Molecular Replacement Example 1

Structure of TPR-rich domain of M. smegmatis EccA3mat

Model preparation, Molecular Replacement and Refinement

Ronan Keegan, CCP4

CCP4

Overview

- This is a simple example demonstrating how to generate and prepare a predicted model for use in Molecular Replacement to help determine the unknown crystal structure
- There is a single copy of a monomer in the asymmetric unit
- We will learn how to generate a predicted model and how to truncate it based on pLDDT scoring
- Phaser will be used to perform the Molecular Replacement step
- We will also refine the resulting solution to further assess whether or not it is correct
- We will start with a introduction to the steps involved in performing MR using CCP4Cloud

CCP4

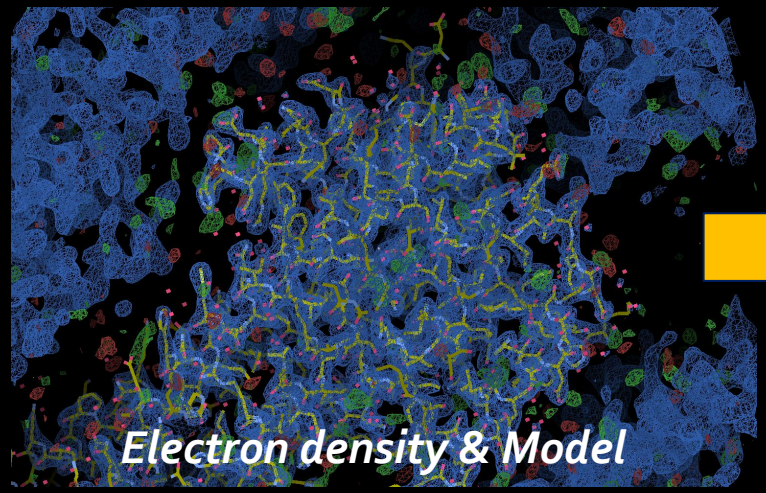
The Phase Problem: Molecular Replacement

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl}|F_o(hkl)|\cos[-2\pi(hx+ky+lz)+\varphi_{hkl}]$$

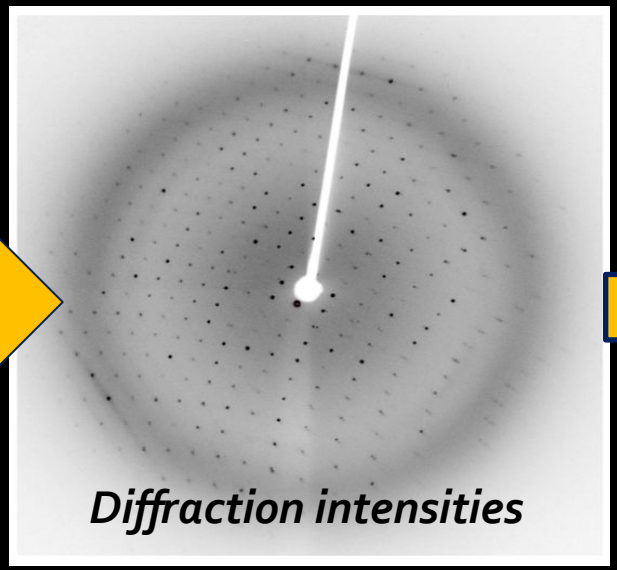
Electron density at (x,y,z)

Amplitudes

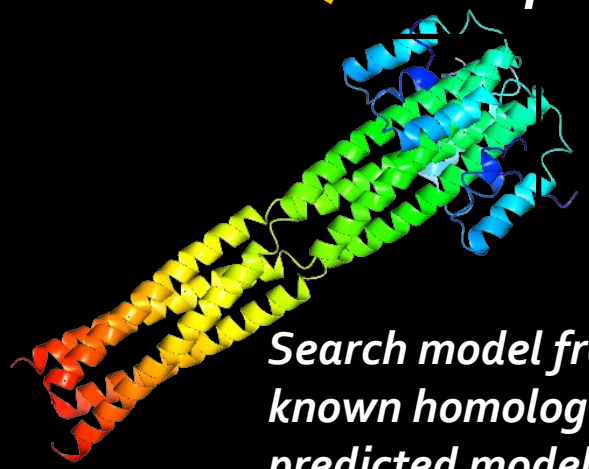
Molecular Replacement



Electron density & Model



Diffraction intensities



Search model from known homologue or predicted model

CCP4

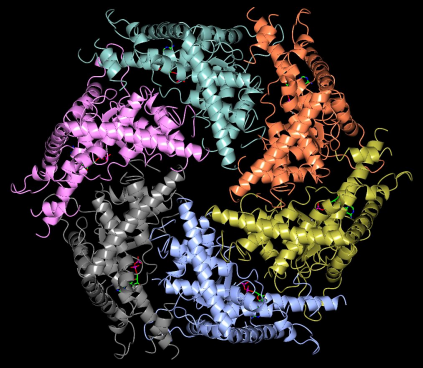
Most common scenarios in Molecular Replacement

Simple

- Single or few molecules in asu
- Good predictions or homologues available

Multimeric

- Multiple copies of monomer or complex in asu
- Good predictions or homologues available

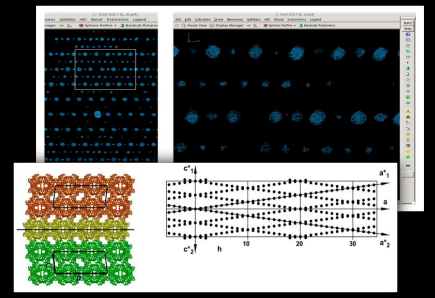


Domain

- Target is made up of structural domains
- Can be apparent from prediction or homologue
- Structural domains can have different relative orientation in crystal

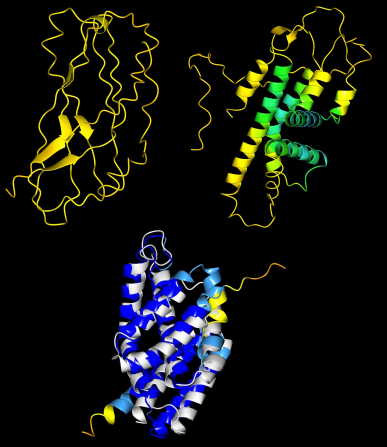
Problematic

- Twinning or tNCS can cause SG uncertainty
- Ambiguity in number of molecules in asu
- Contamination in crystal



Poor Models

- Low confidence predictions
- No close homologues available
- Inaccurate predictions



CCP4

Introduction to Molecular Replacement in CCP4 Cloud

CCP4

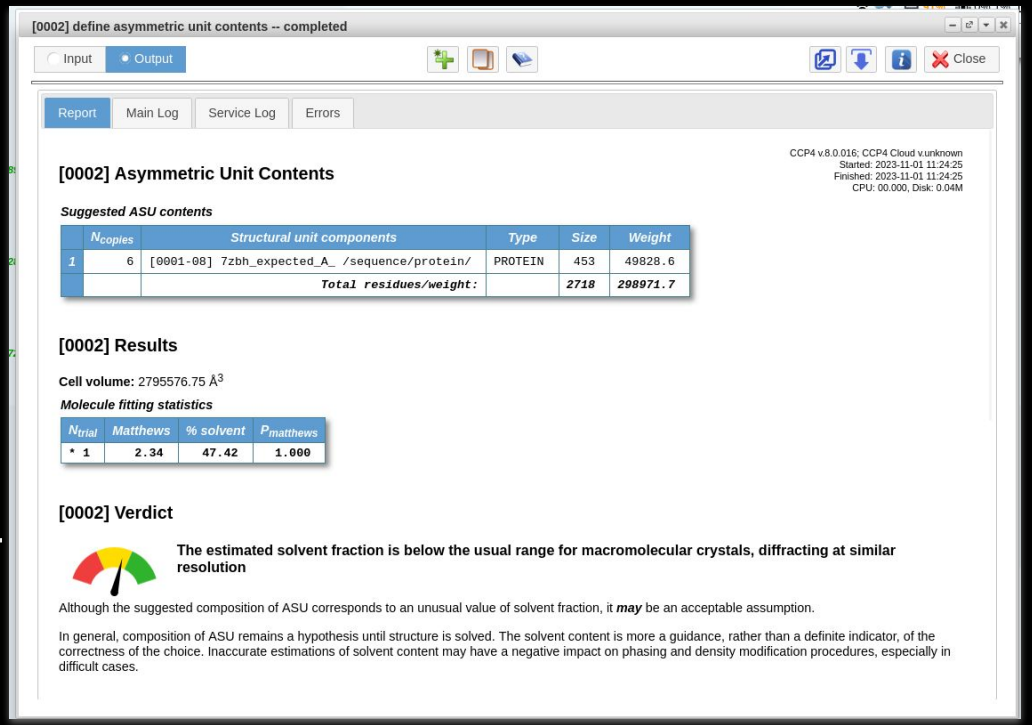
Contents of the Asymmetric Unit

- The first step is to determine how many copies of the molecule we are searching for i.e. how many copies/molecules in the asymmetric unit?

- Estimate this using Matthews Coefficient
  - Assumes roughly 50% of cell contents is solvent
  - Accounts for resolution

- This step in CCP4Cloud also creates the first "revision" of what our structure contains. At this point we know the sequence, the estimated number of copies and the reflection data measurements

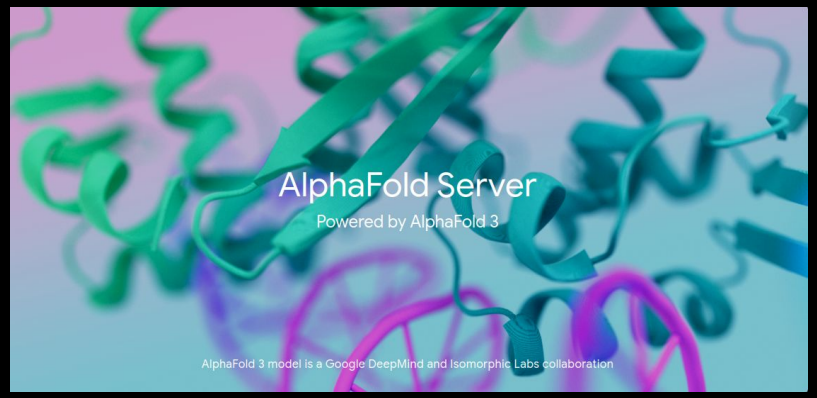
- The number of copies estimated will influence the MR strategy - large numbers may require larger search models

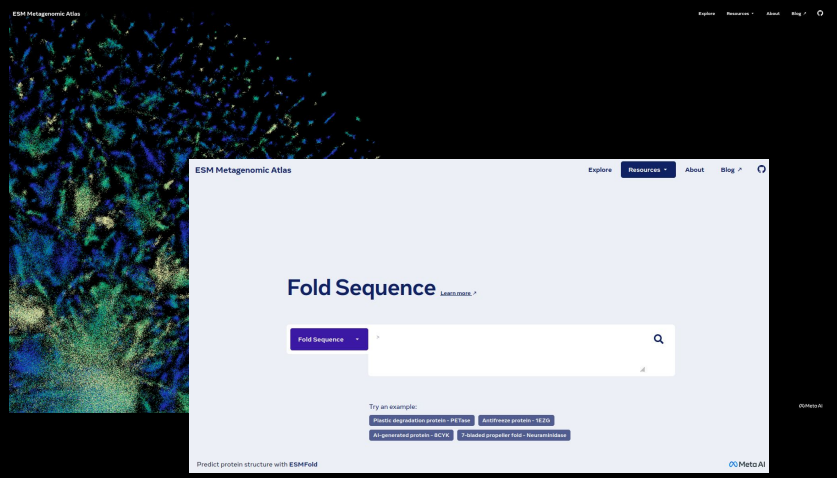


CCP4

Generating or sourcing predicted models

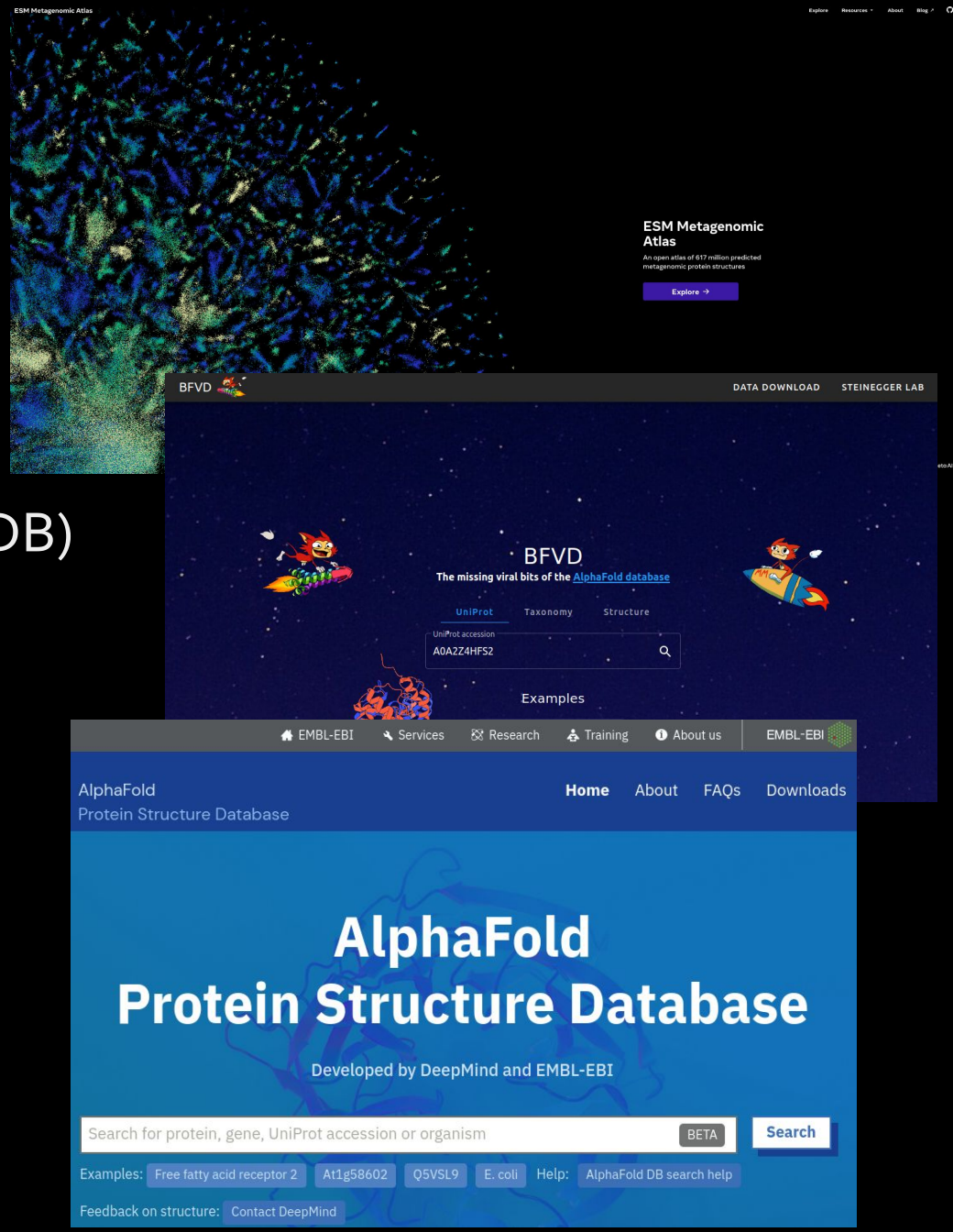
- Generating predictions online:
  - Online servers:
    - AlphaFold3 server
    - Boltz-1 server
    - Chai-1 server
    - Google Colab AlphaFold2 Notebooks
    - ESMFold
    - RoseTTAFold
  - Specialised servers (multimers):
    - Unifold Colab Notebook
    - AlphaFold2 Multimer





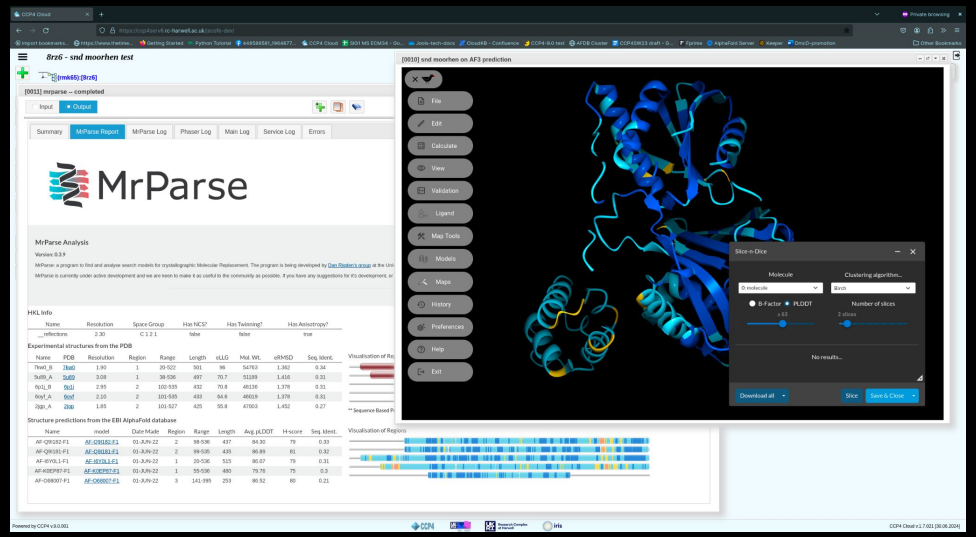
Generating or sourcing predicted models

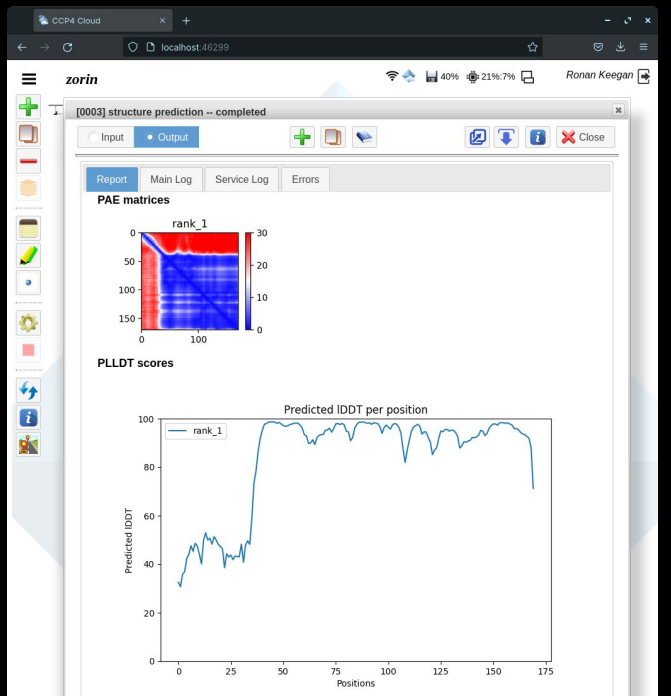
- Online databases:
  - AlphaFold2 Protein Structure Database (AFDB)
    - 214 million predictions
  - Big Friendly Virus Database (BFVD)
    - >300,000 virus structures not present in AFDB
  - ESM Metagenomic Atlas
    - 772 million predictions (using ESMFold)

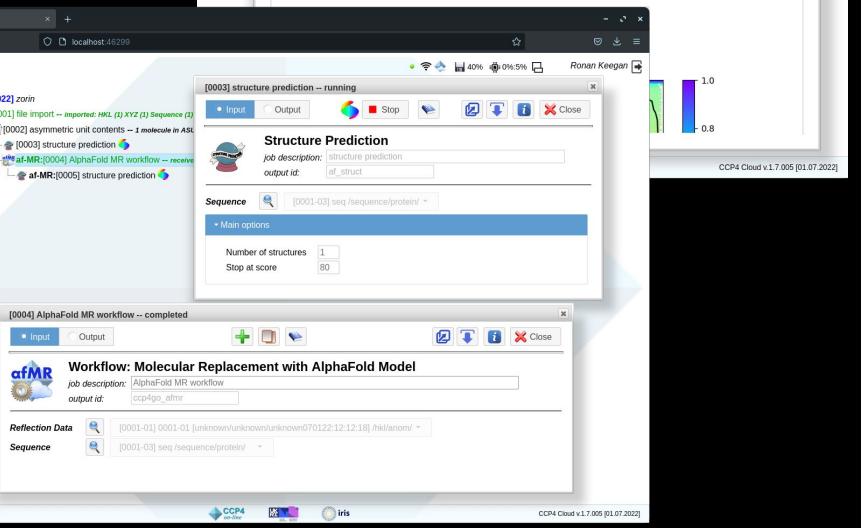


Generating or sourcing predicted models

- CCP4Cloud resources:
  - Structure Prediction task
    - Monomers, multimers and complexes
  - MrParse database search tool
    - Online version searches entire AFDB
    - Also searches PDB database

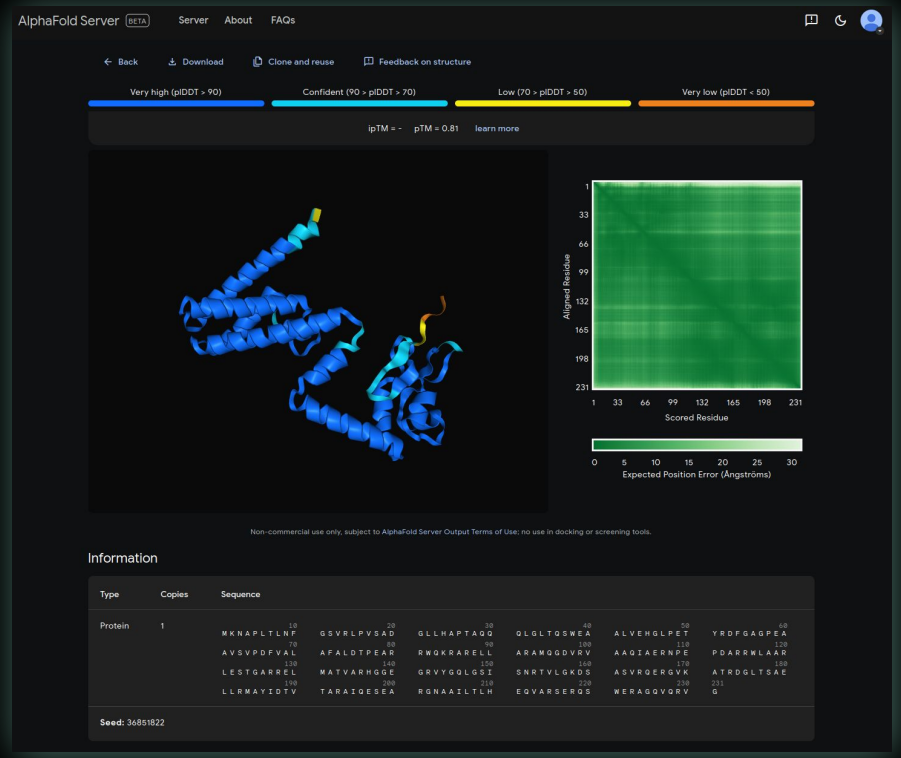






Generating predicted models: AlphaFold3

- ***AlphaFold3*** server has improved accessibility and capability of online prediction:
  a. Quick and easy to use
  b. Can predict multimers and complexes as well as nucleic acids and ligands
  c. Allows for larger numbers of residues than could be predicted with (free) Colab servers
  d. Accuracy of predictions is improvement on ***AlphaFold2***
  e. Limit on number of jobs/day and queue times can be long



Predicted model scoring:

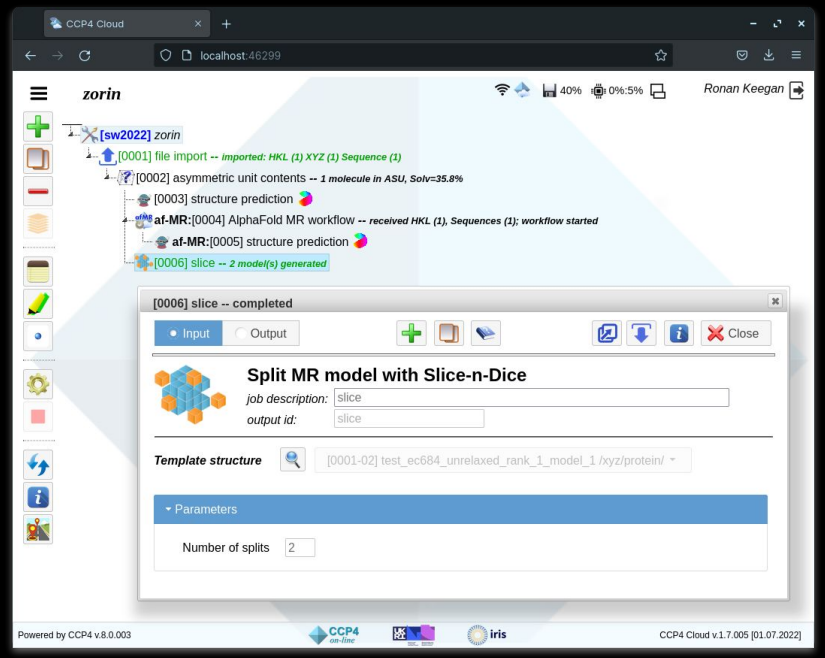
- AlphaFold convention is a per-residue confidence score (pLDDT) between 0 and 100
- Newer convention is per-atom pLDDT
- Other prediction applications use same convention, although you may see fractional values (0 to 1.0) or r.m.s.d estimates

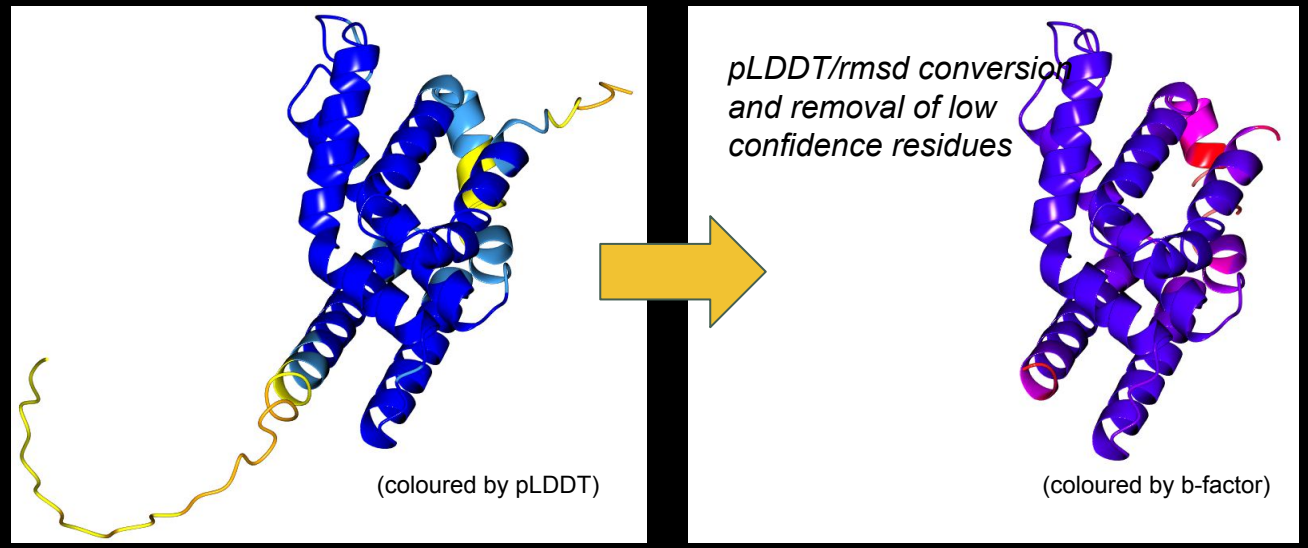
Scoring can be used to eliminate residues unlikely to be present or in same position in crystal structure

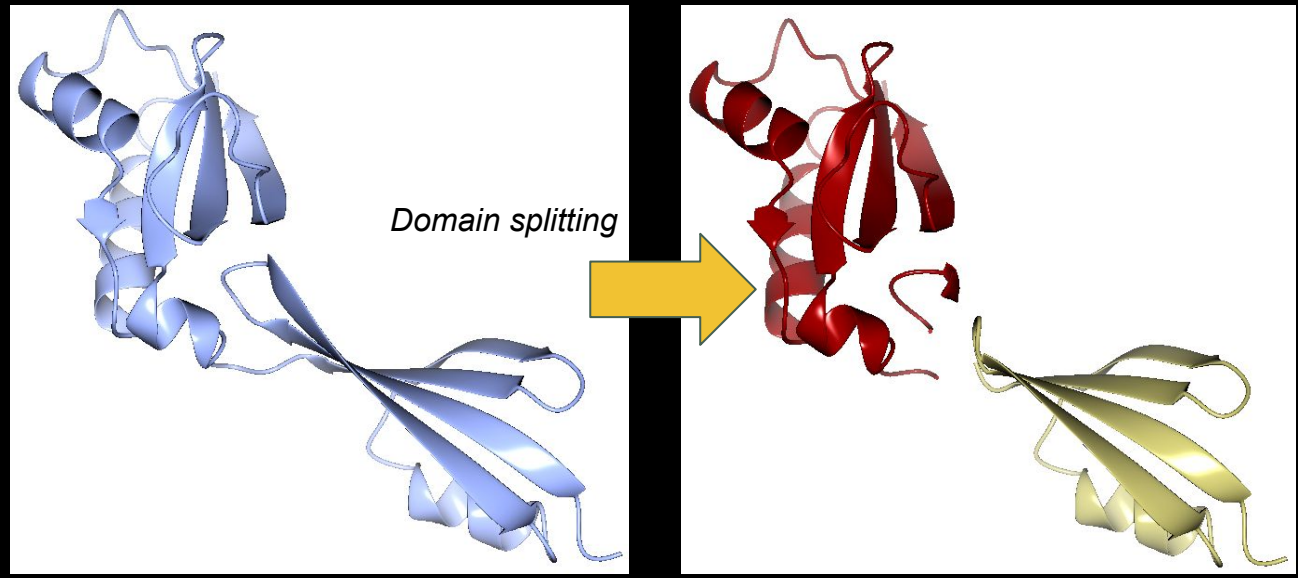
CCP4

"Slice" Task - model preparation

1. Conversion of residue confidence scores - pLDDT (*AlphaFold*) or rmsd (*RoseTTAFold*) into pseudo b-factors
2. Elimination of low confidence predicted residues
3. Identification and splitting of domains/clusters into separate search models







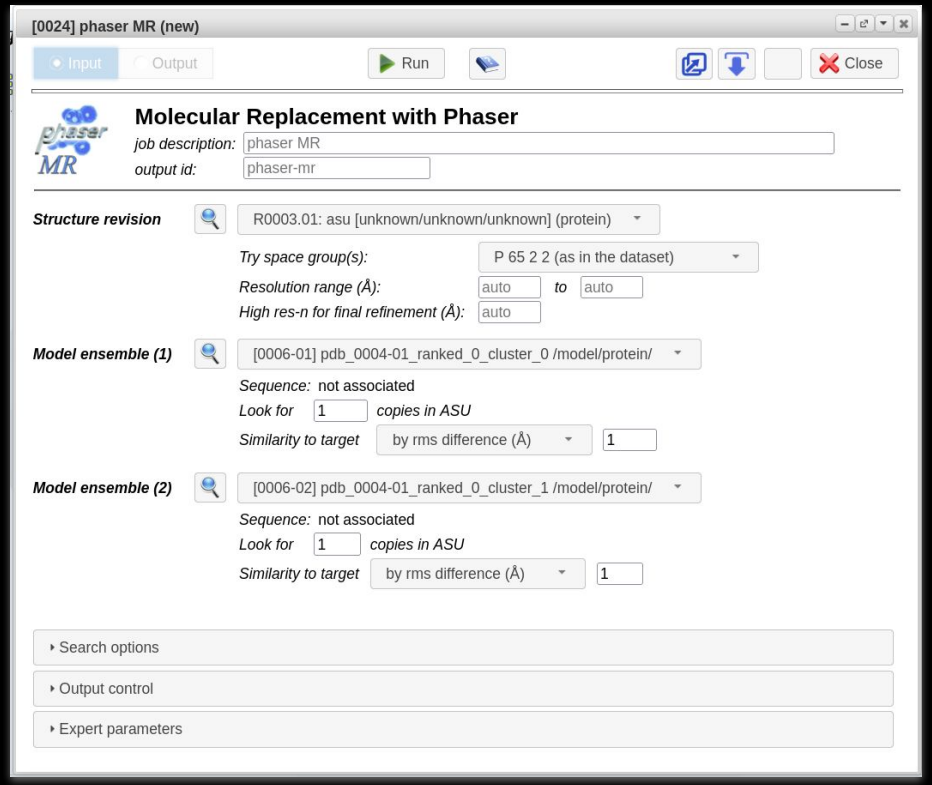
Molecular Replacement in CCP4

- CCP4 Cloud has two programs for doing Molecular Replacement ("MR Solvers")
  - Molrep
    - Automated MR
    - Several useful features e.g. searching a map
  - Phaser
    - Maximum likelihood approach
    - Accounts for potential model errors
    - Best for difficult cases and for correctly positioning fragment search models

CCP4

Molecular Replacement: Phaser

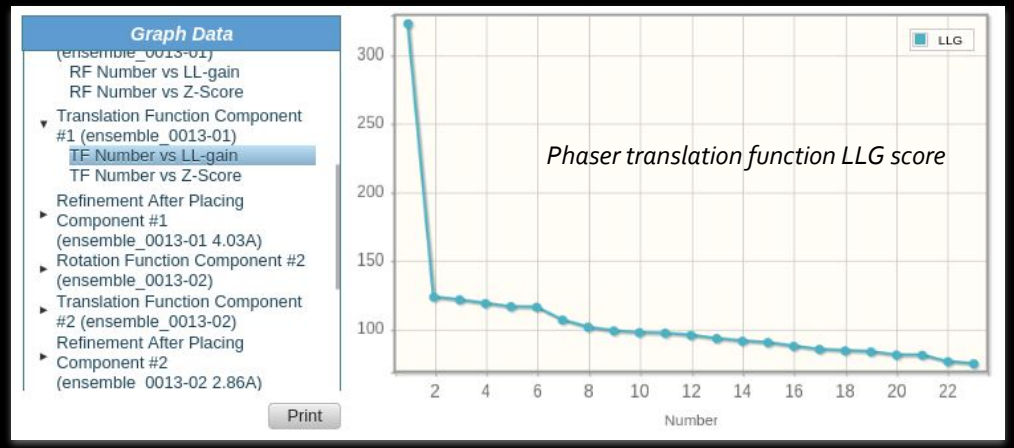
- Important points on using Phaser
  - Phaser accounts for errors in:
    1. Model
       - Provide accurate details of AU composition
    2. Data
       - Provide intensities – internally works out amplitudes accounting for experimental errors
  - Phaser performs clever decision making for automation
    - Provide minimal details and let Phaser make its own decisions e.g. search order, search all possible space groups
    - If it doesn't work take step-by-step approach – 1 copy at a time

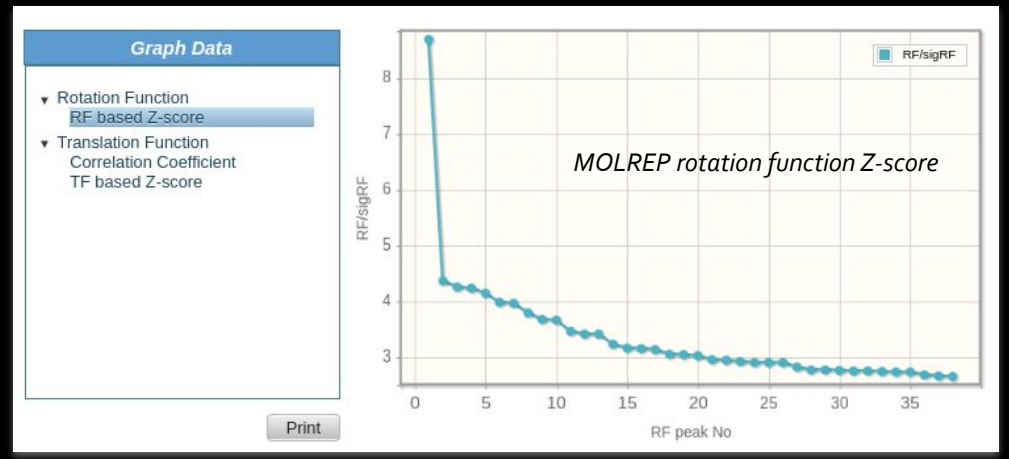


How do I know my MR solution is successful?

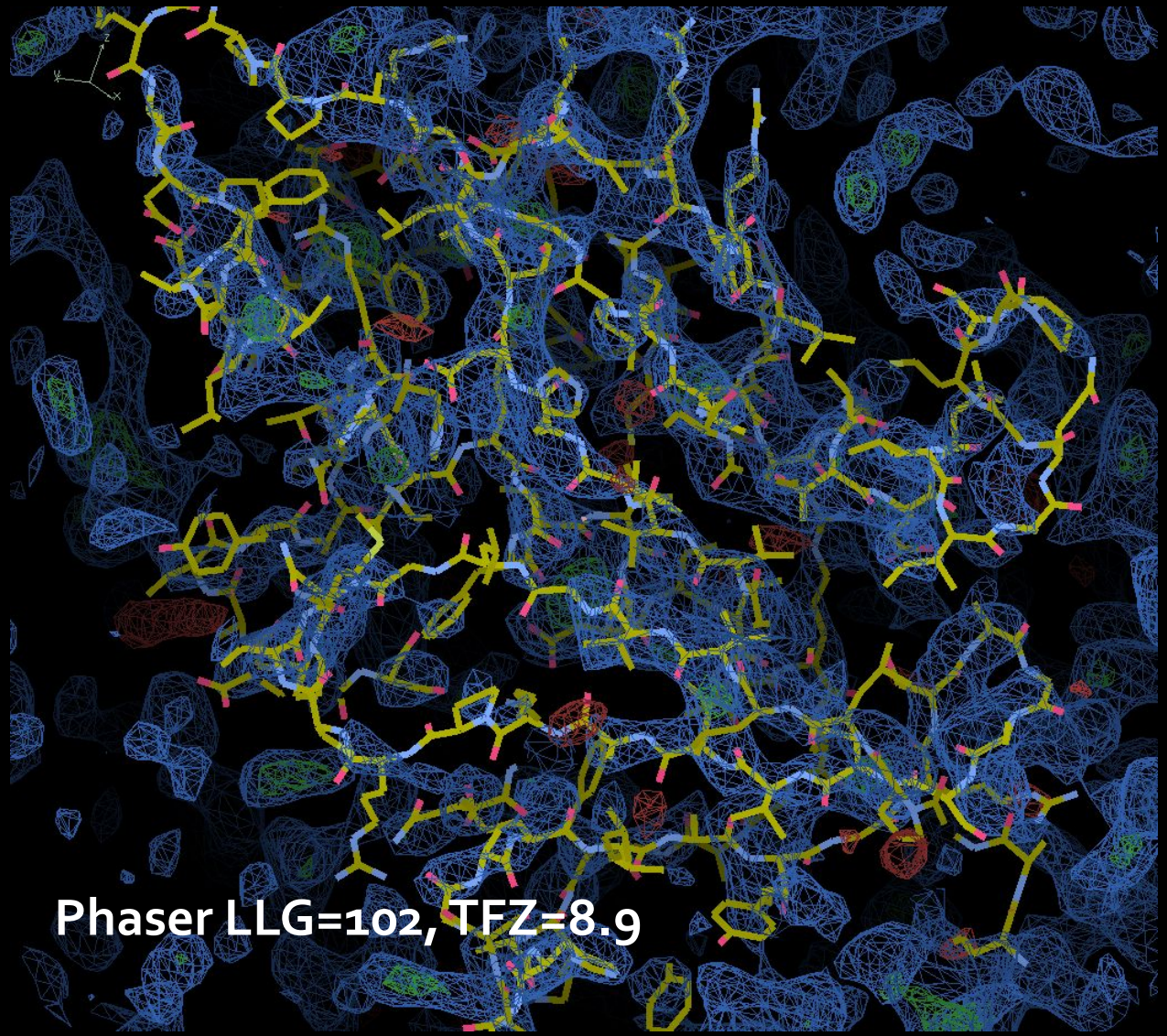
- Rough guide to MR program scoring
  - Phaser scores
    - LLG scores – has it increased by 60 or more after the placement of a new molecule?

      (resolution and space group dependent)
    - TFZ – greater than 8?
    - Few or single solution almost always indicative of success
  - Molrep scores
    - RFZ – rotation search score greater than 5 – is there a clear peak?
    - TFZ – translation search score – is there a clear peak?



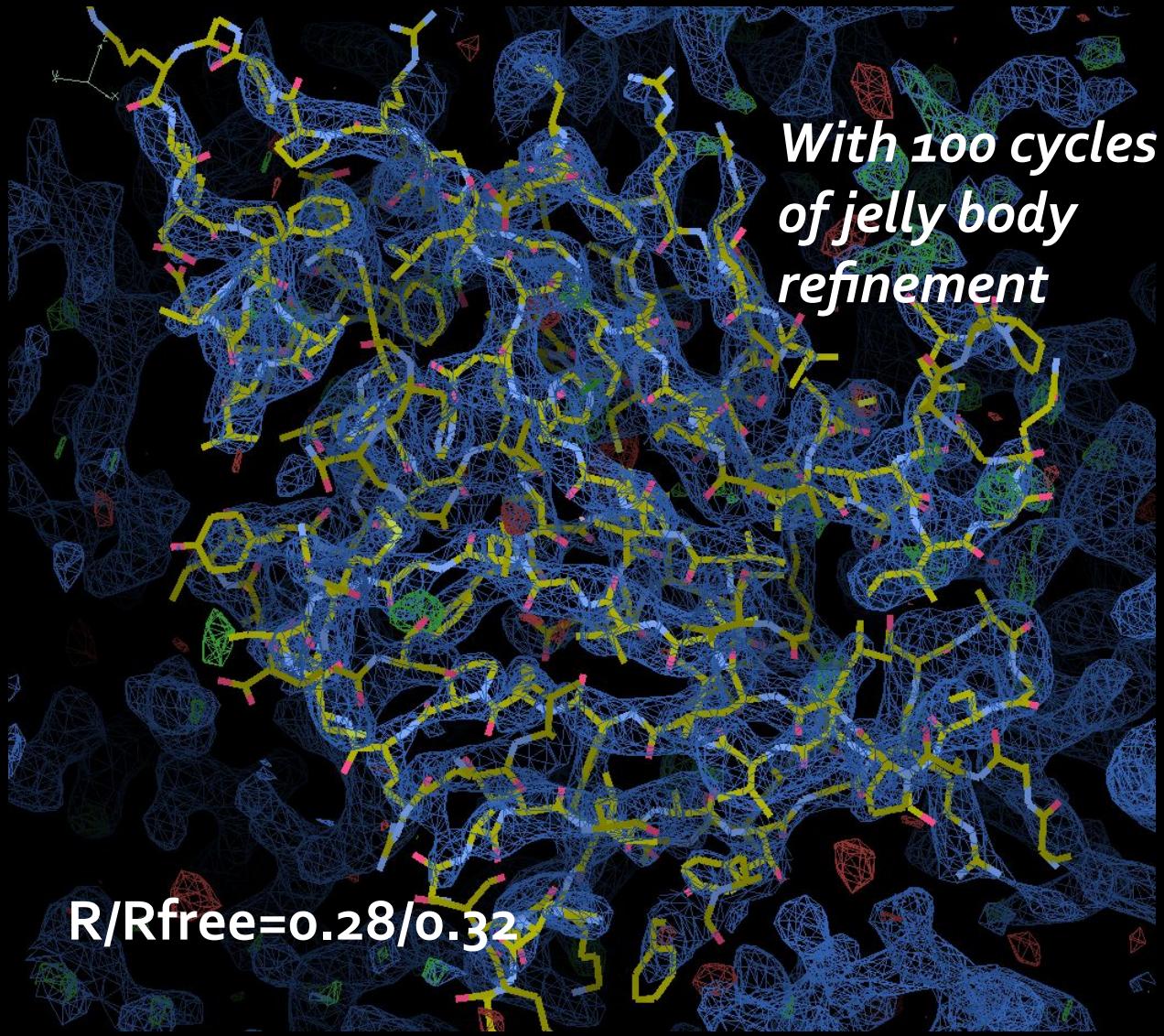


Refinement



- The electron density map is never a good measure of correctness of the solution immediately after MR despite good scores from the MR program

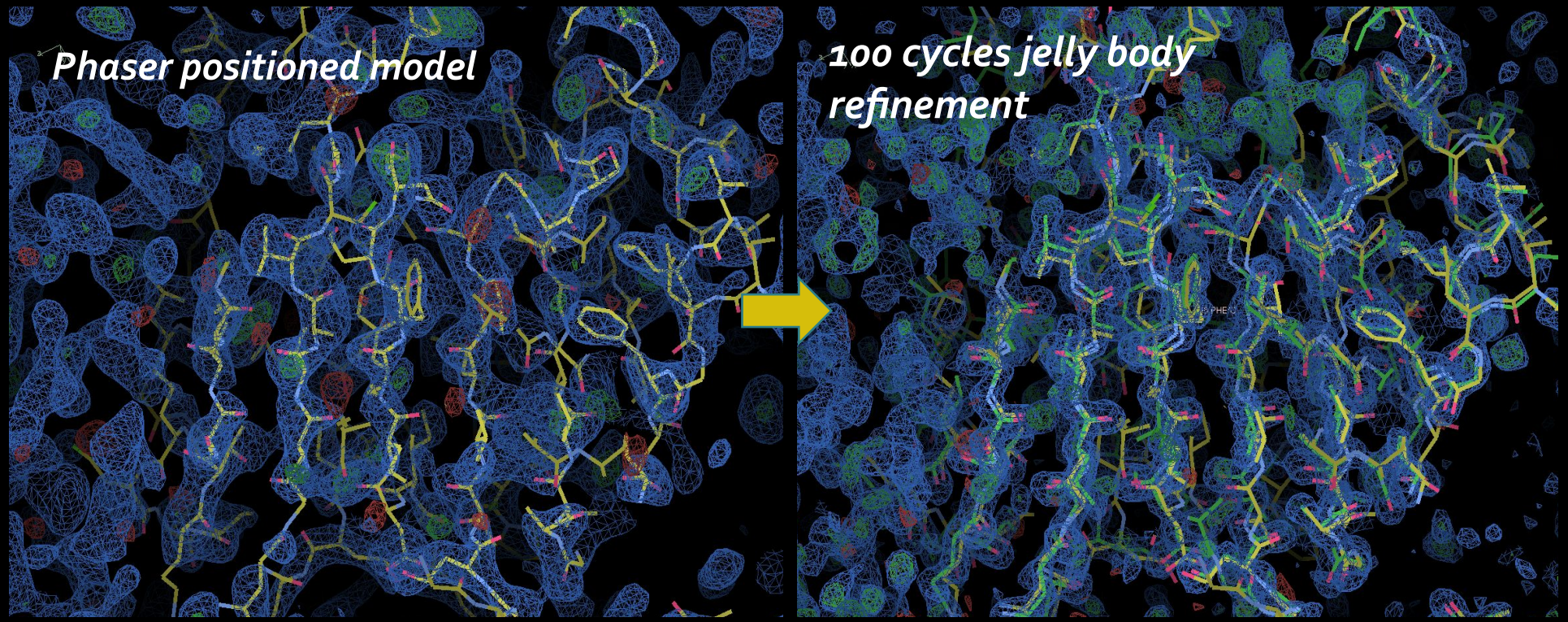
CCP4



- The electron density map is never a good measure of correctness of the solution immediately after MR despite good scores from the MR program
- The model will always require many cycles of refinement to improve the agreement between the model and the data, and hence the correctness of the phases derived from the model

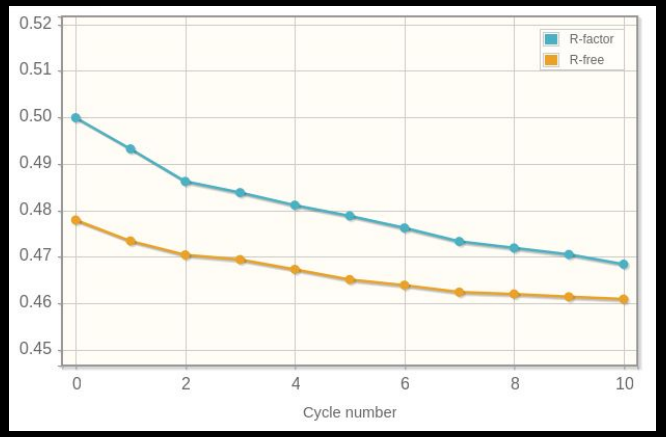
CCP4

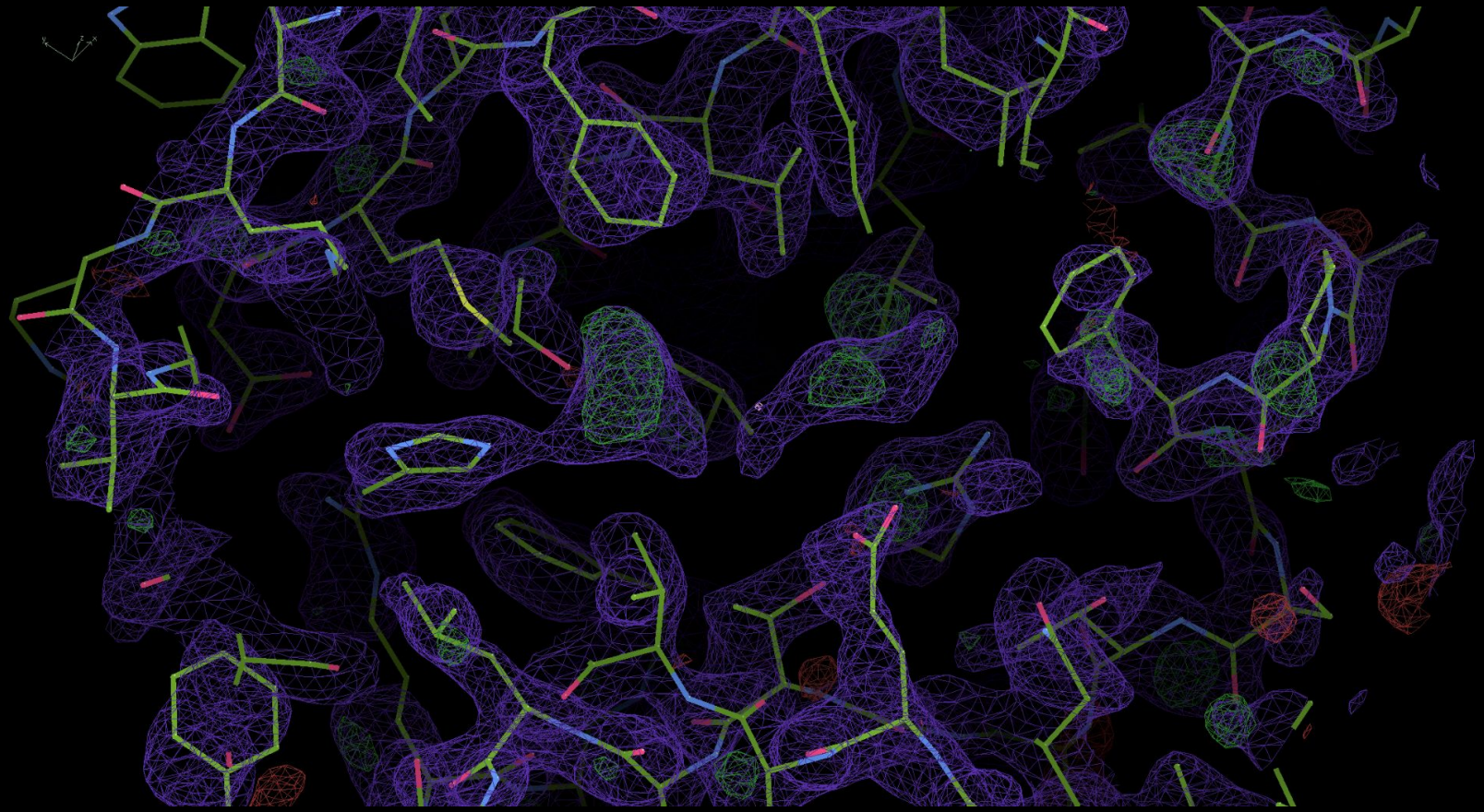
- Refinement
  - Look at Rfactor/Rfree
    - Are they falling? Is Rfree below 0.5?
  - Use 100 cycles of jelly body refinement option in Refmac post MR



Refining predicted models used in Molecular Replacement

- Predicted models (*AlphaFold2*, *Colabfold* etc.) are often significantly different in their main & side chain positioning to the crystal form despite making good MR search models
- They can require lots of cycles of jelly-body refinement in Refmac



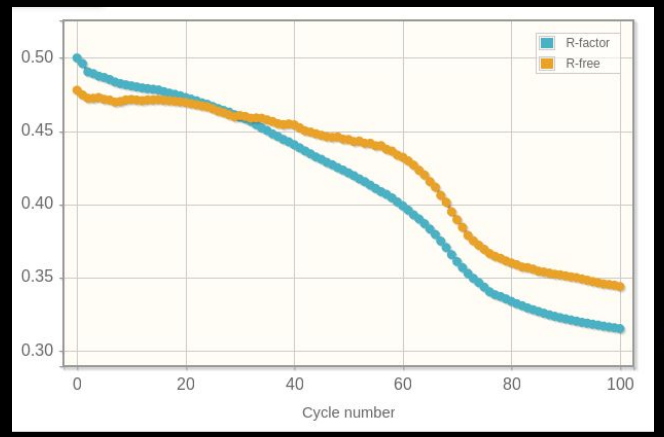


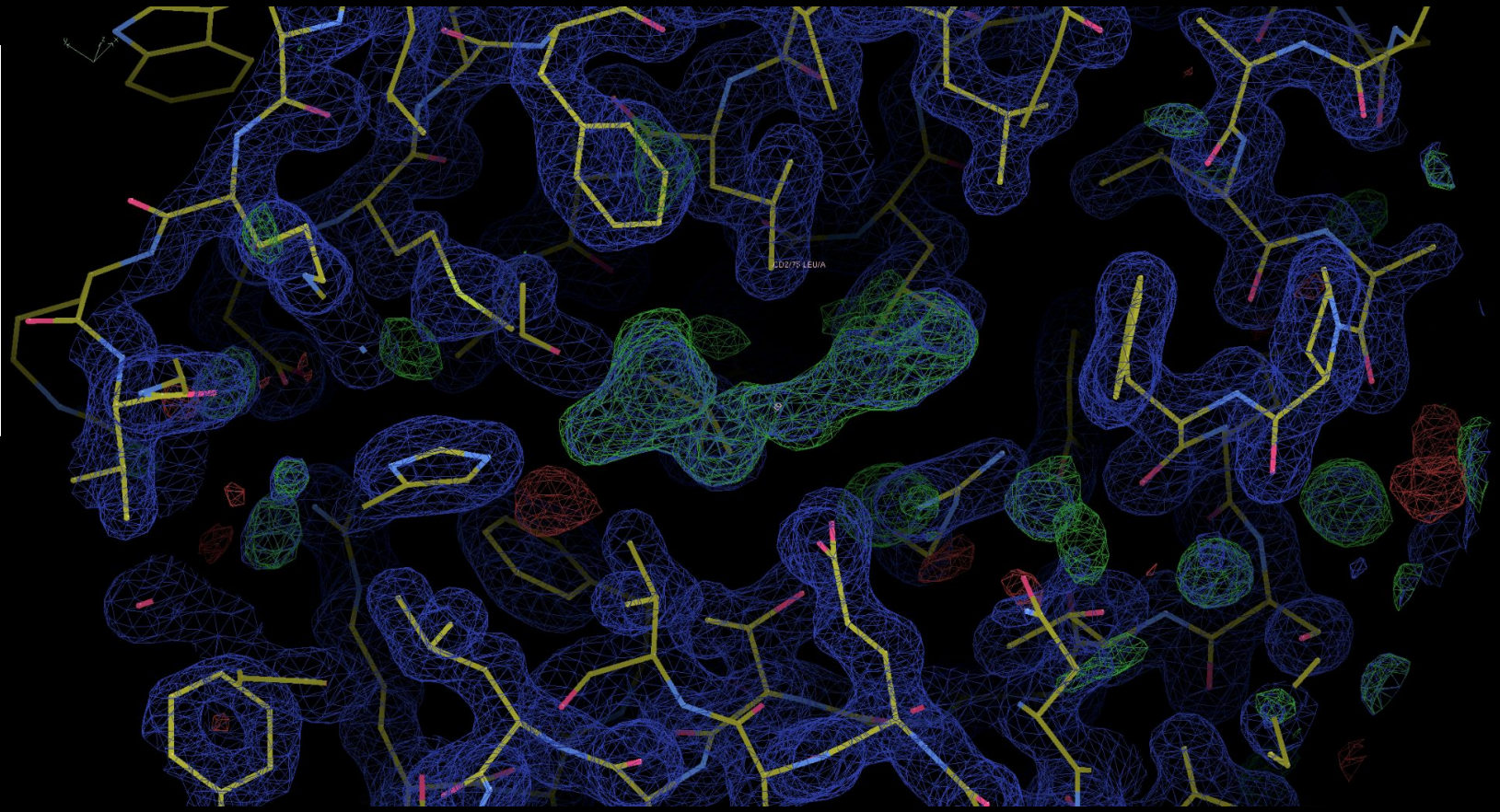
Example: 10 cycles of jelly-body refinement with Refmac
Rfactor = 0.46
Rfree = 0.47

CCP4

Refining predicted models used in Molecular Replacement

- Predicted models (*AlphaFold2*, *Colabfold* etc.) are often significantly different in their main & side chain positioning to the crystal form despite making good MR search models
- They can require lots of cycles of jelly-body refinement in Refmac

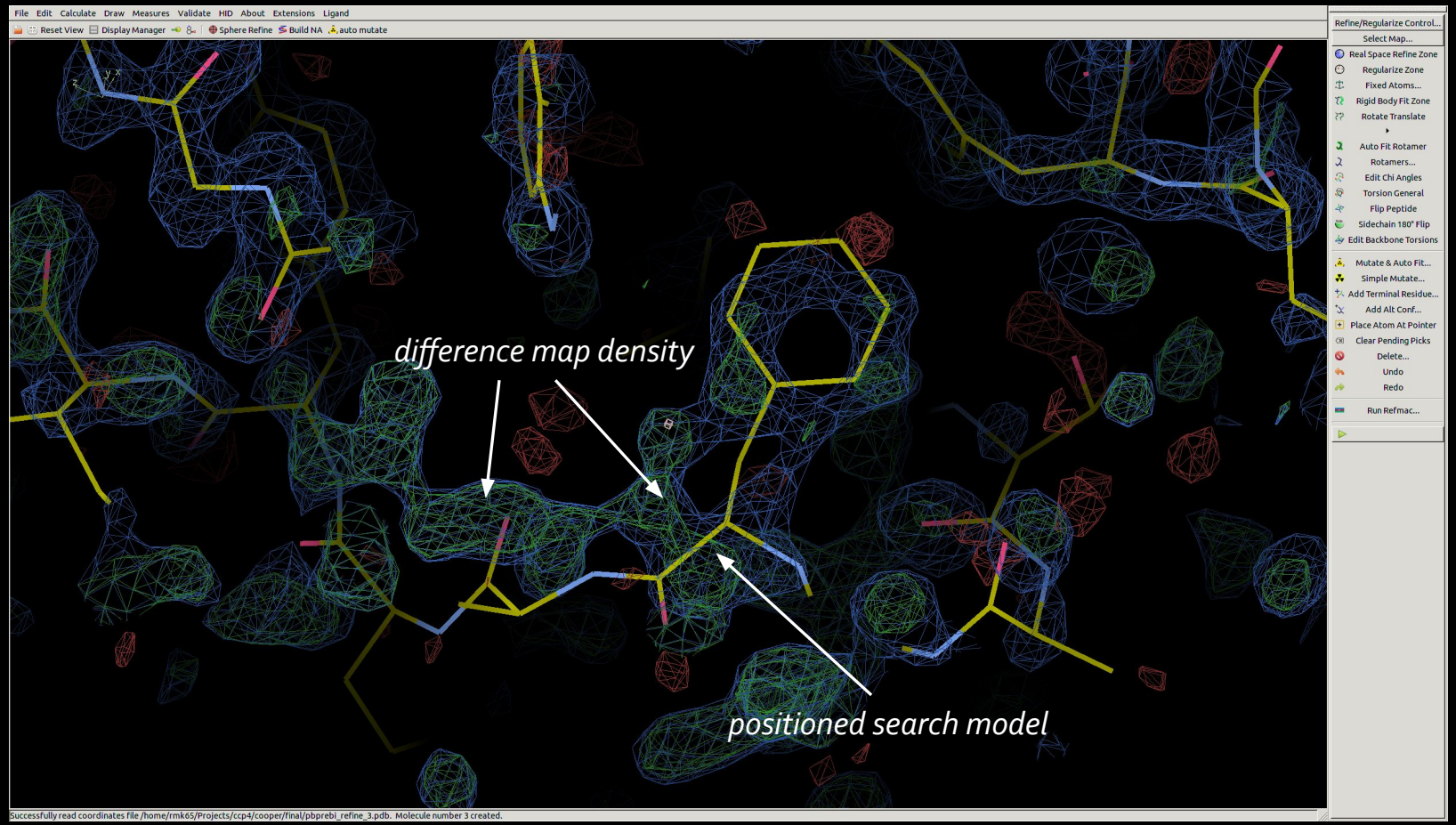




Example: 100 cycles of jelly-body refinement with Refmac
Rfactor = 0.35
Rfree = 0.32

CCP4

- Examine solution by eye (after refinement)
  - Use Coot or Moorhen to examine positioned models & maps



Tutorial

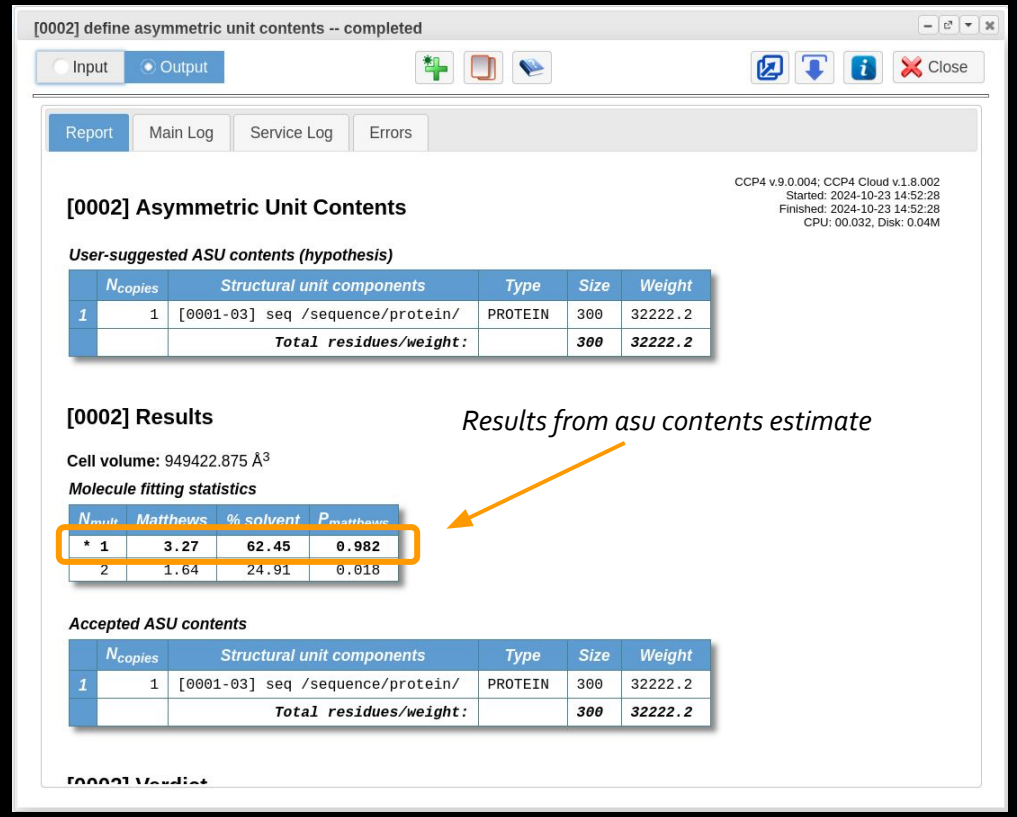
Example 1 instructions

Create a folder "MR tutorials" and new project within it called "MR-example-1" in CCP4Cloud

1. **Import the tutorial data** using the "Import from Cloud" option in "Data import". Select "Tutorials" -> "Data" -> "2_phasing" -> "mr-1-simple". Select all of the files using the Shift key. The import includes sequence, mtz (reflection data) and this pdf document.

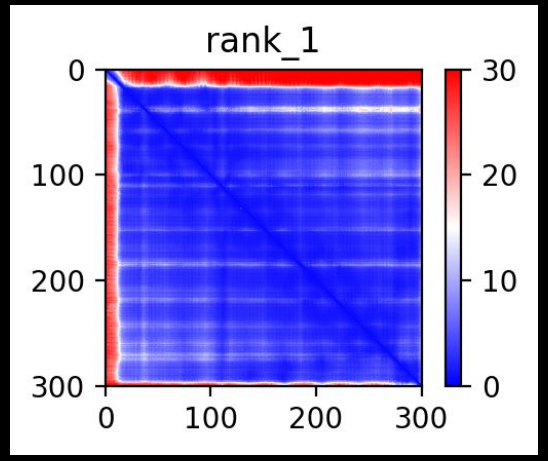
2. **Examine the summary** from the import
   - Find the reflection data details such as spacegroup and resolution
   - We will use the sequence as input to structure prediction to generate a search model for use in MR

3. Next we need to determine the contents of the **asymmetric unit**. Add the "Asymmetric unit contents" task. Run the task using default options. The result should be a prediction of 1 copy in the asymmetric unit with a solvent content of 62.45% and a high probability score (98.2%).

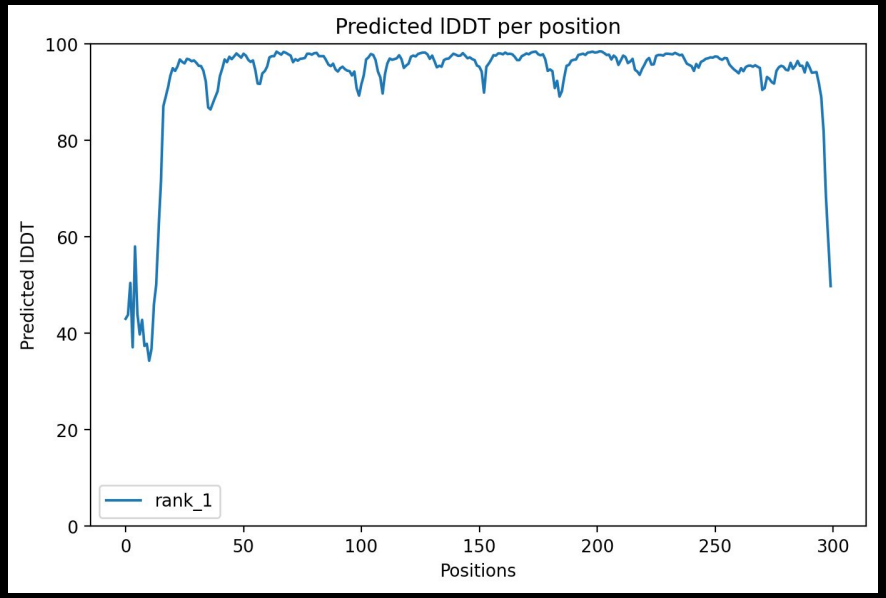


Example 1 instructions ctd..

4. The next step is to **generate a predicted model** for use as a search model in MR. Select the "structure prediction" task from the "Structure prediction" menu. As predicted in the previous task, there is only a single copy of the structure in the asymmetric unit cell. So we will not consider predicting the structure in a complex. Leave "Number of copies in complex" as 1. Leave number of predictions as 1. Click the "Run" button.



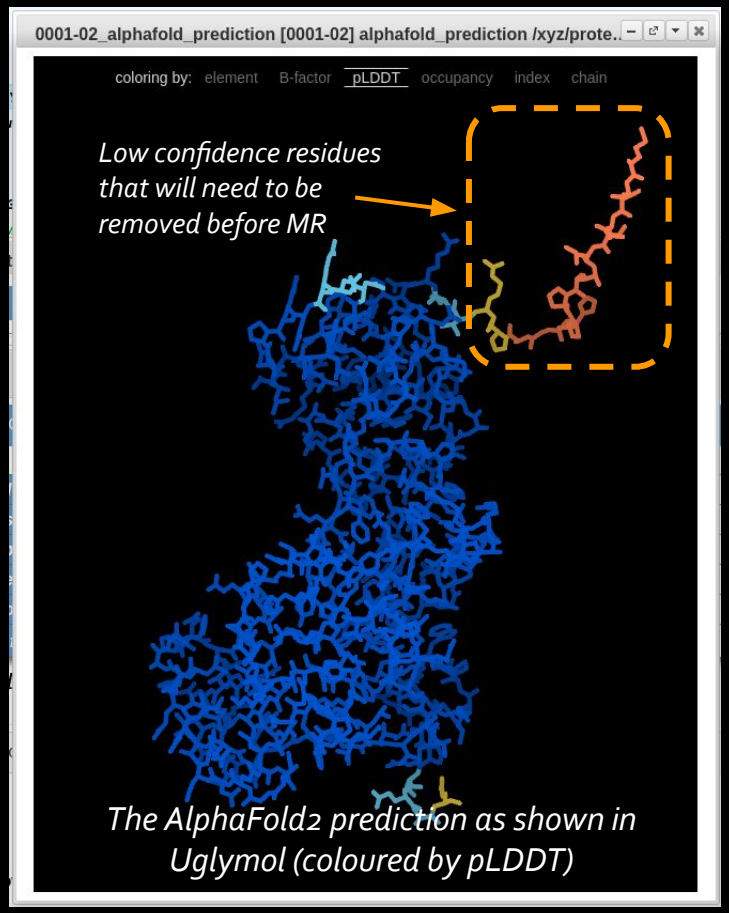
5. The report from the task will show several plots. Two of interest are the **PAE (Predicted Alignment Error) matrix** and the **Predicted LDDT (pLDDT) score**. The PAE matrix shows a score indicative of how well the residues are predicted to align to the true structure. Low scores (blue) show close predicted alignment, high scores (red) show poor alignment. The pLDDT score is a confidence score that shows the per-residue confidence. We can see from the plot that the residues in the N-terminus have low confidence.



Example 1 instructions ctd..

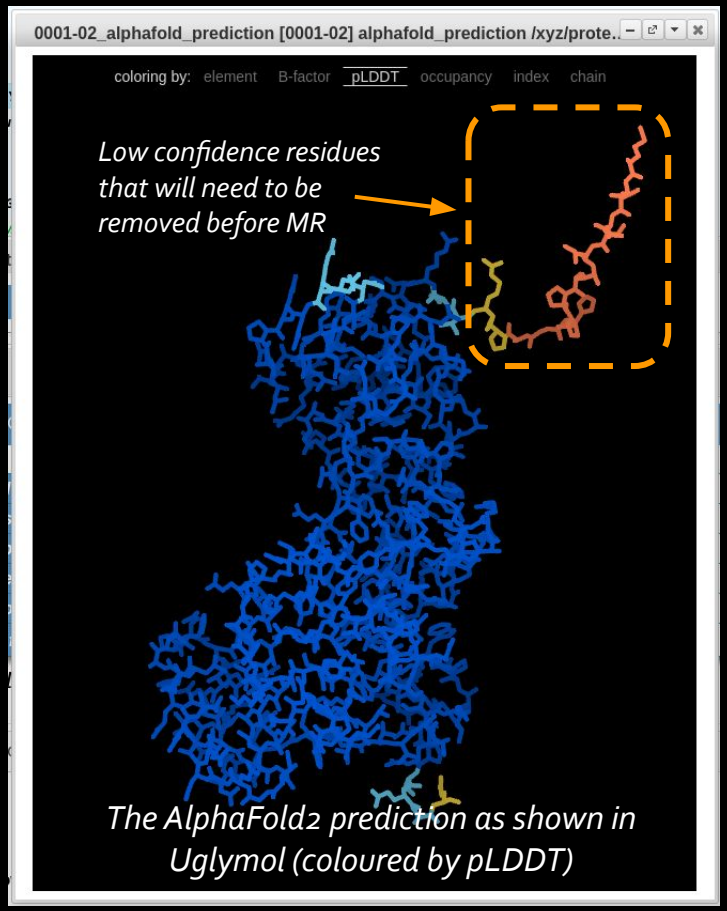
6. Examine the structure using the "**Uglymol**" viewer.
   - Change the colouring using the "C" key. Switch colouring to "pLDDT" to see the confidence scoring for each residue. (*blue is high confidence, yellow/orange is low*)
   - Before we can use the model in MR we'll need to remove the low confidence parts and convert the pLDDT values to a B-factor estimates. Phaser makes use of B-factor information to aid placement of the model in MR.

     (*Note: press "H" for general help in Uglymol*)



Example 1 instructions ctd..

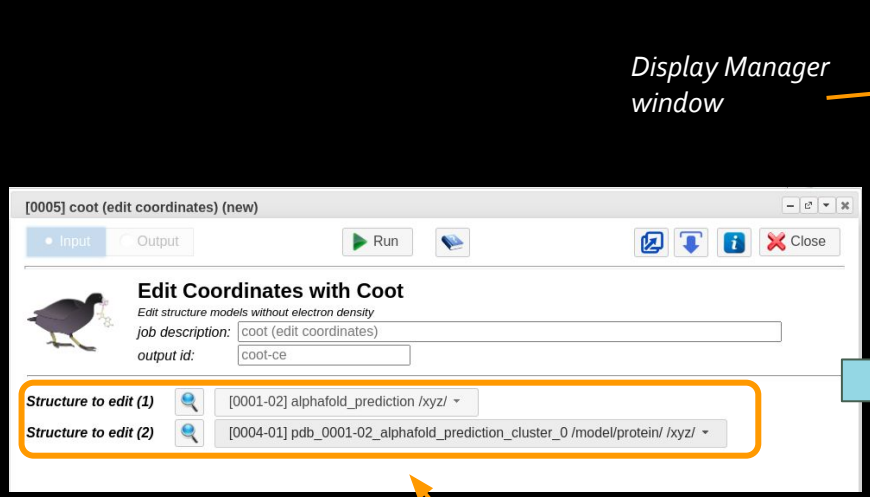
7. The positioning of low confidence residues is likely to differ from what is in the crystal structure, so these should be removed before we attempt MR. To do this we run the "Split MR model with **Slice-n-Dice**" task from the "MR Model Preparation" menu under "Molecular Replacement".
8. pLDDT values are stored in the B-factor column of a coordinate file. The "**Slice-n-Dice**" slice task will remove low confidence residues based on the pLDDT score. This task also converts pLDDT values to B-factor estimates. In addition, this task can "split" a model into domain parts if needed. Often the relative orientation of domains in predicted models will differ from that in a crystal structure and splitting the model will allow each domain to be more easily placed. This will be the subject of another tutorial. Here we have a single globular domain so no splitting is required. Leave "Number of splits" as 1.
9. Run the task and again view the output model in "**Uglymol**". Note that the low confidence residues will have been removed and the pLDDT values will have been converted to B-factor estimates.



The AlphaFold2 prediction as shown in Uglymol (coloured by pLDDT)

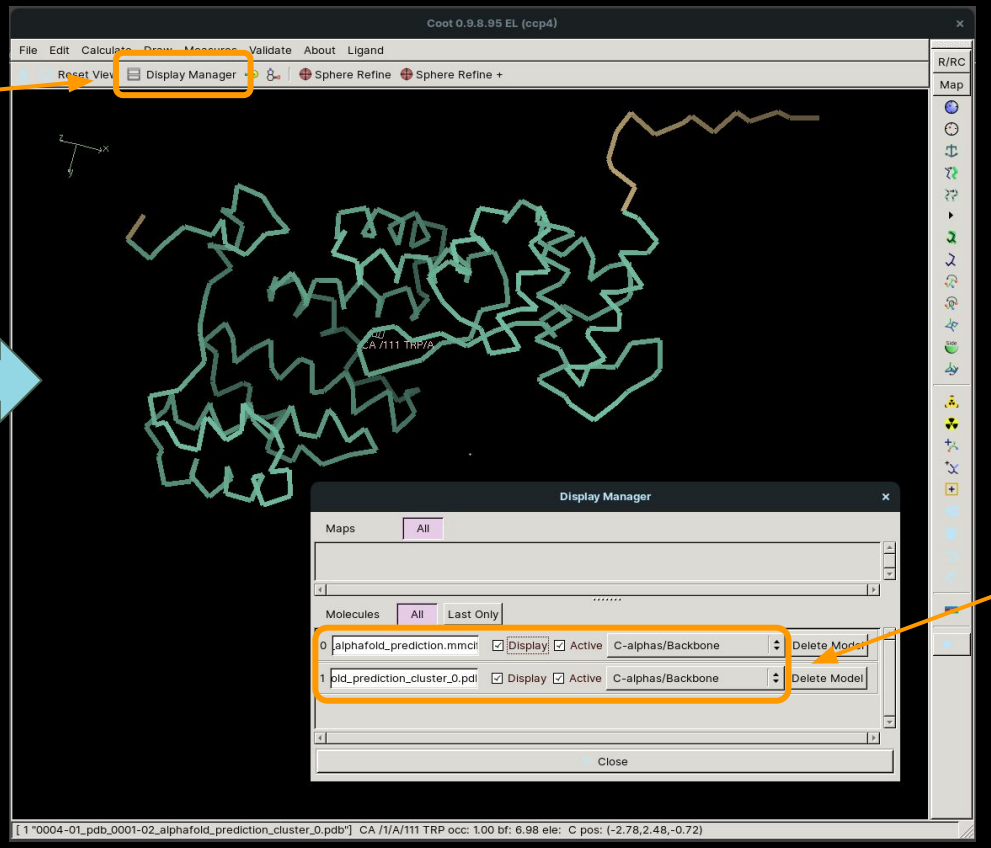
Example 1 instructions ctd..

10. We can also **compare the search model before and after** the process in "Coot" by adding a Coot task and loading both the original and modified predicted models. Switch on and off the model display to see what has changed



Load both models into Coot through the "Edit Coordinates with Coot" task interface

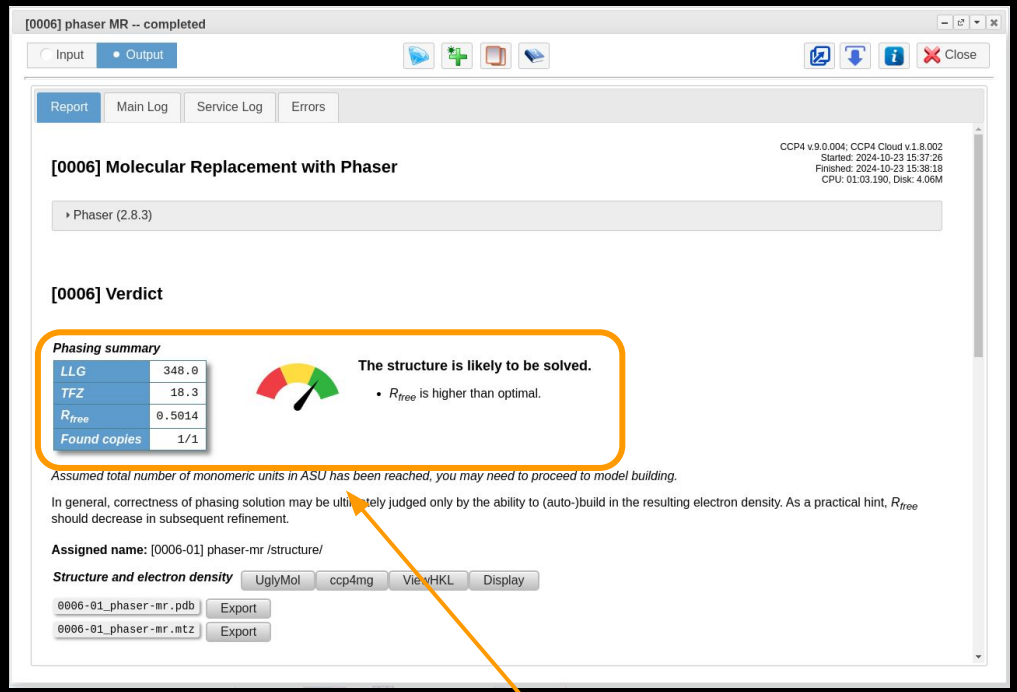
Display Manager window

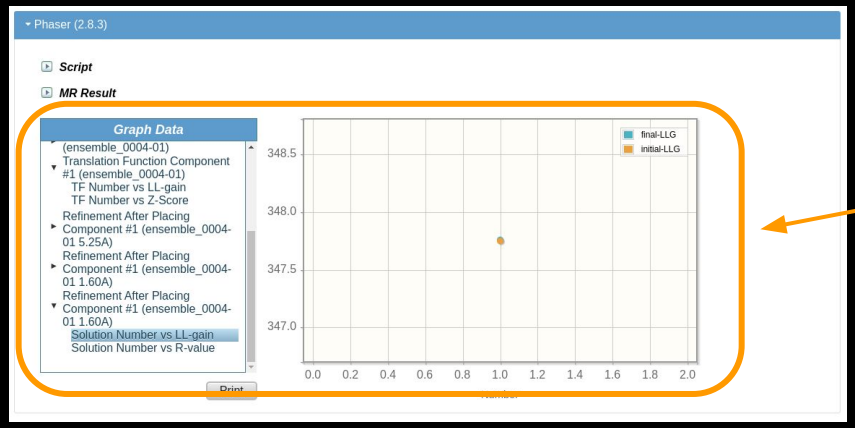


Toggle the display option for the models to see what has changed before and after preparation

Example 1 instructions ctd..

11. The model is now ready for MR. **Add a Phaser task** from the "MR solvers menu"
    - All the data and model are provided to Phaser automatically
    - Run the task and view the output report
    - Examine the increase in the LLG score as each step is taken (rotation function, translation function, refinement steps)
    - We expect the LLG to rise each time with the final refined LLG being greater than 6o for a good solution
    - A translation function Z-score (TFZ) of greater than 8 is also indicative of a good solution.



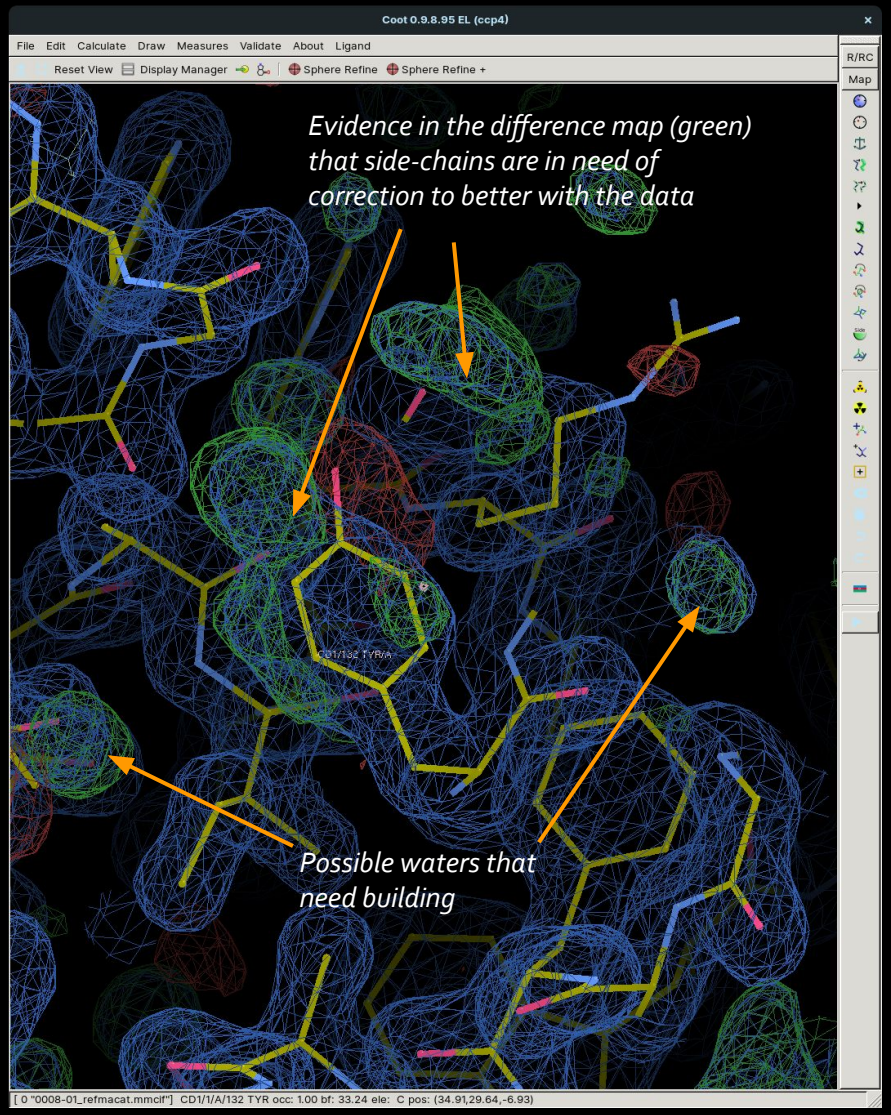


The plot showing the results of the final rigid-body refinement step in Phaser. The LLG value is about 348, well above the expected value for a solution of 6o. A single point solution is also a strong indication of a correct solution. This means that the correct placement of the model stands out from the background of incorrect placements.

The summary from Phaser showing the LLG and TFZ scores for the placed model. The Rfree value at this point is high. We will need to perform jelly-body refinement on the placed model to improve it.

Example 1 instructions ctd..

12. When Phaser has finished follow that task with **refinement using the "Refmacat"** task from the "Refinement" menu
    - Use many cycles (100) of the "jelly-body" refinement option under the "Restraints" section
    - Watch the R/Rfree values decline to a steady state over many cycles
    - The final R/Rfree values should be in the range 0.2->0.3
13. **View the resulting model and map** in Uglymol or Moorhen or Coot
    - Look for clear evidence in the difference map of parts of the model that don't match the data. This could be positive difference density (green) for incorrect side chains or parts of the main chain that need building or water molecules that have not been added
    - At this point the model can be improved through iterative cycles of manual model building in Coot and refinement in Refmacat.



View of refined model in Coot