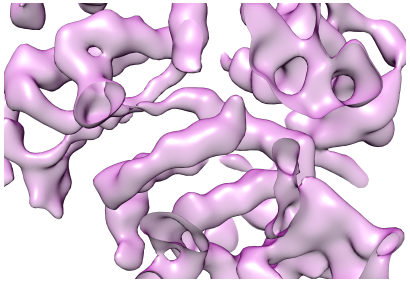
Phenix Tools for Cryo-EM: Validation

Pavel Afonine

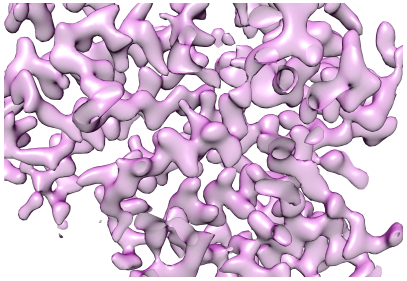
LBNL, Berkeley, California, USA

January 30, 2020

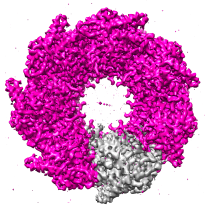
Cryo-EM tools in *Phenix*



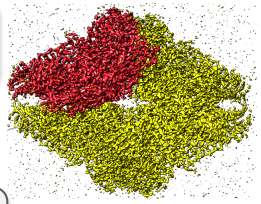
Starting map



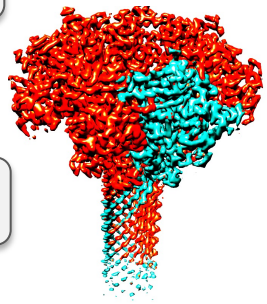
Map improvement



Map symmetry

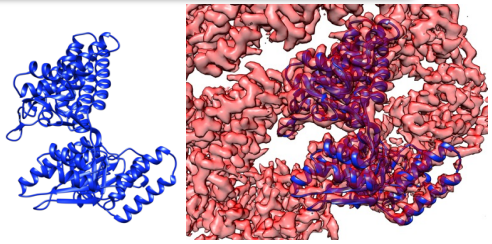


Map manipulation

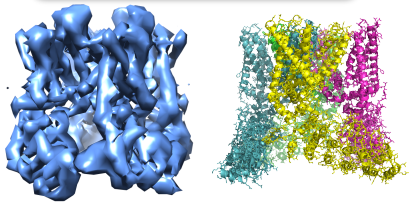


Extract unique part

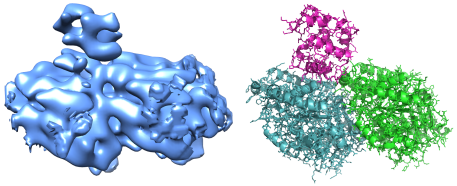
Docking, model building



Refinement

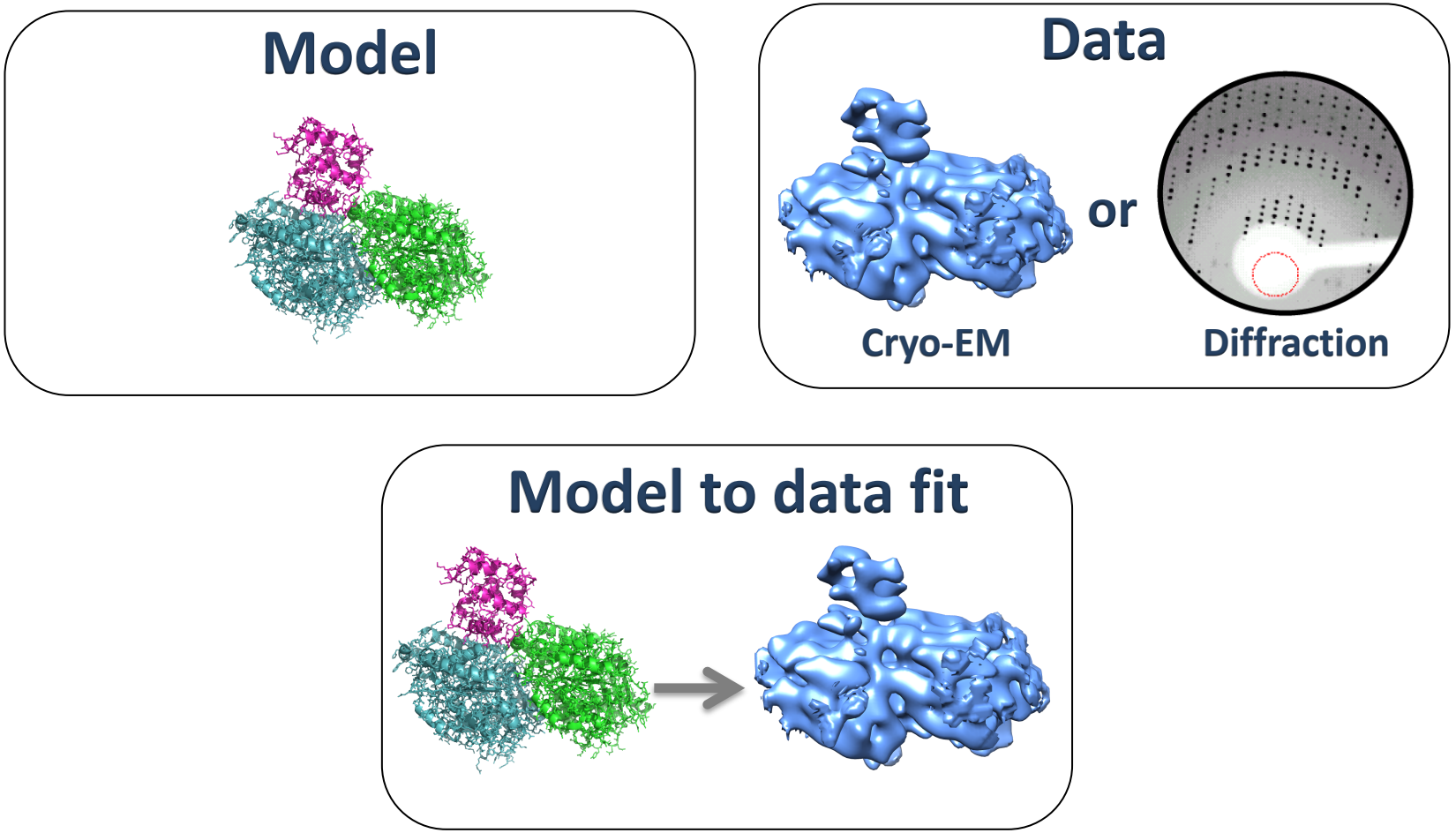


Validation



Complete set of tools for cryo-EM structure solution: from initial reconstruction to final validated model

Validation

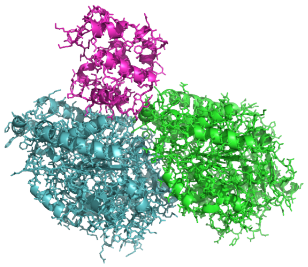


Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations

Validation tools: Crystallography vs Cryo-EM

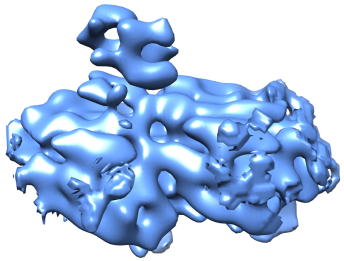
Exact same

Model



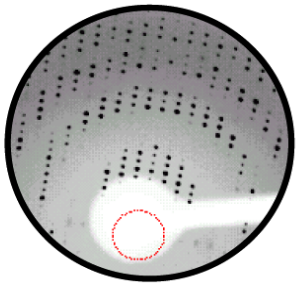
Different

Data



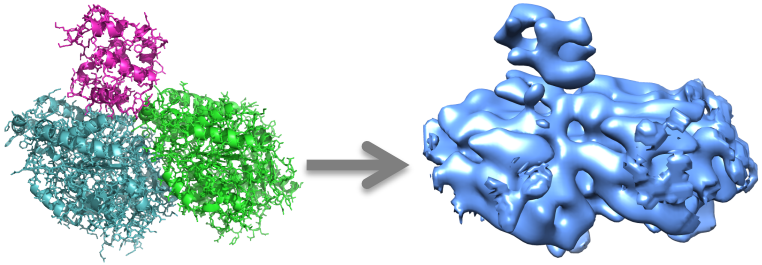
Cryo-EM

or



Diffraction

Model to data fit



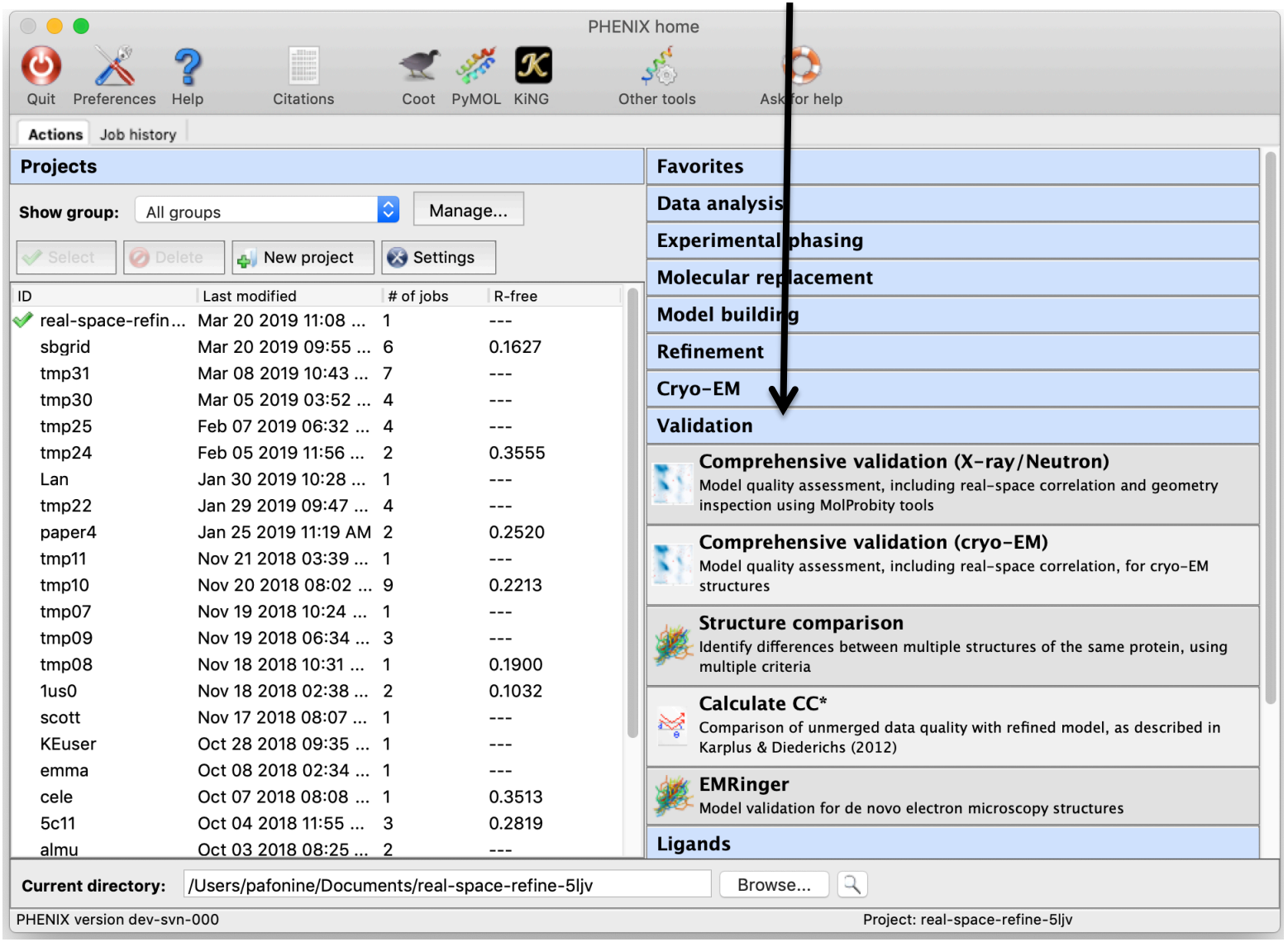
Similar

Validation

- Helps to save time
- Helps to produce better models
- Helps to set correct expectations

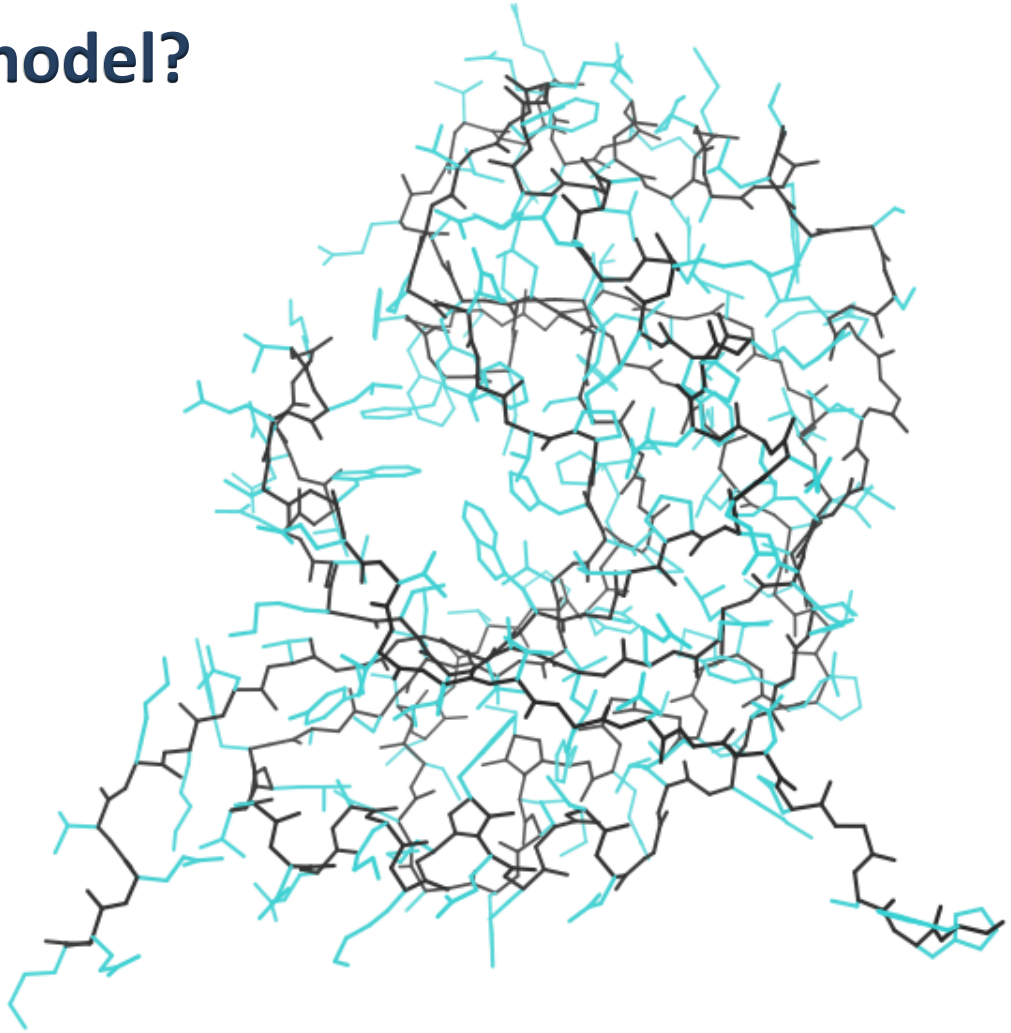
Validation

Validation for crystallography (X-ray, neutron) and cryo-EM



Model validation

Is it a good model?

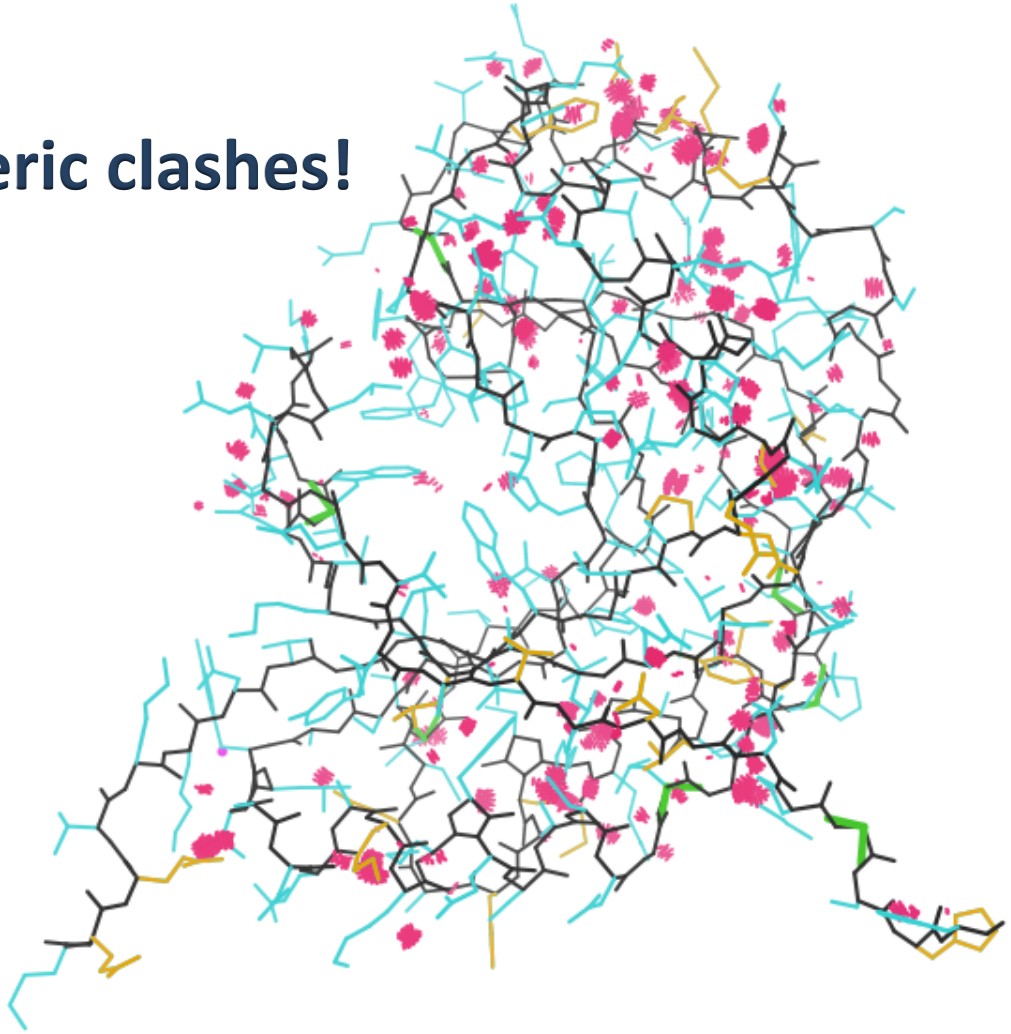


Cyclic Nucleotide Phosphodiesterase (2.4 Å)

Model validation

No.

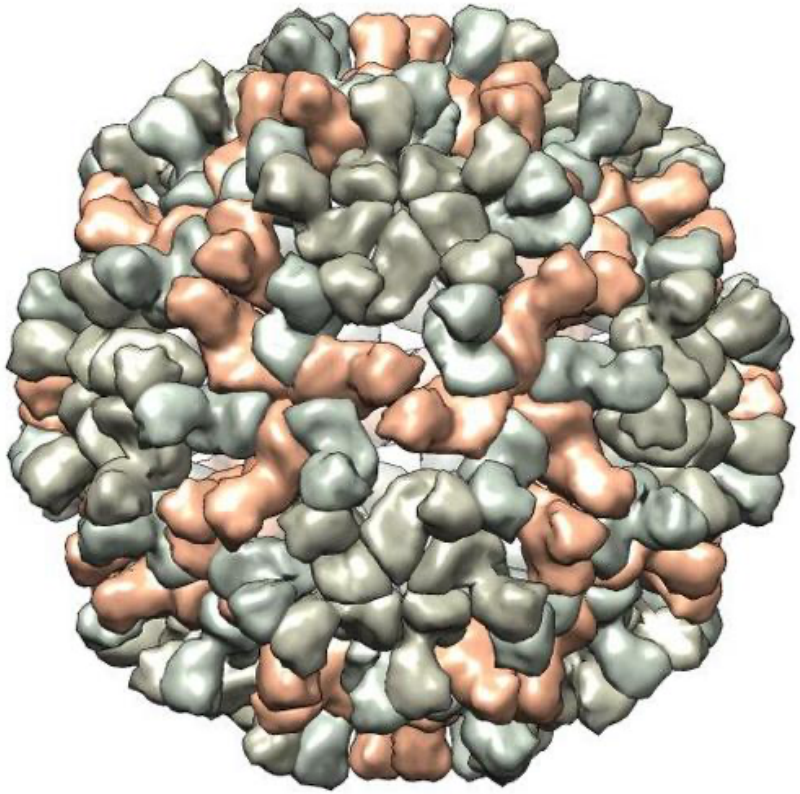
Too many steric clashes!

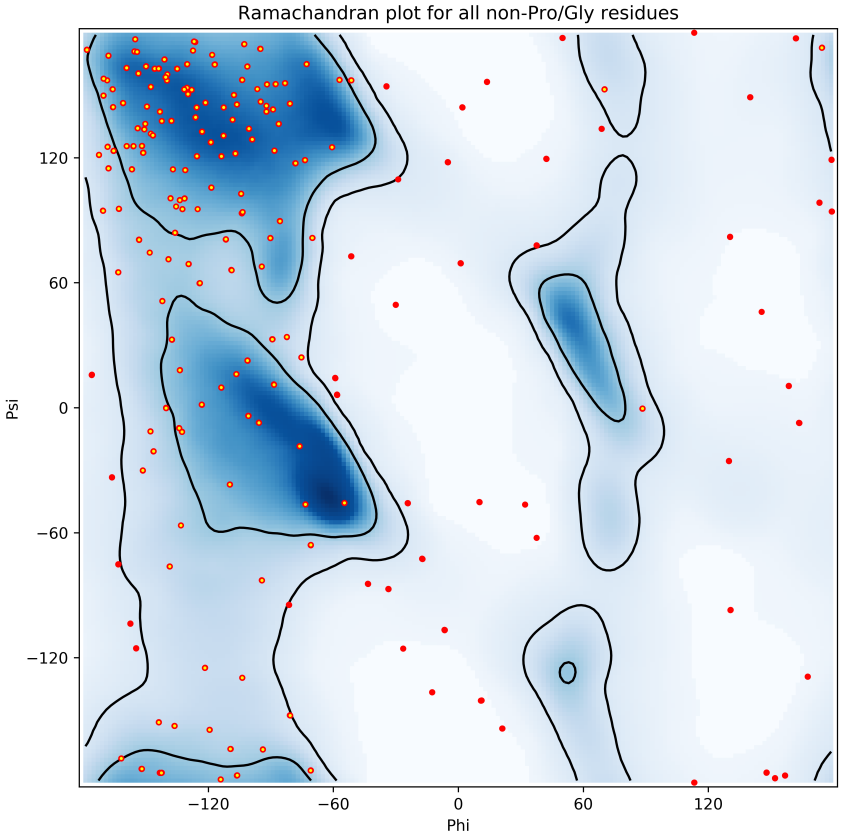


Cyclic Nucleotide Phosphodiesterase (2.4 Å)

Model validation

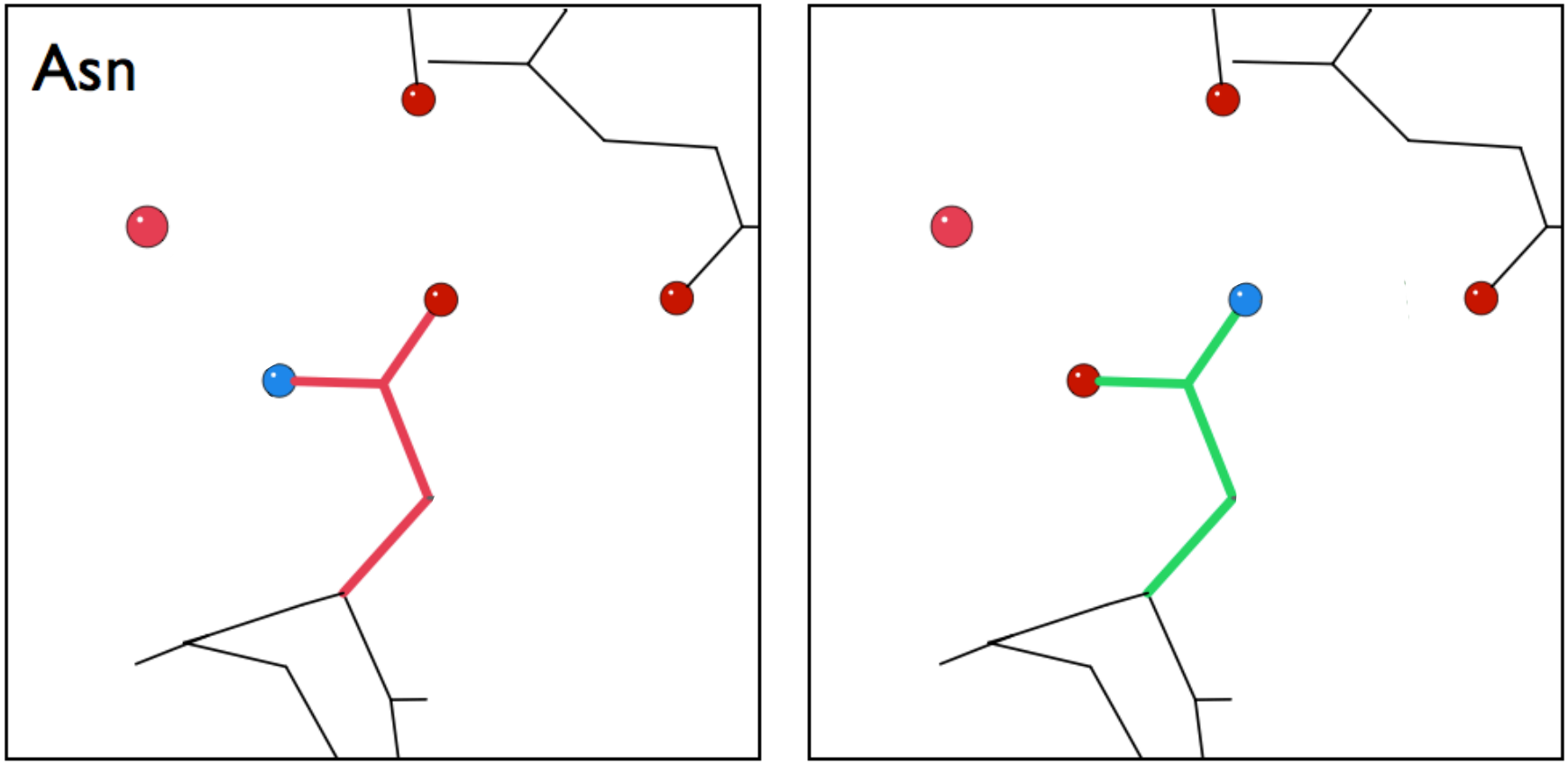
3zx9



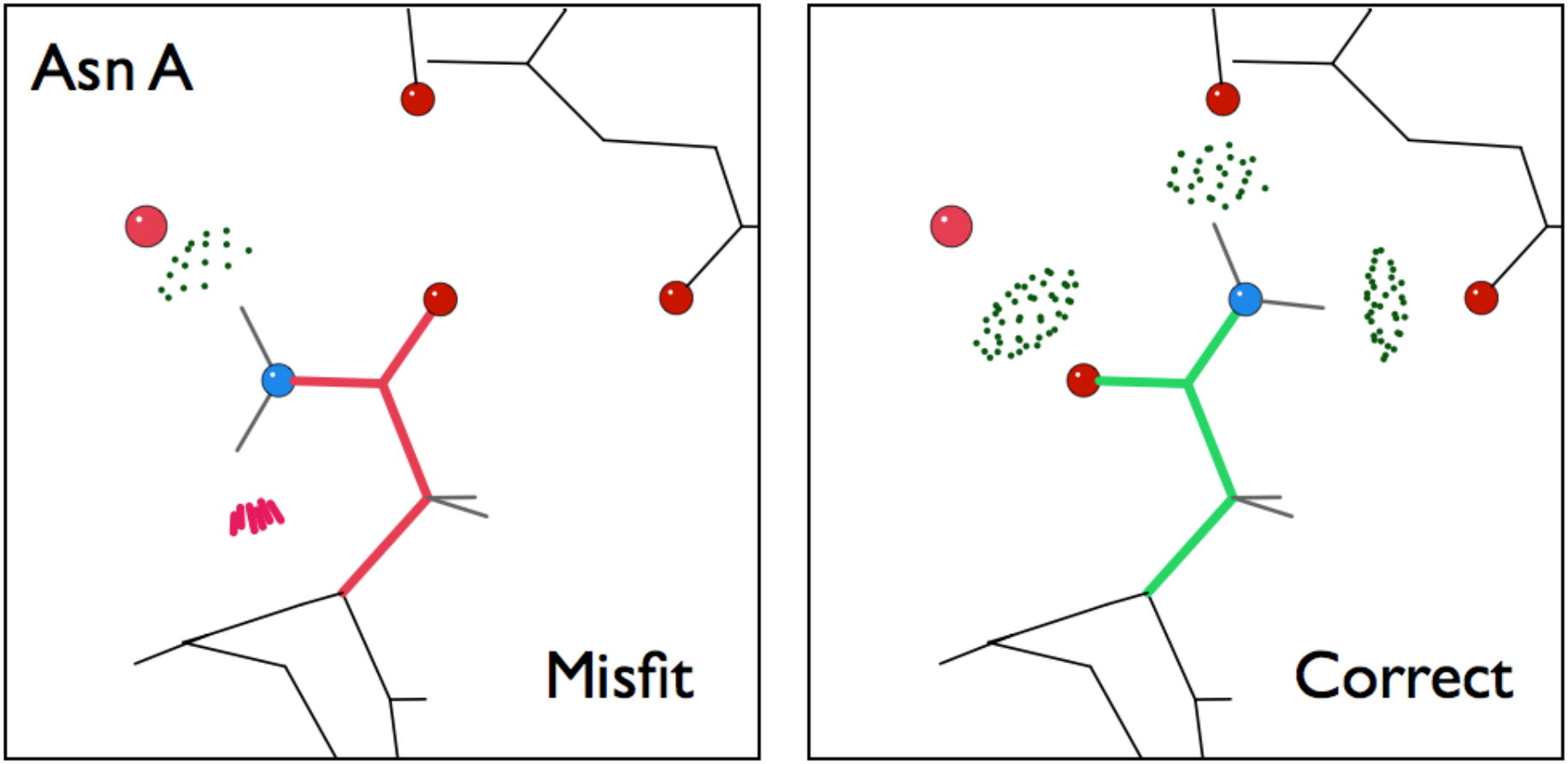


Very unlikely Ramachandran plot!

Clashes – N/Q/H flips

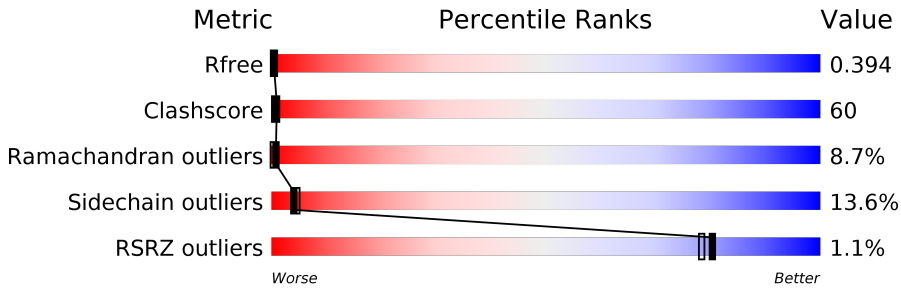


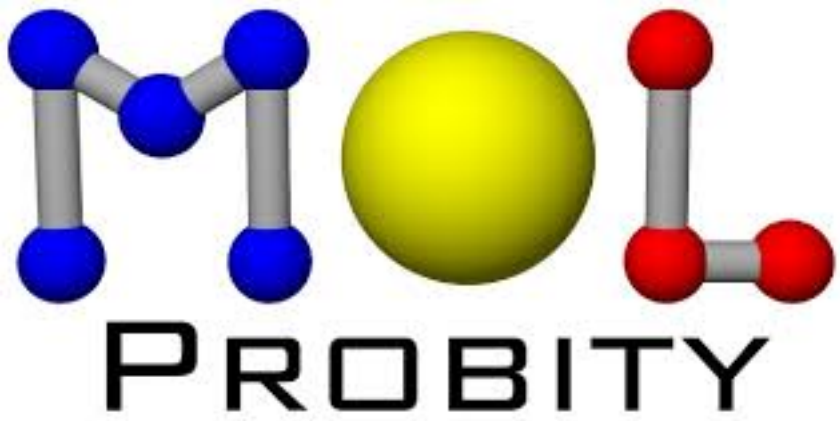
Clashes – N/Q/H flips



Validation – Table 1 (Crystallography)

- **Data**
  - Unit cell parameters & space group
  - Data collection details (T, λ, instrument,…)
  - Resolution & Completeness
  - I/σI
  - Redundancy
  - Wilson B
  - Various CC and R factors
- **Model**
  - Content (macromolecule, ligands, NCS, …)
  - Bond/angle RMSDs
  - 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-to-Data fit**
  - R_{WORK}, R_{FREE}





Validation (cryo-EM)

Comprehensive validation (cryo-EM) (Project: real-space-refine-5ljv_0)

Preferences | Help | Run | Abort | Ask for help

Input/Output | ValidationCryoEM_1

Run status | Summary | MolProbity | Model vs. Data | Data

Open in Coot

Export Table 1

Model

Composition (#)	
Chains	2
Atoms	2500 (Hydrogens: 0)
Residues	Protein: 325 Nucleotide: 0
Water	0
Ligands	MG: 1 ADP: 1
Bonds (RMSD)	
Length (Å) (# > 4σ)	0.029 (146)
Angles (°) (# > 4σ)	2.853 (122)
MolProbity score	3.14
Clash score	19.06
Ramachandran plot (%)	
Outliers	3.10
Allowed	7.12
Favored	89.78
Rotamer outliers (%)	11.57
Cβ outliers (%)	3.68
Peptide plane (%)	
Cis proline/general	5.6/0.0
Twisted proline/general	11.1/0.7
CaBLAM outliers (%)	2.18
ADP (B-factors)	
Iso/Aniso (#)	2500/0
min/max/mean	
Protein	30.26/493.42/109.69
Nucleotide	---
Ligand	57.57/99.69/75.15
Water	---
Occupancy	
Mean	1.00
occ = 1 (%)	100.00
0 < occ < 1 (%)	0.00
occ > 1 (%)	0.00

Data

Box		
Lengths (Å)	50.92, 68.34, 83.08	
Angles (°)	90.00, 90.00, 90.00	
Supplied Resolution (Å)	3.6	
Resolution Estimates (Å)	Masked	Unmasked
d FSC (half maps; 0.143)	---	---
d 99 (full/half1/half2)	3.7/---/---	3.1/---/---
d model	3.7	3.7
d FSC model (0/0.143/0.5)	3.4/3.5/3.6	3.4/3.6/3.9
Map min/max/mean	-0.42/0.80/0.03	

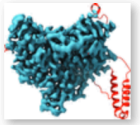
Model vs. Data

CC (mask)	0.83
CC (box)	0.55
CC (peaks)	0.34
CC (volume)	0.83
Mean CC for ligands	0.86

Idle | Project: real-space-refine-5ljv_0

RESEARCH PAPERS

Acta Cryst. (2018). D**74**, 814-840
https://doi.org/10.1107/S2059798318009324
Cited by 71
OPEN



New tools for the analysis and validation of cryo-EM maps and atomic models

P. V. Afonine, B. P. Klaholz, N. W. Moriarty, B. K. Poon, O. V. Sobolev, T. C. Terwilliger, P. D. Adams and A. Urzhumtsev

Recent advances in the field of electron cryomicroscopy (cryo-EM) have resulted in a rapidly increasing number of atomic models of biomacromolecules that have been solved using this technique and deposited in the Protein Data Bank and the Electron Microscopy Data Bank. Similar to macromolecular crystallography, validation tools for these models and maps are required. While some of these validation tools may be borrowed from crystallography, new methods specifically designed for cryo-EM validation are required. Here, new computational methods and tools implemented in *PHENIX* are discussed, including d_{99} to estimate resolution, *phenix.auto_sharpen* to improve maps and *phenix.mtriage* to analyze cryo-EM maps. It is suggested that cryo-EM half-maps and masks should be deposited to facilitate the evaluation and validation of cryo-EM-derived atomic models and maps. The application of these tools to deposited cryo-EM atomic models and maps is also presented.

Keywords: cryo-EM; **atomic models**; **model quality**; **data quality**; **validation**; **resolution**.

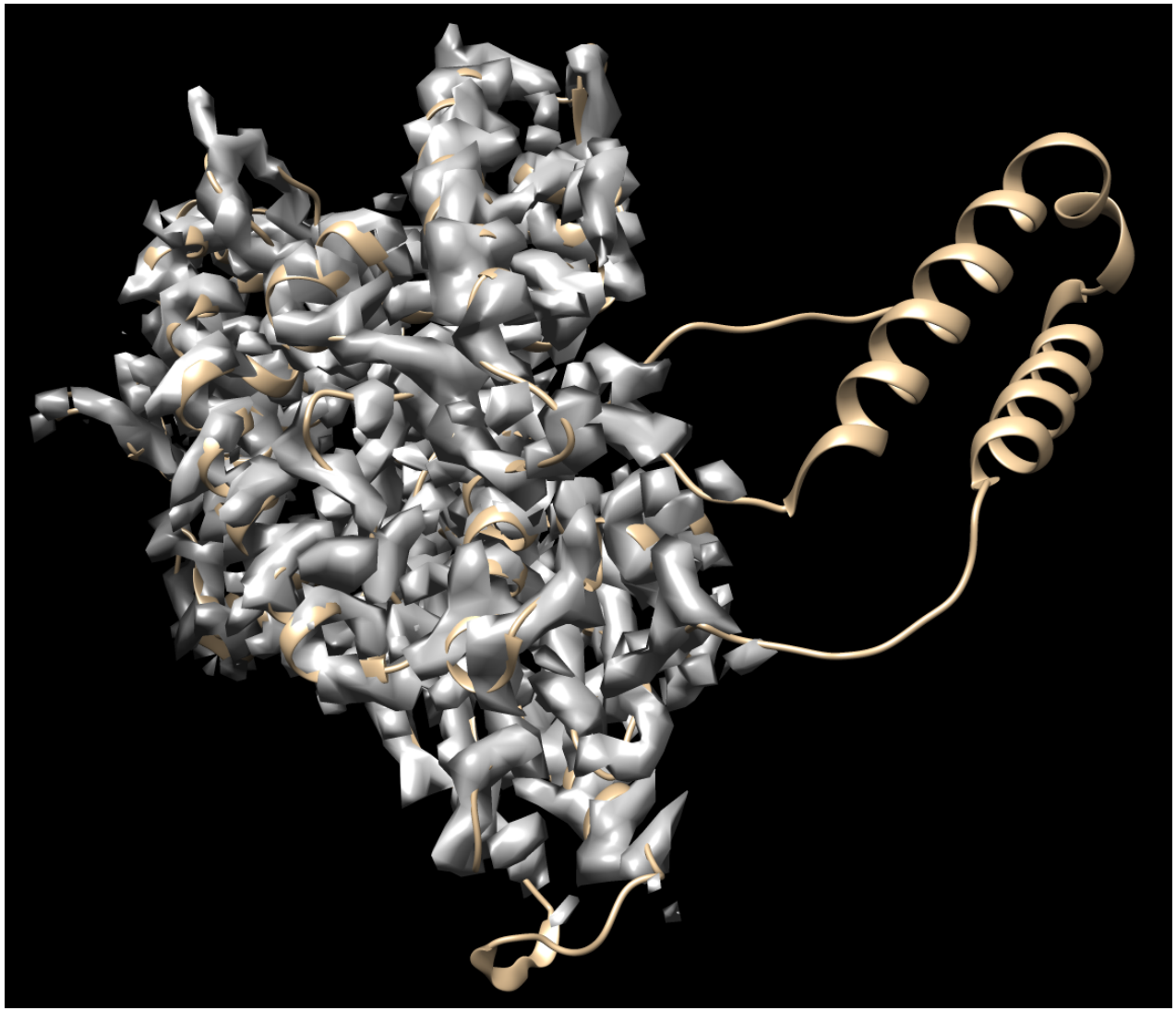
Read article **Similar articles**

Validation: model-to-map fit

3a5x (emd_1641) | 4.0Å | CC ≈ 0

Validation: model-to-map fit

3j9e (emd_6240) | 3.3Å | CC = 0.85 | Year: 2015



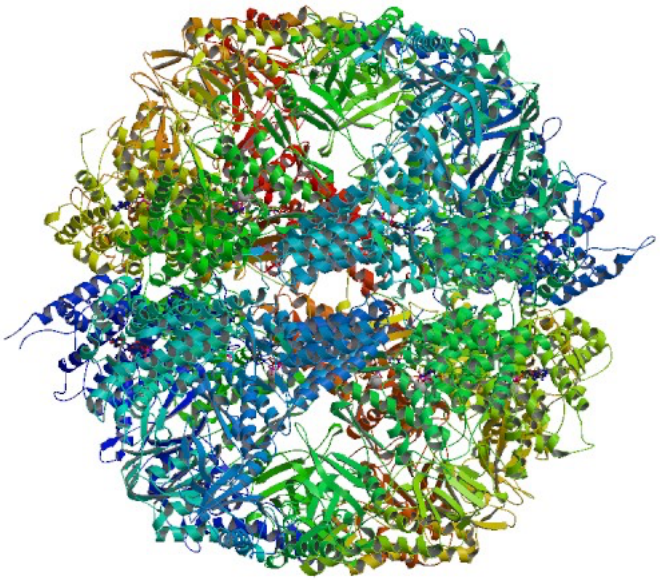
Latest trends, developments and… issues

Latest trends and developments

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

Latest trends and developments

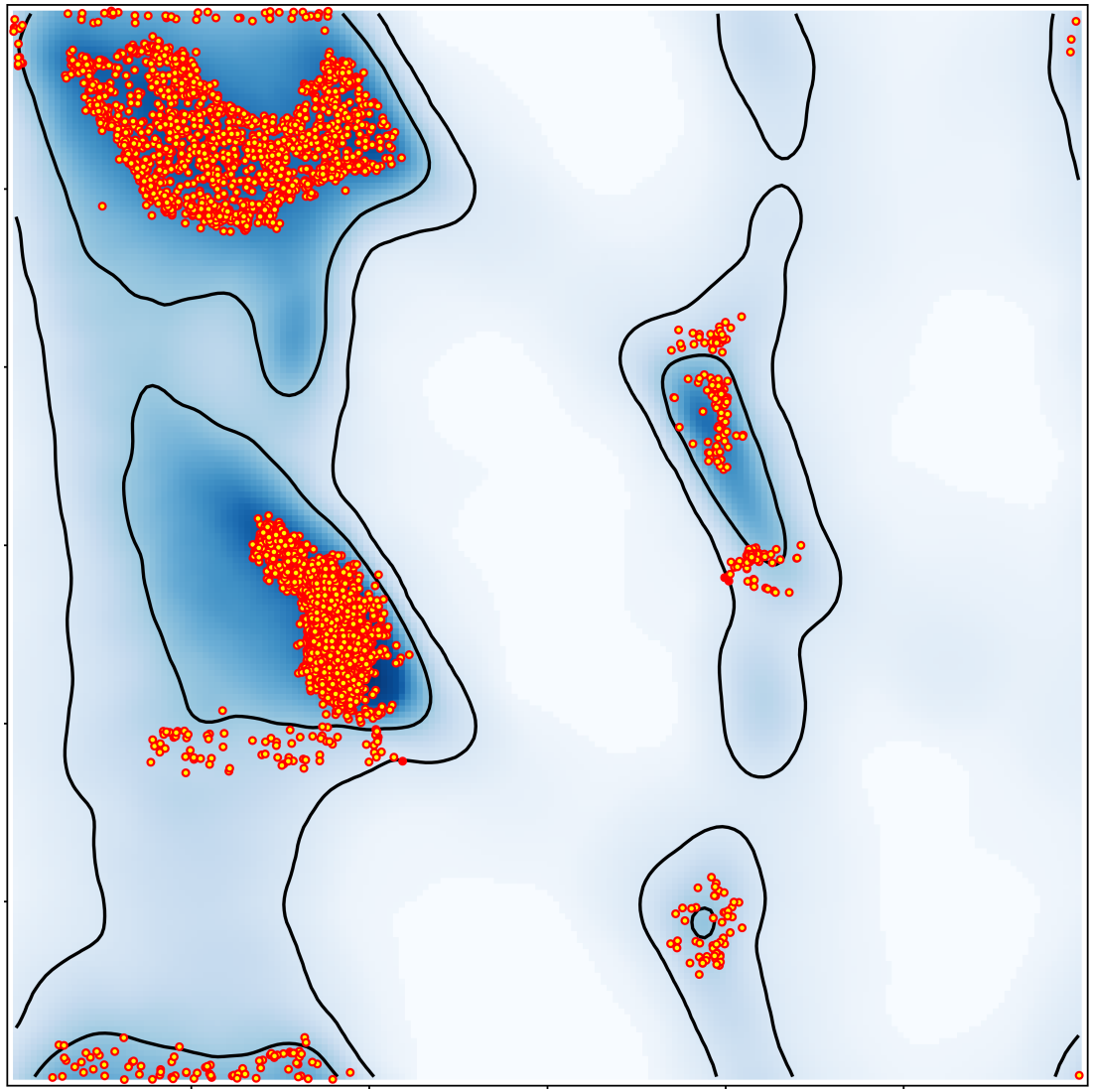
PNAS, 2019 116 (39) 19513-19522



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

- Perfect statistics

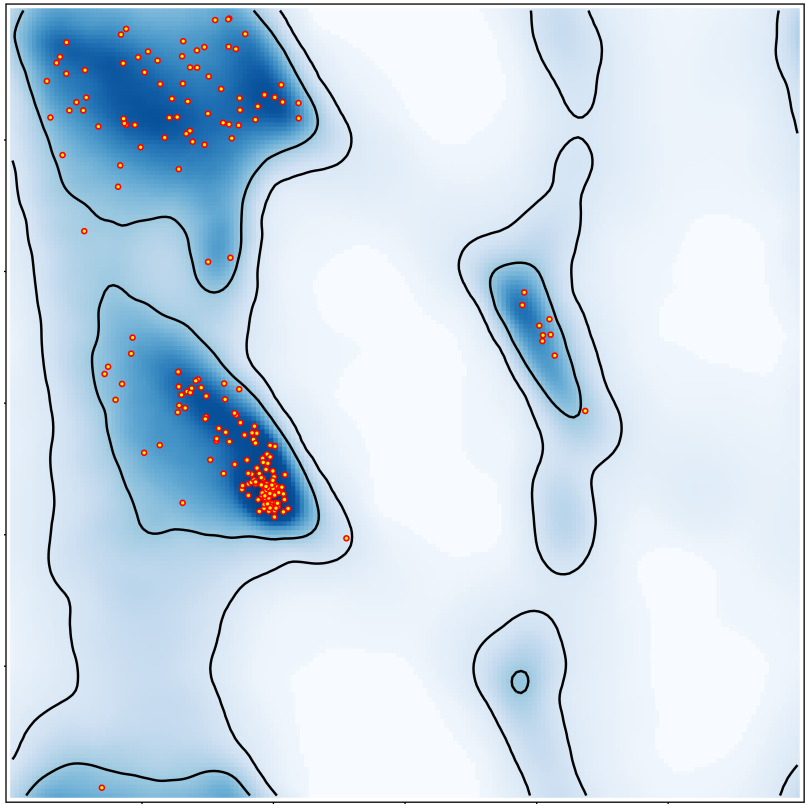
Latest trends and developments



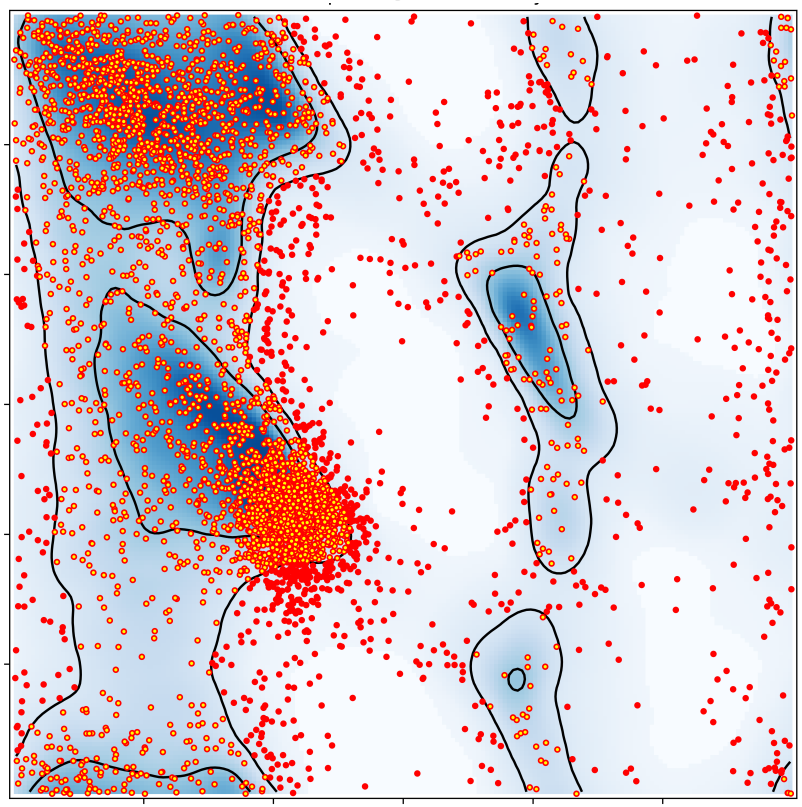
- Is it a good plot?

Example: good vs bad plots

1us0 | 0.66 Å



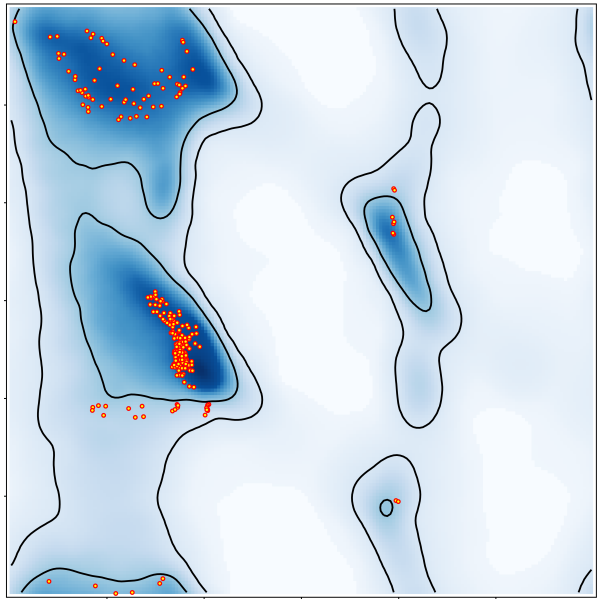
5a9z | 4.7 Å



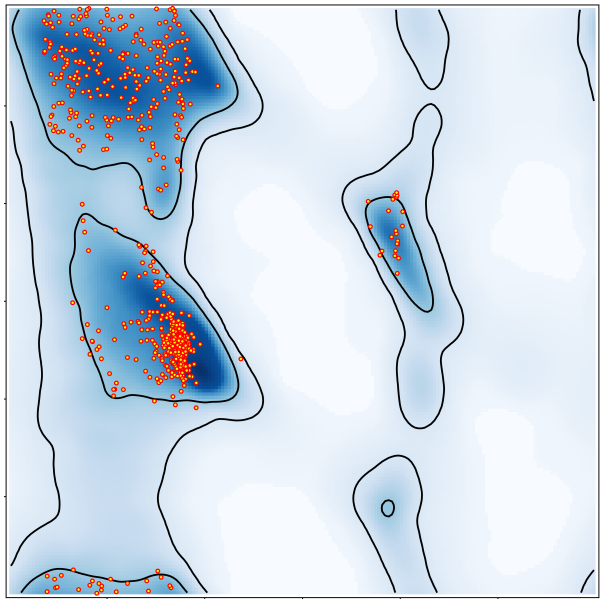
- A trained eye can distinguish good and bad plots

Example: poor plots

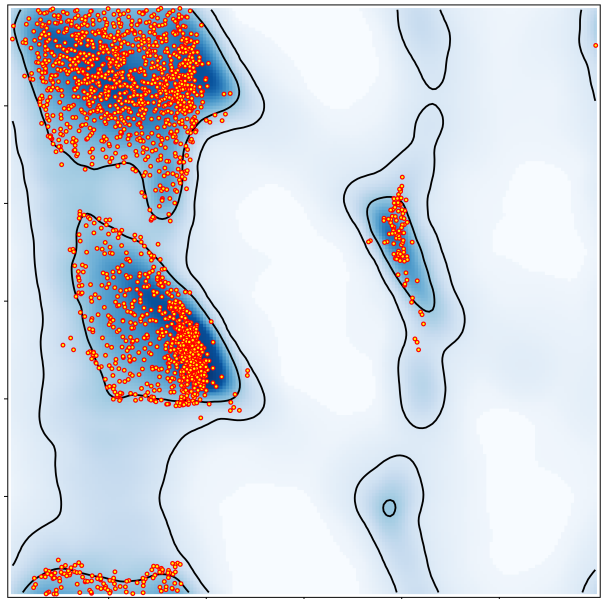
6bu9 | 6.8 Å



6dzv | 4.2 Å



6cs1 | 4.6 Å



- Overall Ramachandran plot counts (favored/outlier/allowed) will not flag this
- A trained eye required to appreciate the issue

- How to tell good vs bad without looking at the plot?

Ramachandran plot Z-score

CABIOS

Vol. 13 no. 4 1997
Pages 425–430

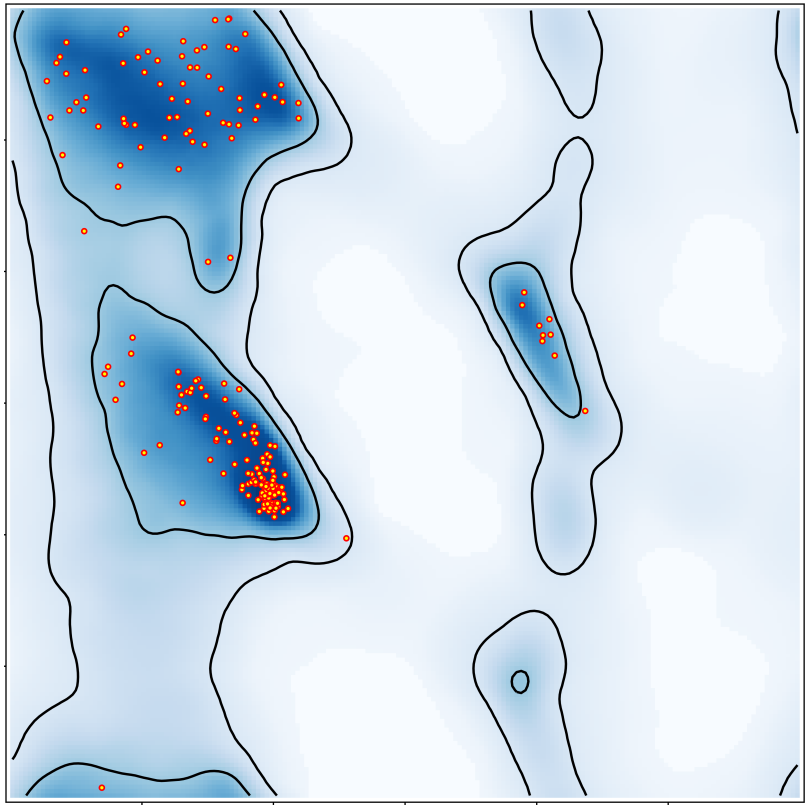
Objectively judging the quality of a protein structure from a Ramachandran plot

Rob W.W.Hooft, Chris Sander and Gerrit Vriend

- Ramachandran Z-score is good at identifying odd-looking Ramachandran plots!
  - Used in PDBREDO and WhatCheck. Implemented in *Phenix* (Oleg Sobolev)
    - One number, simple criteria:
      - Z<-3: Poor
      - -3 < Z < -2: Suspicious
      - Z > -2: Good

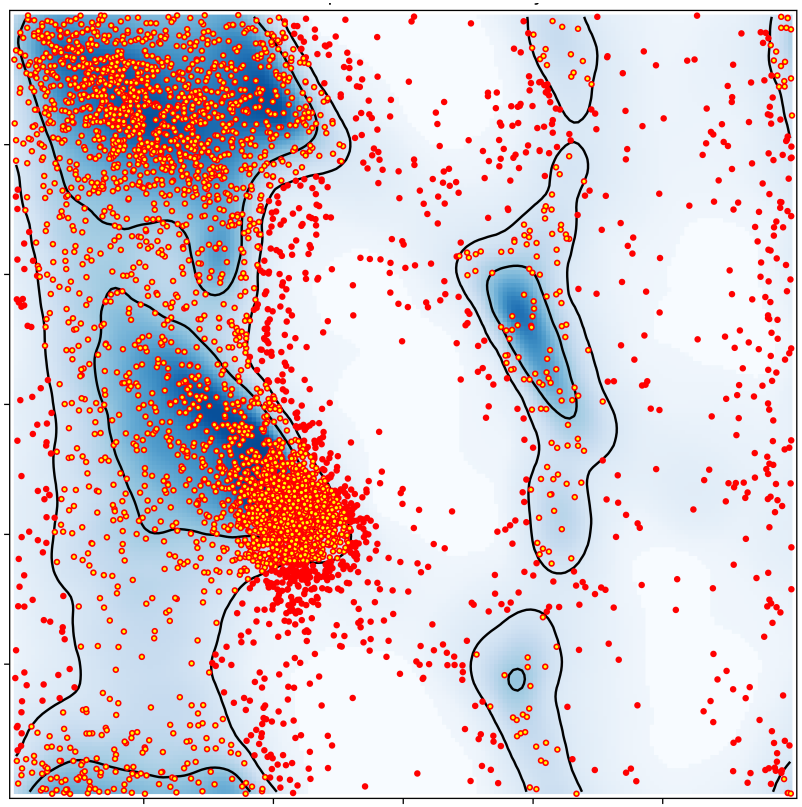
Example: good vs bad plots

1us0 | 0.66 Å



RamaZ = -0.5

5a9z | 4.7 Å

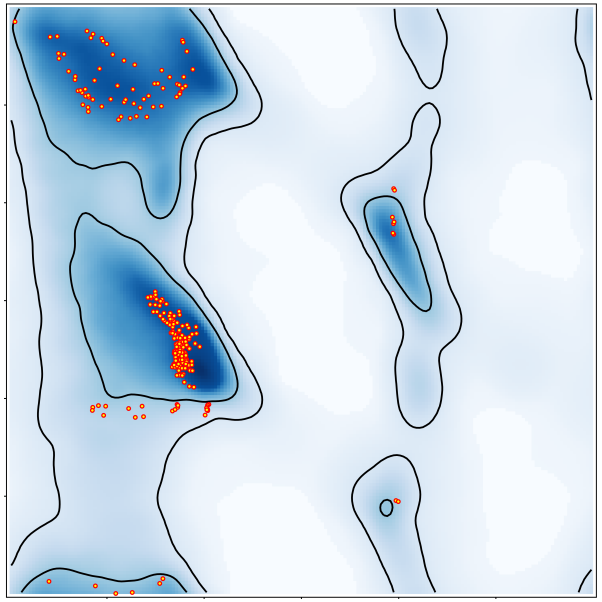


RamaZ = -7.7

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

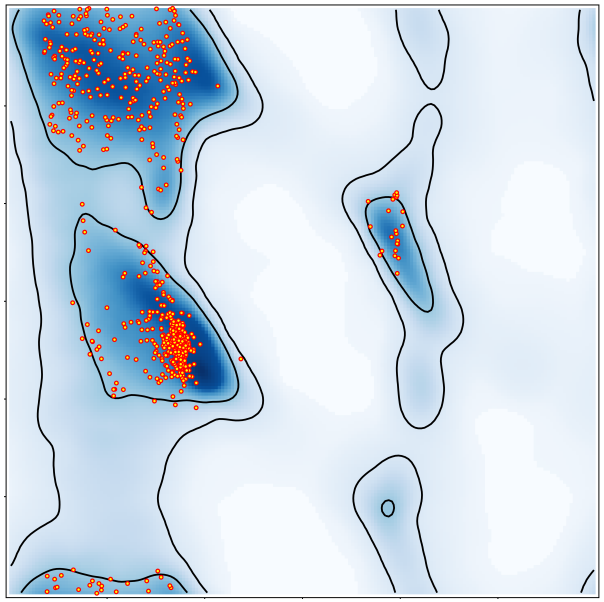
Example: odd plots

6bu9 | 6.8 Å



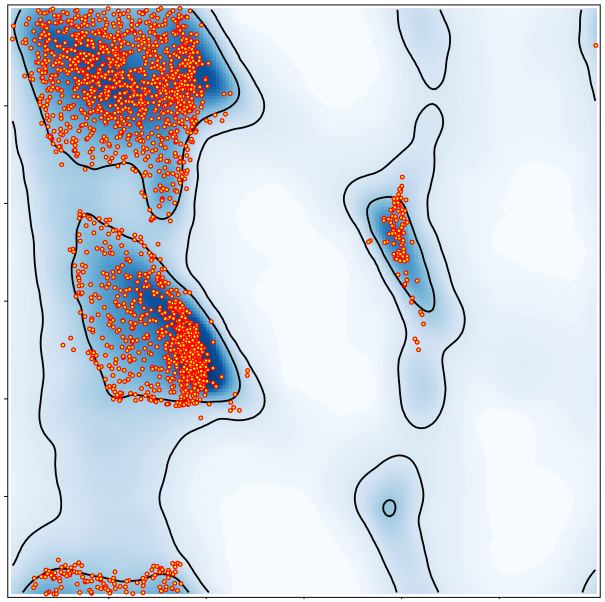
RamaZ = -5.0

6dzv | 4.2 Å



RamaZ = -4.1

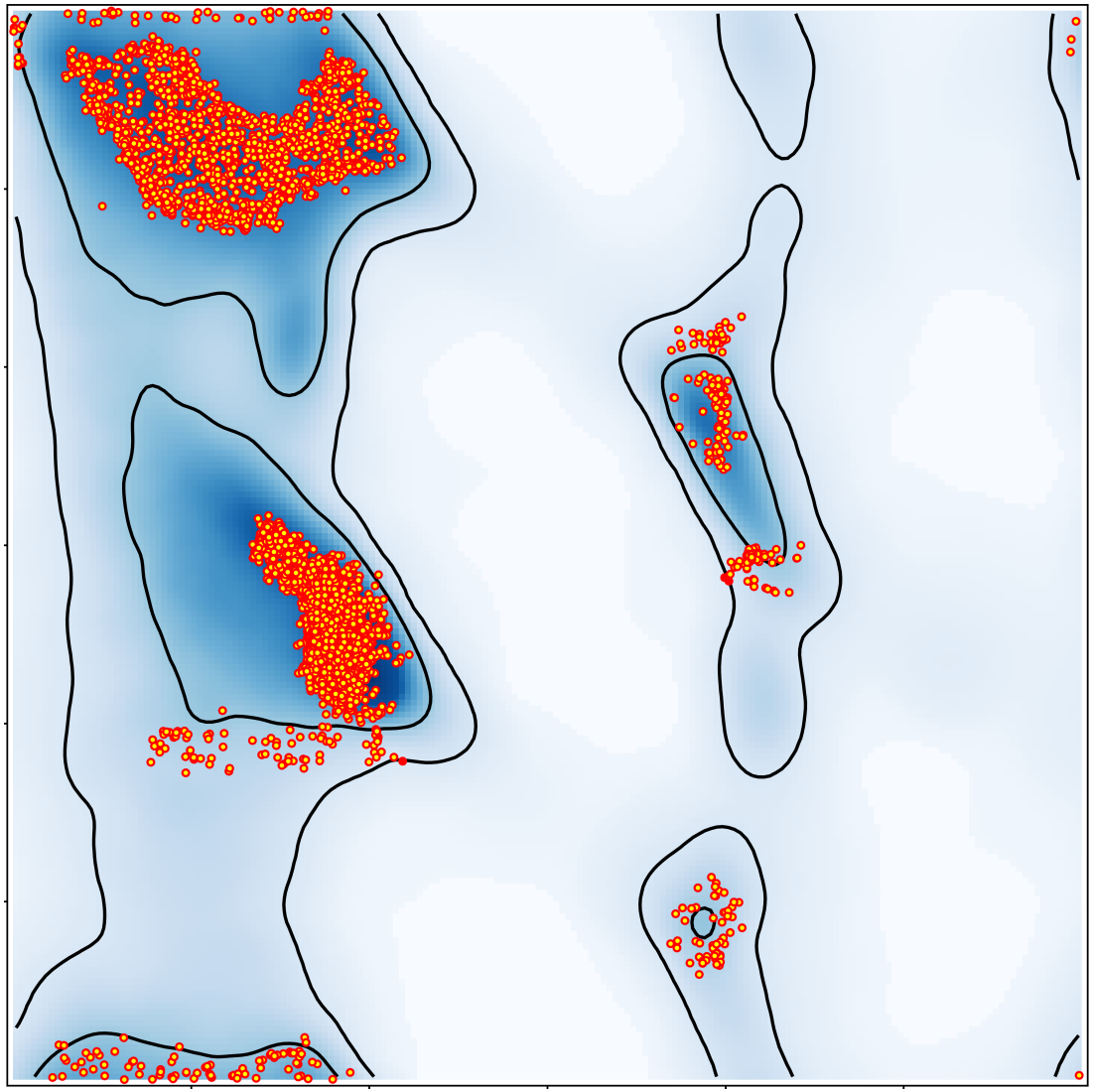
6cs1 | 4.6 Å



RamaZ = -4.2

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

6KS6

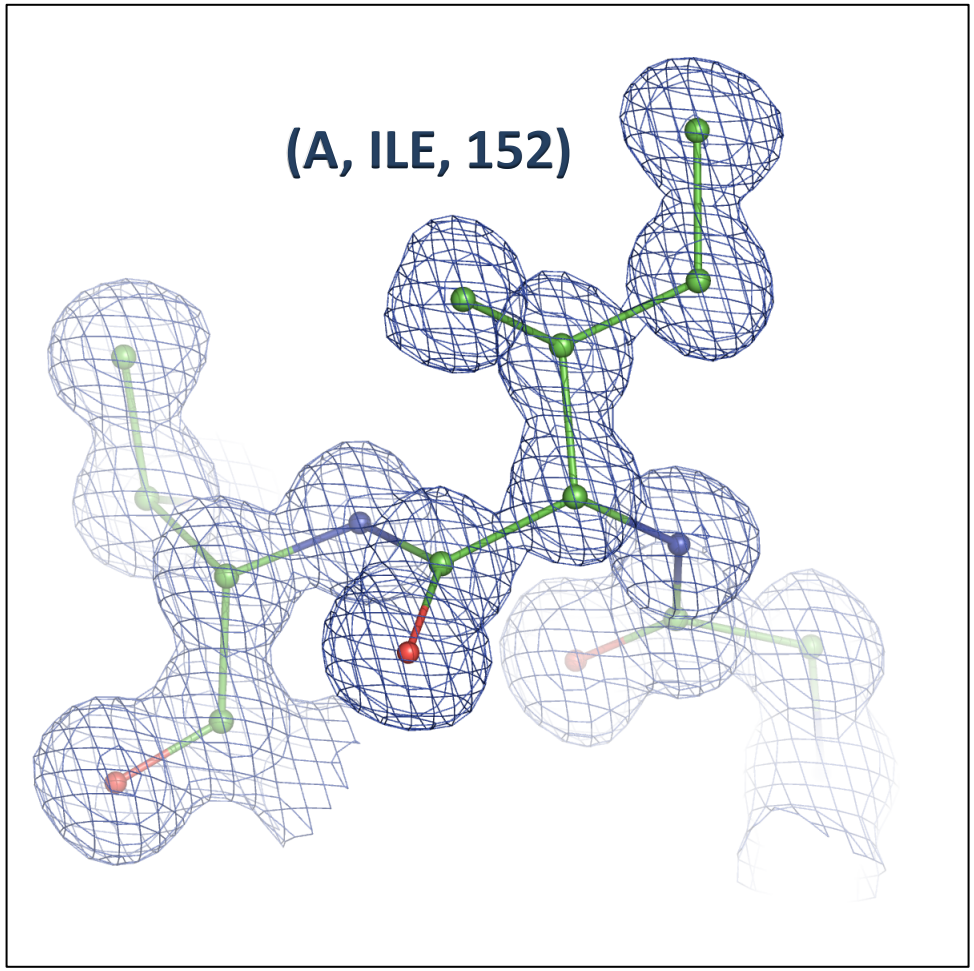


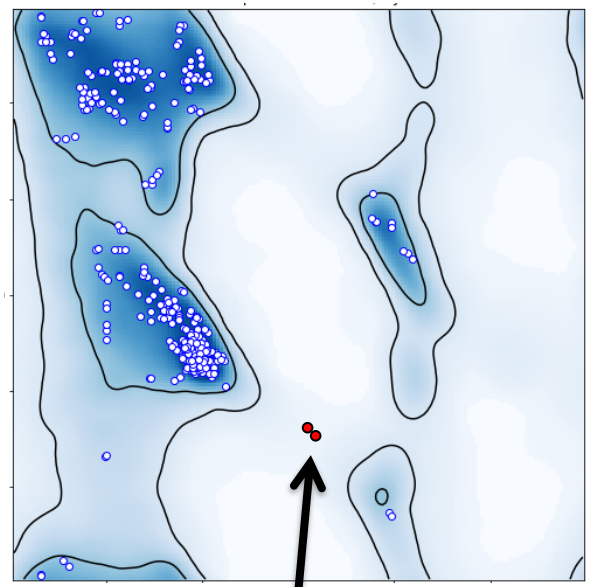
RamaZ = -3.3

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

Ramachandran plot

PDB code 3NOQ, 1 Å





Outliers:

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

Valid Ramachandran plot outliers: justified by the data (density map)

Ramachandran plot facts

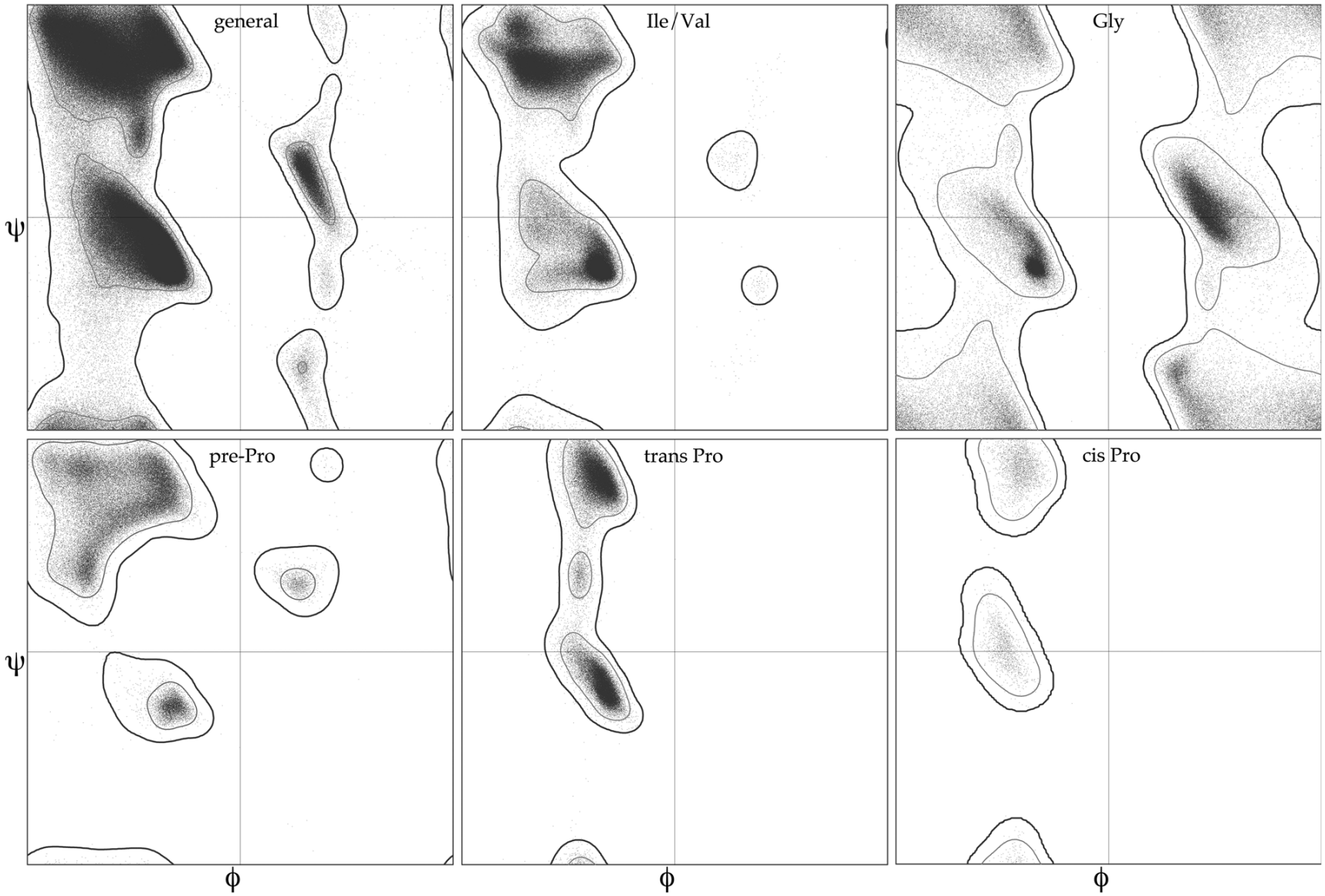
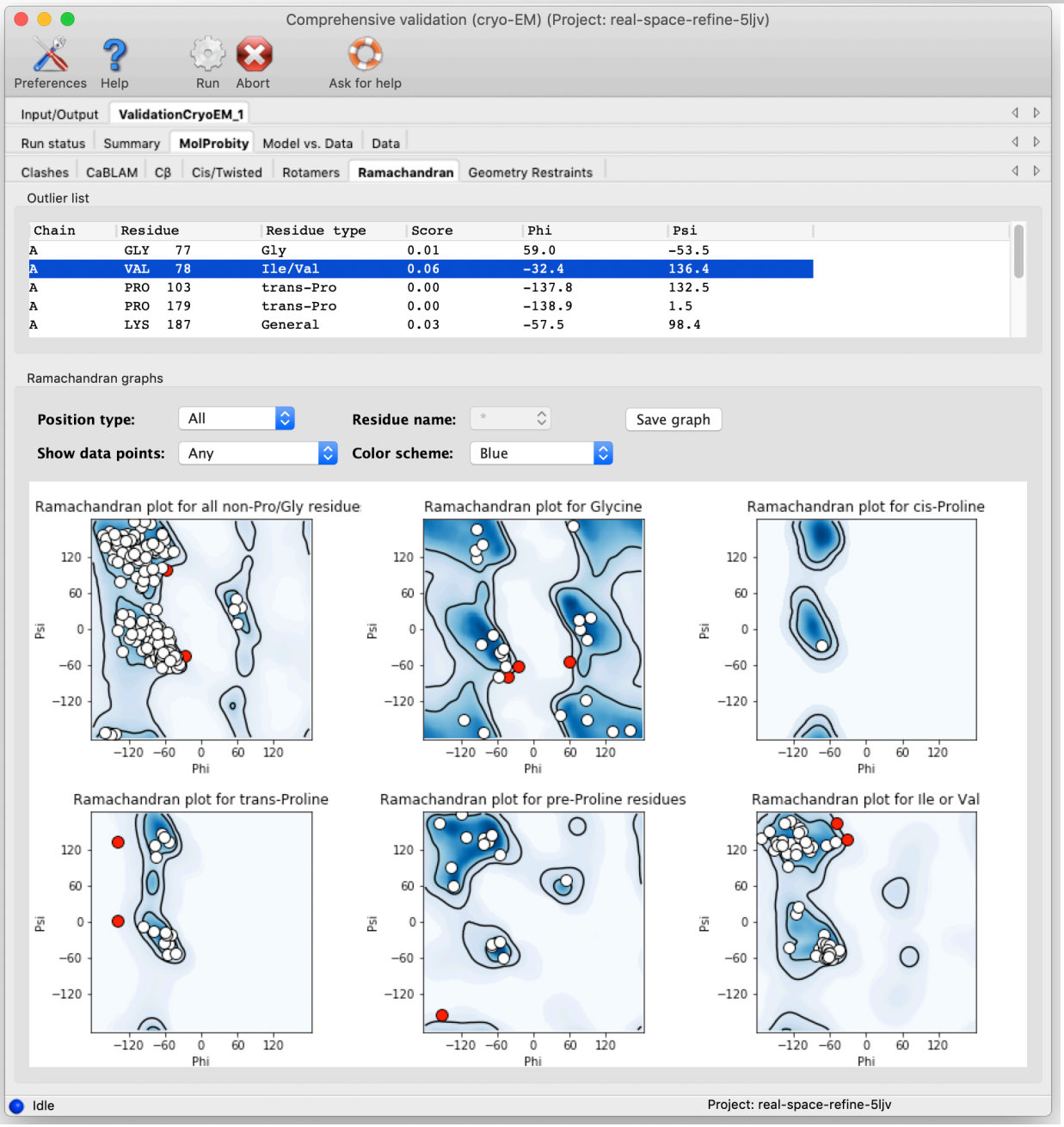


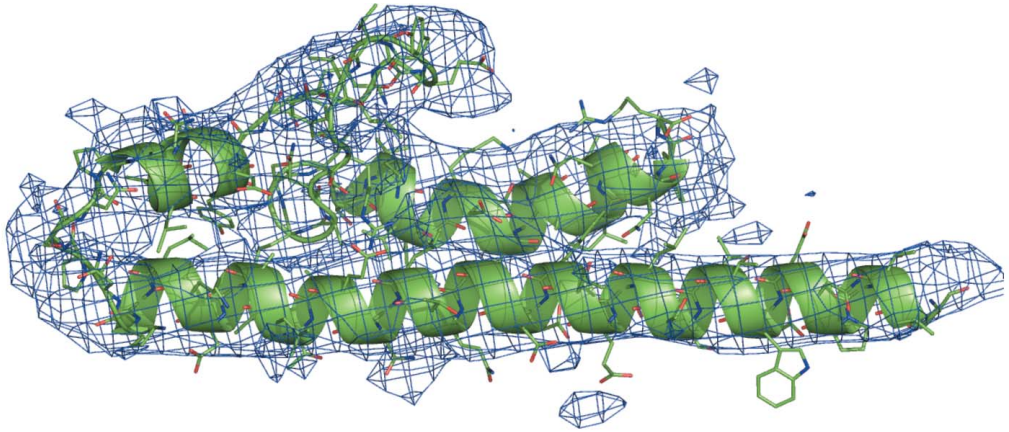
Image from Jane and David Richardson, Duke University

Ramachandran plot

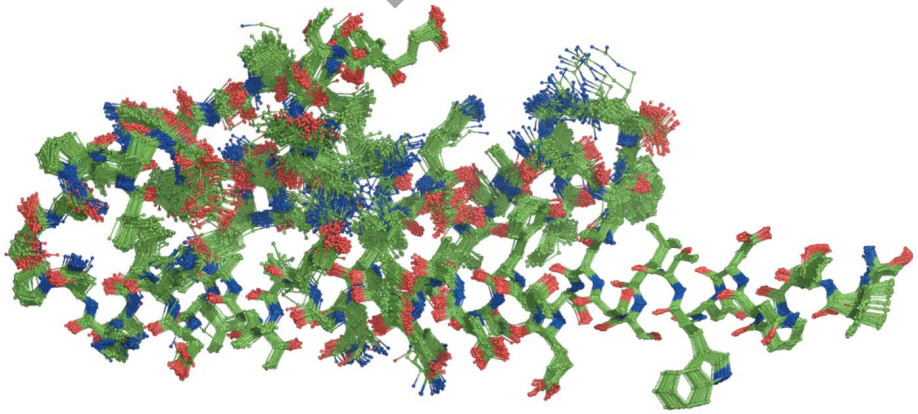


Multiple interpretation of low-res maps

- Low-resolution maps allow non-unique interpretation



phenix.mia



Model-map correlation coefficient (CC)

- Definition
  - With or w/o subtracting mean

$$CC(\rho_1,\rho_2)=\left(\sum_{\mathbf{n}}\left(\rho_1(\mathbf{n})\right)^2\right)^{-1/2}\left(\sum_{\mathbf{n}}\left(\rho_2(\mathbf{n})\right)^2\right)^{-1/2}\left(\sum_{\mathbf{n}}\rho_1(\mathbf{n})\rho_2(\mathbf{n})\right)$$

$$CC(\rho_1,\rho_2)=\left(\sum_{\mathbf{n}}\left(\rho_1(\mathbf{n})-\langle\rho_1\rangle\right)^2\right)^{-1/2}\left(\sum_{\mathbf{n}}\left(\rho_2(\mathbf{n})-\langle\rho_2\rangle\right)^2\right)^{-1/2}\left(\sum_{\mathbf{n}}\left(\rho_1(\mathbf{n})-\langle\rho_1\rangle\right)\left(\rho_2(\mathbf{n})-\langle\rho_2\rangle\right)\right)$$

- How model map is calculated
  - Approximation (e.g. N-gaussian)
  - Form-factors (electron vs crystallographic, eg. X-ray)
  - Fourier map
    - Box or sphere of Fourier map coefficients
- Region in the map used to calculate CC
  - Whole box
  - Mask around atoms
    - Atom radius