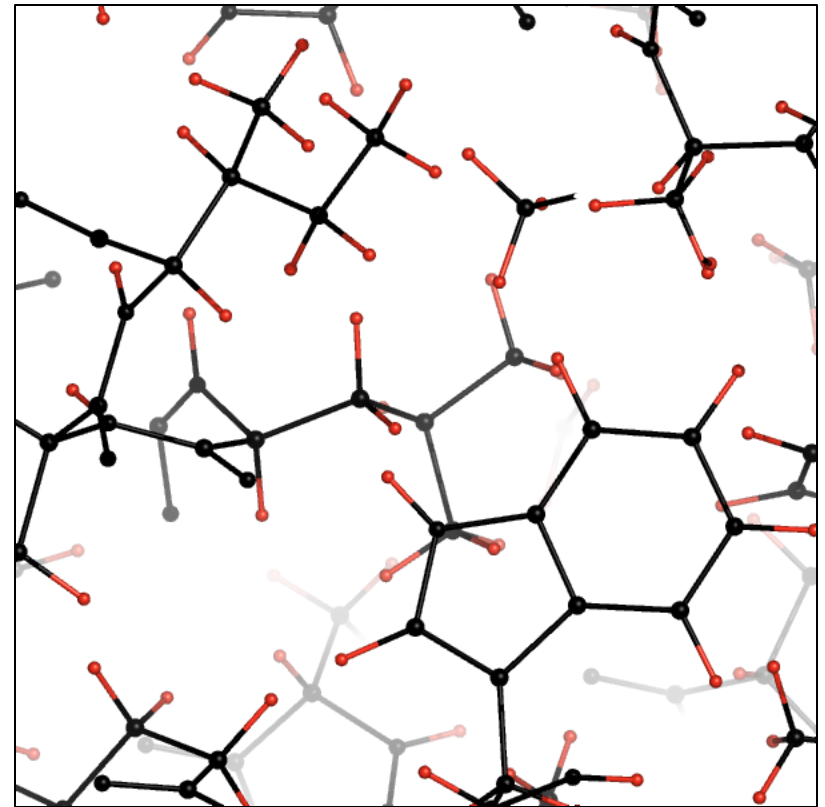
Phenix refinement (real- and reciprocal-space) use rotamer-specific restraints on torsion chi-angles

Model validation: clashes

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



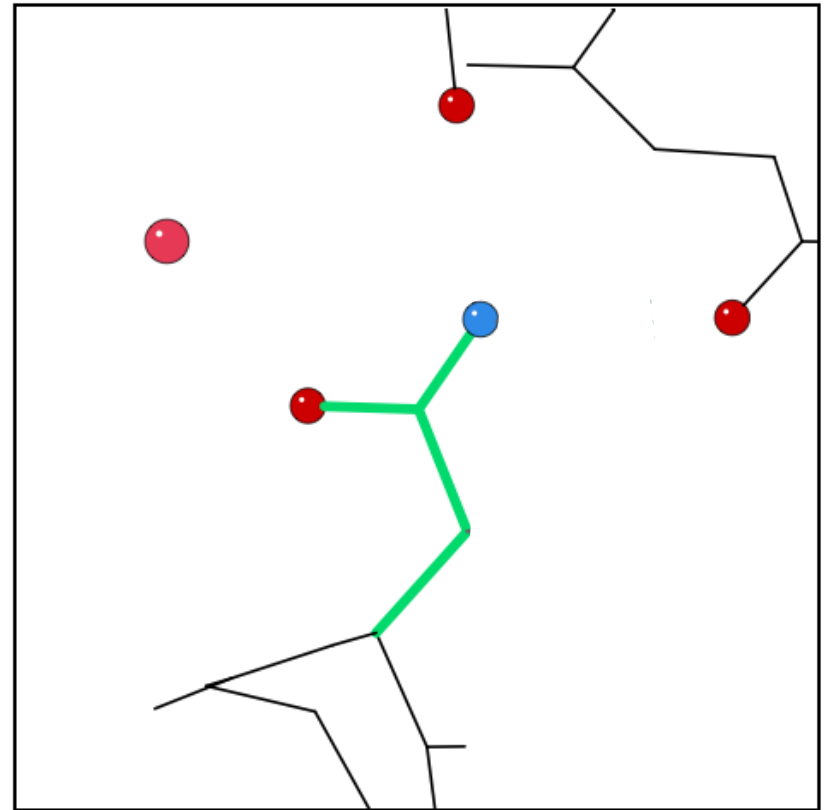
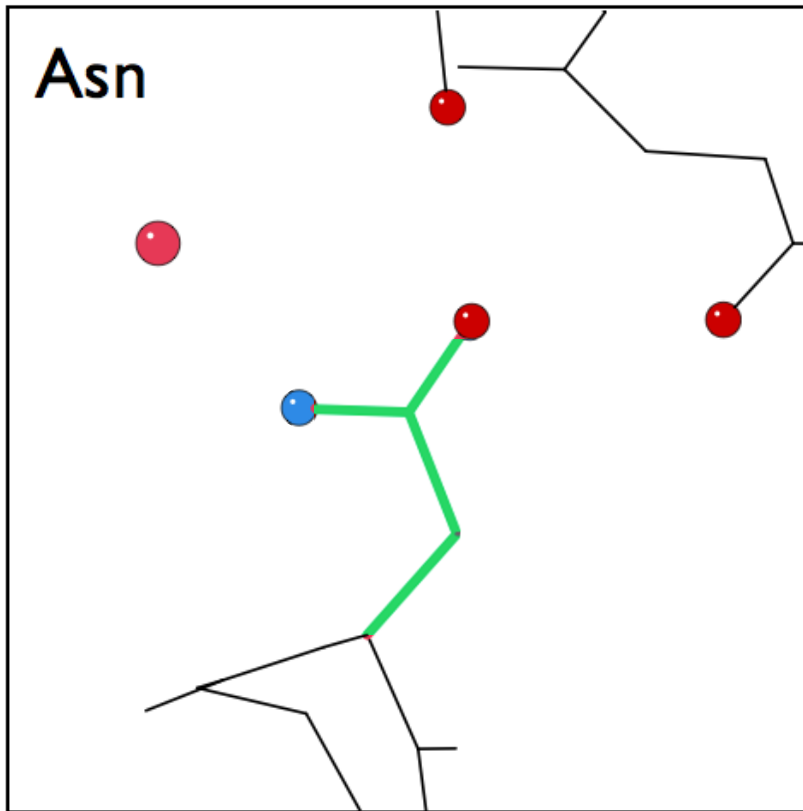
No H atoms



H atoms added

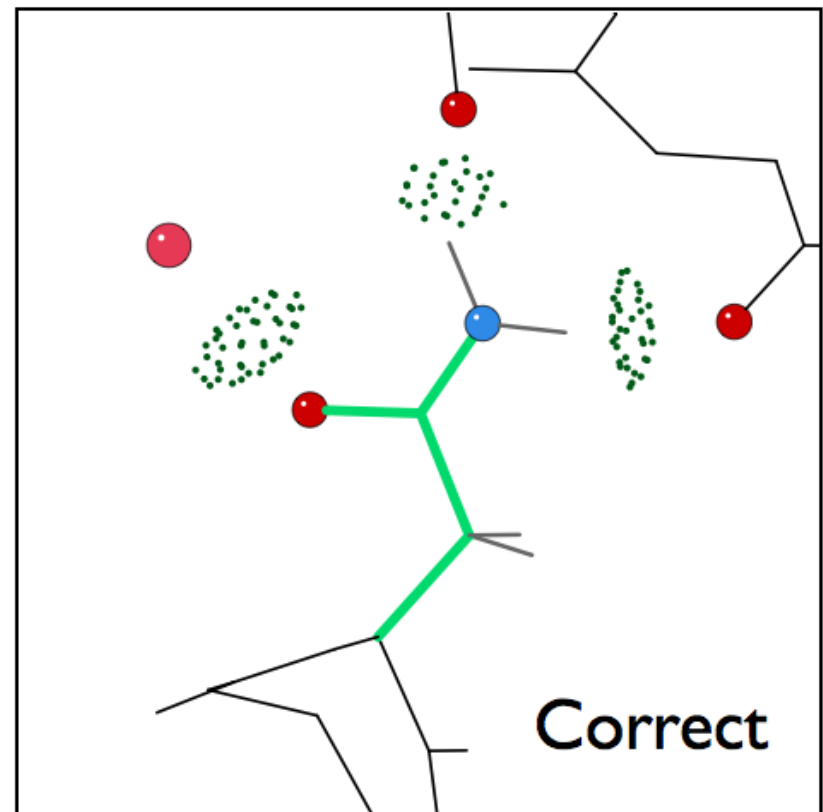
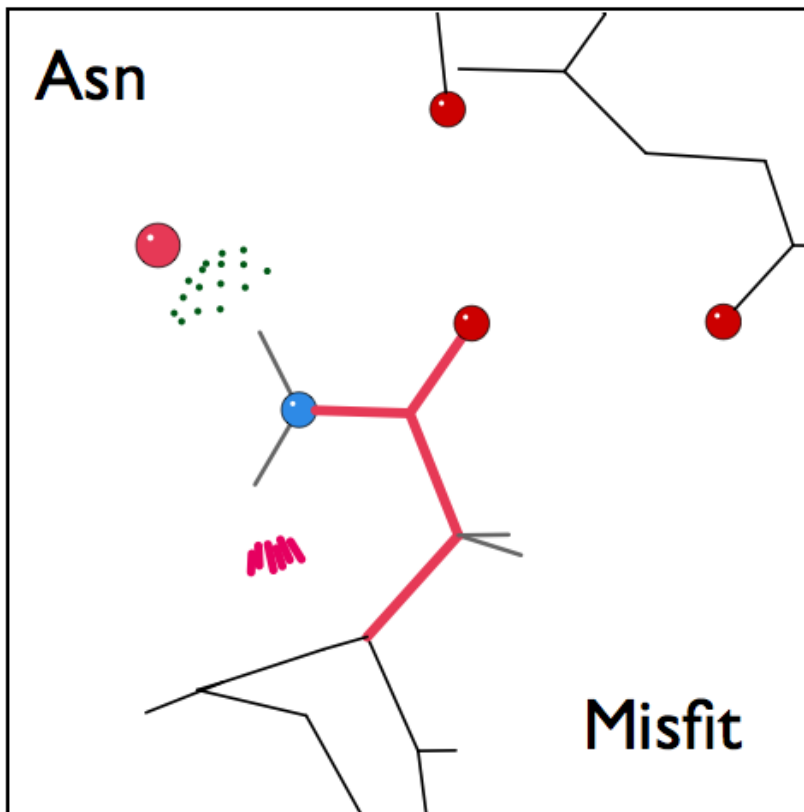
Model validation: clashes

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



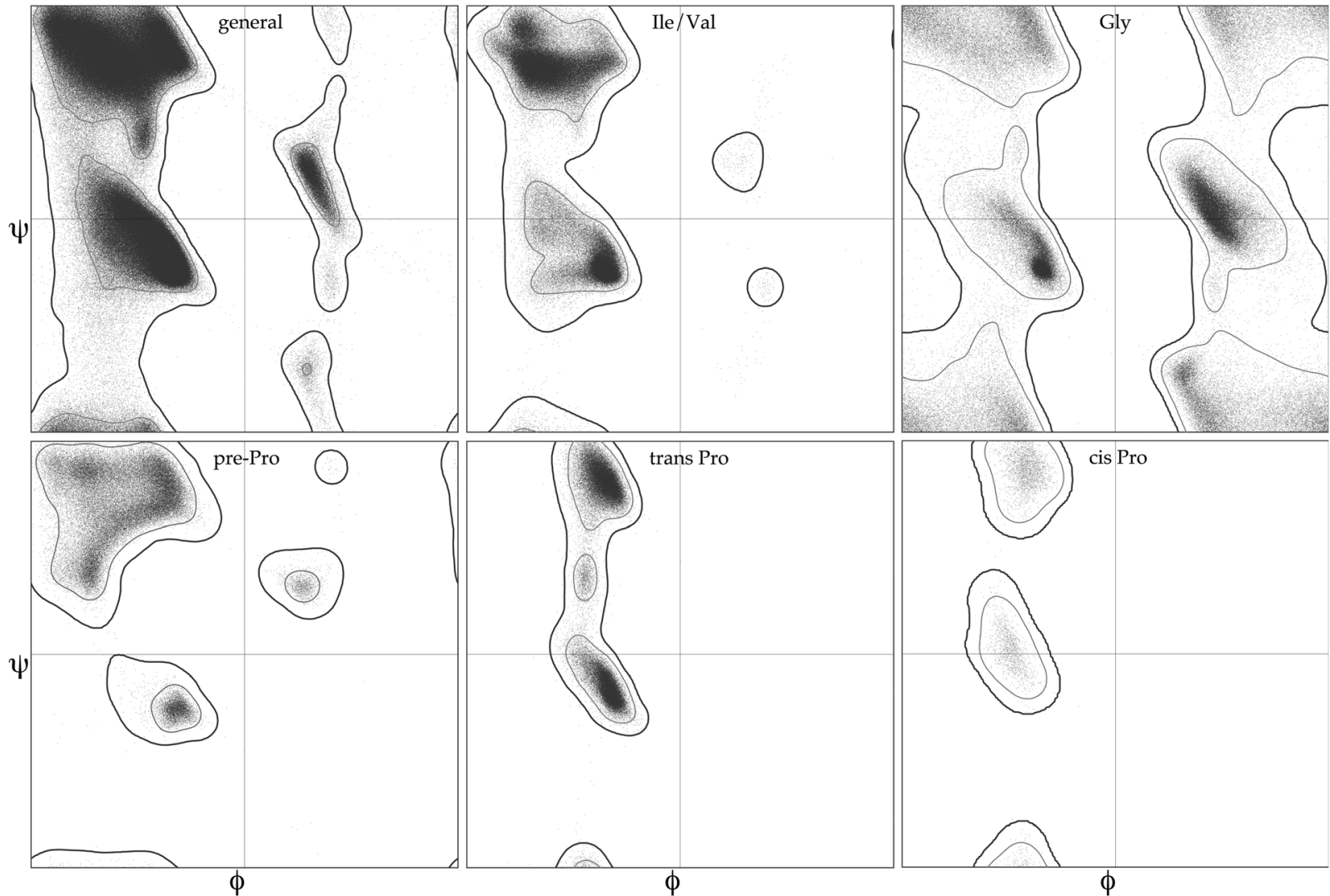
Model validation: clashes

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

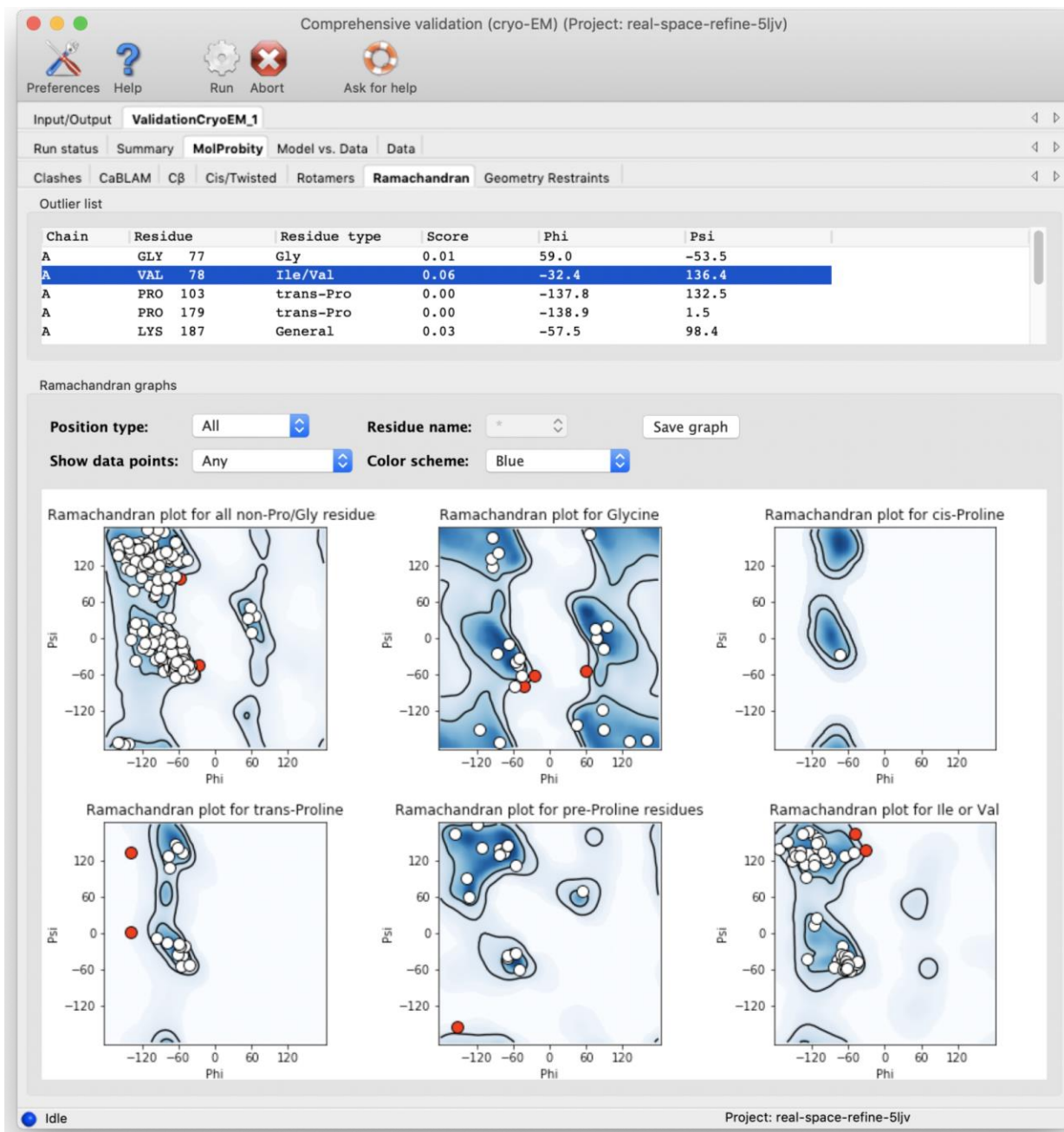


Model validation: Ramachandran plot

- Different plots for different classes of residues



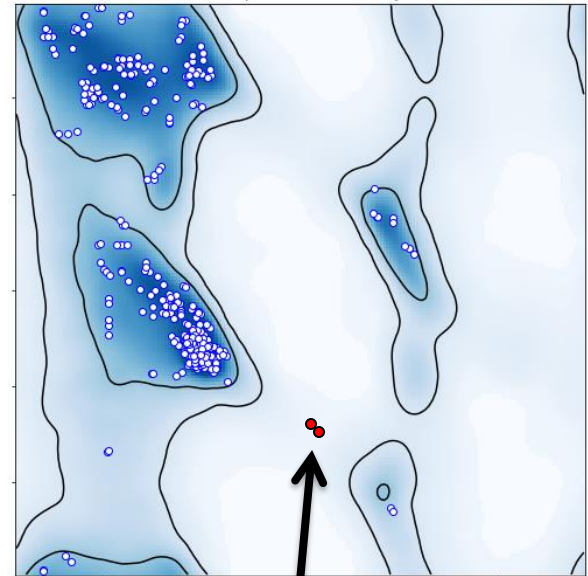
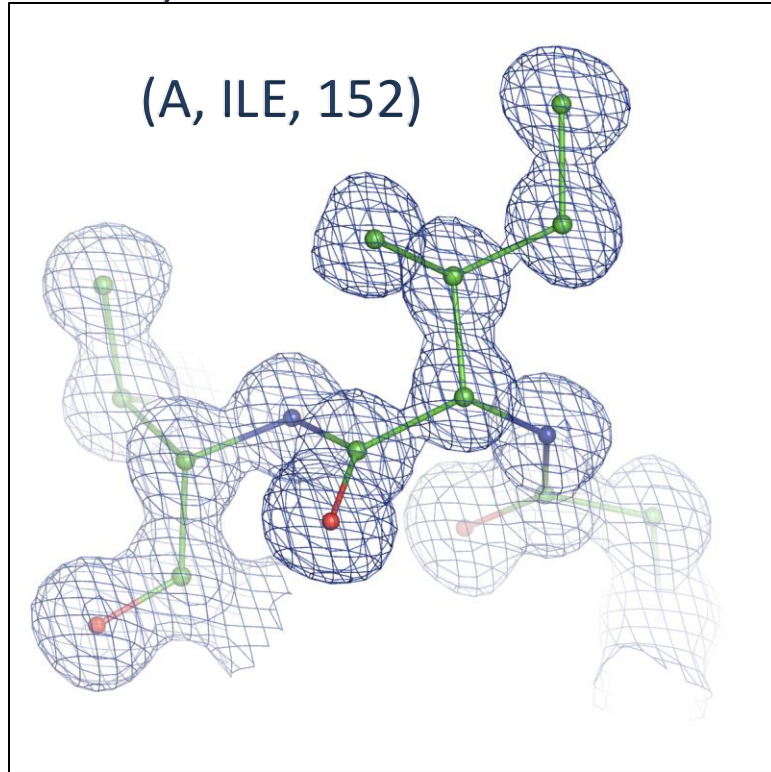
Model validation: Ramachandran plot



Model validation: Ramachandran plot

- A Ramachandran plot outlier \neq wrong

3NOQ, 1 Å



Outliers:

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

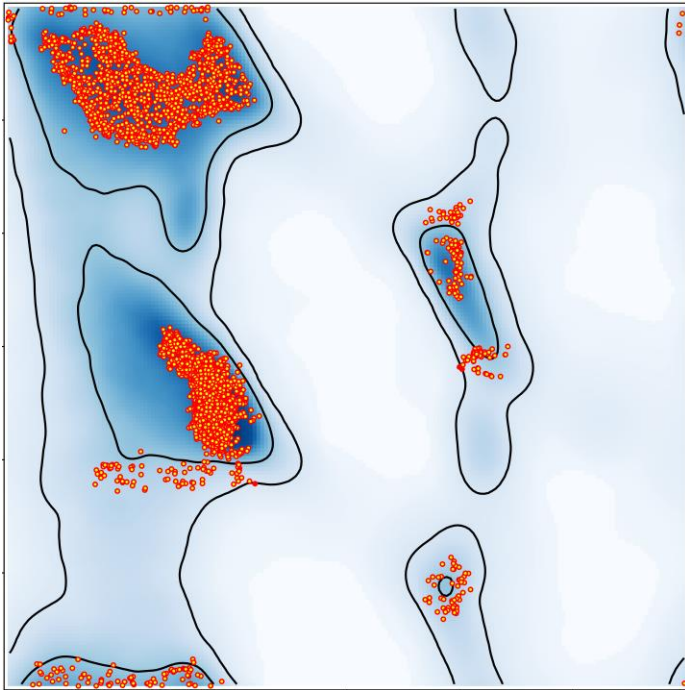
- All outliers need to be explained (supported by the data)

Validation and Refinement **"conflict"**

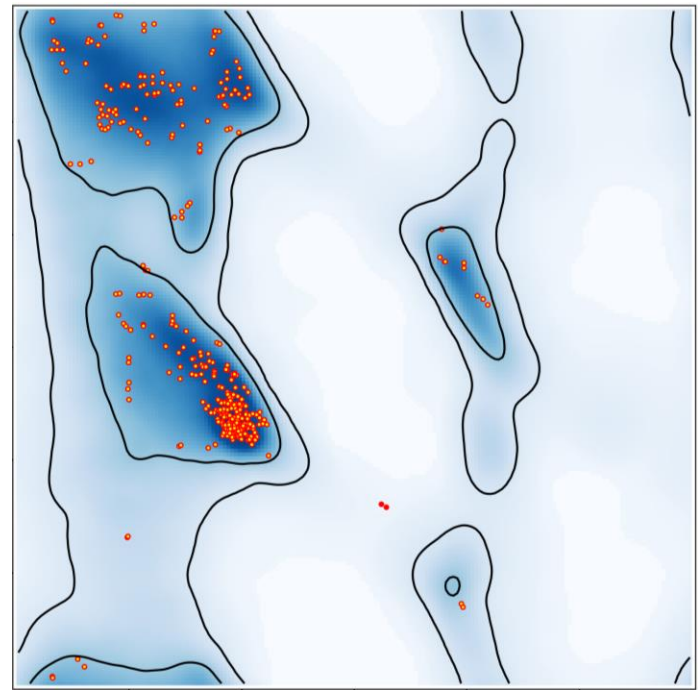
- Validation metrics progressively become refinement goals
 - Ramachandran plot restraints
 - C β deviation restraints
 - Secondary structure restraints
 - Restraints on χ angles of amino-acid side-chain rotamers
- As result, validation becomes less capable of catching problems

Example

Q: How we know the plot looks wrong?



A: Because we know how good plot looks like!



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

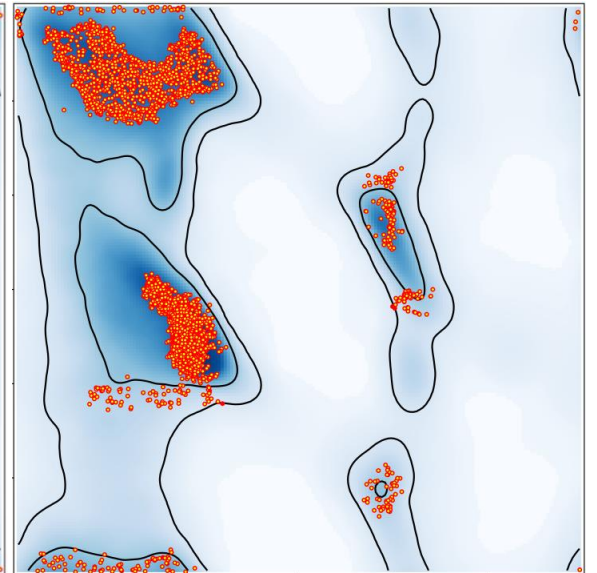
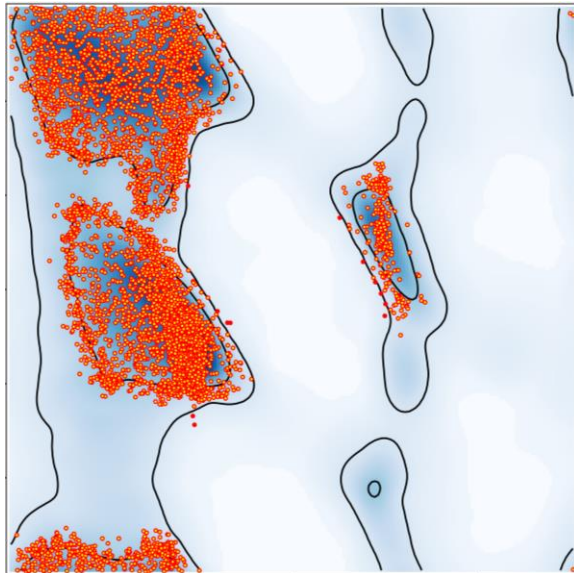
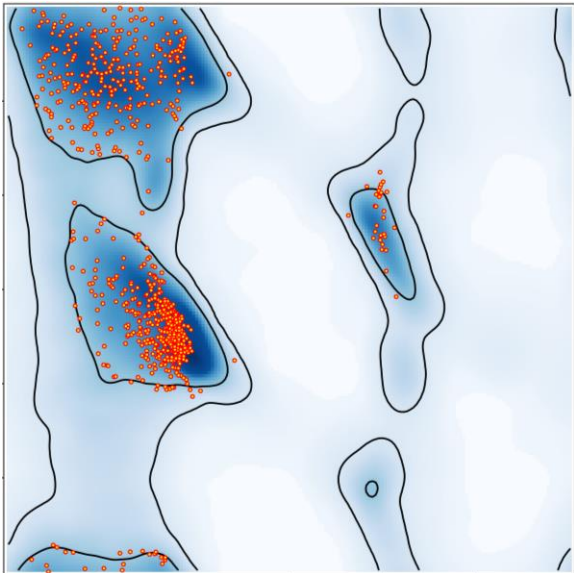
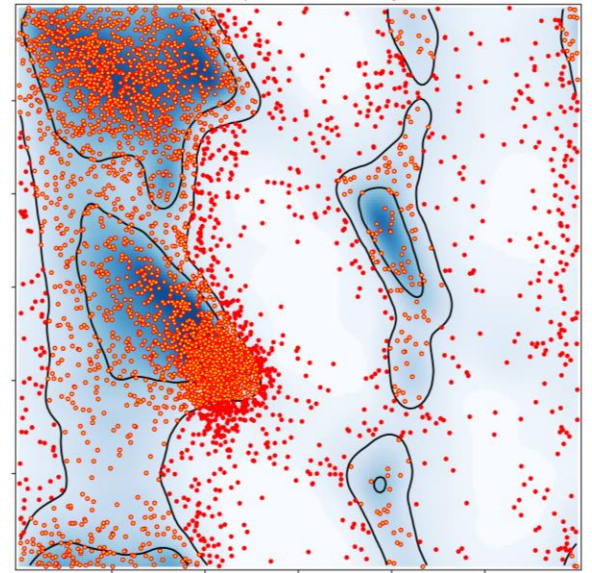
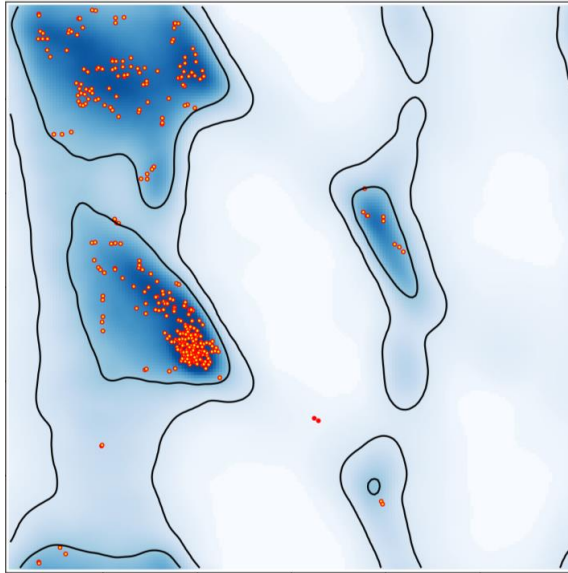
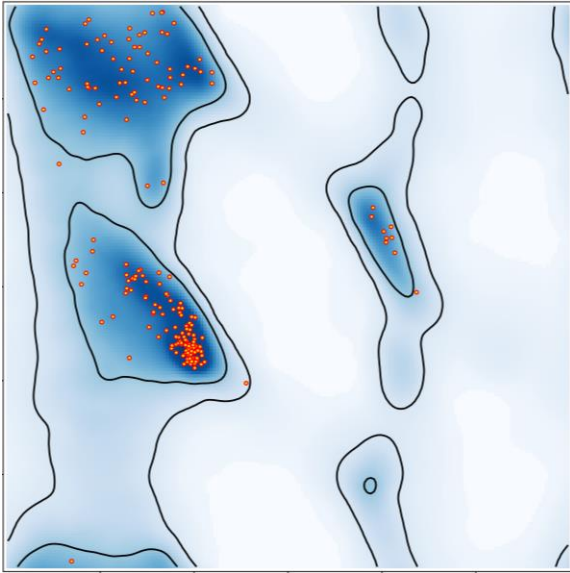
 **CellPress**

Resource

A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry

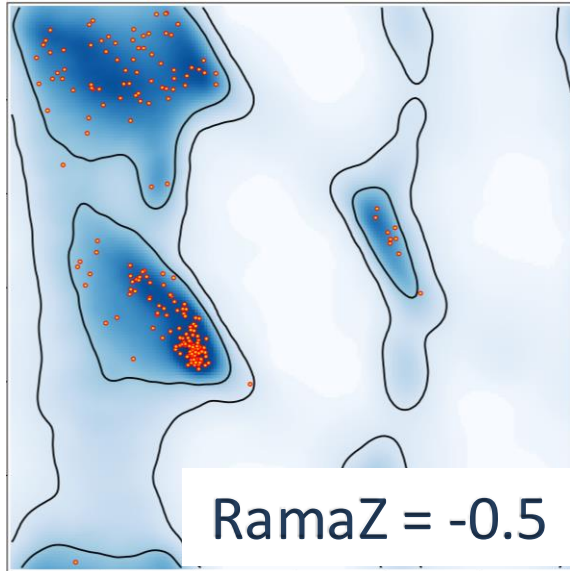
Oleg V. Sobolev,^{1,5,*} Pavel V. Afonine,¹ Nigel W. Moriarty,¹ Maarten L. Hekkelman,^{2,3} Robbie P. Joosten,^{2,3,*} Anastassis Perrakis,^{2,3} and Paul D. Adams^{1,4}

How you can tell good vs bad plot?

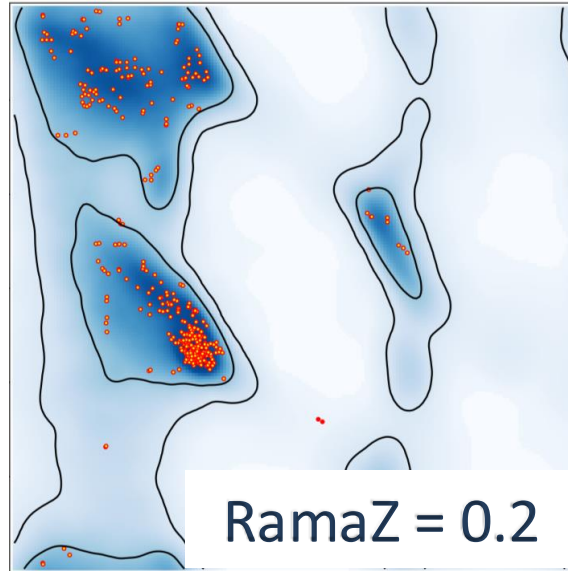


Model validation: *Ramachandran plot Z-score*

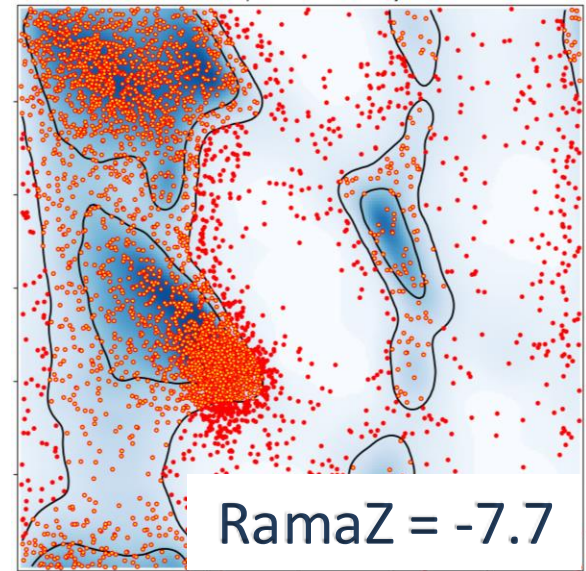
Good



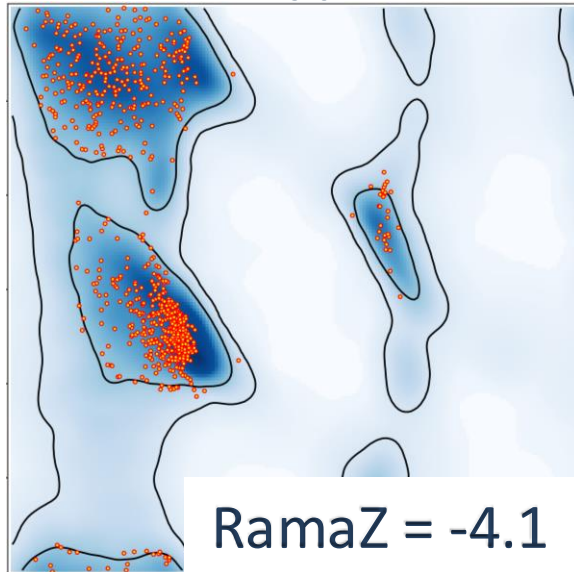
Good



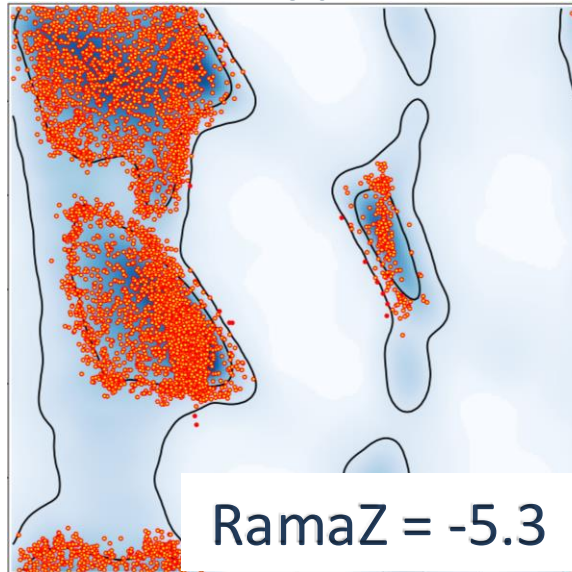
Bad



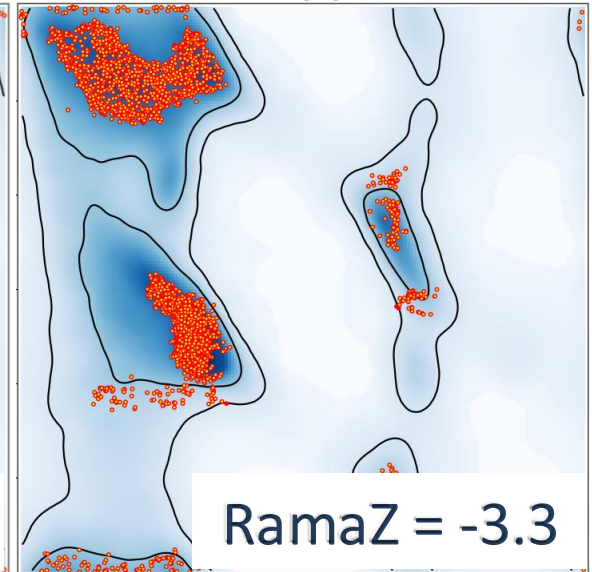
Bad



Bad

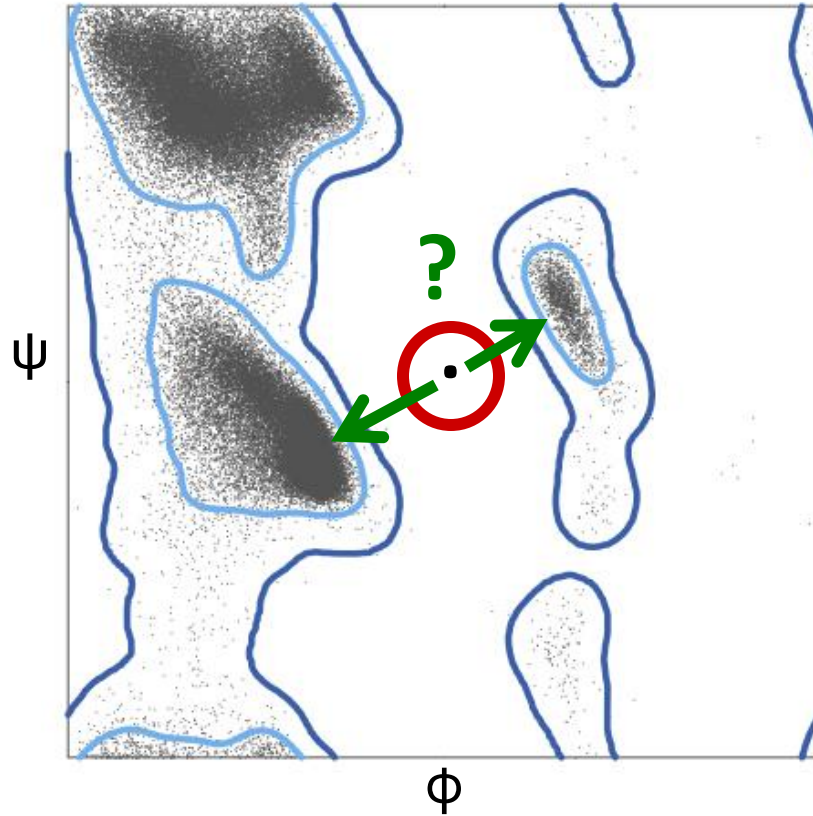


Bad



How did that happen?

$$E = \sum w * (\phi_{\text{model}} - \phi_{\text{target}})^2 + \sum w * (\psi_{\text{model}} - \psi_{\text{target}})^2$$

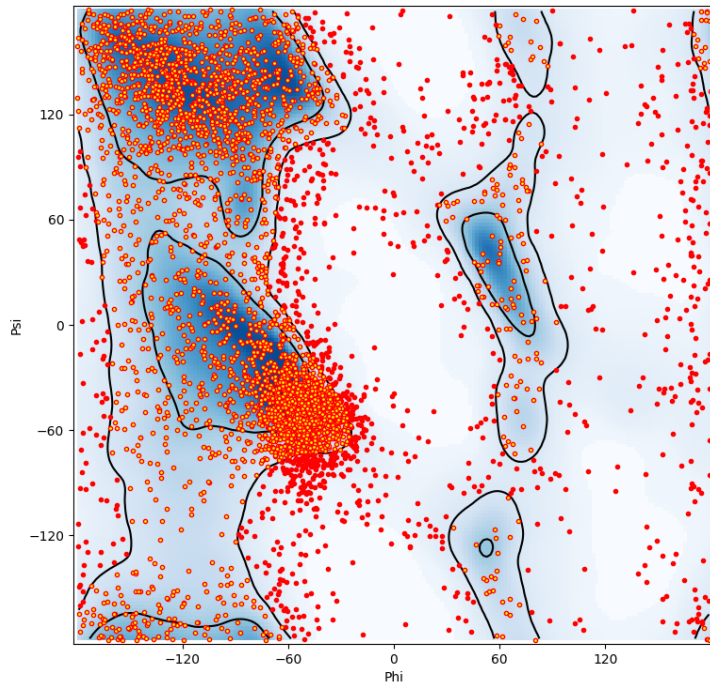


- Setting up Ramachandran plot, secondary structure, etc, restraints can be ambiguous and is error prone!

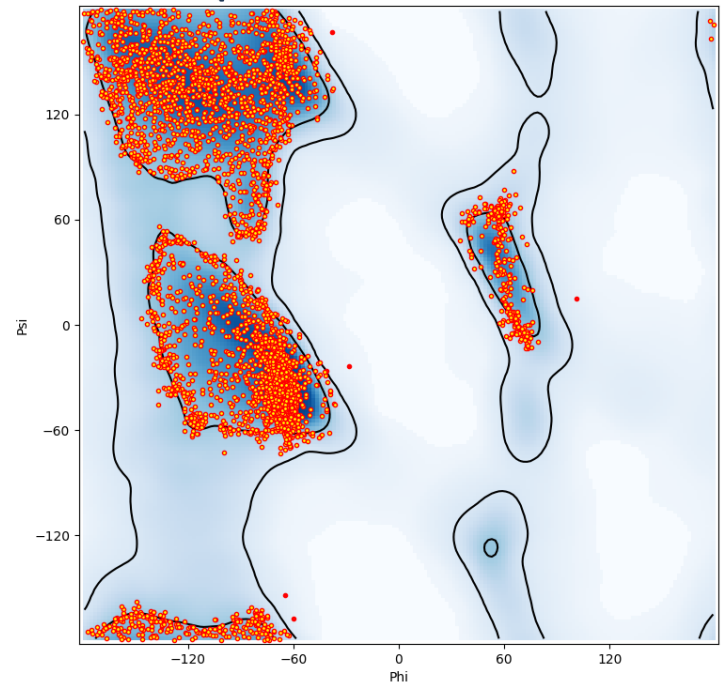
How did that happen?

PDB code: 5a9z

Original



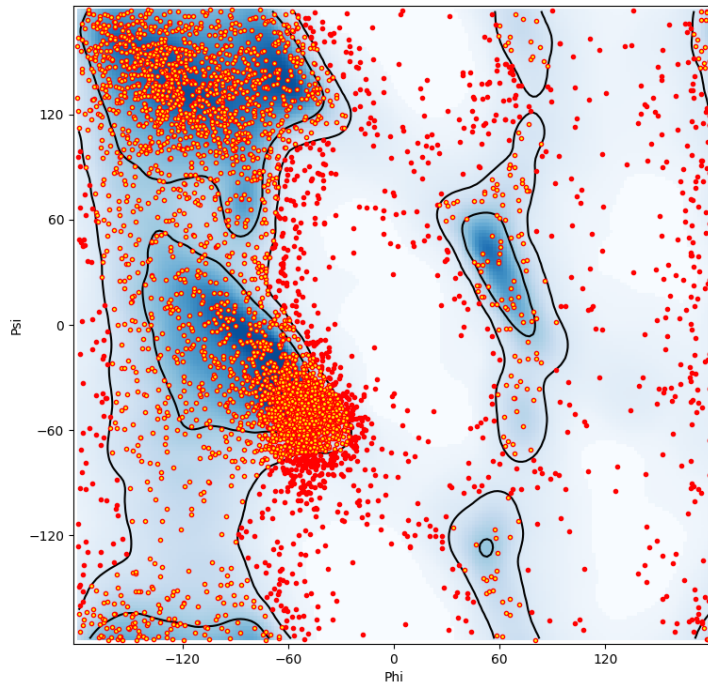
Refined with Ramachandran
plot restraints



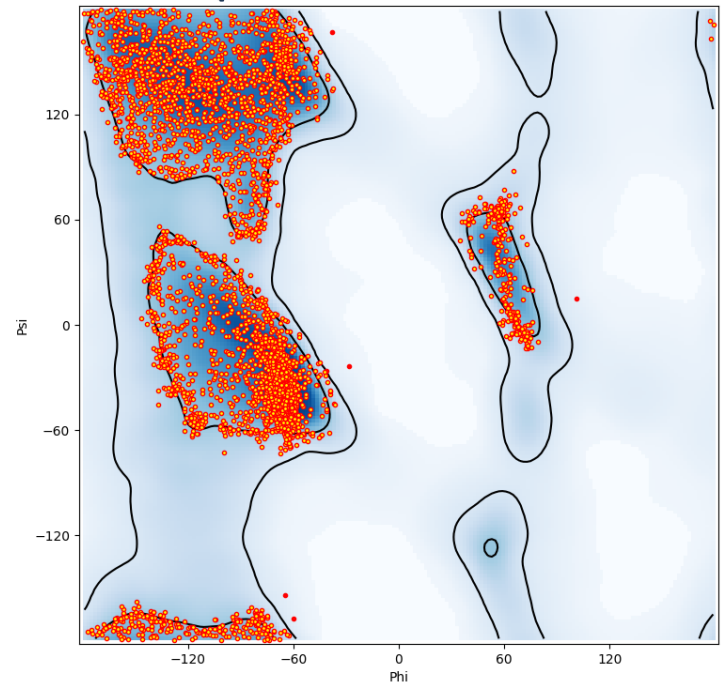
How did that happen?

PDB code: 5a9z

Original



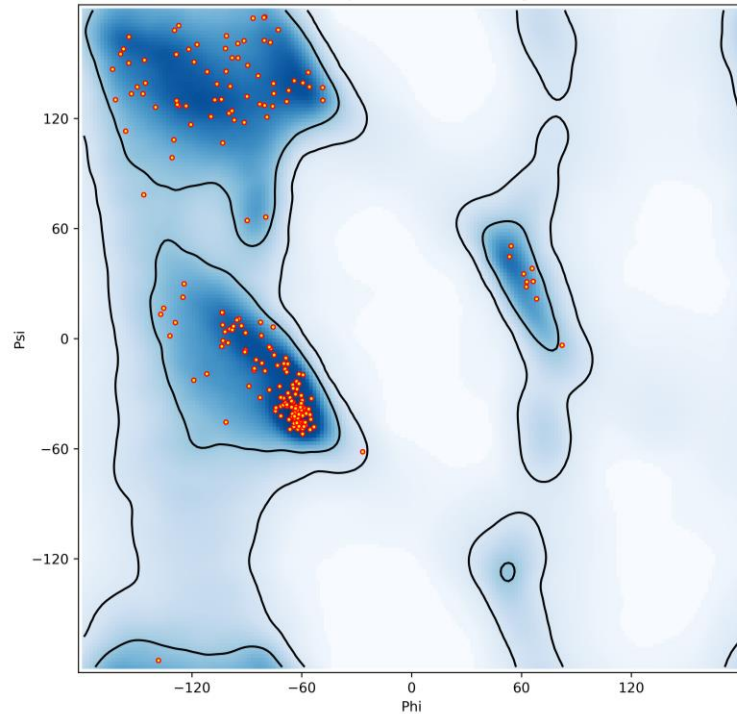
Refined with Ramachandran
plot restraints



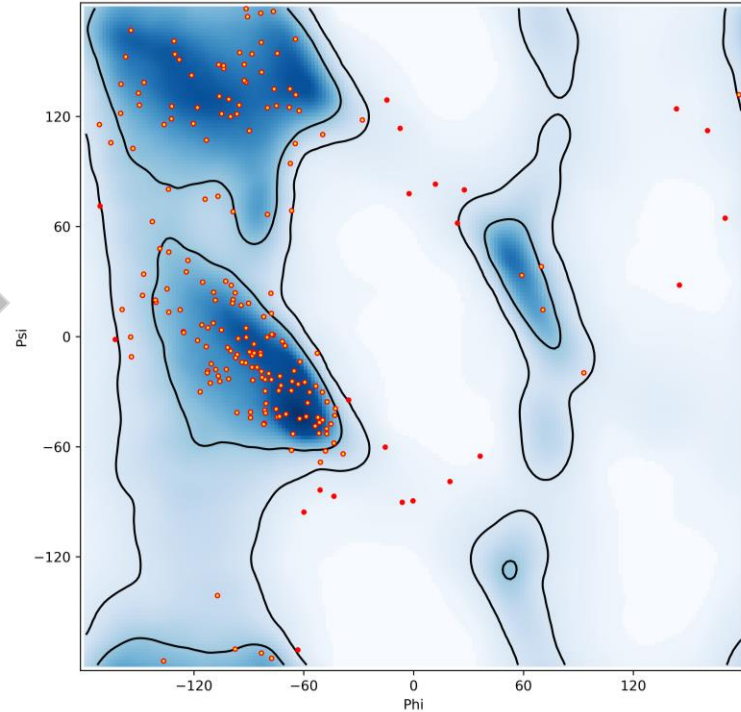
Don't use Ramachandran plot restraints to remove outliers!

Ramachandran plot restraints

Before refinement



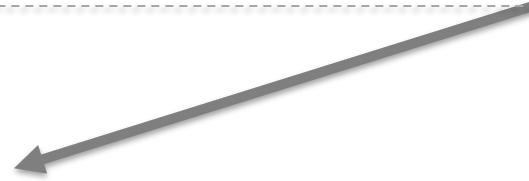
After refinement
(No Ramachandran plot restraints)



Use Ramachandran plot restraints to stop outliers from occurring!

Restraints and limitations

$$T = T_{\text{DATA}} + W * T_{\text{RESTRAINTS}}$$



$$T_{\text{RESTRAINTS}} = T_{\text{BOND}} + T_{\text{ANGLE}} + T_{\text{DIHEDRAL}} + T_{\text{PLANE}} + T_{\text{REPULSION}} + T_{\text{CHIRALITY}}$$

- Restraints are too limited:
 - No attraction terms (electrostatics, etc)
 - Not using information about protein structure (secondary structure, rotamers)
 - Limited to tabulated entities in the libraries (e.g., Monomer Library, GeoStd)

A better solution: restraints from QM

T

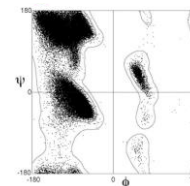
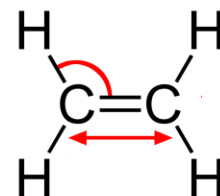
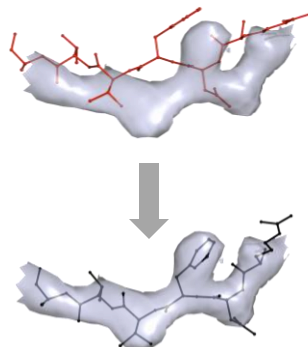
=

T_{DATA}

+

w * T_{RESTRAINTS}

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



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

Replace with
energies/gradients
from QM calculations

NEW: AQuaRef – QM based refinement in *Phenix*

Bonus content

Pandemic of junk or lack of validation in action

Despite all efforts to popularize (and enforce) the validation in recent years, poorly scoring models are still getting into databases **now**

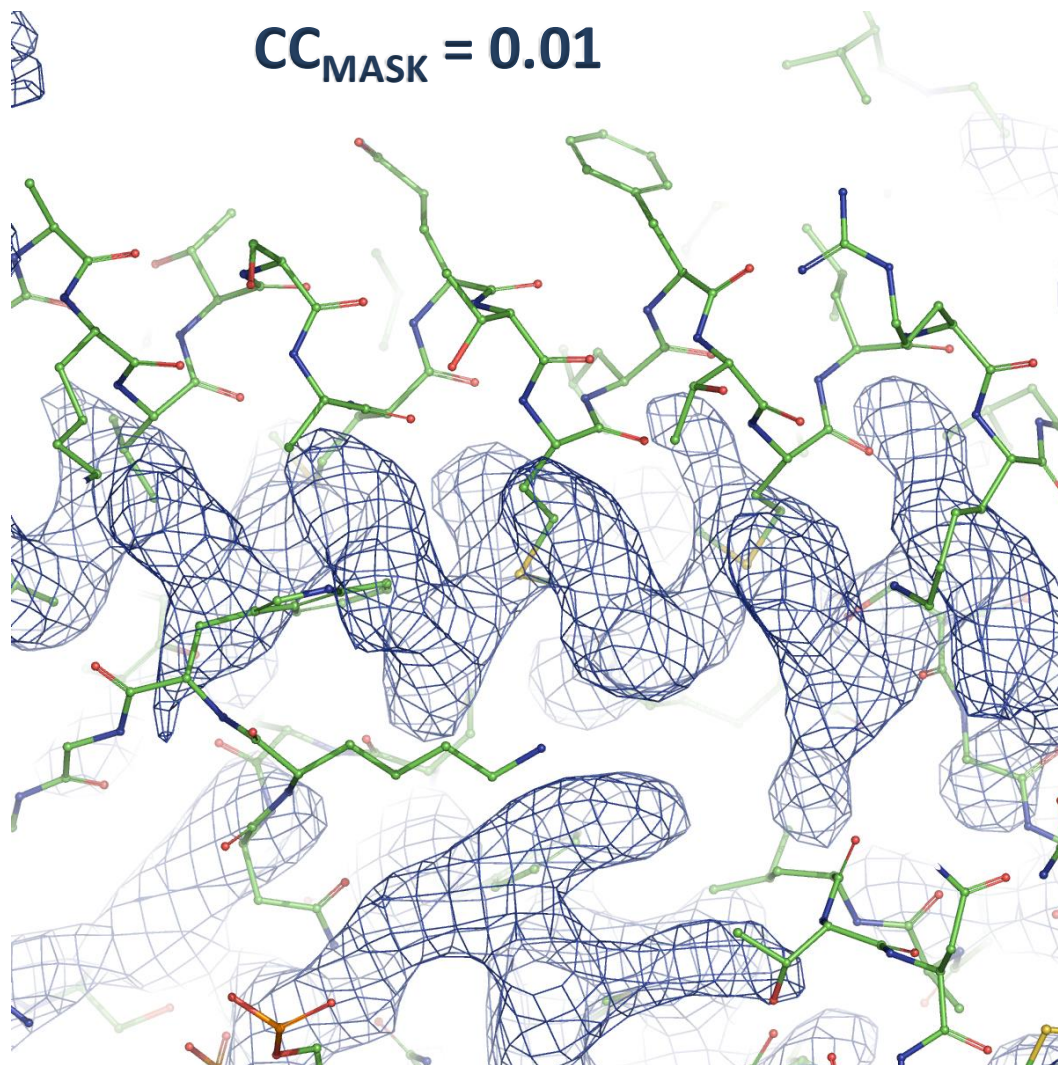
Examples (recent years)!

Model does not fit the map

PDB: 8gwb | EMD-3: 34308 | 2.8 Å | Cell (2022) 185: 4347-4360

Chain	CC_{MASK}
-------	--------------------

A	0.01
B	0.02
C	0
D	0.01
I	0.04
J	0
F	0.12
E	0.08
G	0.1
M	0.16
A	0
F	-0.13
E	0.16
A	0.1
G	0.15
M	0.19



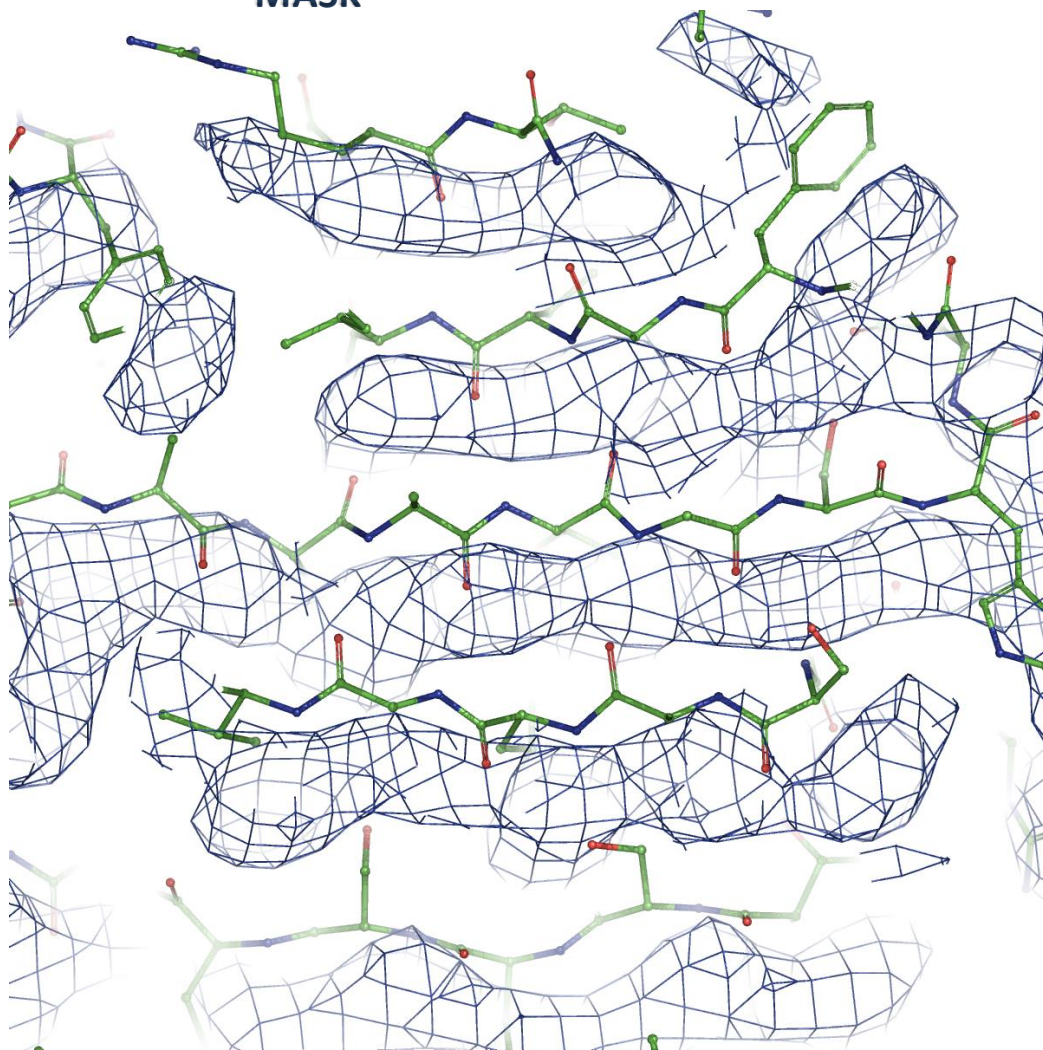
Model does not fit the map

PDB: 7xov | EMDB: 33360 | 3 Å | Cell Discov (2022) 8: 55-55

Chain CC_{MASK}

A	0.04
B	-0.01
G	0.18
N	0.06
R	0.03
R	-0.02

$CC_{\text{MASK}} = 0.02$



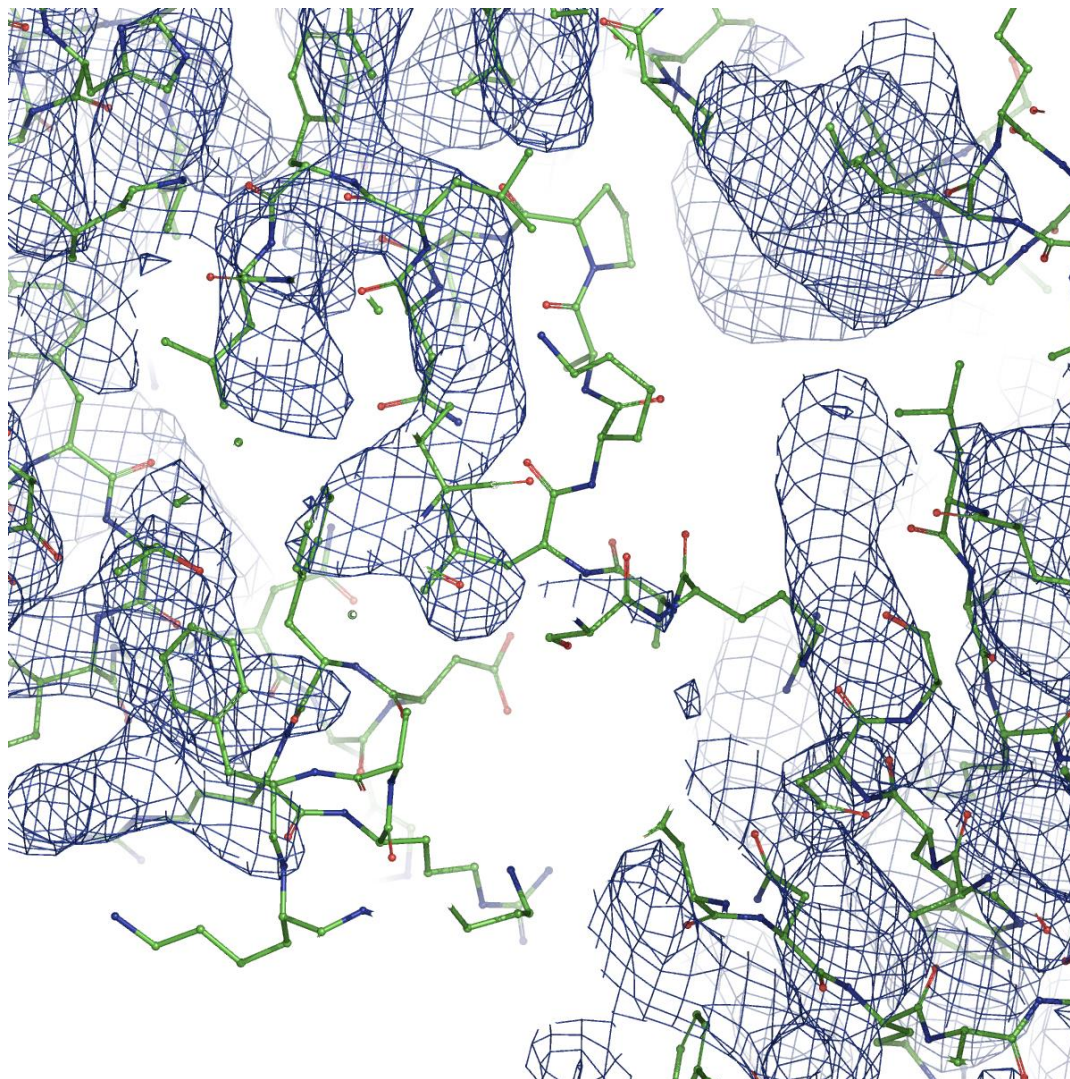
Model does not fit the map

PDB: 7w6p | EMDDB: 32331 | 3.5 Å | Science (2022) 377: 7065-7065

Chain	CC_{MASK}
-------	--------------------

A	0.09
B	0.11
G	0.12
H	0.07
R	0.16
R	-0.08

$CC_{\text{MASK}} = 0.1$



Model does not fit the map

PDB: 8V85 | EMDDB: 43023 | 2.9 Å | Nat Commun (2024) 15: 3296-3296

$CC_{\text{MASK}} = 0.15$



More in 3D: [Structure](#) | [Sequence](#) | [Annotations](#) | [Electron Density](#) | [Refinement Report](#)

Symmetry: Asymmetric - C1
Stoichiometry: Monomer - A1

[Similar Assemblies](#)

 **8V85** | **pdb_000**

60S ribosome biogenesis intermediate
(pass filtered locally refined map)

PDB DOI: <https://doi.org/10.2210/pdb8V85/pdb> |

Classification: **RNA BINDING PROTEIN**

Organism(s): *Saccharomyces cerevisiae* BY4741

Mutation(s): No

Deposited: 2023-12-04 Released: 2024-05-01

Deposition Author(s): Cruz, V.E., Weirich, C.S., Pe

Funding Organization(s): National Institutes of Health (NIH/NIGMS), Robert A. Welch Foundation, Cancer Research and Biotechnology Program (CPRIT)

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY

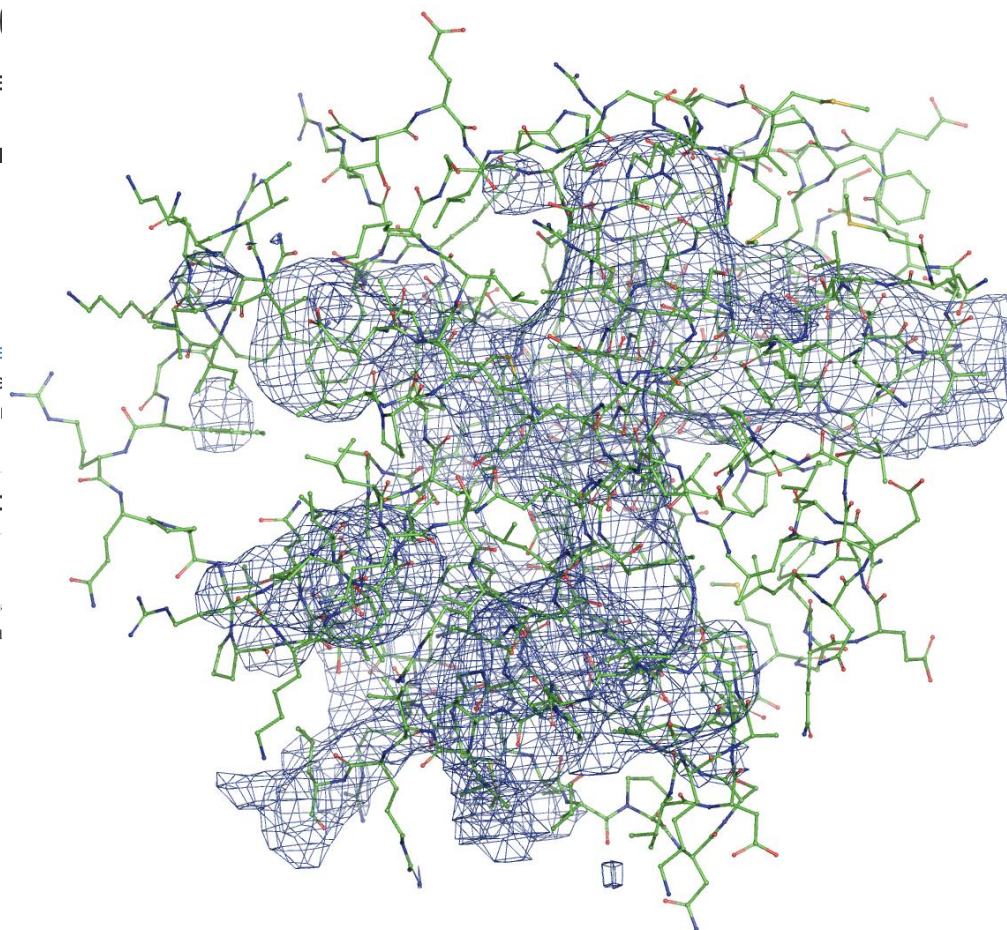
Resolution: 2.90 Å

Aggregation State: PARTICLE

Reconstruction Method: SINGLE PARTICLE

wwPDB

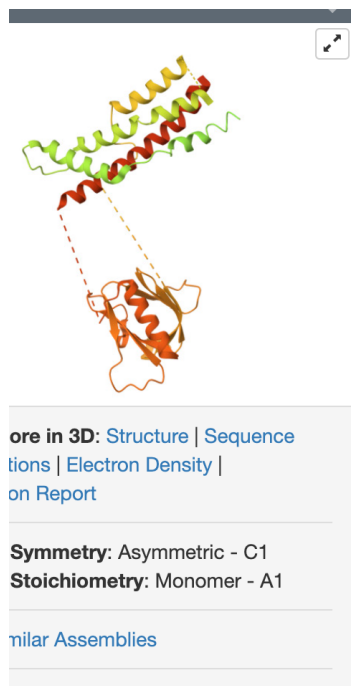
Ramach
Sid



Model does not fit the map

PDB: 8SZ7 | EMDB: 40902 | 2.8 Å | Dev Cell (2024) 59: 1783

$CC_{\text{MASK}} = 0.19$



 **8SZ7** | **pdb_00**

Cryo-EM of the GDP-bound human membrane in the super constricted :

PDB DOI: <https://doi.org/10.2210/pdb8SZ7/pdb>

Classification: [HYDROLASE](#)

Organism(s): [Homo sapiens](#)

Expression System: [Escherichia coli 'BL21-Go](#)

Mutation(s): Yes

Deposited: 2023-05-27 **Released:** 2024-05-01

Deposition Author(s): [Jimah, J.R.](#), [Canagarajah](#)

Funding Organization(s): National Institutes of Health and Kidney Disease (NIH/NIDDK), National Institutes of Health (NIH/NIGMS)

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY

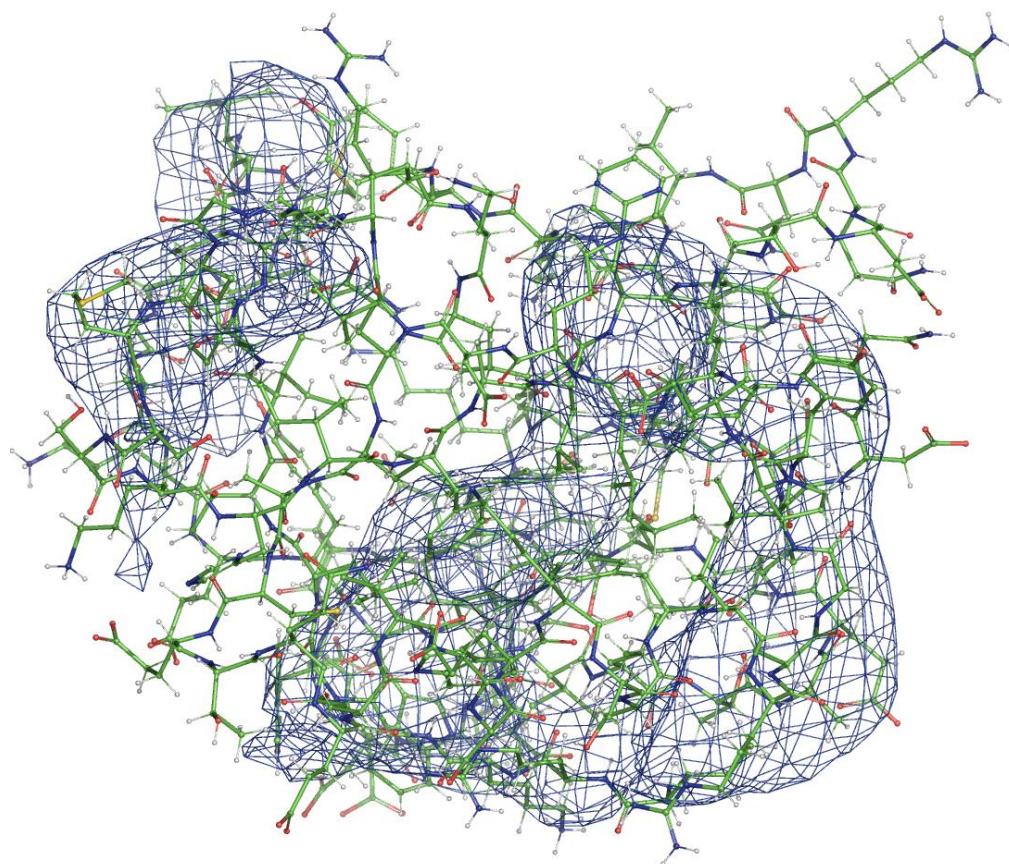
Resolution: 2.84 Å

Aggregation State: FILAMENT

Reconstruction Method: HELICAL

ww

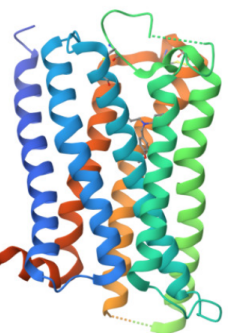
Ran



Model does not fit the map

PDB: 8x63 | EMDB: 38078 | 3.2 Å | Nat Commun (2024) 15: 84-84

$CC_{\text{MASK}} = 0.13$



More in 3D: [Structure](#) | [Sequence](#) | [Electron Density](#) | [Ligand Interaction \(Y5E\)](#) | [Stoichiometry](#) | [Membrane](#)

Symmetry: Asymmetric - C1
Stoichiometry: Monomer - A1

[Similar Assemblies](#)

 **8X63** | **pdb_000083**

CryoEM structure of the histamine H1 receptor with mepyramine

PDB DOI: <https://doi.org/10.2210/pdb8X63/pdb> EM Ma

Classification: **MEMBRANE PROTEIN**

Organism(s): [Homo sapiens](#), [Escherichia coli](#)

Expression System: [Spodoptera frugiperda](#)

Mutation(s): Yes

Membrane Protein: [Yes](#) [PDBTM](#) [MemProtMD](#) [mpstruc](#)

Deposited: 2023-11-20 **Released:** 2024-01-17

Deposition Author(s): [Wang, D.D.](#), [Guo, Q.](#)

Funding Organization(s): [National Natural Science Found](#)

Experimental Data Snapshot

Method: ELECTRON MICROSCOPY

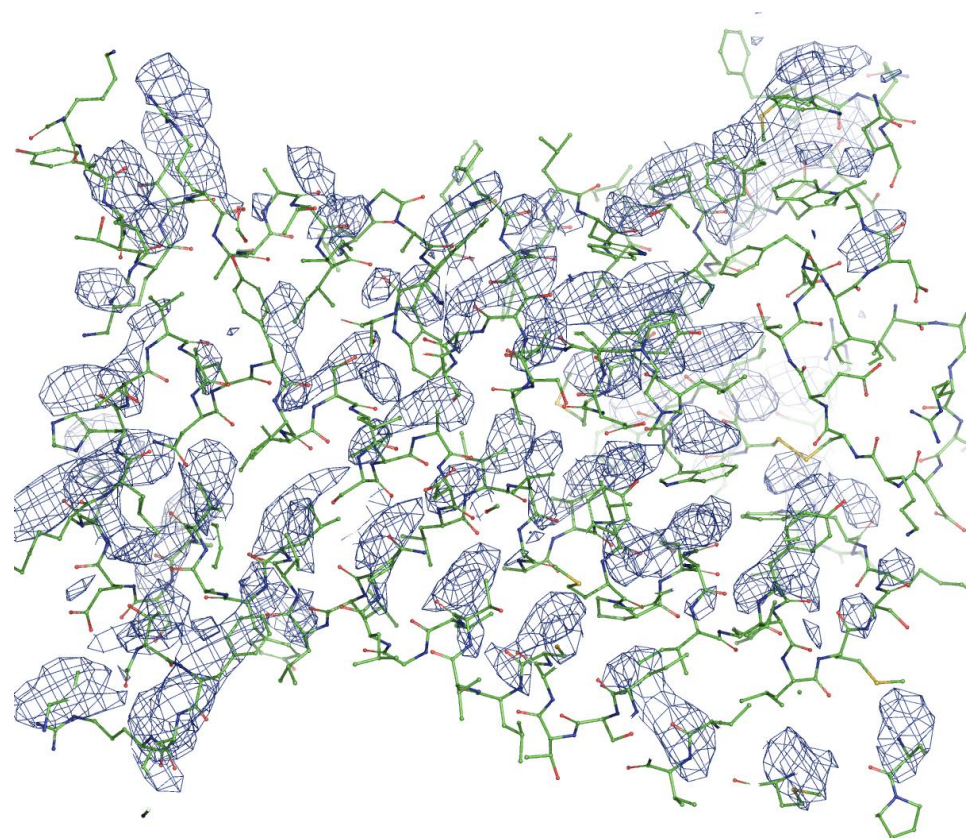
Resolution: 3.20 Å

Aggregation State: PARTICLE

Reconstruction Method: SINGLE PARTICLE

wwPDB Validation

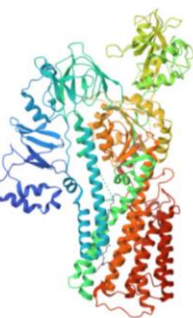
[Class](#)
[Ramachandran plot](#)
[Sidechain outliers](#)



Model does not fit the map

PDB: 8iEN | EMDB: 35387 | 3.25 Å | Nat Commun (2023) 14: 1978-1978

$CC_{\text{MASK}} = 0.0$



Explore in 3D: [Structure](#) | [Sequence](#)
[Annotations](#) | [Electron Density](#) |
[Validation Report](#) |
[Ligand Interaction \(SPM\)](#) |
[Contact Membrane](#)

Crystallographic Symmetry: Asymmetric - C1
Crystallographic Stoichiometry: Monomer - A1

8iEN | **pdb_00008iEN**

Cryo-EM structure of ATP13A2 in the E2-Pi state

PDB DOI: <https://doi.org/10.2210/pdb8iEN/pdb> EM Map EMD-35387

Classification: **TRANSPORT PROTEIN**

Organism(s): *Homo sapiens*

Expression System: *Homo sapiens*

Mutation(s): No

Membrane Protein: Yes **PDBTM** **MemProtMD**

Deposited: 2023-02-15 **Released:** 2023-12-20

Deposition Author(s): Liu, Z.M., Mu, J.Q., Xue, C.Y.

Funding Organization(s): National Science Foundation (NSF, China)

Experimental Data Snapshot

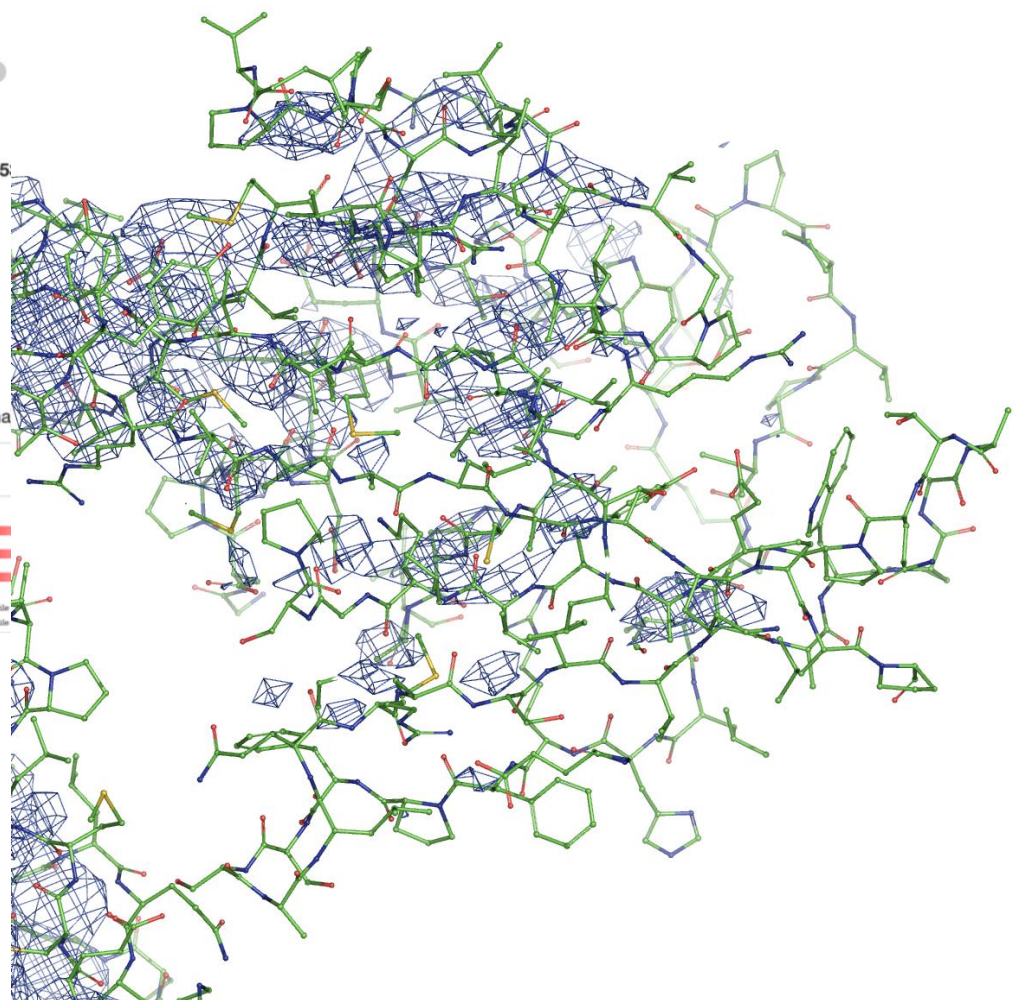
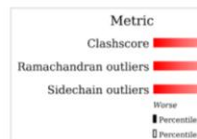
Method: ELECTRON MICROSCOPY

Resolution: 3.25 Å

Aggregation State: PARTICLE

Reconstruction Method: SINGLE PARTICLE

wwPDB Validation



Model does not fit the map

PDB: 9c91 | EMDB: 45359 | 2.78 Å | Nat Commun (2025) 16: 2955

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [nature communications](#) > [articles](#) > [article](#)

Article | [Open access](#) | Published: 26 March 2025

Structure of dimerized assimilatory NADPH-dependent sulfite reductase reveals the minimal interface for diflavin reductase binding

[Behrouz Ghazi Esfahani](#), [Nidhi Walia](#), [Kasahun Neselu](#), [Yashika Garg](#), [Mahira Aragon](#), [Isabel Askenasy](#), [Hui Alex Wei](#), [Joshua H. Mendez](#) & [M. Elizabeth Stroupe](#) ✉

[Nature Communications](#) **16**, Article number: 2955 (2025) | [Cite this article](#)

1343 Accesses | 1 Altmetric | [Metrics](#)



 **9C91** | **pdb_00**

Assimilatory NADPH-dependent sulfite reductase

PDB DOI: <https://doi.org/10.2210/pdb9C91/pdb>

Classification: **FLAVOPROTEIN**

Organism(s): **Escherichia coli**

Expression System: **Escherichia coli**

Mutation(s): No

Deposited: 2024-06-13 Released: 2025-02-12

Deposition Author(s): [Ghazi Esfahani, B.](#), [Walia, Nidhi](#), [Mendez, J.H.](#), [Stroupe, M.E.](#)

Funding Organization(s): National Science Foundation

Experimental Data Snapshot

wwPDB

Method: ELECTRON MICROSCOPY

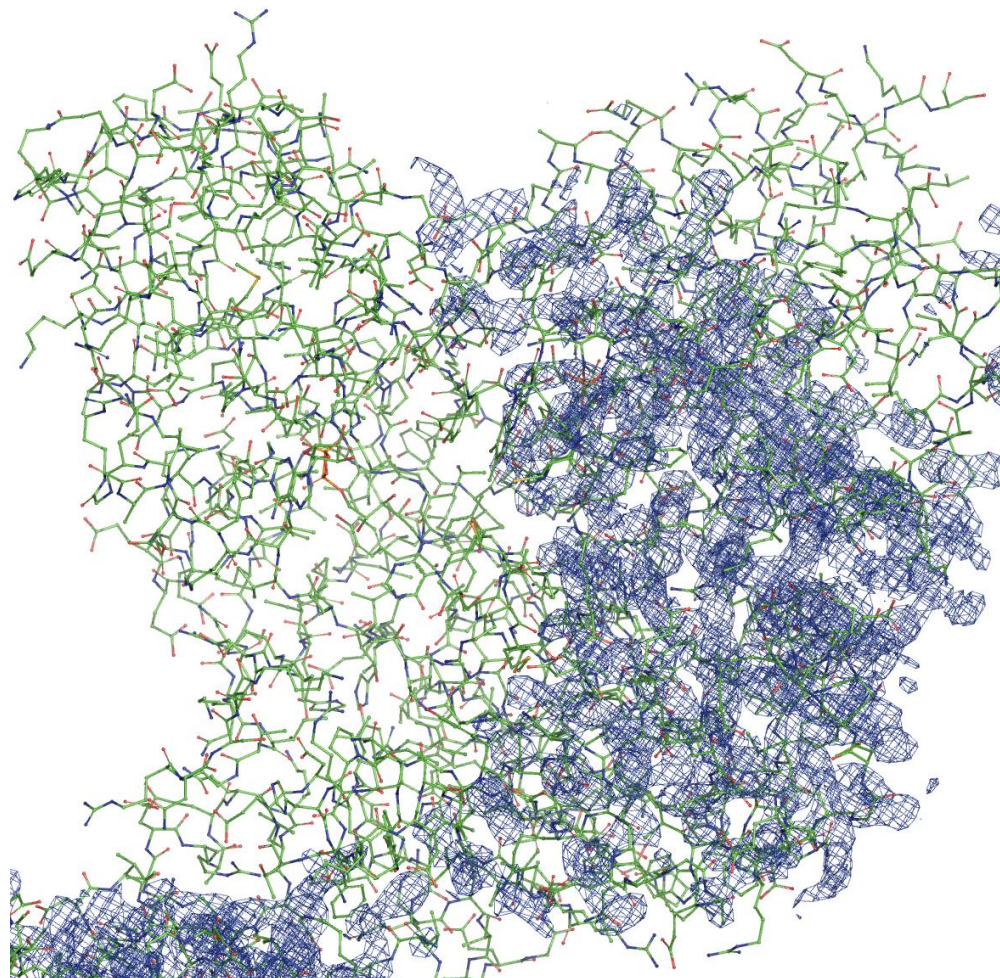
Resolution: 2.78 Å

Aggregation State: PARTICLE

Reconstruction Method: SINGLE PARTICLE

Ran

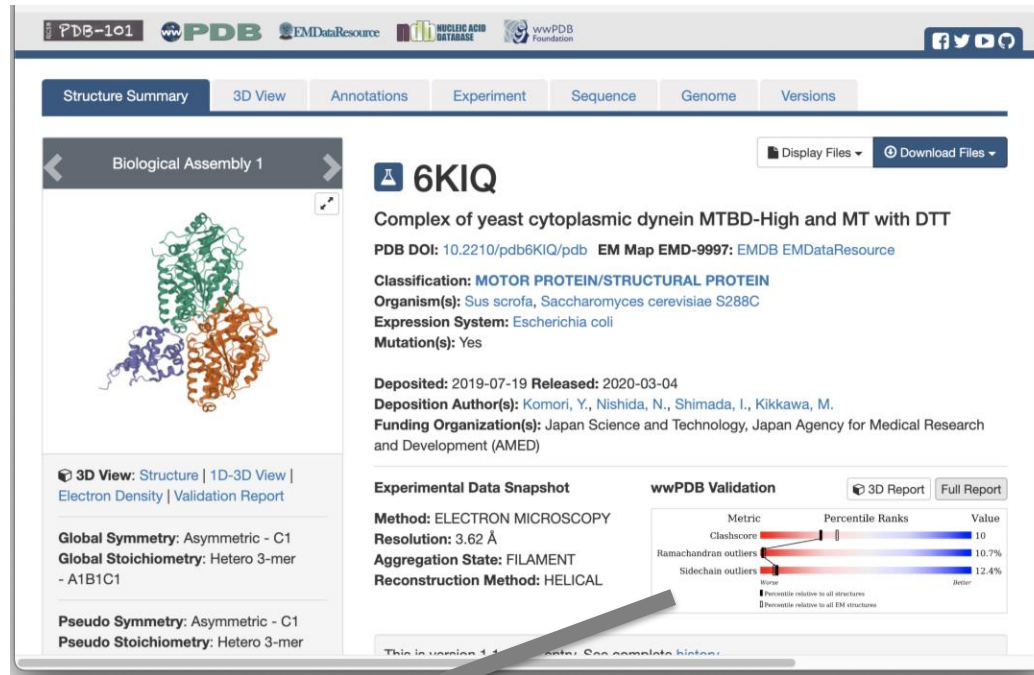
$CC_{\text{MASK}} = 0.0$



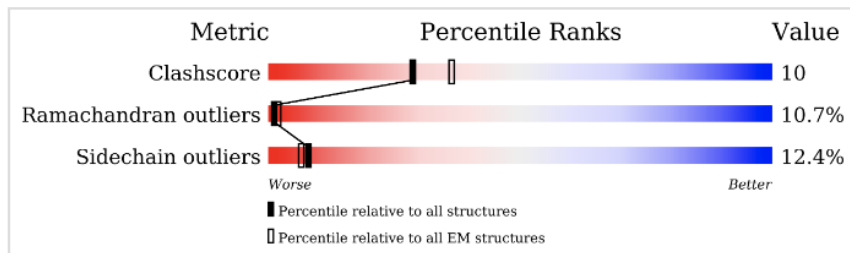
in 3D: [Structure](#) | [Sequence](#)
[Electron Density](#) |
[Report](#) |
[Reaction](#) (SRM)

metry: Asymmetric - C1
chometry: Hetero 2-mer -

Validation reports (RCSB)



wwPDB Validation



Page 34

Full wwPDB EM Validation Report

EMD-

3D Report

Full Report

9.5 Map-model fit summary ⓘ

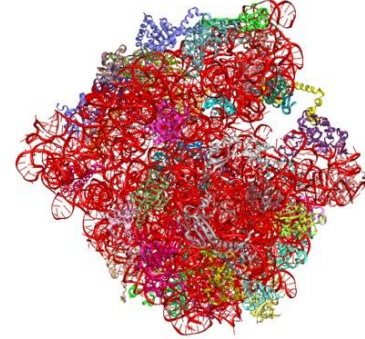
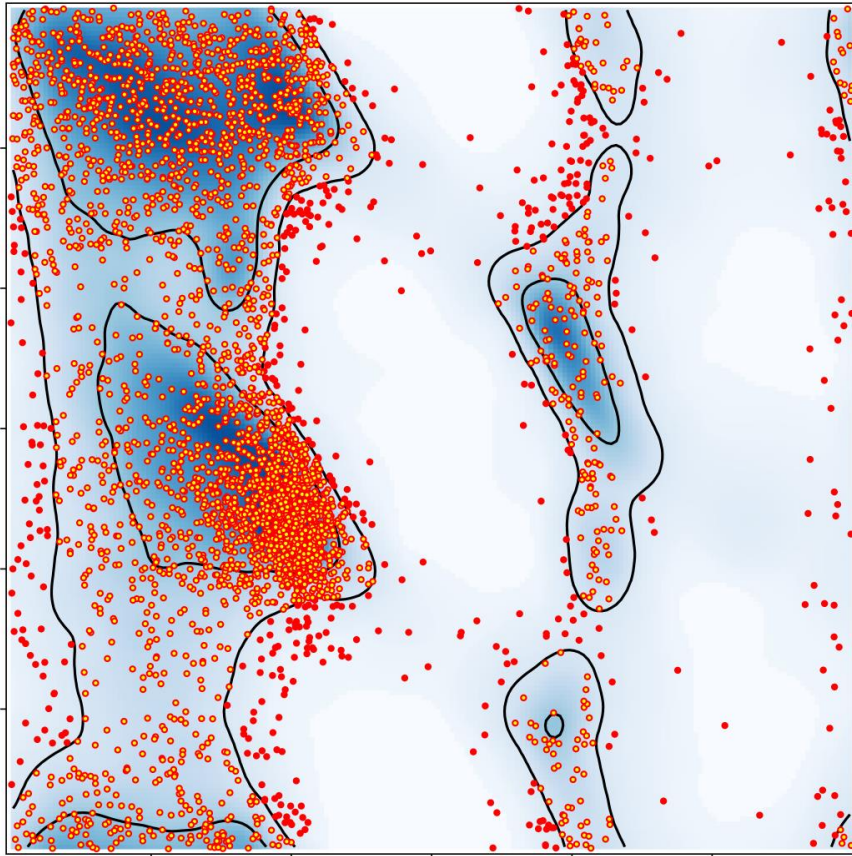
The table lists the average atom inclusion at the recommended contour level (0.125) for the entire model and for each chain.

Chain	Atom inclusion	Q-score
All	0.9062	0.4550
M	0.5810	0.3210
a	0.9659	0.4790
b	0.9656	0.4730

Lack of (useful) model-to-map fit statistics!

Poor model geometry

(2019) Nature 570: 400-404 | PDB: 6o9j | EMDB: 0661 | 3.9Å



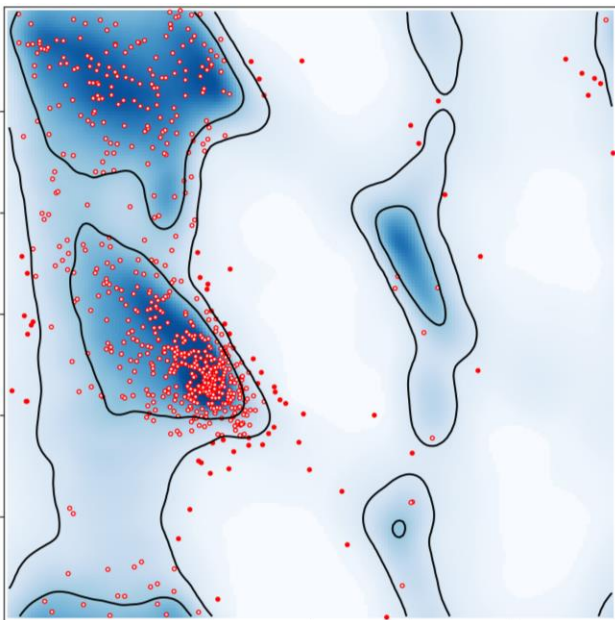
Metric	6o9j	Expected
Clashscore	70	Less than 10
Ramachandran favored, %	59	More than 98
Ramachandran outliers, %	15	0
Rotamer outliers, %	23	0
C _β deviations, %	0.5	0

Poor model geometry

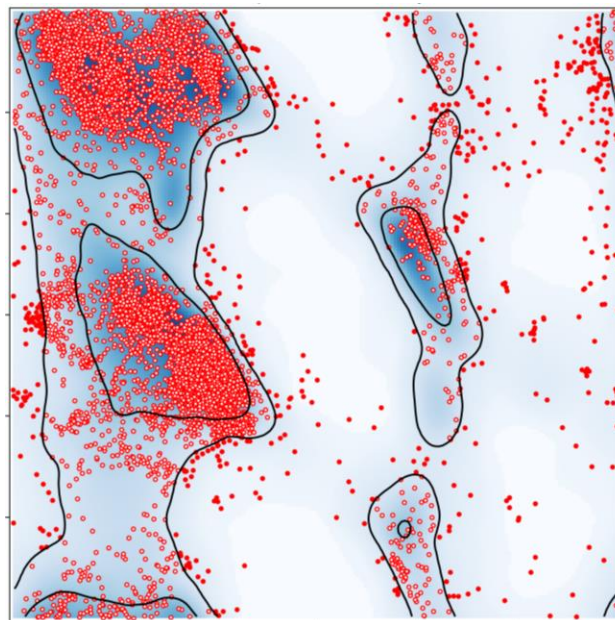
	6kio	6kiq	6o9j	7ase
Bond/angle	0.04/3.4	0.04/3.7	0.01/1.3	0.02/2.2
Clashscore	11	12	55	9
Rotamer outl., %	8	15	23	3
Cb deviations, %	5	16	0.5	1.4
Ramachandran, % favored outliers	74	70	59	79
	7	11	15	7
Resolution (Å)	3.9	3.6	3.9	3.3
Published in	Nature Comm.	Nature Comm.	Nature	Cell
Year	2020	2020	2019	2020

Poor model geometry

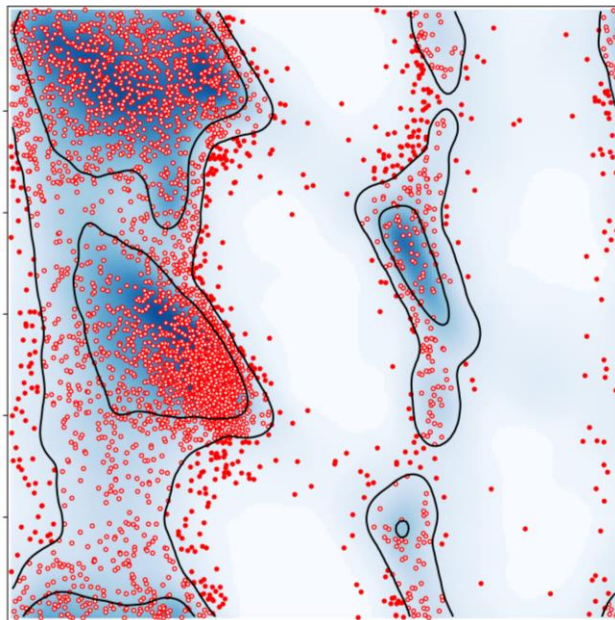
6kiq



7ase



6o9j



6kio

