Science and Technology Facilities Council
Scientific Computing

Cross-Validation for Measuring and Preventing Overfitting in Cryo-EM

Jess Huntley
Mathematical Research Software Engineer,
Science and Technology Facilities Council
jessica.huntley@stfc.ac.uk

Science and Technology Facilities Council
Ada Lovelace Centre

Agenda

- **Introduction to Cryo-EM**
- Project Motivation
- Measuring Overfitting
- The Cross-Validation Workflow
- Project Extensions

What is Cryo-EM?

- Biological macromolecules imaged in electron microscope
- Molecules rapidly frozen in a thin layer of vitreous ice – captures molecules in their native state.
- Randomly oriented and positioned in ice layer.

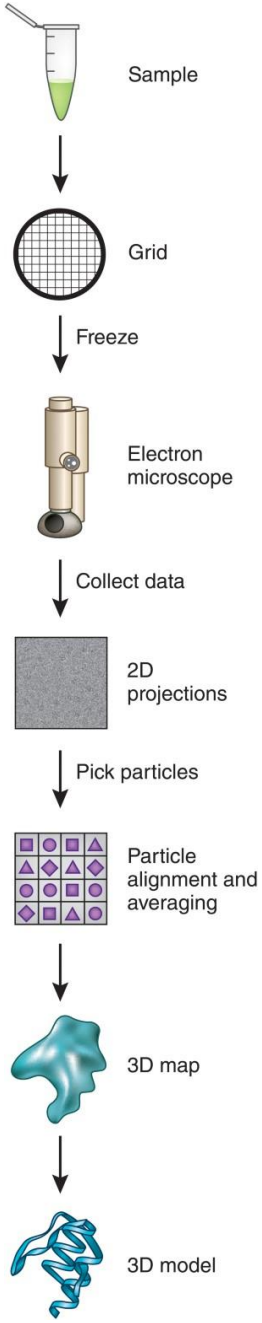
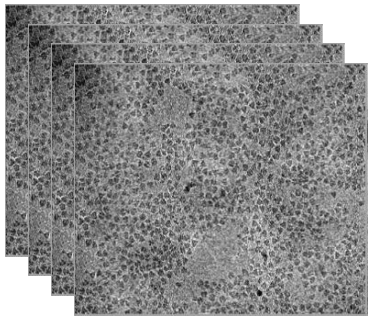


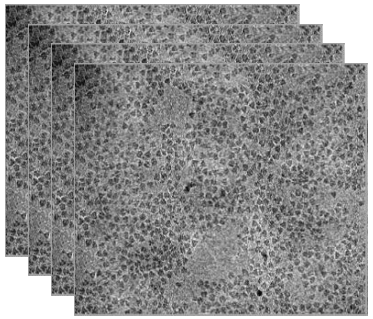
Figure: Doerr, A. Single-particle cryo-electron microscopy. Nat Methods 13, 23 (2016). https://doi.org/10.1038/nmeth.3700

Cryo-EM Image Processing

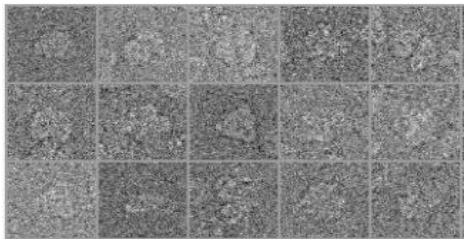


Movie Frames

Cryo-EM Image Processing



Movie Frames

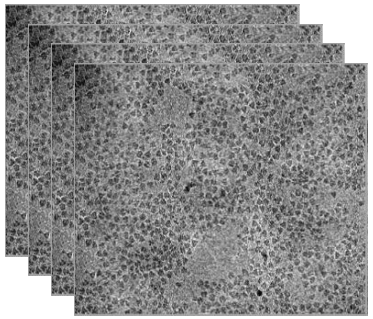


Particle picking
and extraction

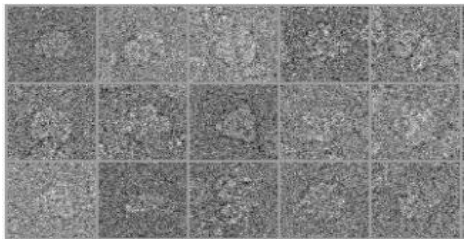
Science and Technology Facilities Council

Scientific Computing

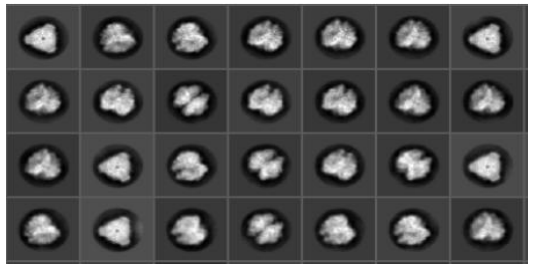
Cryo-EM Image Processing



Movie Frames



Particle picking
and extraction

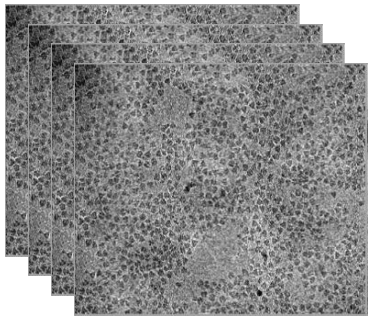


Particle alignment
and classification

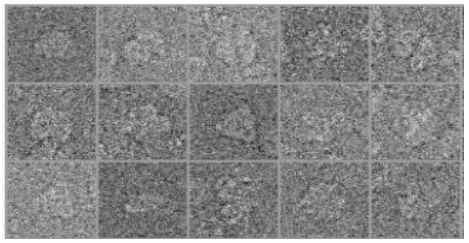
Science and Technology Facilities Council

Scientific Computing

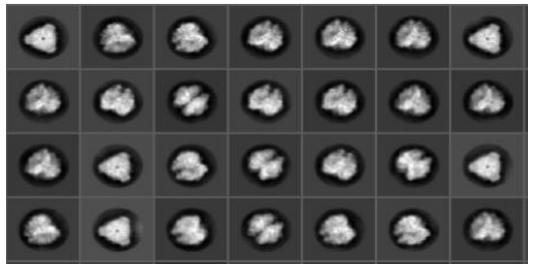
Cryo-EM Image Processing



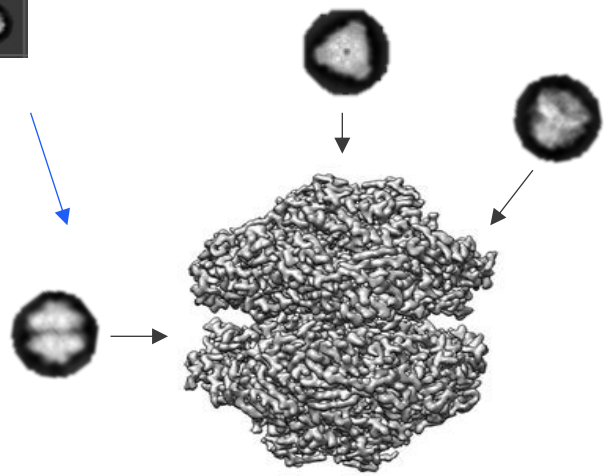
Movie Frames



Particle picking
and extraction



Particle alignment
and classification

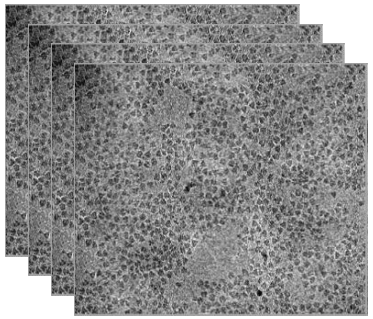


3D Reconstruction

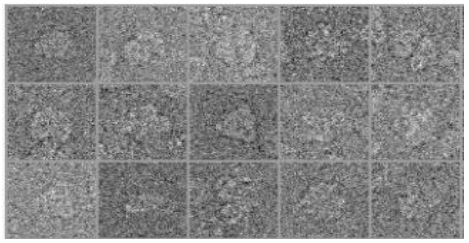
Science and Technology Facilities Council

Scientific Computing

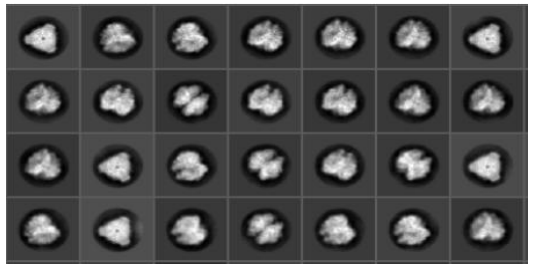
Cryo-EM Image Processing



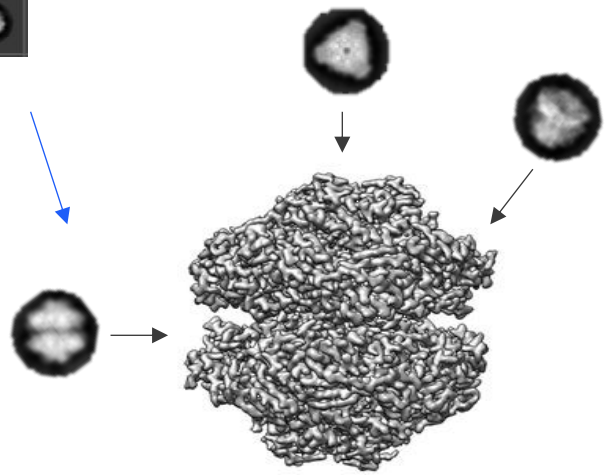
Movie Frames



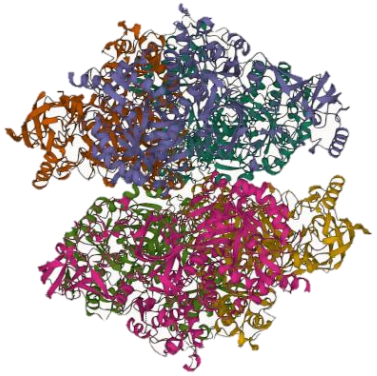
Particle picking and extraction



Particle alignment and classification



3D Reconstruction



Model Building

Science and Technology Facilities Council

Scientific Computing

Agenda

- Introduction to Cryo-EM
- **Project Motivation**
- Measuring Overfitting
- The Cross-Validation Workflow
- Project Extensions

Project Motivation

- Images are noisy!

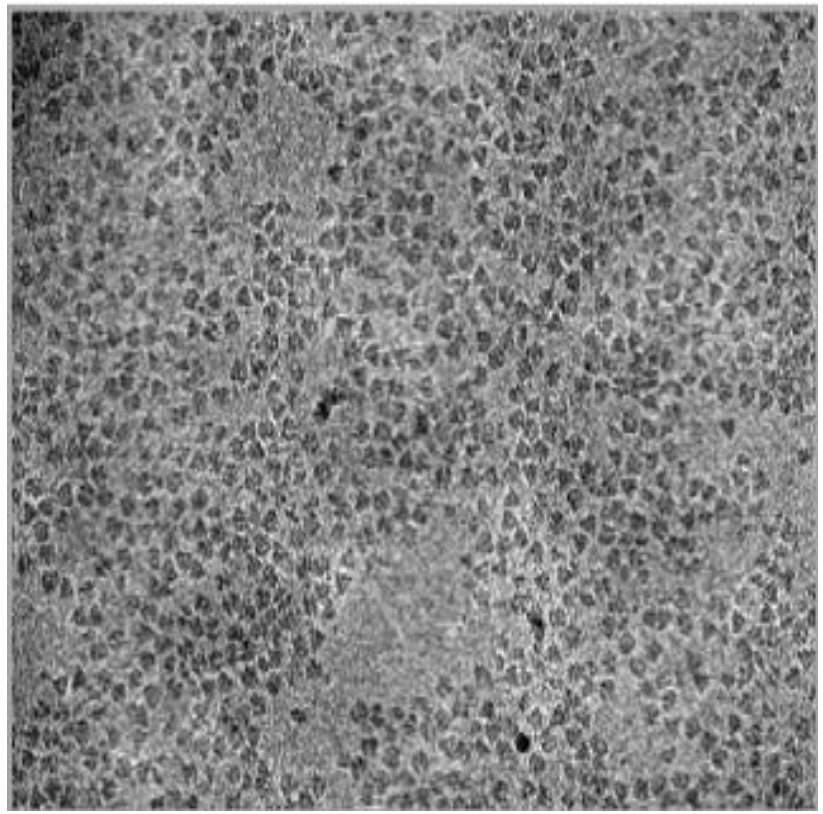


Image Formation Model

Usually, all movie frames are summed together to calculate a 3D reconstruction. Instead, consider a “single-frame reconstruction”.

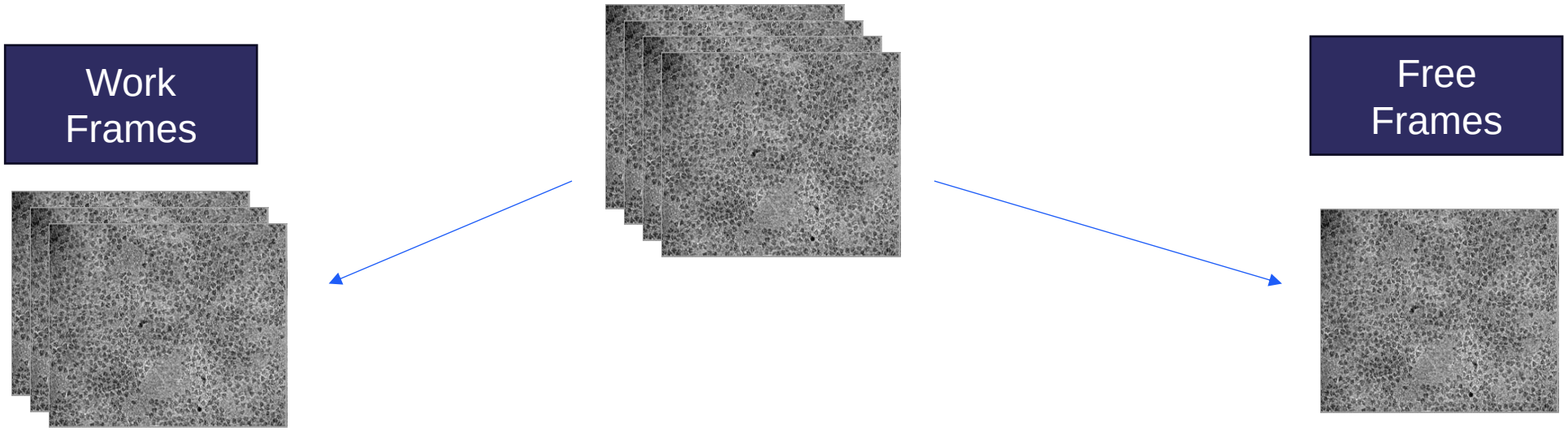
In reciprocal space, each structure factor F_i can be written as

$$F_i = F_S + N_i$$

for (randomly-chosen) frame with index i

F_S is “true” signal shared by all frames, and N_i is the noise contribution from that particular frame

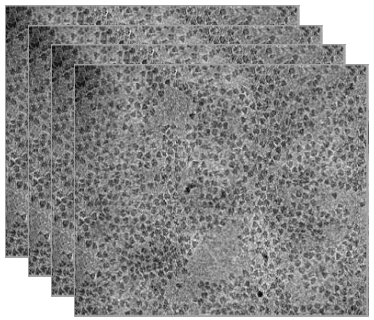
Work, Free, and Test Sets



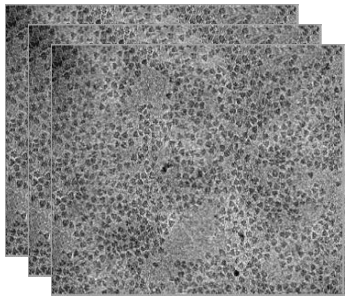
$$F_w = n_w F_s + N_w = n_w F_s + \sum_i^{n_w} N_i$$

$$F_f = n_f F_s + N_f = n_f F_s + \sum_{n_w+1}^{n_w+n_f} N_i$$

Work, Free, and Test Sets

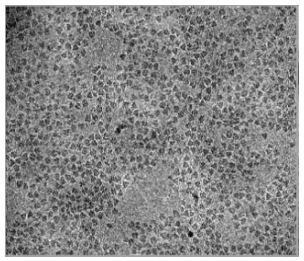


Work Frames



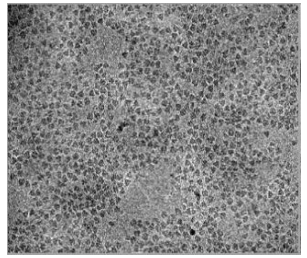
$$F_w = n_w F_s + N_w = n_w F_s + \sum_i^{n_w} N_i$$

Test Frames



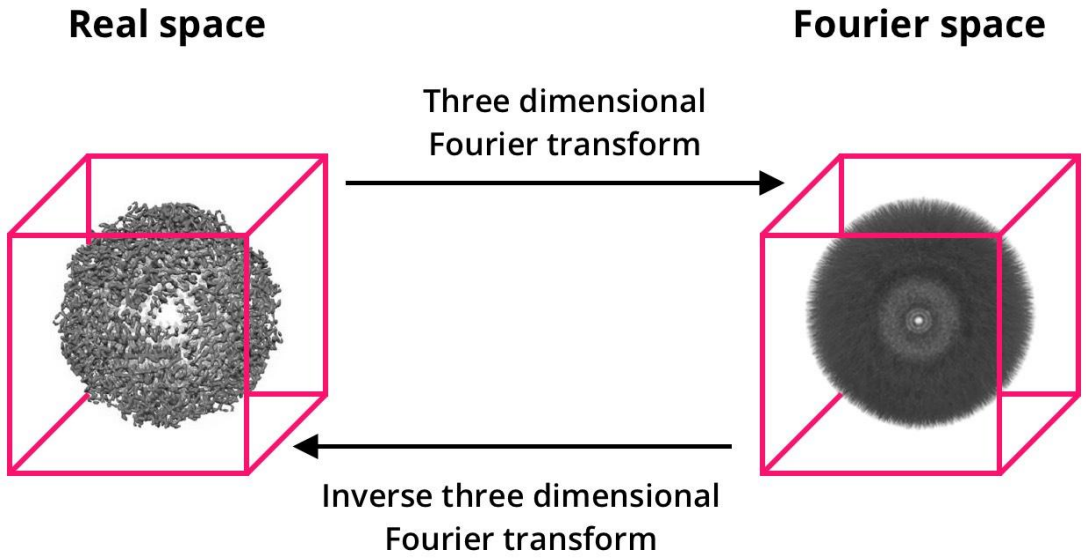
$$F_t = n_f F_s + N_t = n_f F_s + \sum_i^{n_f} N_i$$

Free Frames



$$F_f = n_f F_s + N_f = n_f F_s + \sum_{n_w+1}^{n_w+n_f} N_i$$

Fourier Shell Correlation

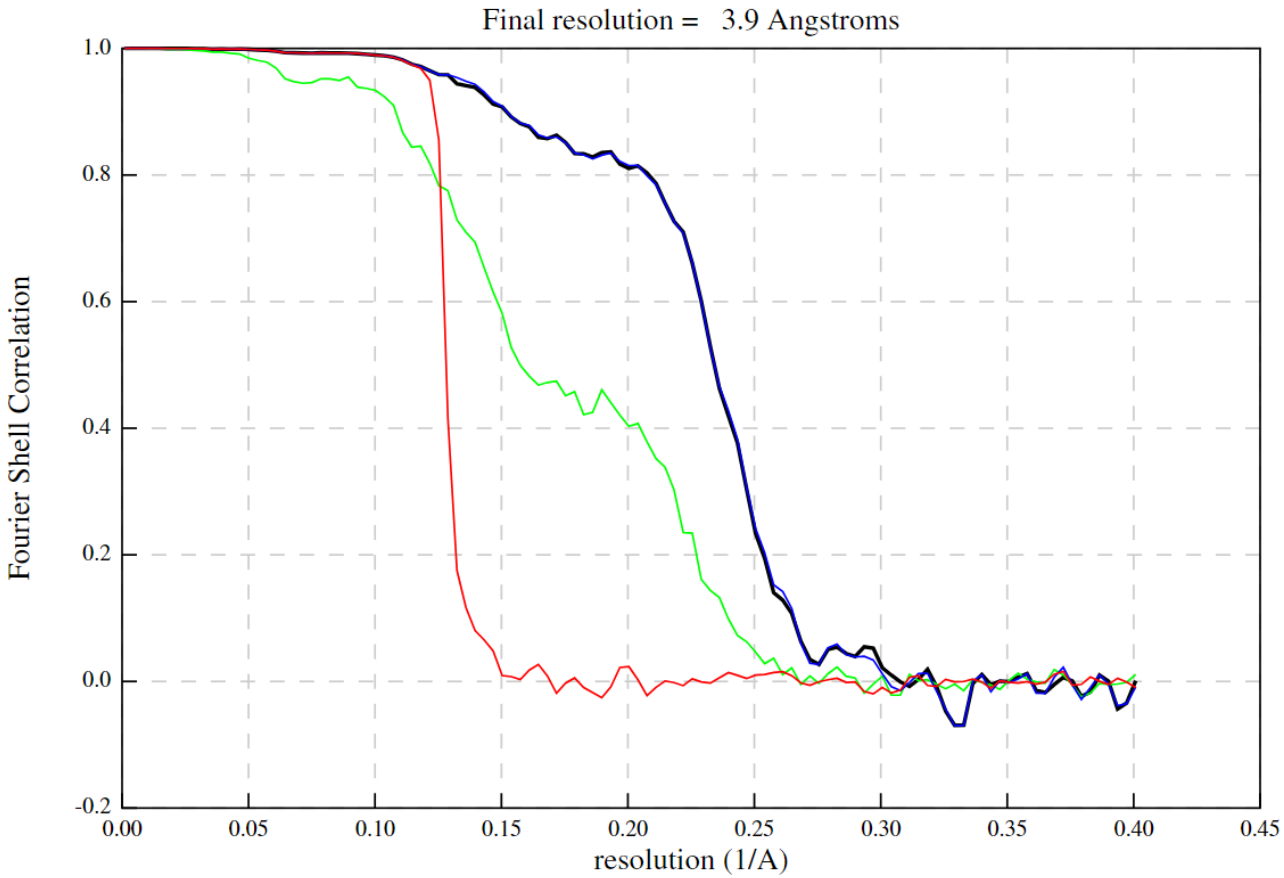


- Widely used method to assess agreement between 3D volumes in cryo-EM
- Defined as the Pearson Correlation Coefficient between two complex variables:

$$FSC(r) = Corr(F_A, F_B)$$
$$= \frac{Cov(F_A, F_B)}{\sqrt{Var(F_A)Var(F_B)}}$$

Figure: https://myscope.training/CRYO_Why_do_Fourier_transforms

Fourier Shell Correlation



Agenda

- Introduction to Cryo-EM
- Project Motivation
- **Measuring Overfitting**
- The Cross-Validation Workflow
- Project Extensions

Estimating FSC curves

- There are several FSC curves that can be directly calculated between structure factors from an atomic model (F_c) and the work, free, and test maps.
- Can use these to estimate what we want to know:
  - How closely does the fitted model match the real structure observed in the data, i.e. FSC_{cs}
  - How much has the model been over-fitted to noise, i.e. FSC_{cn}

Assumptions

- All frames are equally well fitted to the final reconstruction, so all common shared signal is included in F_s and noise N_w is uncorrelated with the signal.
- Assume that the power of the noise in each single frame reconstruction is the same and noise variables N_i are uncorrelated to each other

Estimating FSC curves

Starting from Fisher Z-transform, $z = \frac{1}{2}\ln(\frac{1+r}{1-r})$

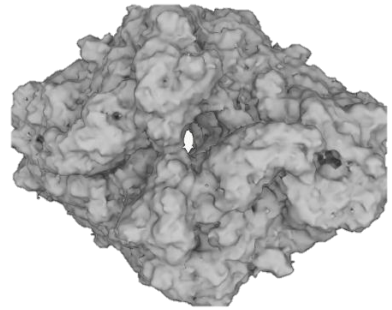
$$FSC_{cs} = \tanh(\frac{1}{2}\ln(\frac{\sqrt{FSC_{tf}}+FSC_{cf}}{\sqrt{FSC_{tf}}-FSC_{cf}}))$$

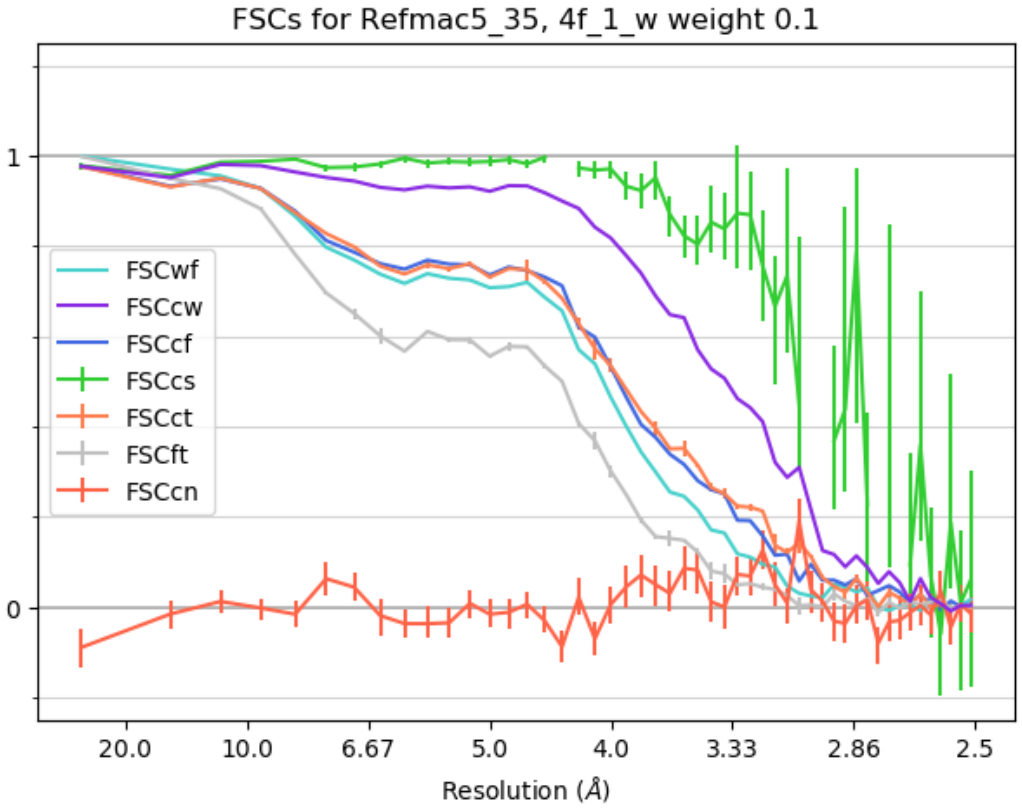
$$FSC_{cn} = \tanh(\frac{1}{2}\ln(\frac{\sqrt{1-FSC_{tf}}+\sqrt{\frac{n_w}{n_f}}(FSC_{ct}-FSC_{cf})}{\sqrt{1-FSC_{tf}}-\sqrt{\frac{n_w}{n_f}}(FSC_{ct}-FSC_{cf})}))$$

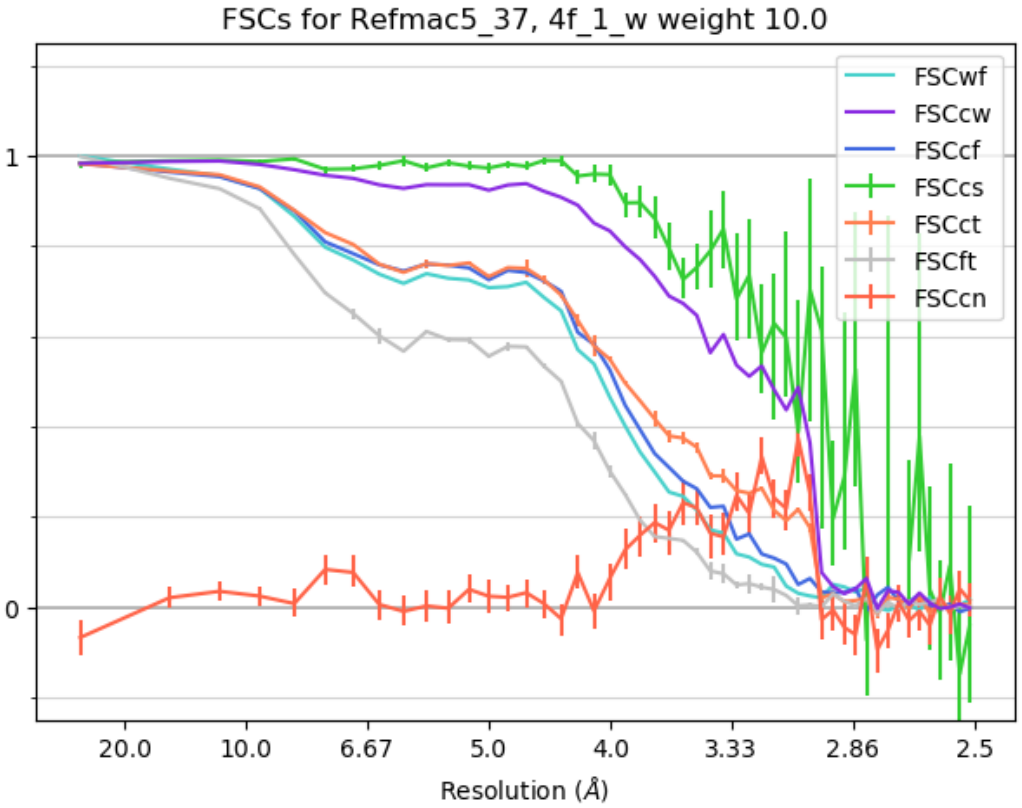
Uncertainty Quantification

- Bootstrapping used estimate 95% confidence intervals
  - Repeatedly sample Fourier components for each Fourier shell **with replacement**
- Need many bootstrap iterations (~1000) for a good estimate of underlying distribution
  - Slow to compute but embarrassingly parallel

Beta-Galactosidase Dataset







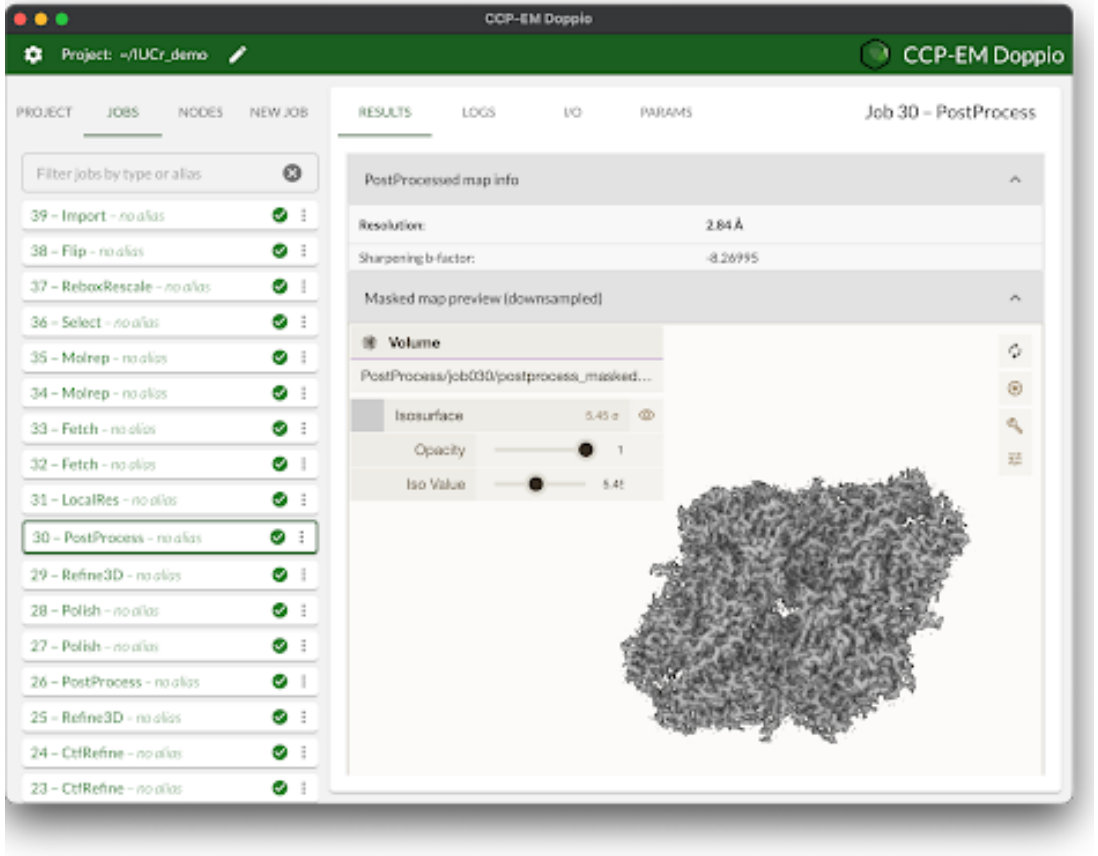
Science and Technology Facilities Council

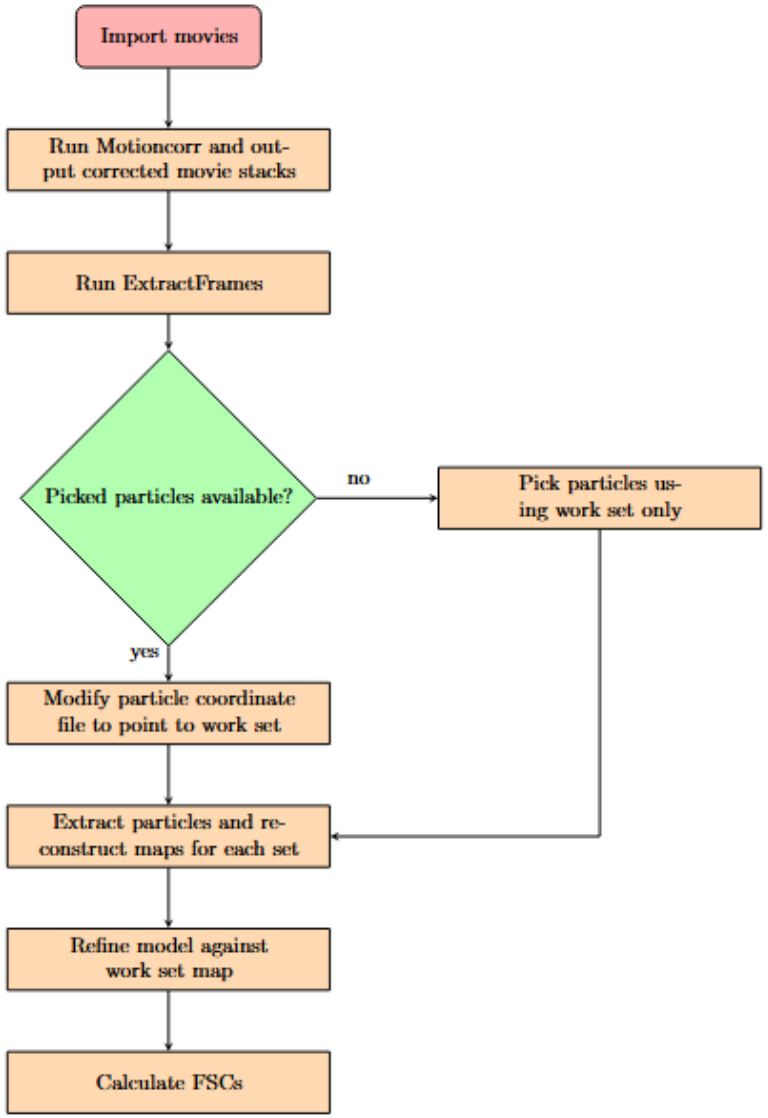
Scientific Computing

Agenda

- Introduction to Cryo-EM
- Project Motivation
- Measuring Overfitting
- **Software Implementation**
- Project Extensions

The Cross-Validation Workflow

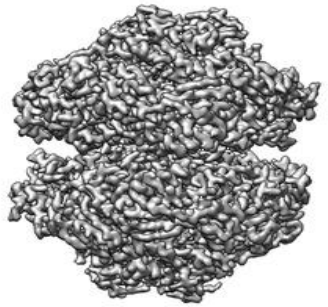


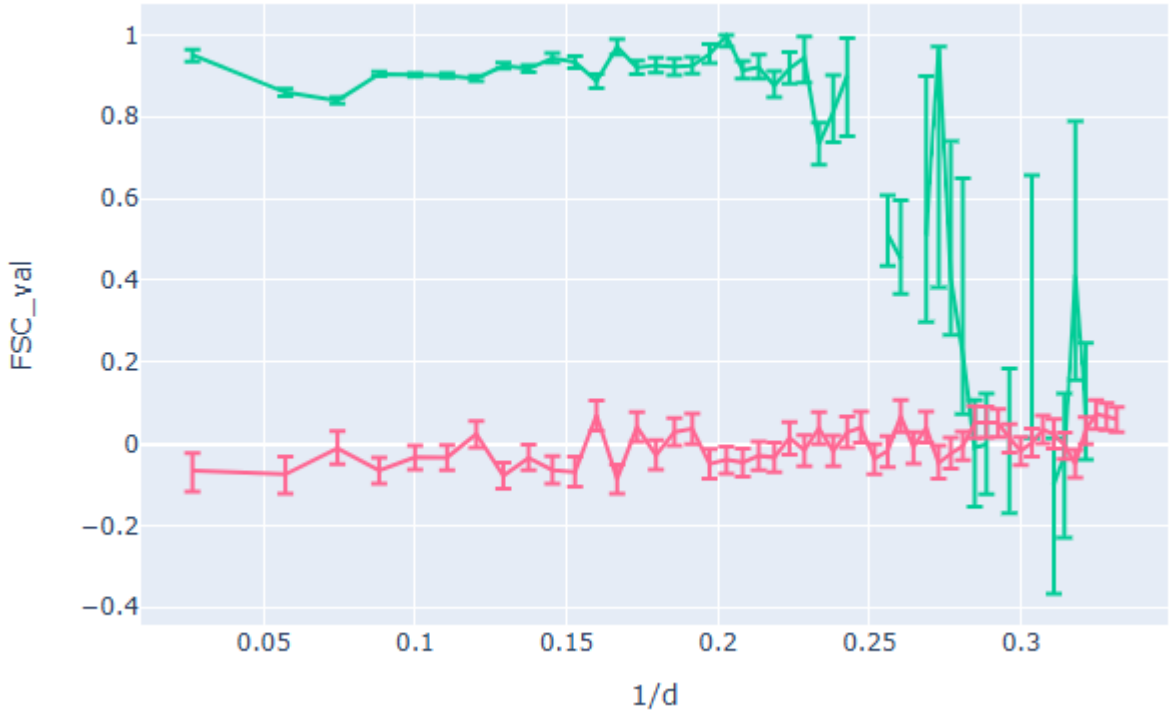


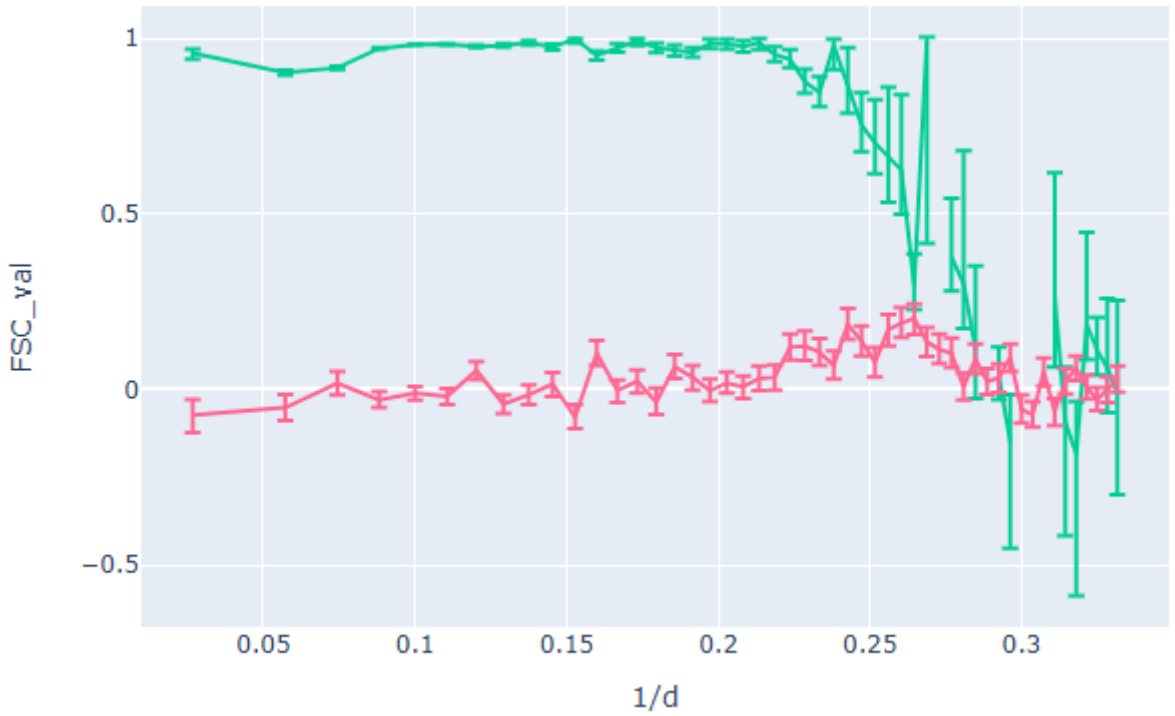
Science and Technology Facilities Council

Scientific Computing

Jack Bean Urease Dataset





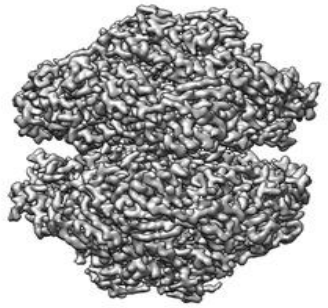


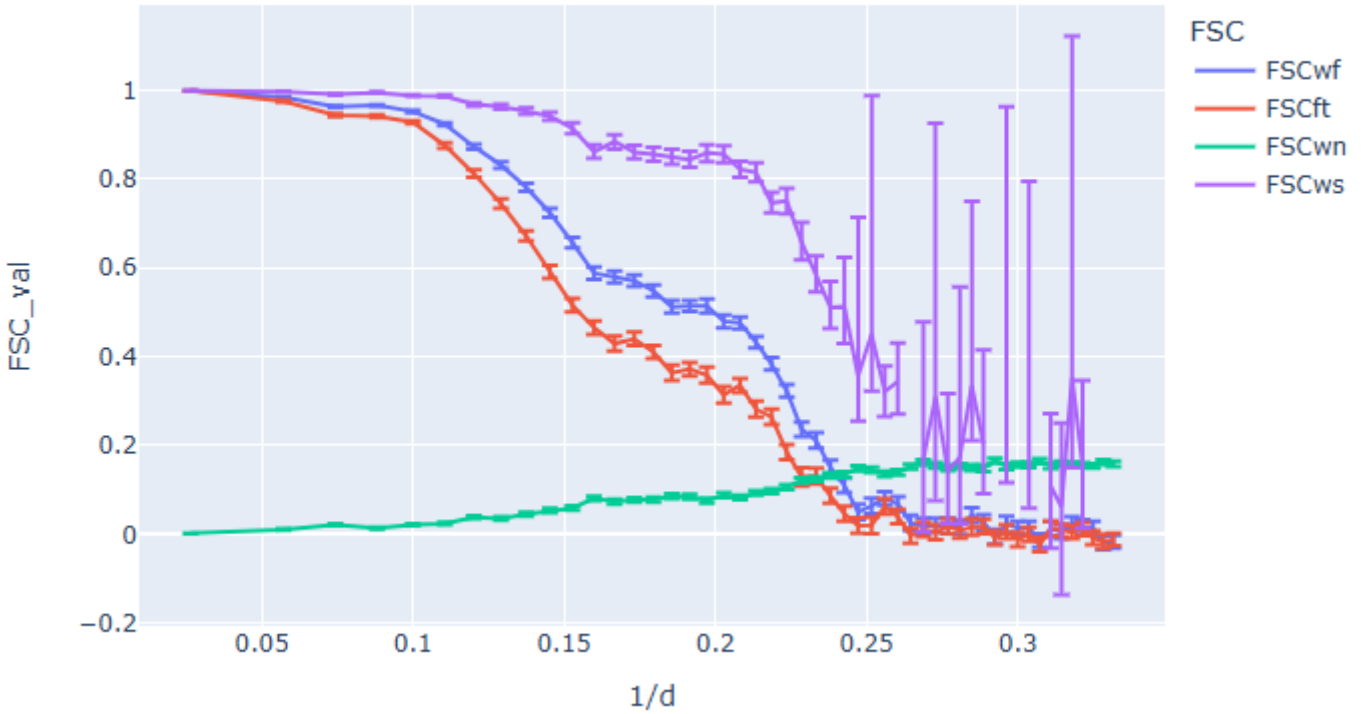
FSCcs FSCcn

Science and Technology Facilities Council

Scientific Computing

Jack Bean Urease Dataset

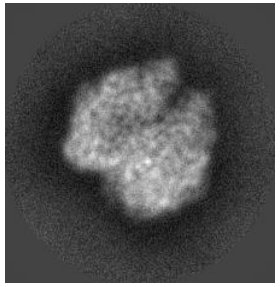


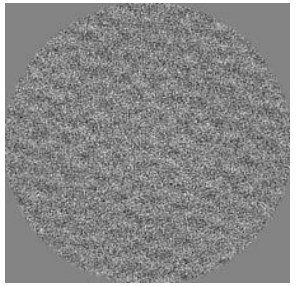


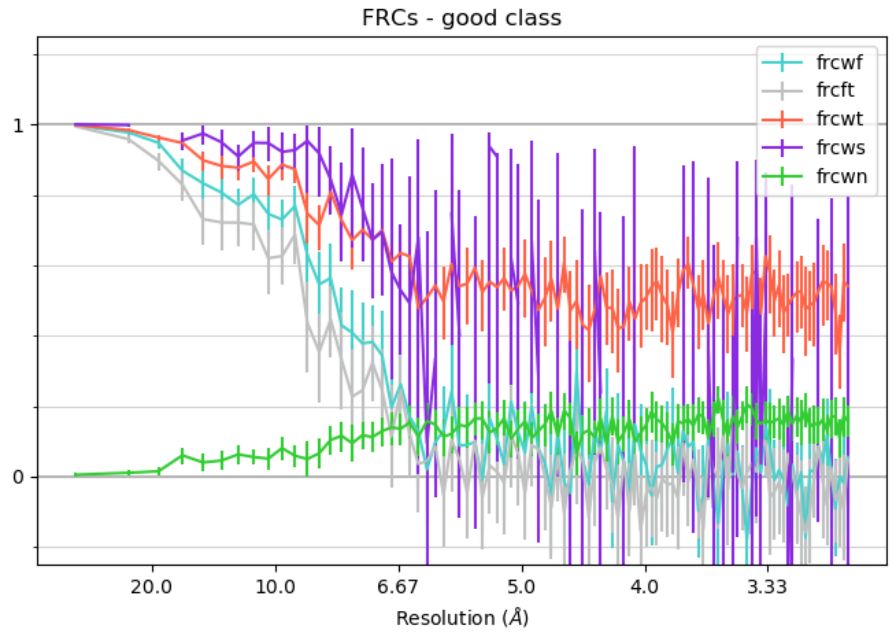
Science and Technology Facilities Council

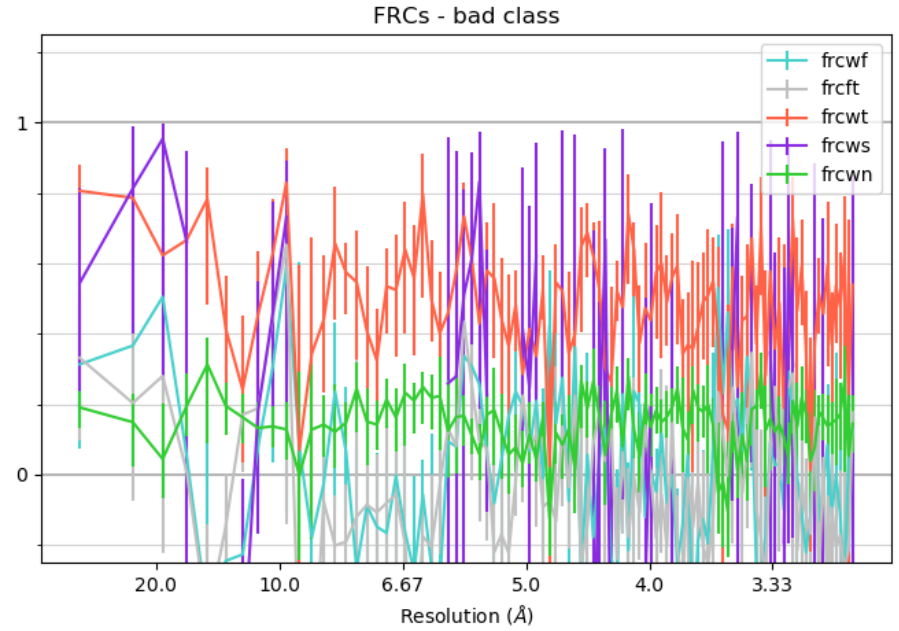
Scientific Computing

Validating 2D Class Averages









What's Next?

- More testing!
  - Involvement of electron Bio-Imaging Centre (eBIC) scientists
  - How to effectively summarise 2D class average FRCs?
  - Can we account for electron damage across movie frames?

- Method Refinements
  - Properly accounting for degrees of freedom for bootstrapping
  - How to account for variability of free set sampling?
  - How to handle noisy data points?

Science and Technology Facilities Council

Scientific Computing

Thank you

jessica.huntley@stfc.ac.uk

scd.stfc.ac.uk @SciComp_STFC