Validation: data analysis

Pavel Afonine

phenix-online.org

lbl.gov

qrefine.com

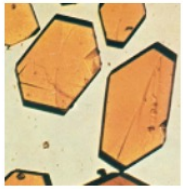
Missoula, MT

June 27th 2024

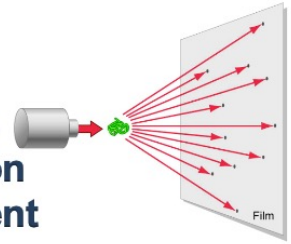
Solving structure by crystallography

Pure sample

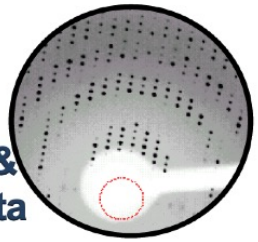
Crystals



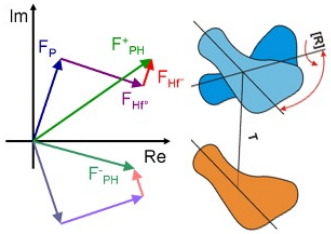
Diffraction experiment



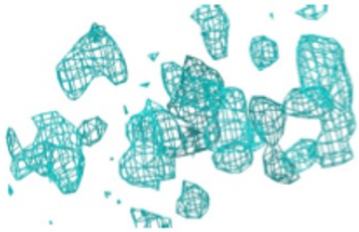
Process & check data



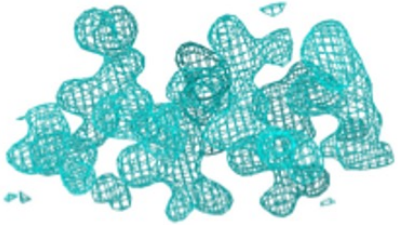
Solve phase problem



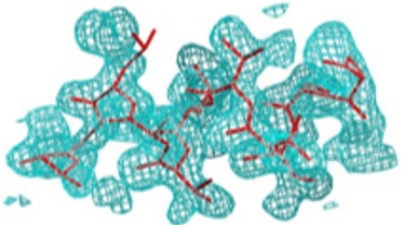
Initial map



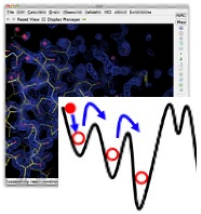
Improved map



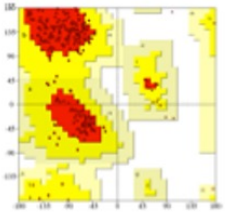
Build model



Refine model



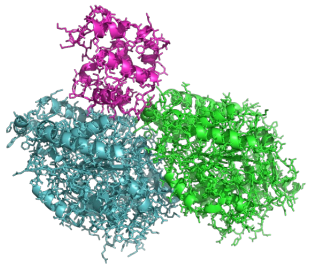
Validate model



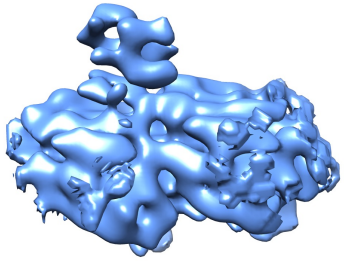
Publish & deposit

Validation

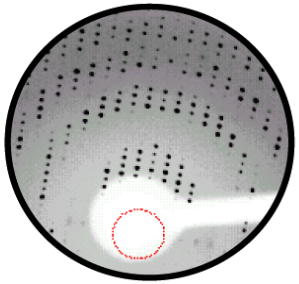
Model



Data



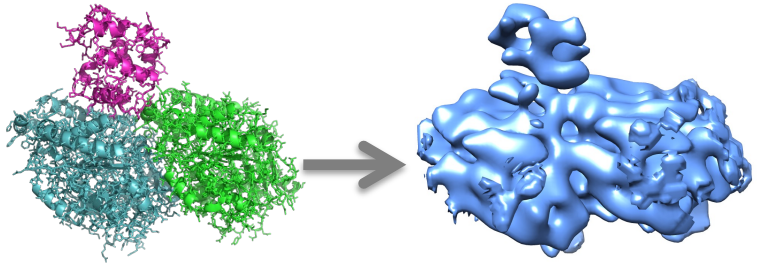
or



Cryo-EM

Diffraction

Model to data fit

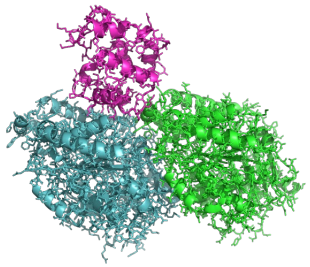


Validation = checking model, data and model-to-data fit are all make sense and obey to prior expectations

Validation tools: *Crystallography vs Cryo-EM*

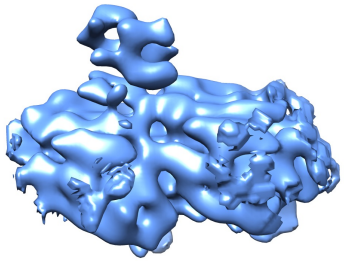
Exact same

Model

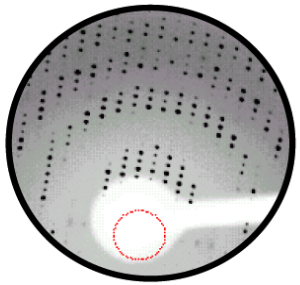


Different

Data



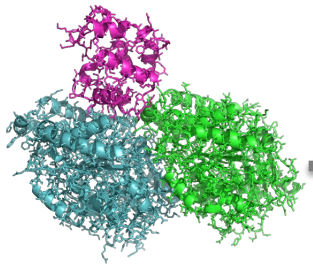
or

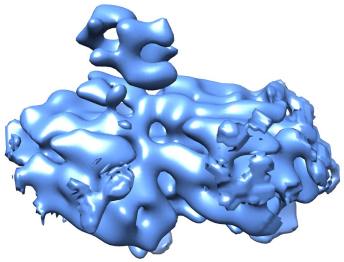


Cryo-EM

Diffraction

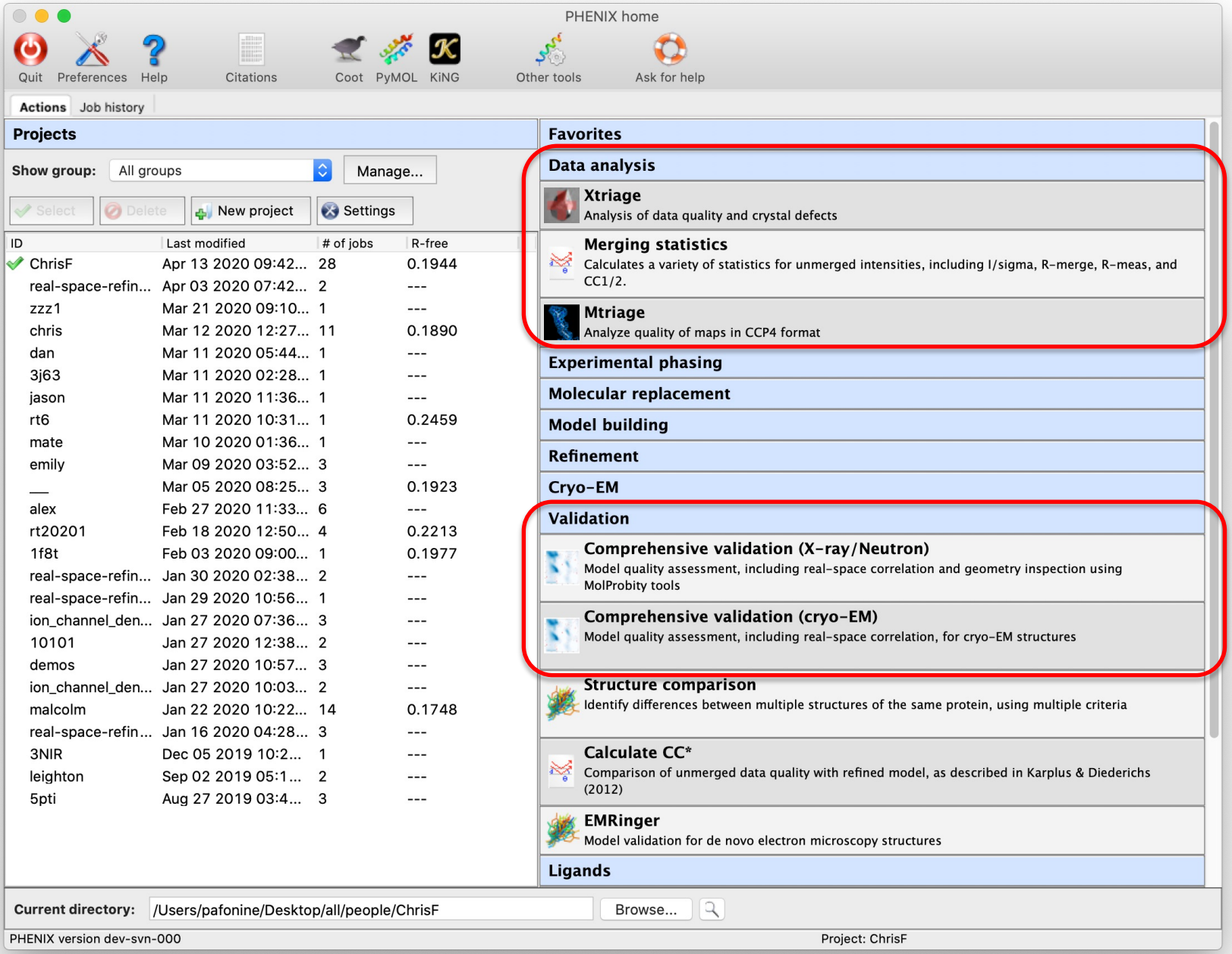
Model to data fit





Similar

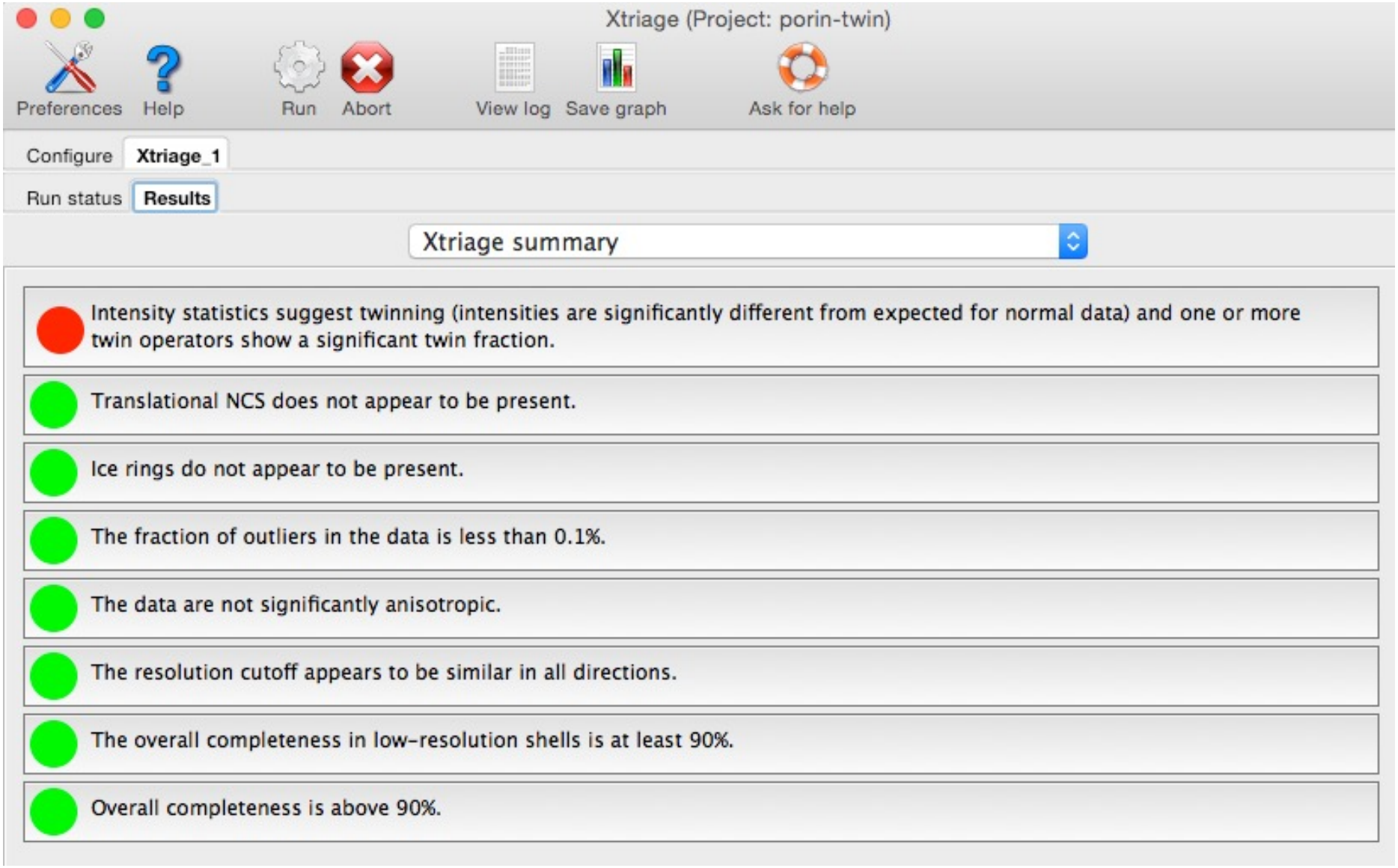
Validation tools in Phenix



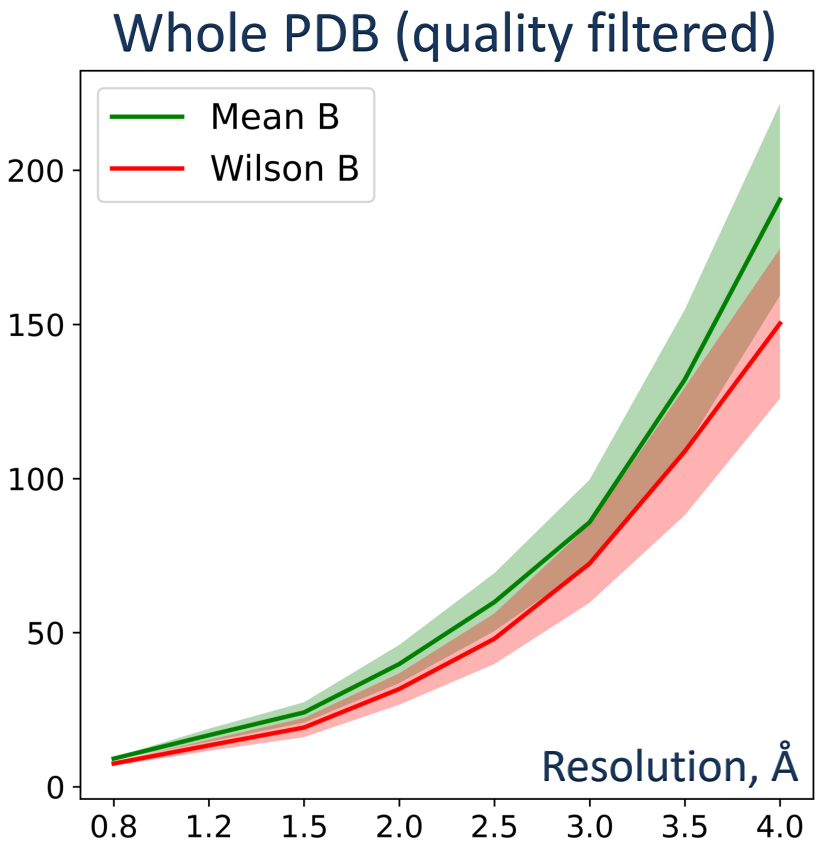
Xtriage: all about your diffraction data

- Matthews coefficient probabilities
- Completeness by resolution
- Wilson plot sanity
- Detection of translational NCS (tNCS)
- Analysis of systematic absences and combination of tNCS with current space group
- 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


Xtriage (Project: porin-twin)
Preferences
Help
Run
Abort
View log
Save graph
Ask for help
Configure
Xtriage_1
Run status
Results
Xtriage summary
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%.

Wilson B

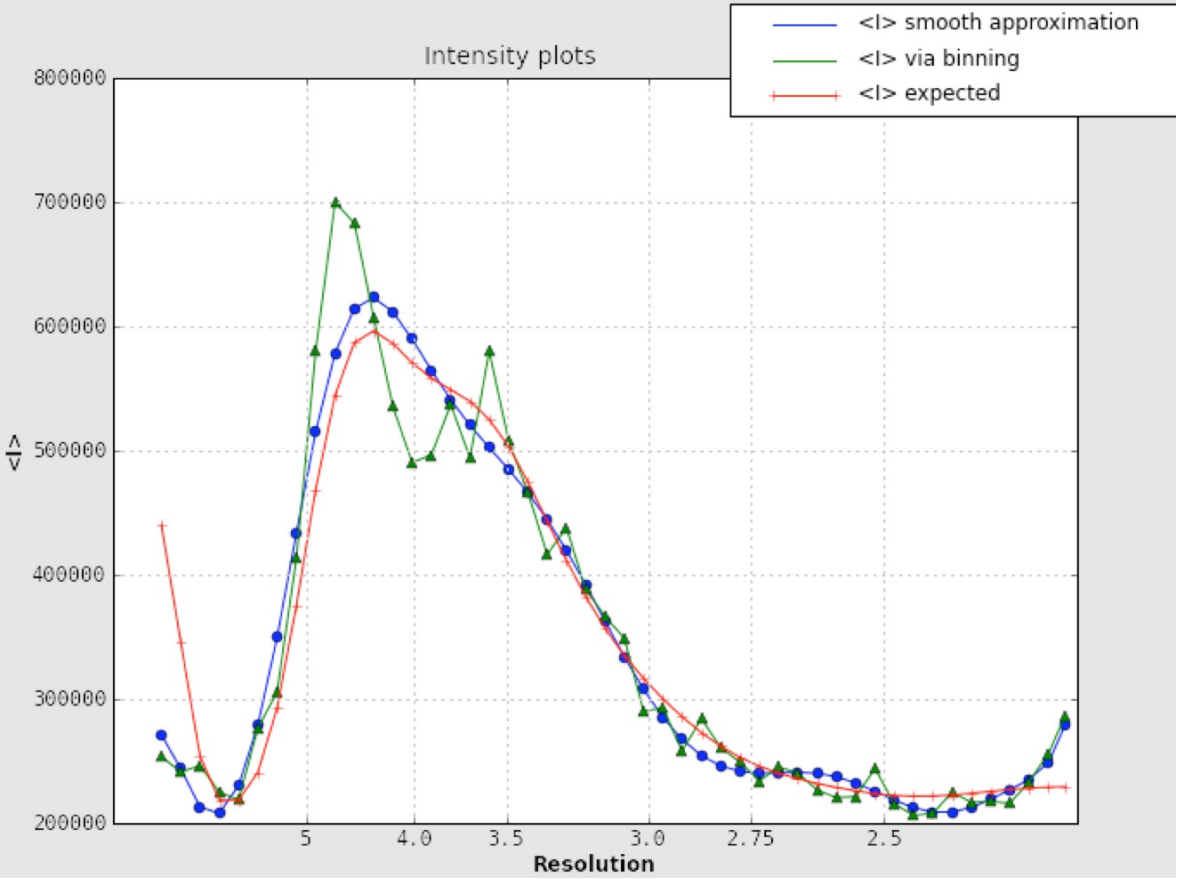


Wilson statistics assumes atoms of the same kind are randomly distributed in the unit cell and have the same isotropic B-factors

- 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

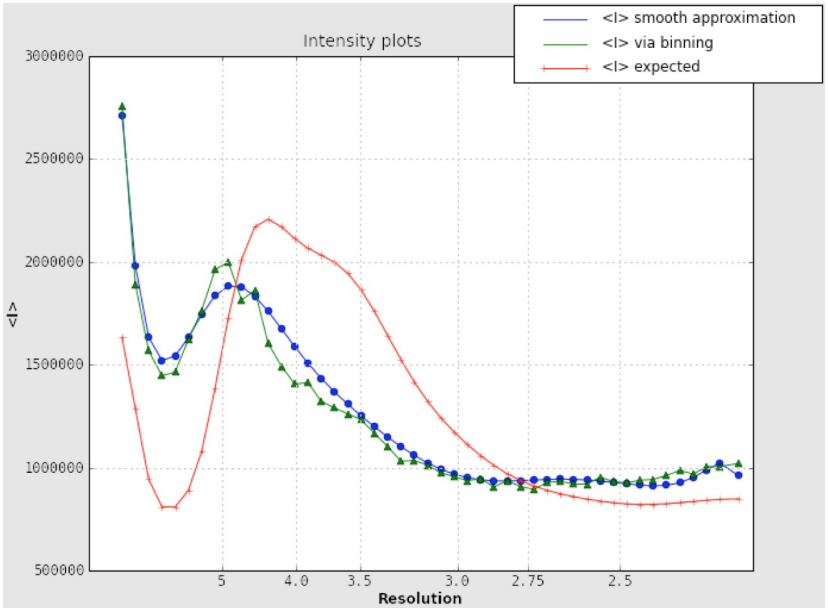
Wilson plot (mean intensity vs resolution)

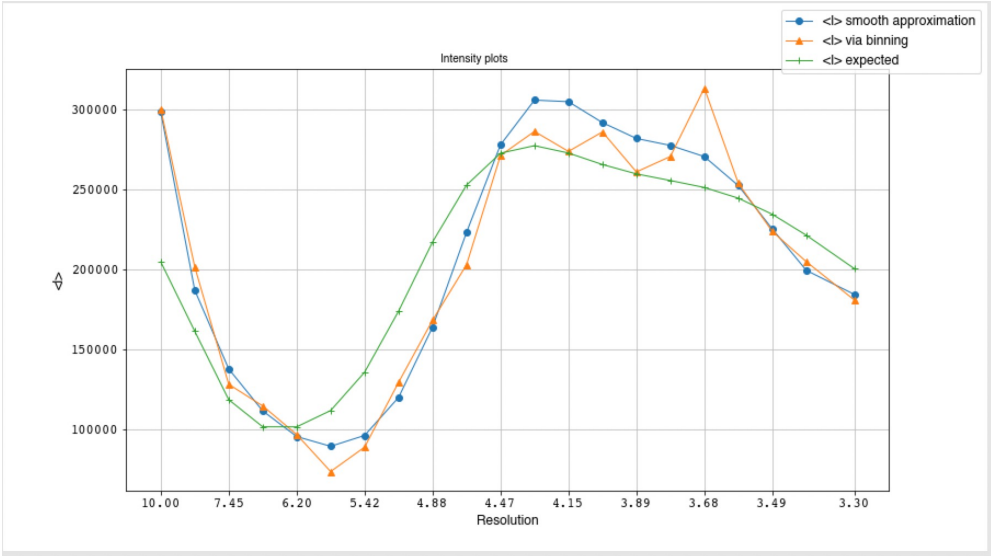
- The Wilson plot looks at mean intensity of diffraction by resolution, a curve which has a predictable shape



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)

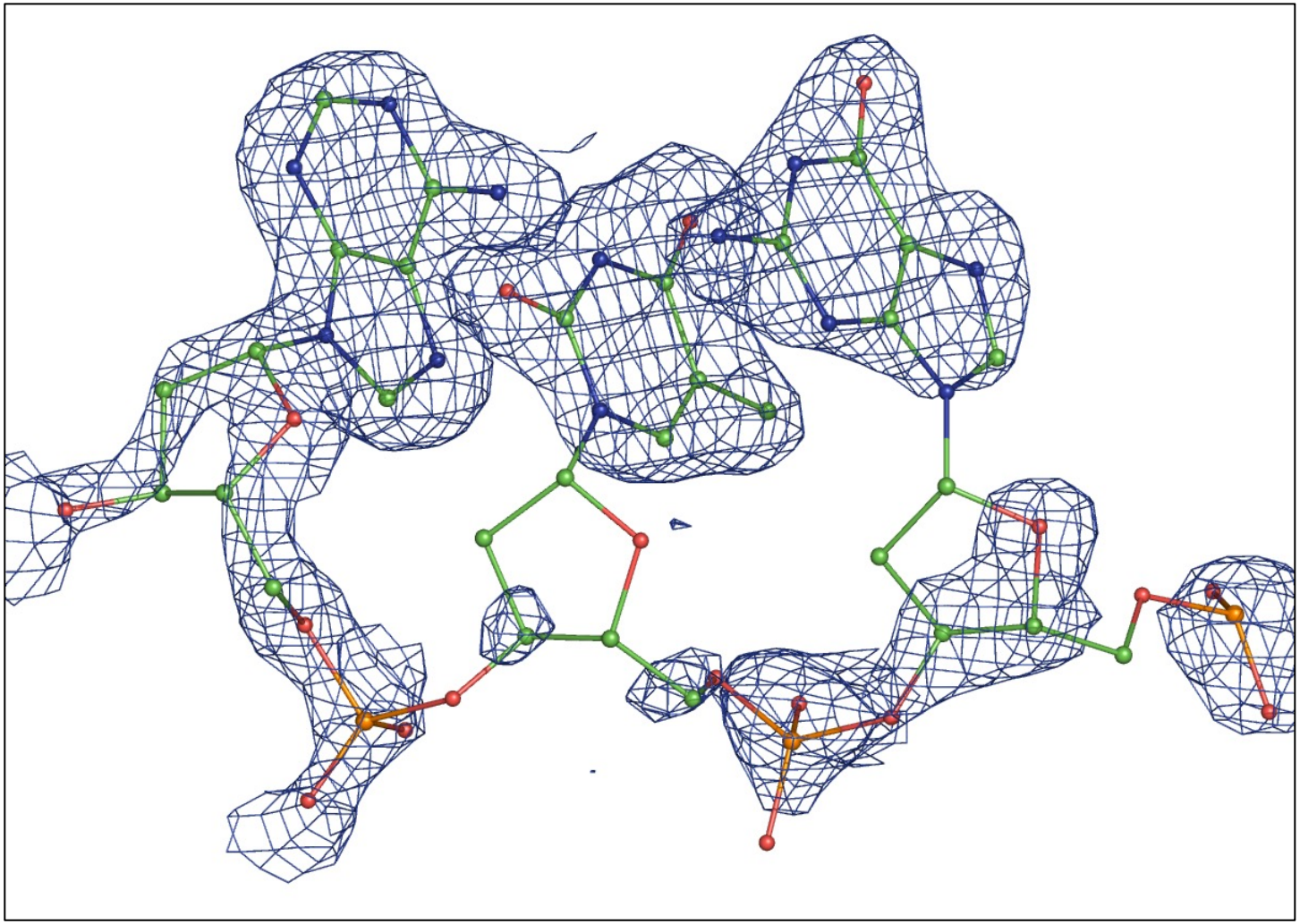




Data completeness

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

2mFo-DFc , 1σ

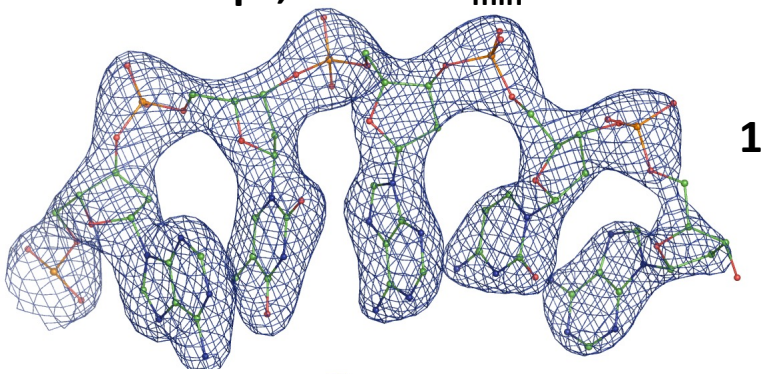


Data completeness

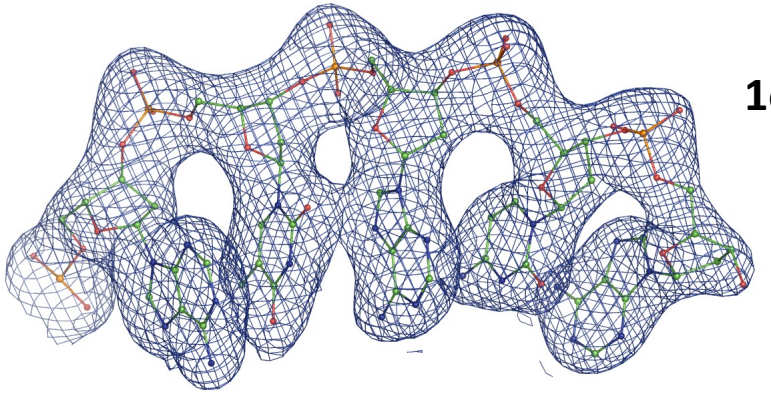
```
Completeness by resolution:
 19.9274 -  3.2441 0.78
  3.2441 -  2.5767 0.99
  2.5767 -  2.2515 1.00
  2.2515 -  2.0459 1.00
  2.0459 -  1.8993 0.99
```

Overall completeness in d_{min}-inf: 0.95

Fcalc maps, full set d_{min}-inf

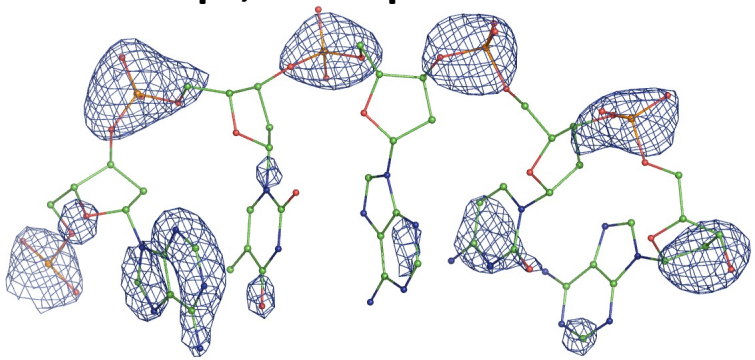


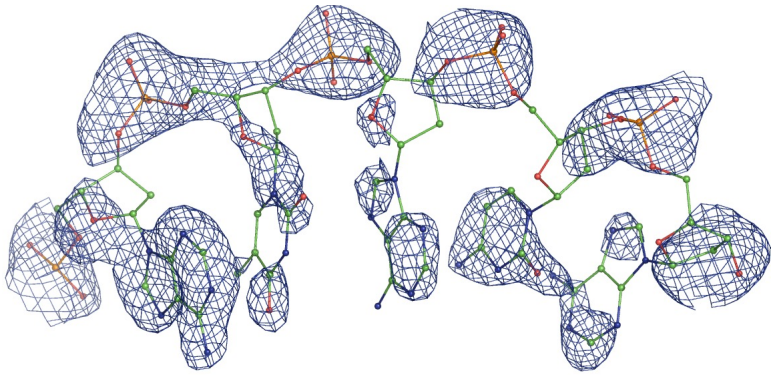
1.5σ map cutoff



1σ map cutoff

Fcalc maps, incomplete set

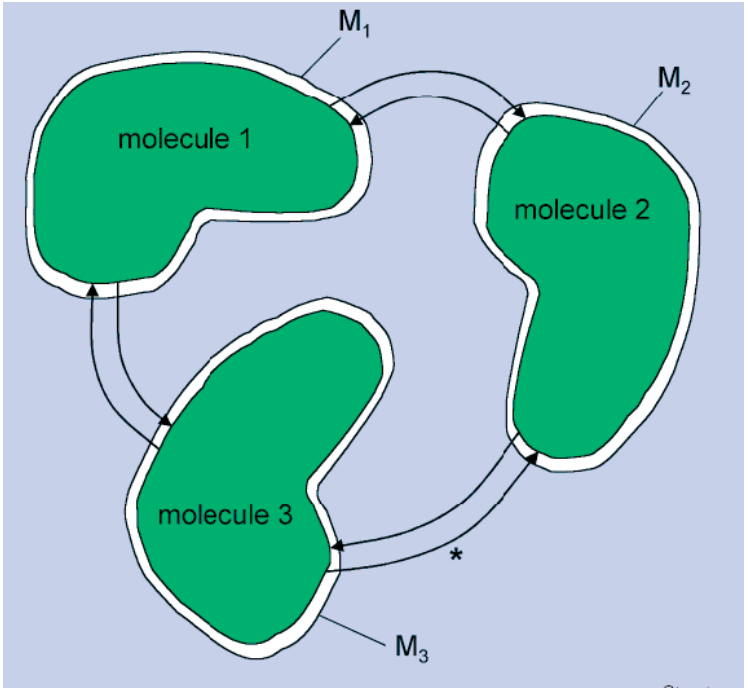




Systematic data incompleteness can distort maps

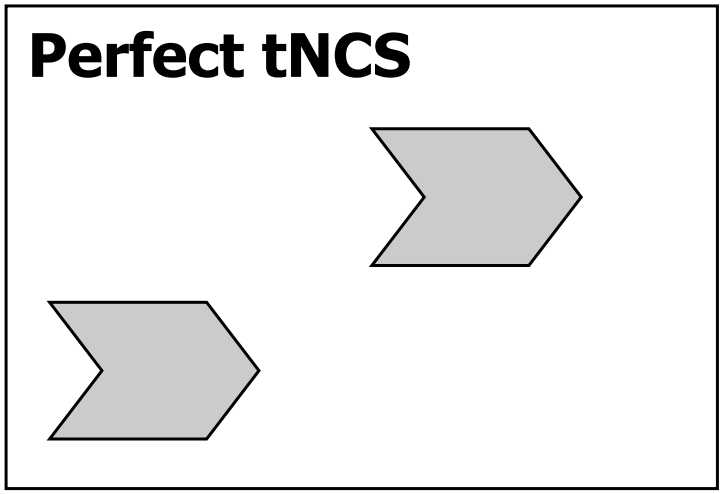
Non-crystallographic symmetry NCS

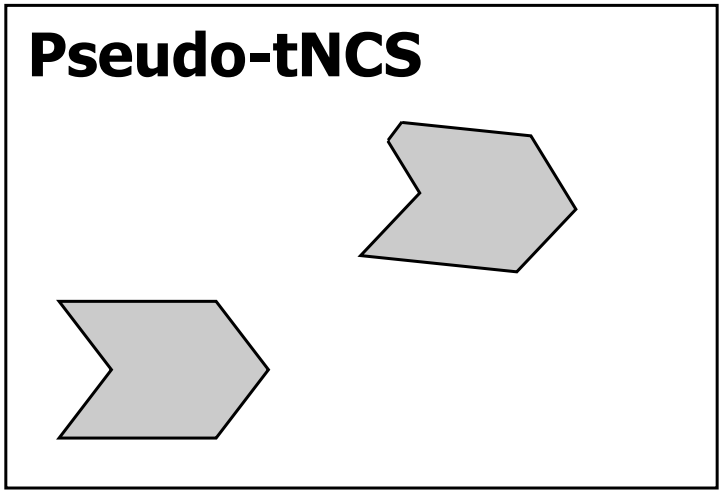
- Two or more molecules in the ASU related by rotation-translation
- NCS is found in about 1/3 to 1/2 of crystal structures
- Usually helps solving/refining models at medium-to-low resolution
- A special case of NCS, translational NCS (tNCS) leads to complications



Translational NCS (tNCS)

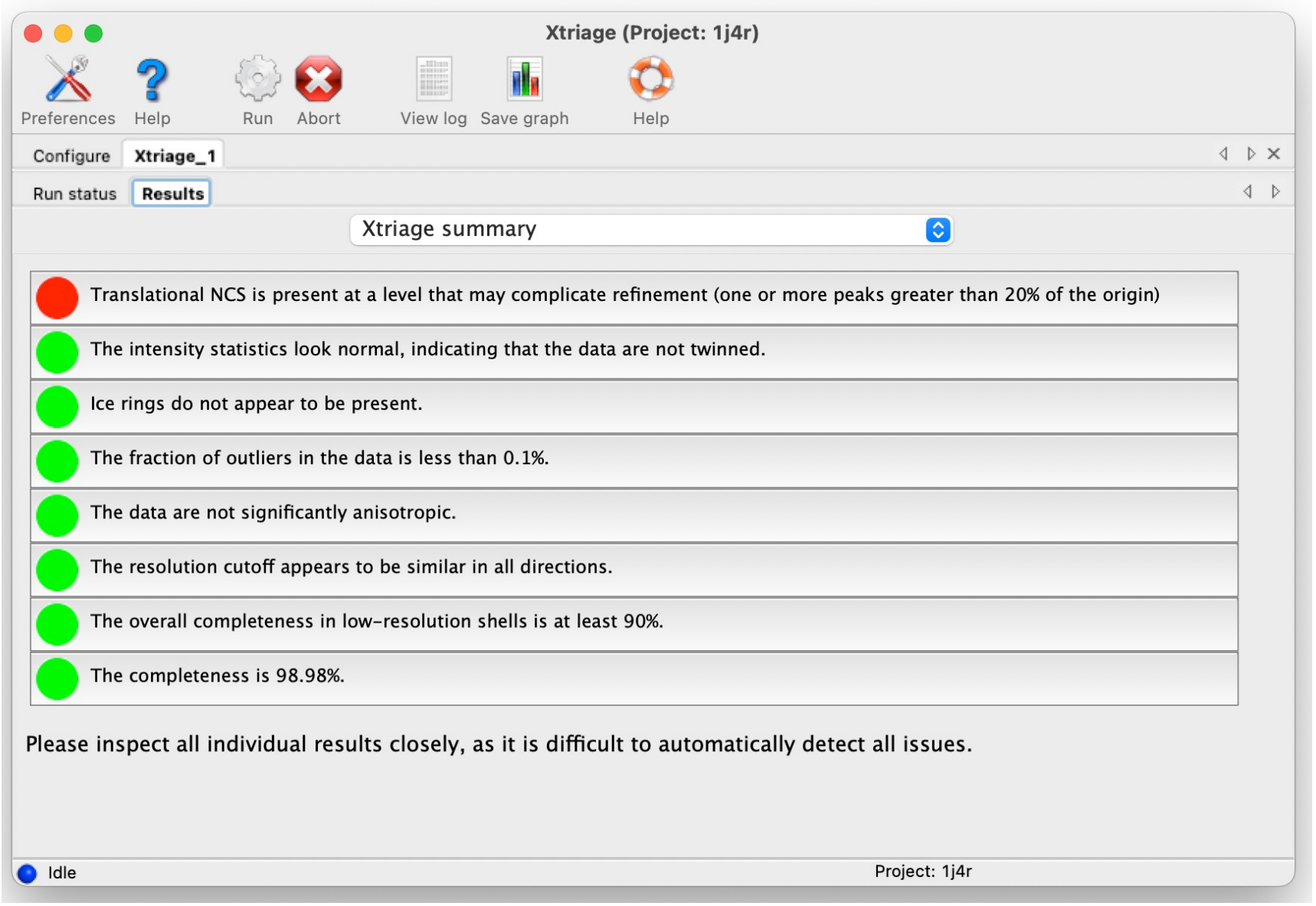
- tNCS arises when the ASU contains components that are oriented in (nearly) the same way and can be superimposed by a translation that does not correspond to any symmetry operation in the space group.



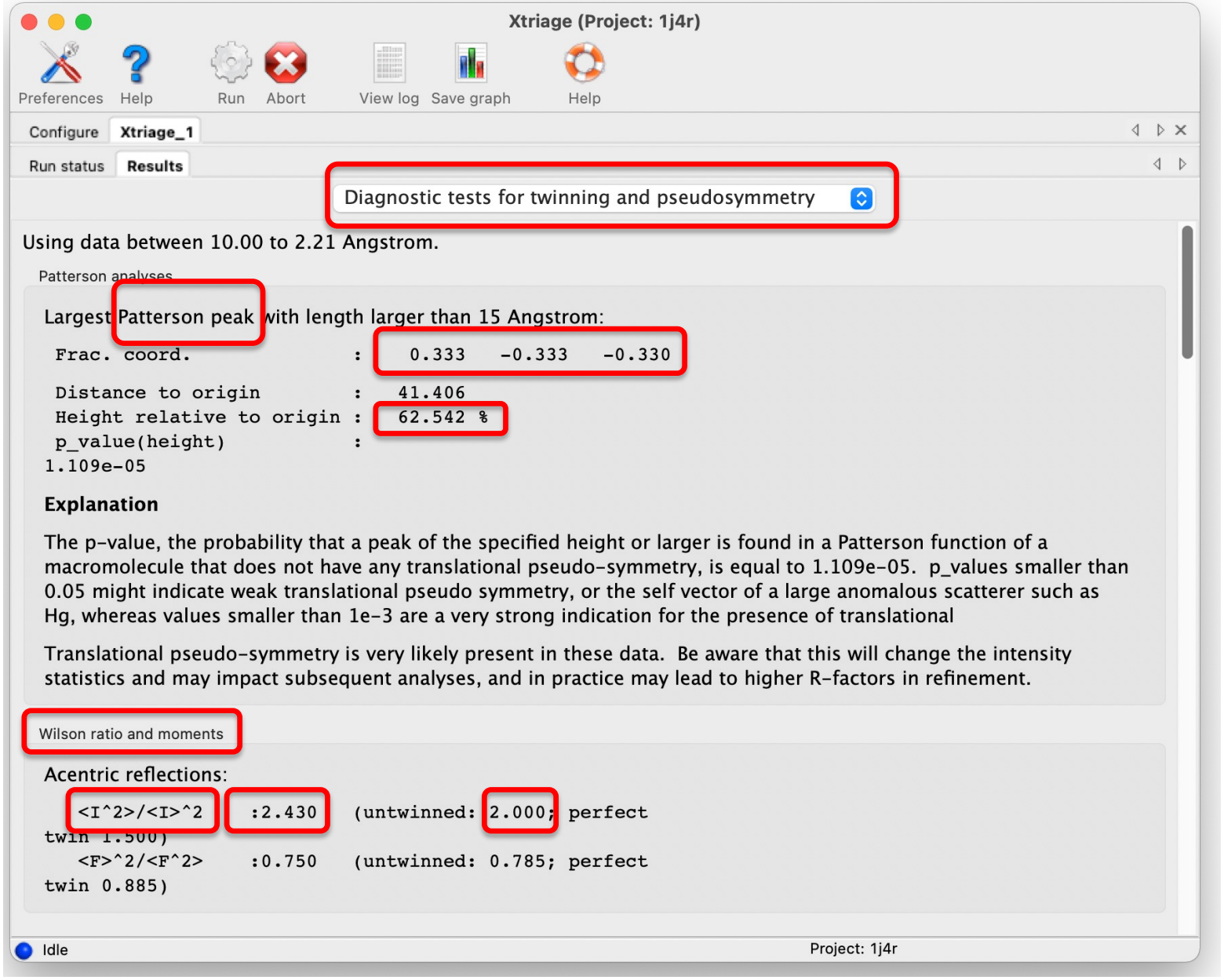


- Used to complicate MR (Phaser now can deal with it!)
- Risk to bias OMIT map

Translational NCS (tNCS)

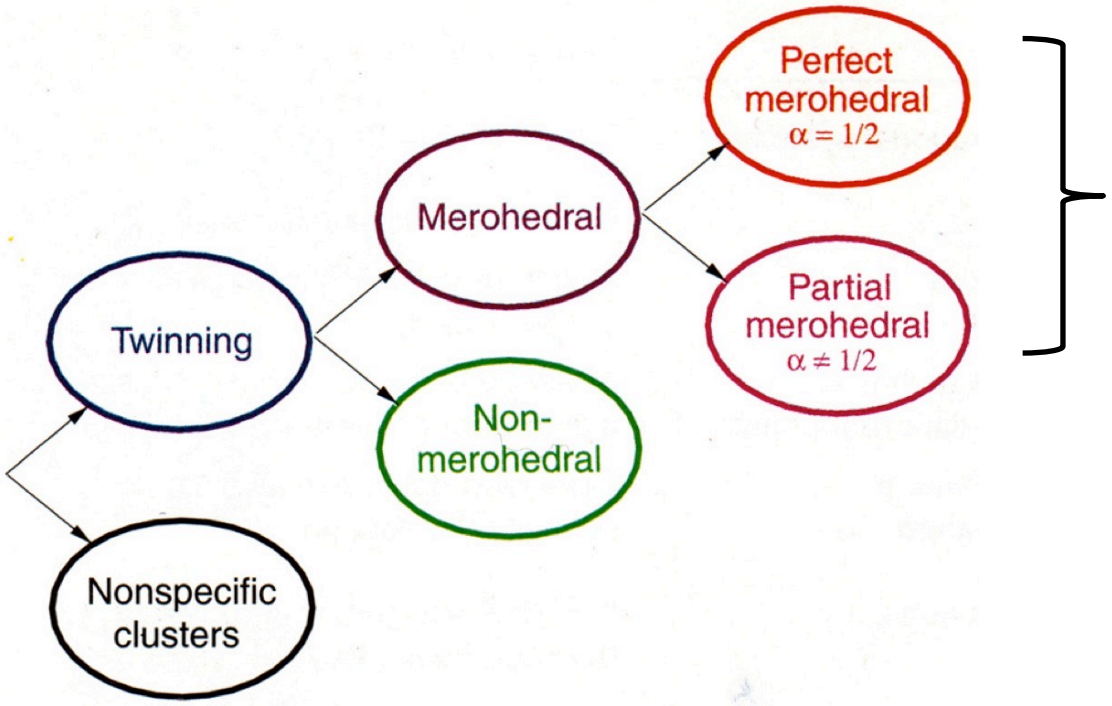


Translational NCS (tNCS) and twinning



Twinning

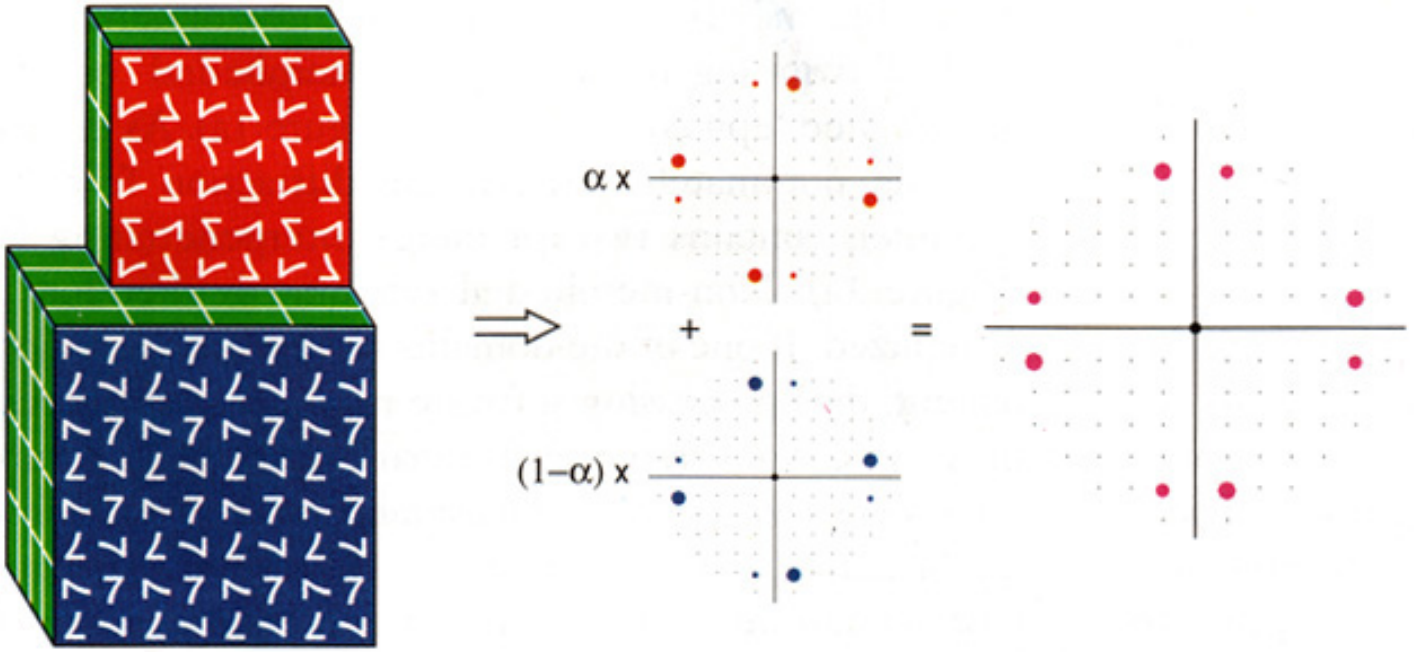
- Twinning is a crystal growth disorder



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

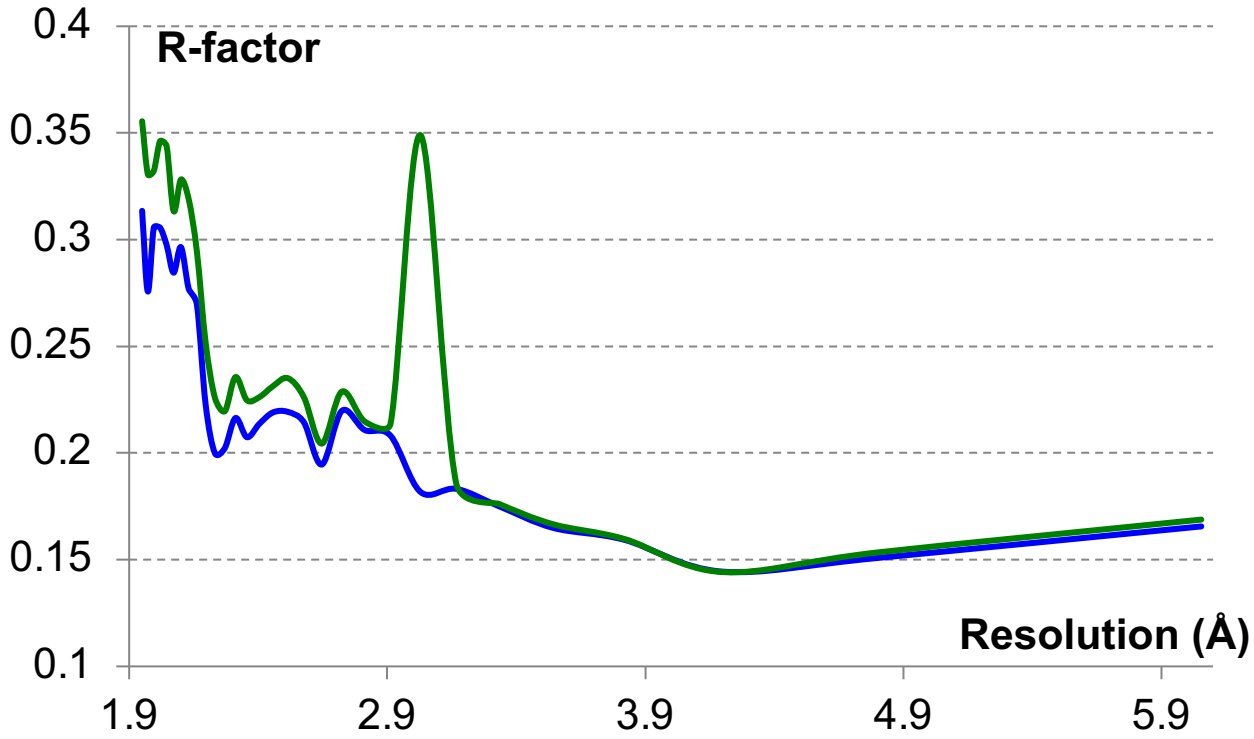
$$I_{\mathrm{OBS}}(\mathbf{h}) = \alpha_1 I(\mathbf{h}) + \ldots + \alpha_N I(\mathbf{T}_N \mathbf{h})$$

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

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
```