Validation: data analysis

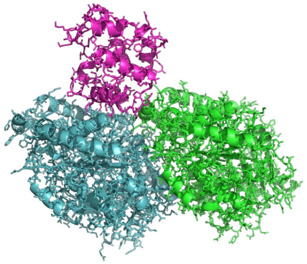
Pavel Afonine

Lawrence Berkeley National Laboratory (LBNL)

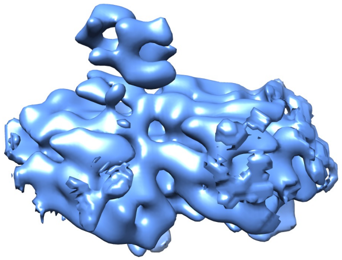
September 26th, 2024
BNL

Validation

Model

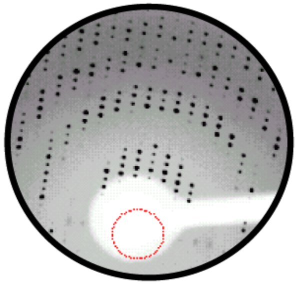


Data



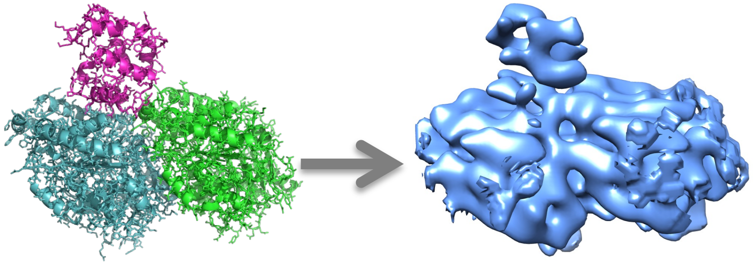
Cryo-EM

or



Diffraction

Model to data fit

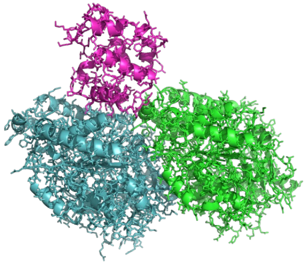


Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations

Validation tools: *Crystallography vs Cryo-EM*

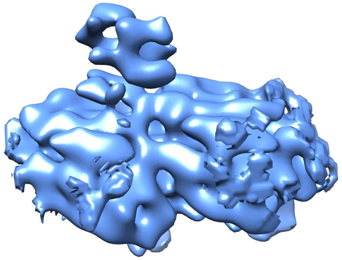
Exact same

Model



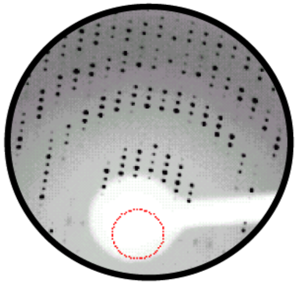
Different

Data



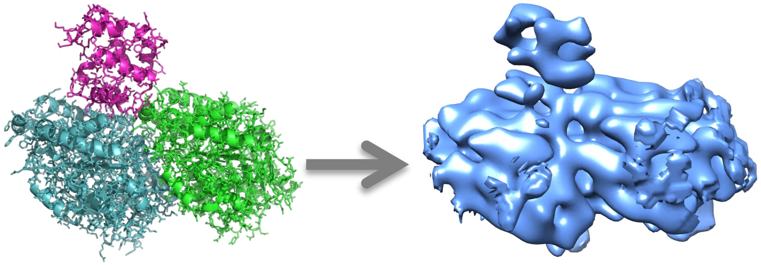
Cryo-EM

or



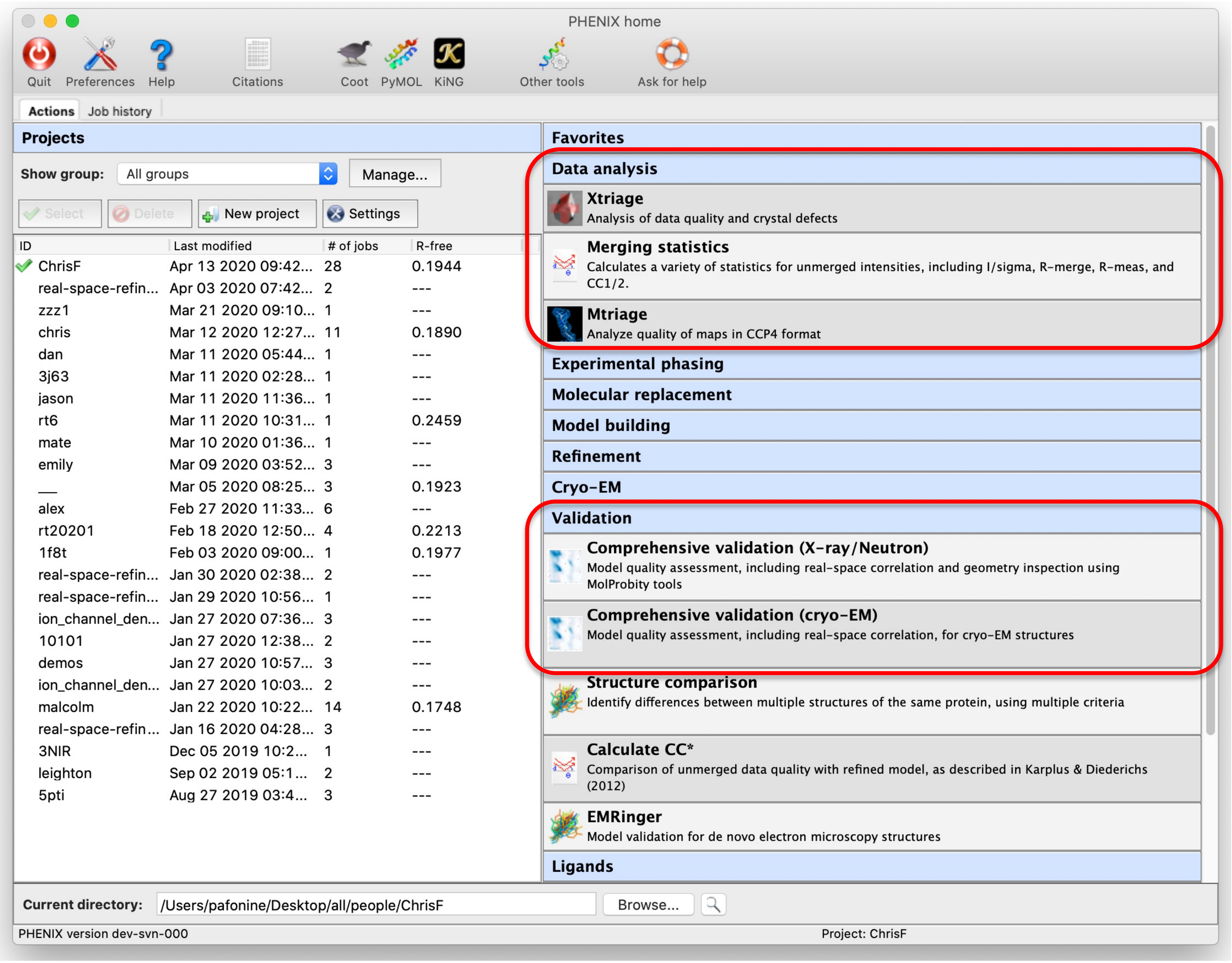
Diffraction

Model to data fit



Similar

Validation tools in Phenix



Projects

Show group: All groups

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

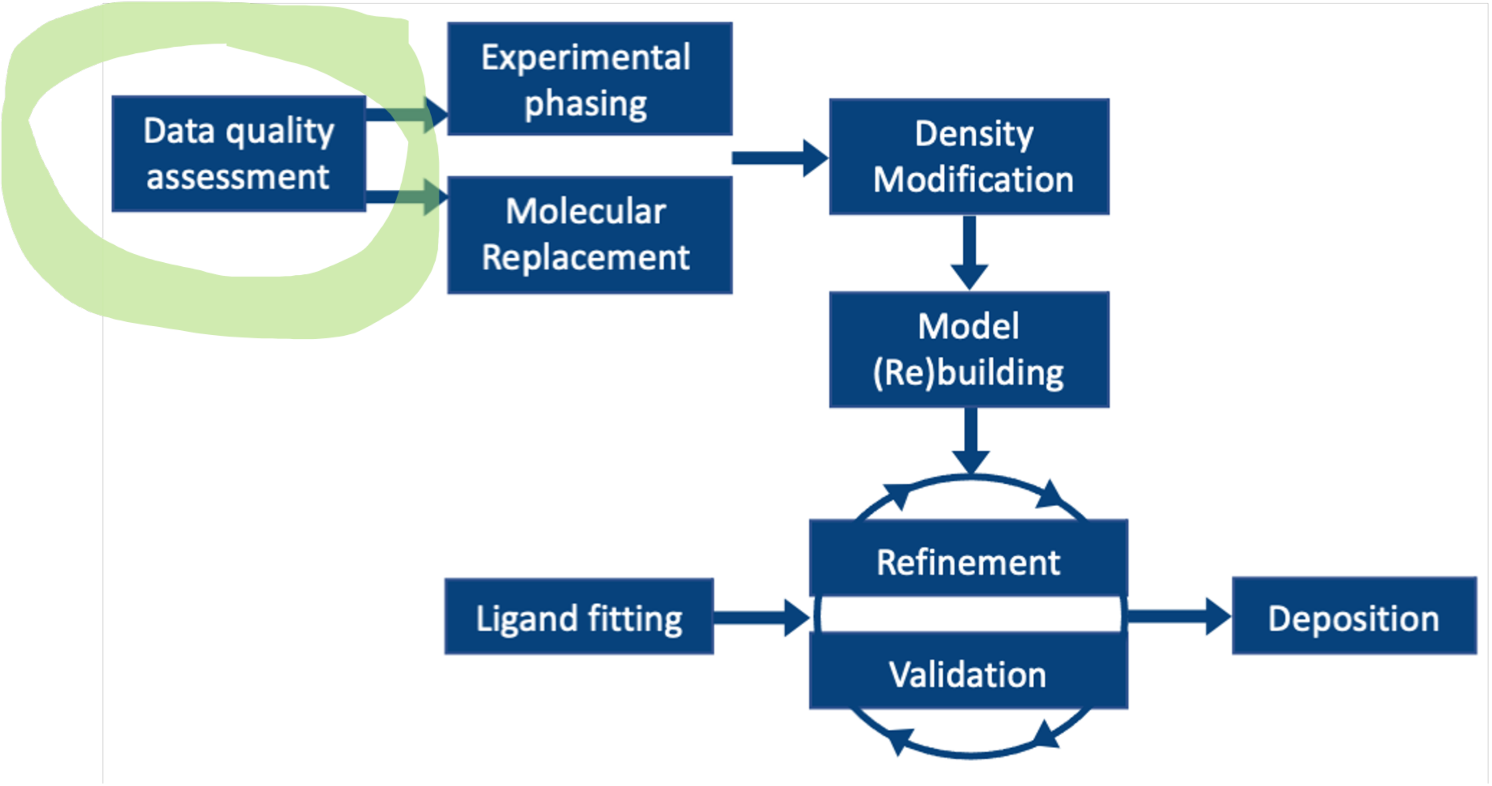
Current directory: /Users/pafonine/Desktop/all/people/ChrisF

PHENIX version dev-svn-000

Project: ChrisF

Xtriage: all about your Xtal data

Before doing anything else, you should validate your data!



Data quality assessment

Experimental phasing

Molecular Replacement

Density Modification

Model (Re)building

Refinement

Validation

Ligand fitting

Deposition

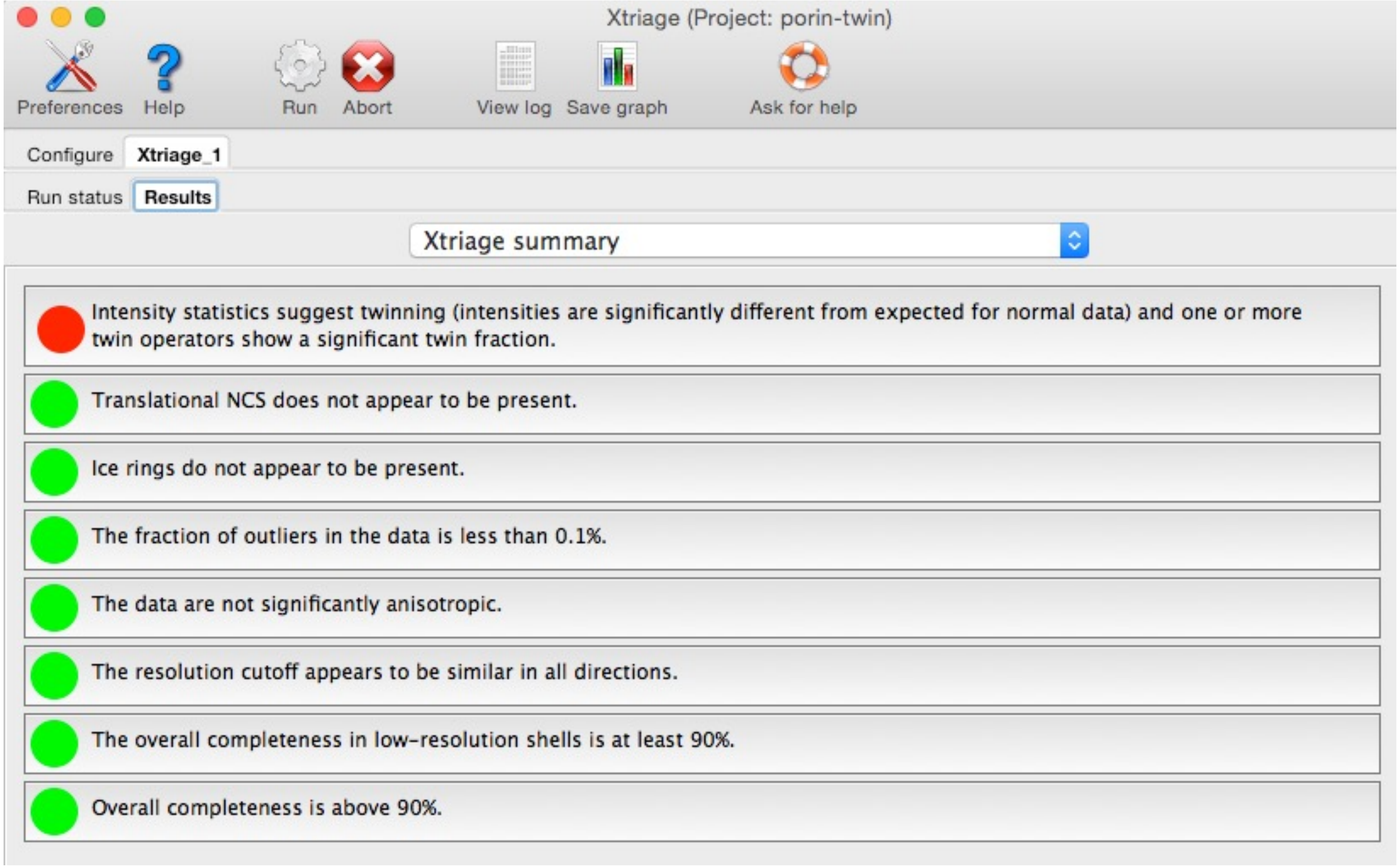
Possible experimental X-ray data problems

- Twining
- Translational NCS
- Wrong crystal
- Wrong space group
- Ice rings
- Data completeness
- Data anisotropy
- Resolution (overall, effective)
- Anomalous signal
- …

Xtriage: all about your Xtal data

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

Xtriage



🔴 Intensity statistics suggest twinning (intensities are significantly different from expected for normal data) and one or more twin operators show a significant twin fraction.

🟢 Translational NCS does not appear to be present.

🟢 Ice rings do not appear to be present.

🟢 The fraction of outliers in the data is less than 0.1%.

🟢 The data are not significantly anisotropic.

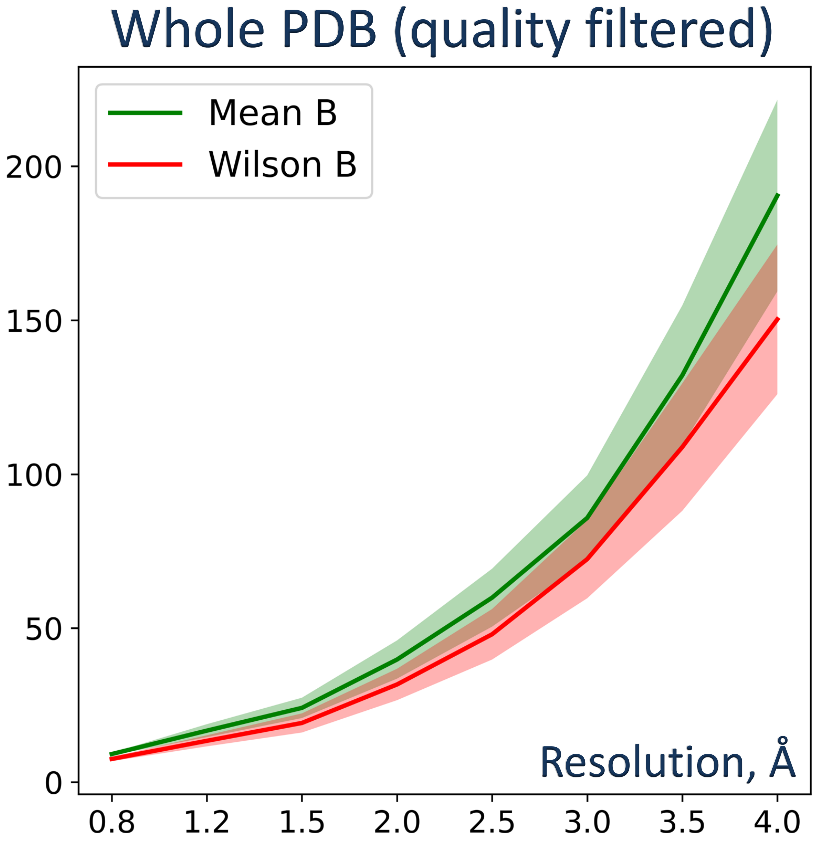
🟢 The resolution cutoff appears to be similar in all directions.

🟢 The overall completeness in low-resolution shells is at least 90%.

🟢 Overall completeness is above 90%.

Xtriage performs diagnostics for pathologies and data properties

Wilson B

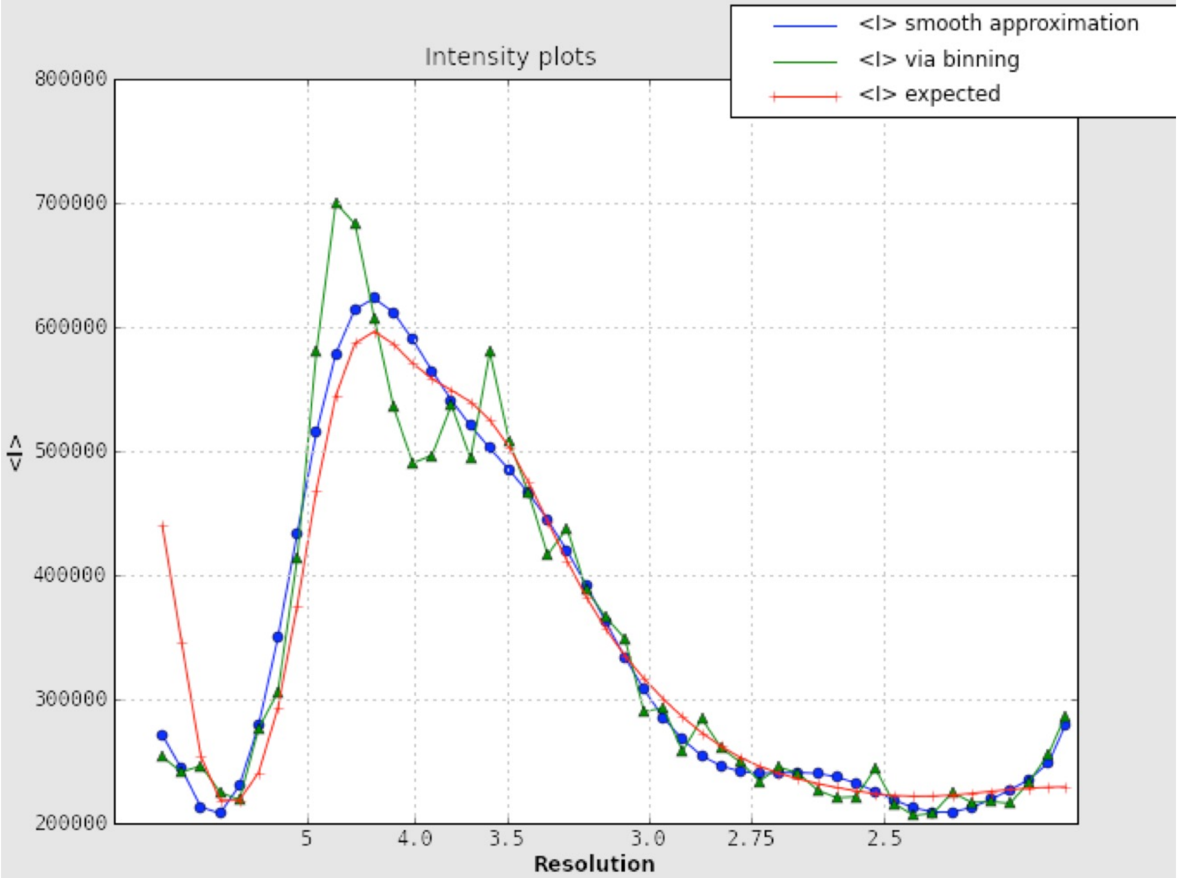


Wilson statistics assumes atoms of the same kind are randomly distributed in the unit cell and have the same isotropic B-factors

- 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

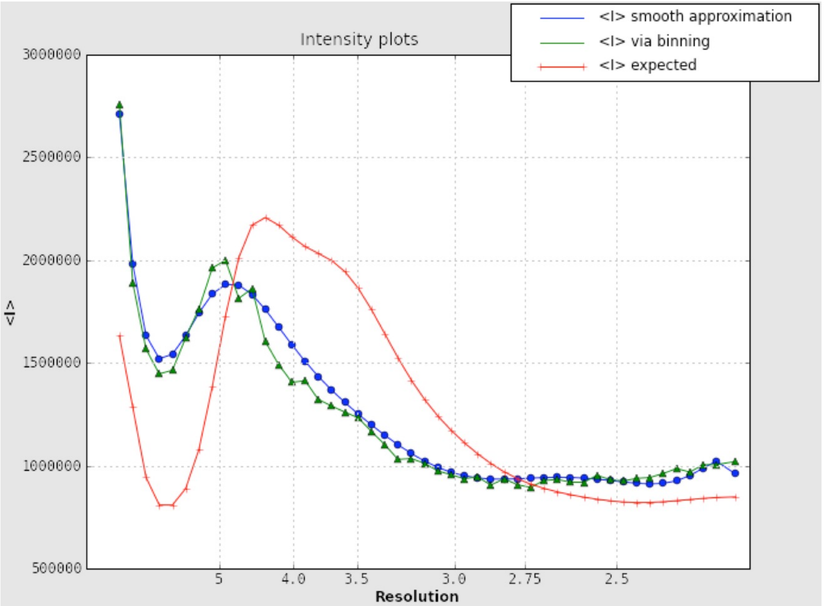
Wilson plot (mean intensity vs resolution)

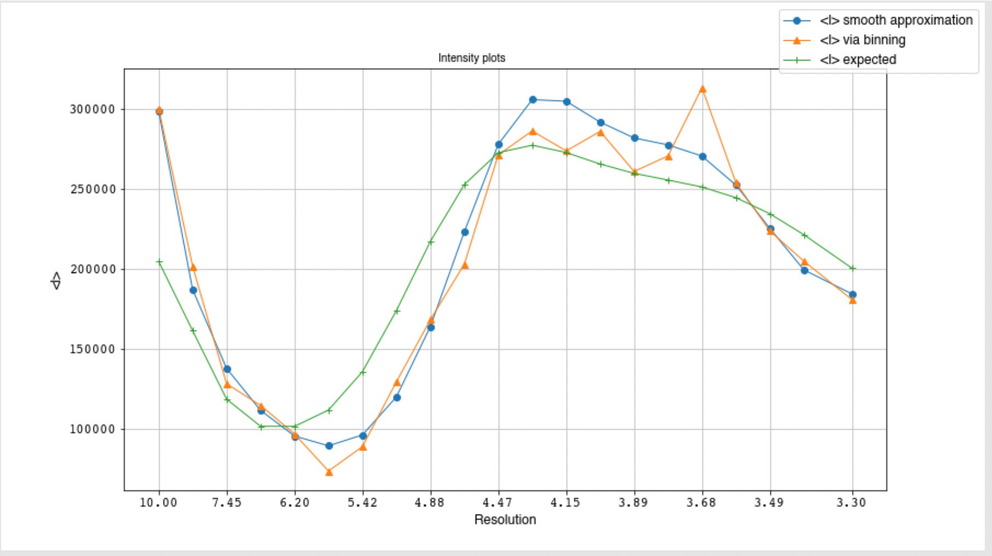
- The Wilson plot looks at mean intensity of diffraction by resolution, a curve which has a predictable shape



Wilson plot (mean intensity vs resolution)

- Main reasons for deviations from expected distribution
  - Bad data (e.g., ice rings or poor data processing
  - Macromolecule that doesn’t look like the average protein
  - Looking at only a part of the plot (e.g., low-resolution data)

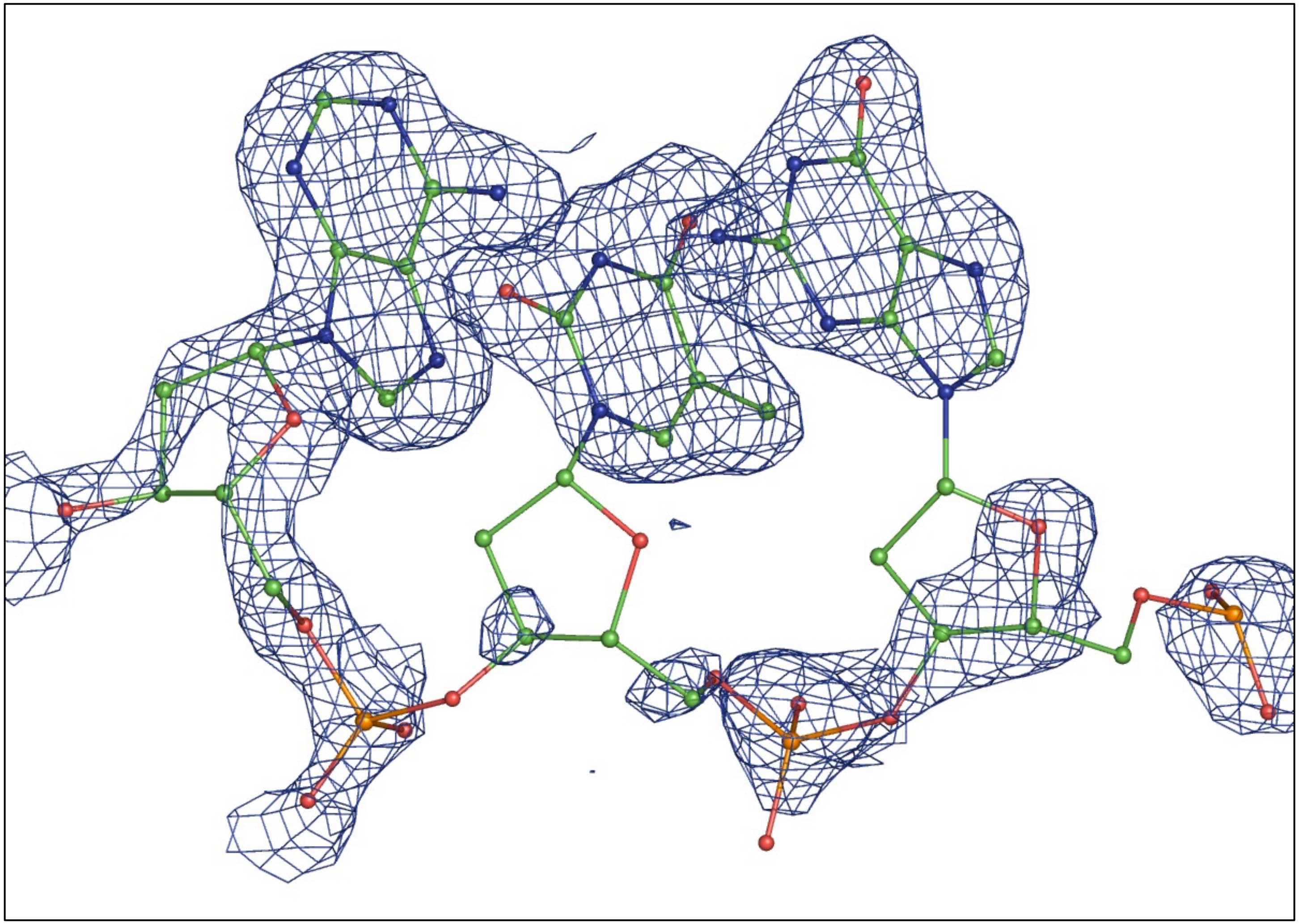




Data completeness

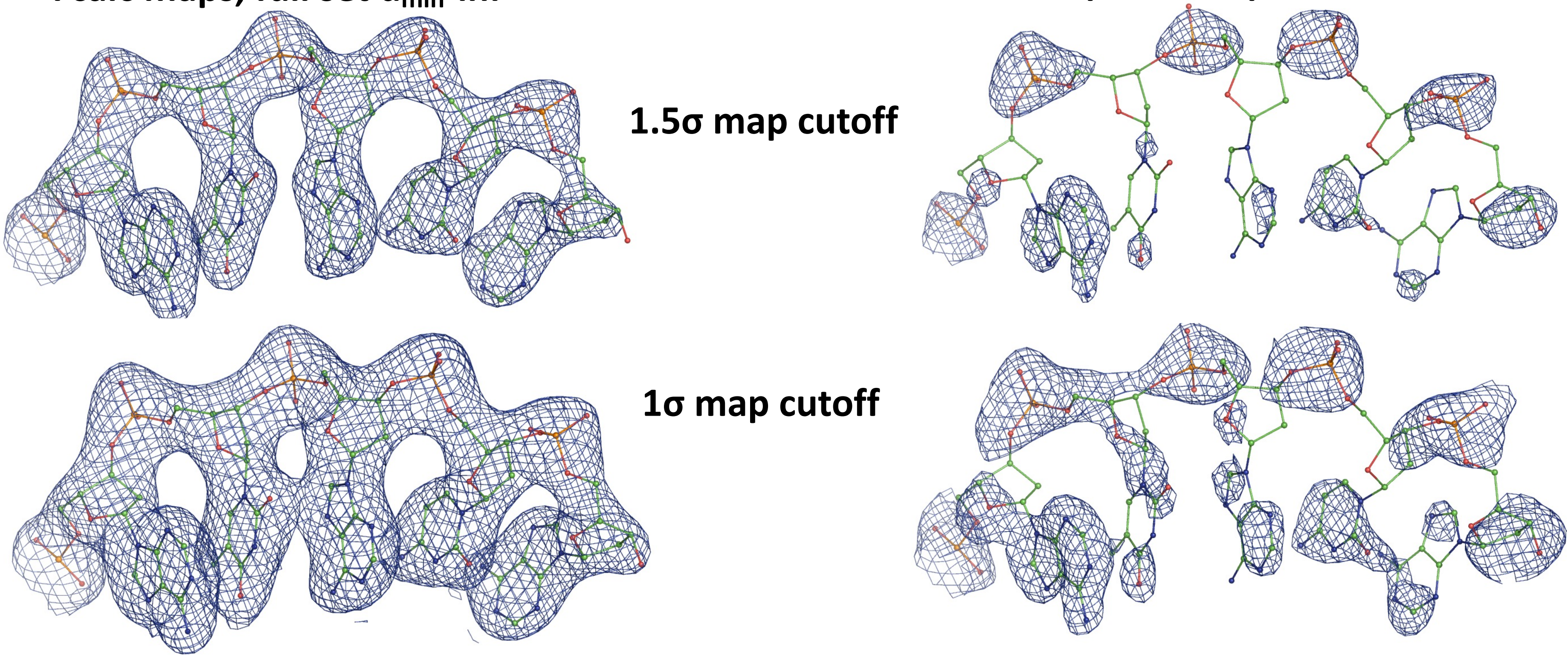
- PDB code: 1NH2, resolution 1.9Å, showing E6-E8

2mFo-DFc , 1σ



Data completeness

```
Completeness by resolution:
 19.9274 -  3.2441 0.78
  3.2441 -  2.5767 0.99
  2.5767 -  2.2515 1.00
  2.2515 -  2.0459 1.00
  2.0459 -  1.8993 0.99
Overall completeness in dmin-inf: 0.95
```



Fcalc maps, full set d_{min}-inf

Fcalc maps, incomplete set

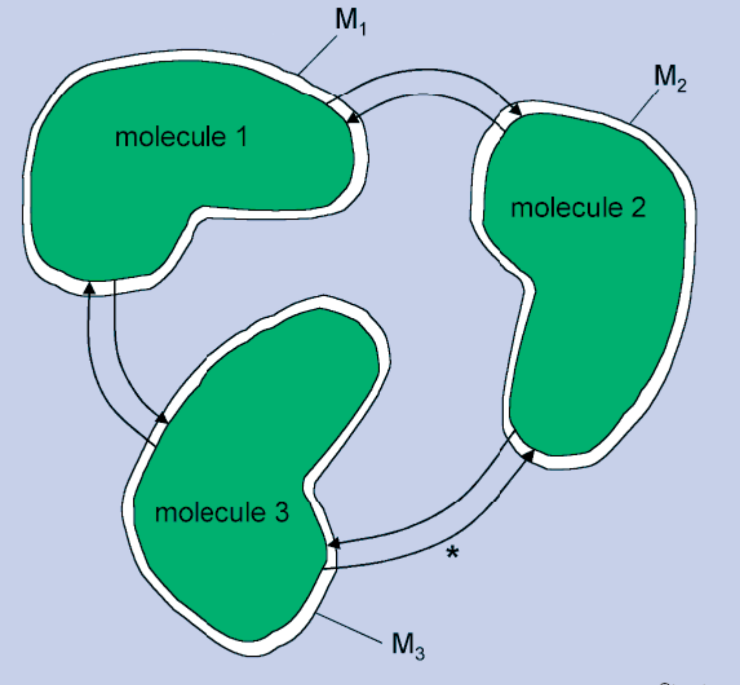
1.5σ map cutoff

1σ map cutoff

Systematic data incompleteness can distort maps

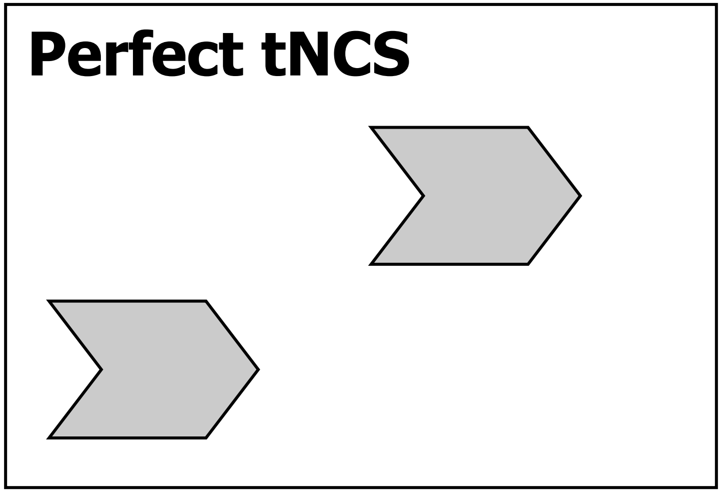
Non-crystallographic symmetry NCS

- Two or more molecules in the ASU related by rotation-translation
- NCS is found in about 1/3 to 1/2 of crystal structures
- Usually helps solving/refining models at medium-to-low resolution
- A special case of NCS, translational NCS (tNCS) leads to complications

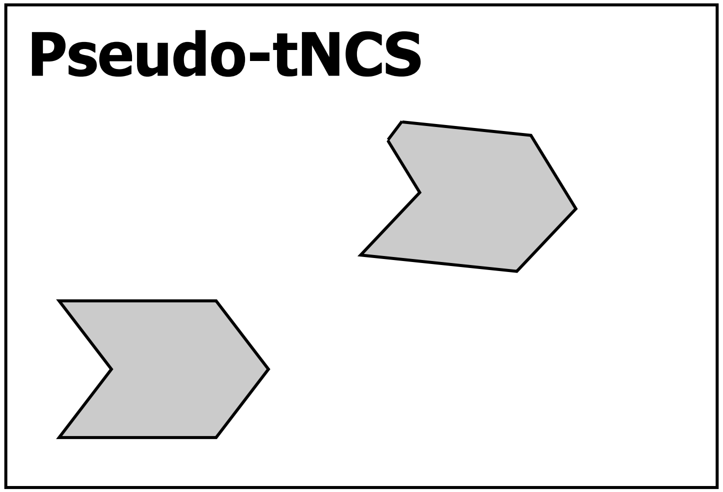


Translational NCS (tNCS)

- tNCS arises when two or more crystallographically independent copies are in the same (or nearly the same) orientation in the unit cell and can be superimposed by a translation that does not correspond to any symmetry operation in the space group.



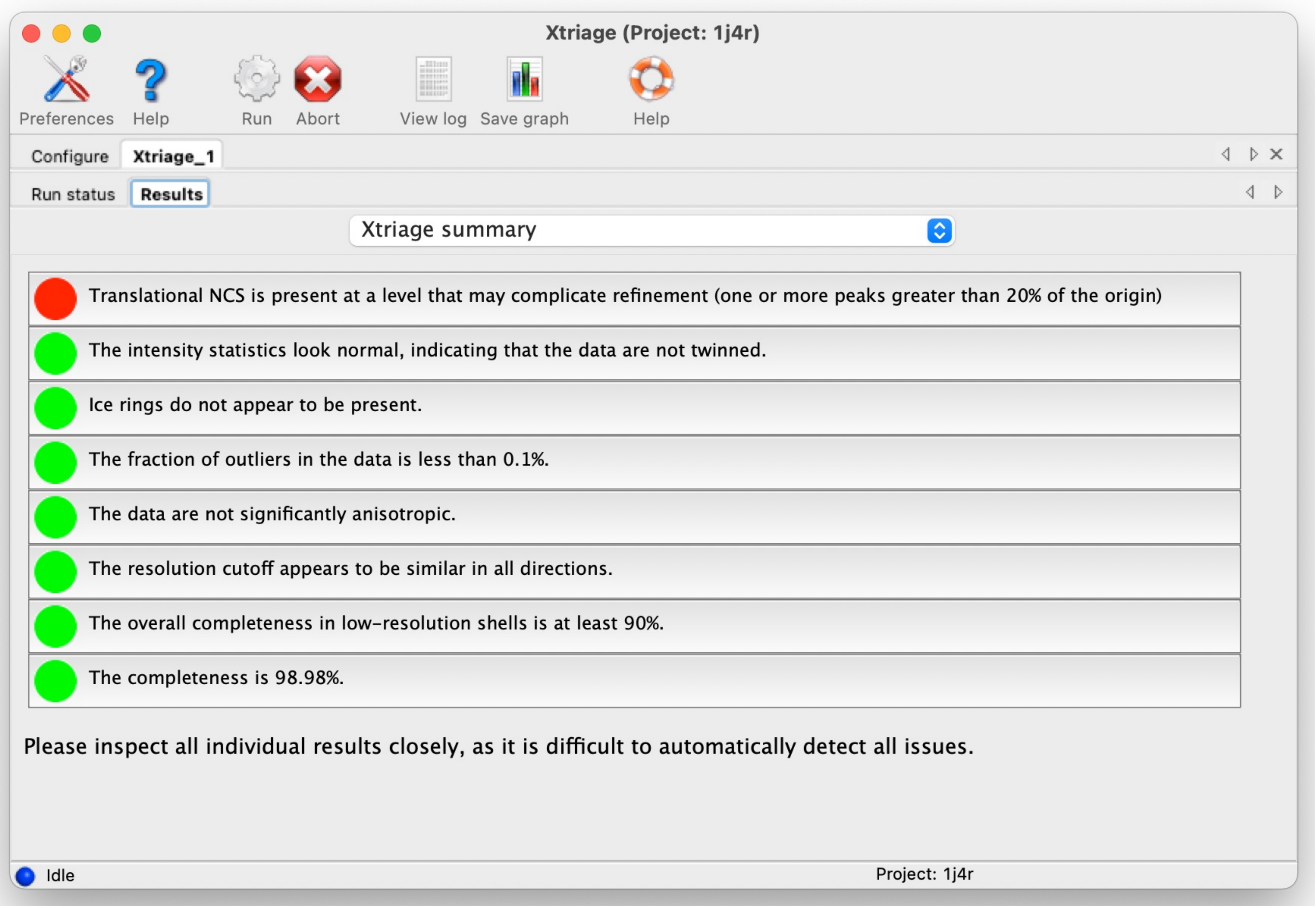
Perfect tNCS



Pseudo-tNCS

- Used to complicate MR (no it is taken care of)
- Risk to bias OMIT map

Translational NCS (tNCS)



🟥 Translational NCS is present at a level that may complicate refinement (one or more peaks greater than 20% of the origin)

🟩 The intensity statistics look normal, indicating that the data are not twinned.

🟩 Ice rings do not appear to be present.

🟩 The fraction of outliers in the data is less than 0.1%.

🟩 The data are not significantly anisotropic.

🟩 The resolution cutoff appears to be similar in all directions.

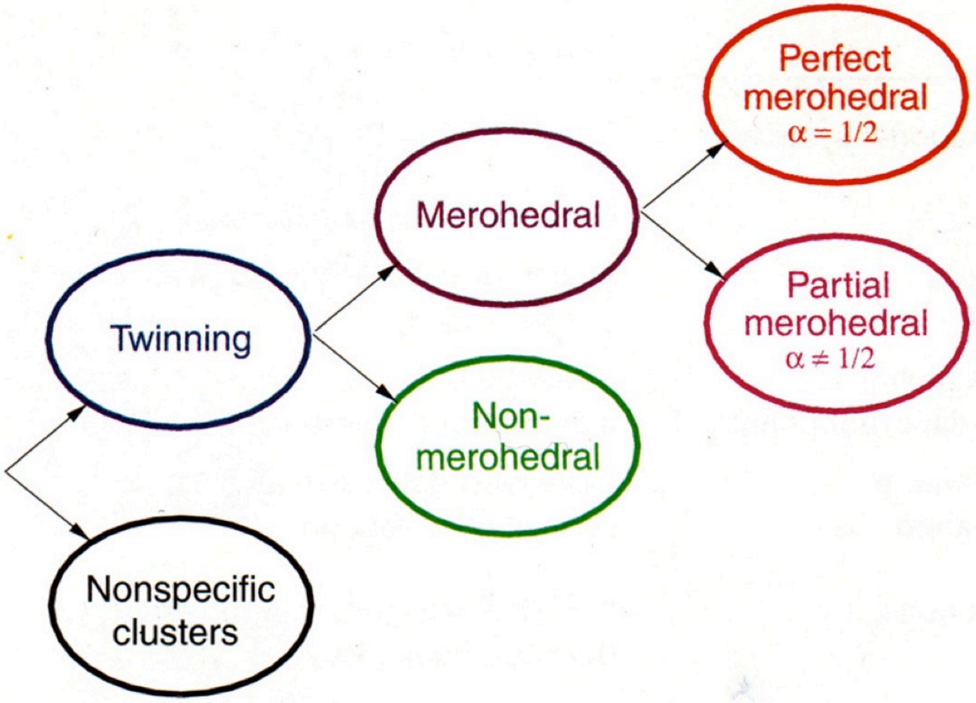
🟩 The overall completeness in low-resolution shells is at least 90%.

🟩 The completeness is 98.98%.

Please inspect all individual results closely, as it is difficult to automatically detect all issues.

Twinning

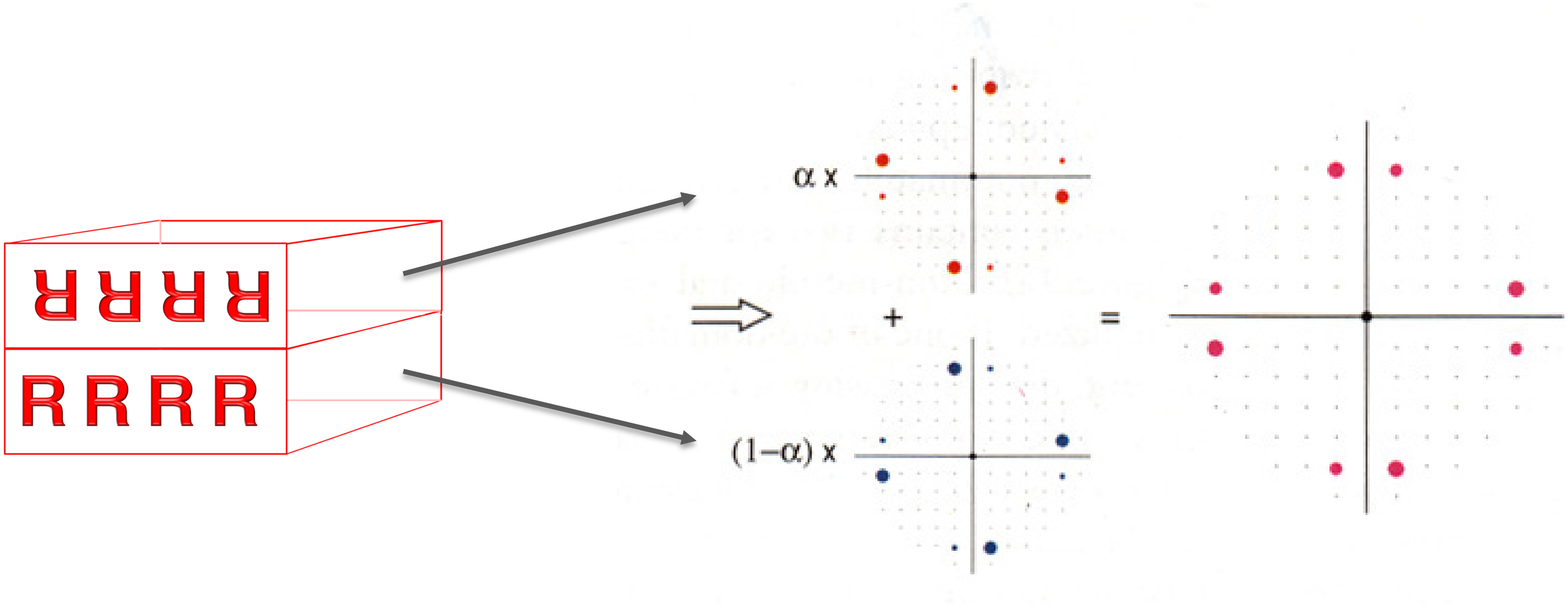
- Twinning is a crystal growth disorder



Typically only merohedral twinning is dealt with in a meaningful way in macromolecules

Twinning

- Merohedral twining occurs when your crystal is composed of identical but rotated crystals combined together such that their lattices matching



- Observed intensity is a weighted sum of individual intensities:

$$I_{\mathrm{OBS}}(\mathbf{h}) = \alpha_1 I(\mathbf{h}) + \ldots + \alpha_N I(\mathbf{T}_N \mathbf{h})$$

$$\alpha_1 + \ldots + \alpha_N = 1$$

Twinning

- Twinning parameterization
  - **Twin law** describes orientation of different species relative to each other (rotation matrix T that transforms hkl indices of one species into the other)
  - **Twin fraction (α):** fractional contribution of each component
    - Estimated by Xtriage
    - Refined by phenix.refine

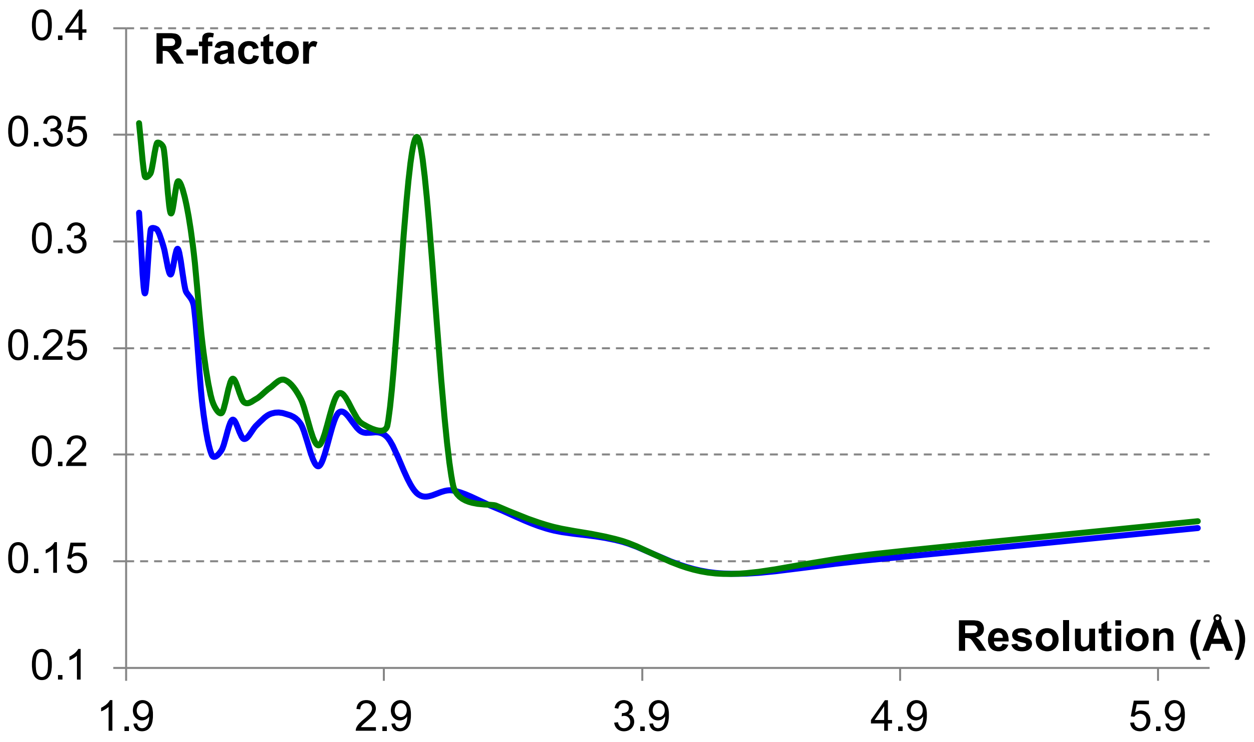
$$I_{\mathrm{OBS}}(\mathbf{h}) = \alpha_1 I(\mathbf{h}) + \ldots + \alpha_N I(\mathbf{T}_N \mathbf{h})$$

$$\alpha_1 + \ldots + \alpha_N = 1$$

Twinning

- tNCS can mask effects of twinning
- If both are present, intensity distributions may look like normal
  - First check for tNCS and use different test for twinning (L-test)
- If crystal is twinned, you have lost information
- Maps going to have model bias that is worse than usual
- Experimental phasing may be difficult
- False symmetry may appear

Watch for outliers



- 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
```