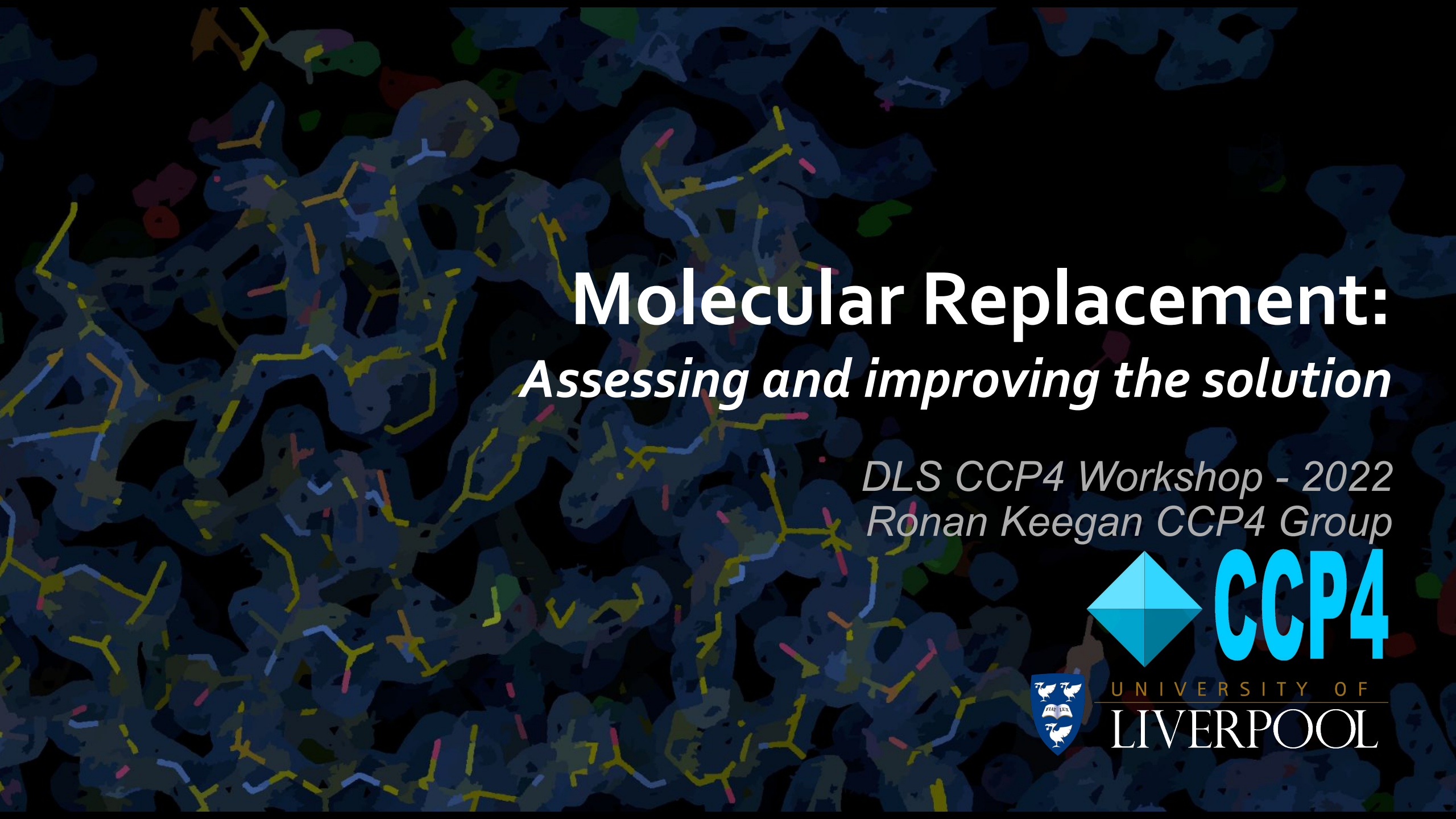


Molecular Replacement:

Assessing and improving the solution

DLS CCP4 Workshop - 2022

Ronan Keegan CCP4 Group

CCP4

UNIVERSITY OF LIVERPOOL

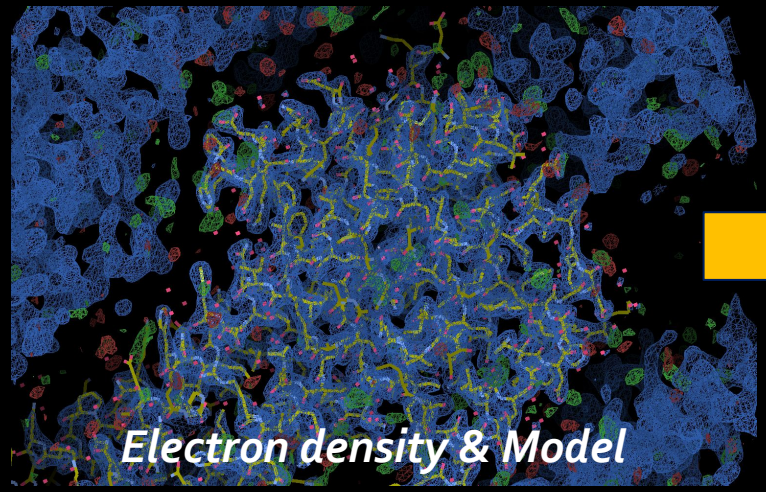
The Phase Problem

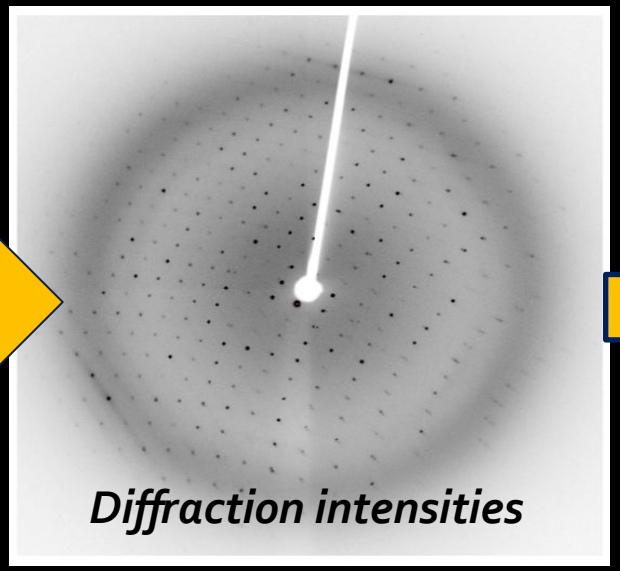
$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} |F_o(hkl)| \cos[-2\pi(hx + ky + lz) + \varphi_{hkl}]$$

Electron density at (x,y,z)

Amplitudes

Phases





?

CCP4

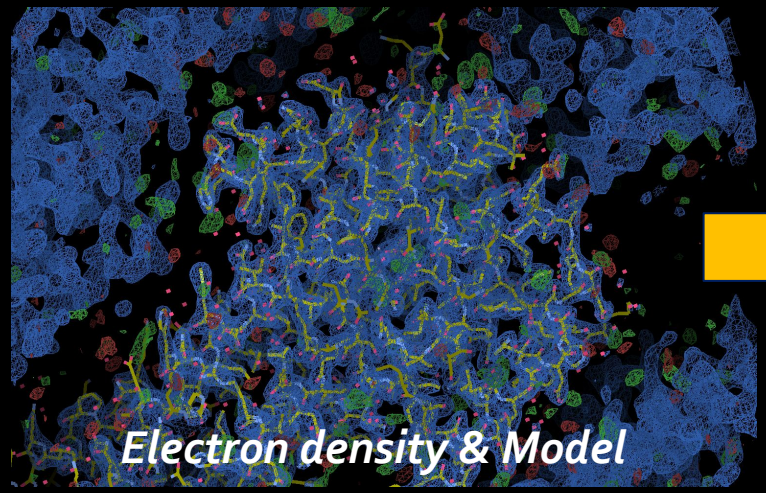
The Phase Problem: Experimental Phasing

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

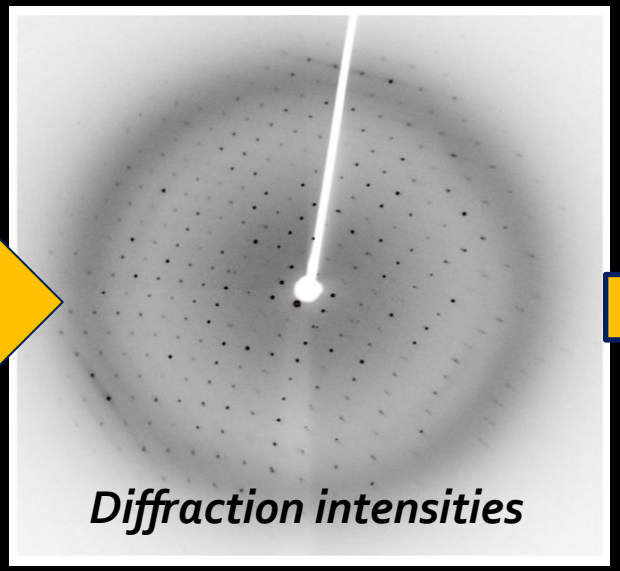
Electron density at (x,y,z)

Amplitudes

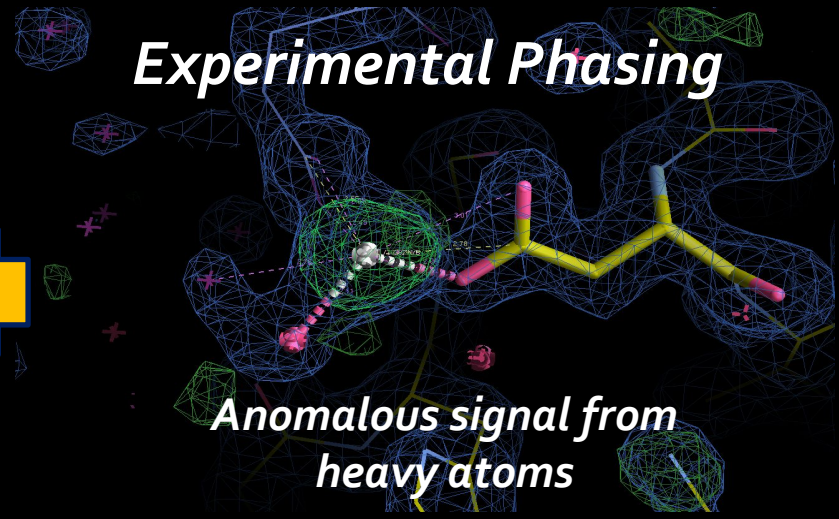
Phases



Electron density & Model



Diffraction intensities



Experimental Phasing

Anomalous signal from heavy atoms

CCP4

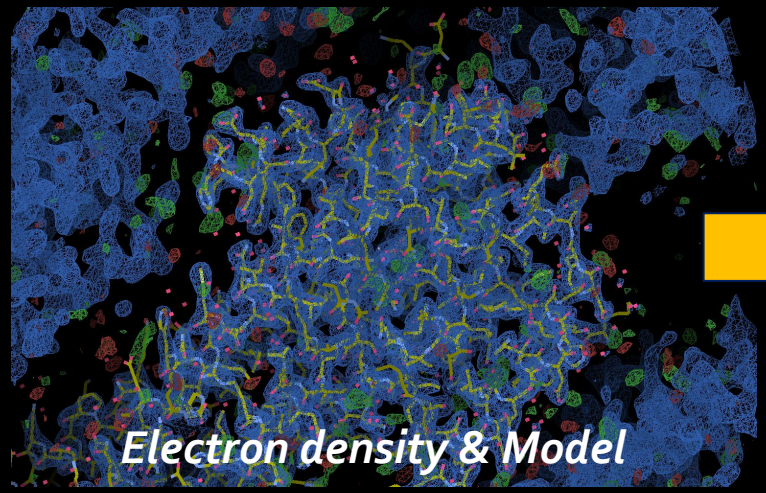
The Phase Problem: Molecular Replacement

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

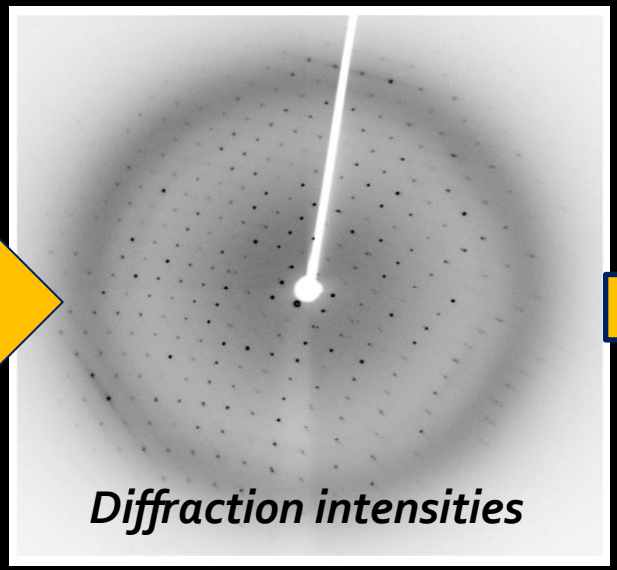
Electron density at (x,y,z)

Amplitudes

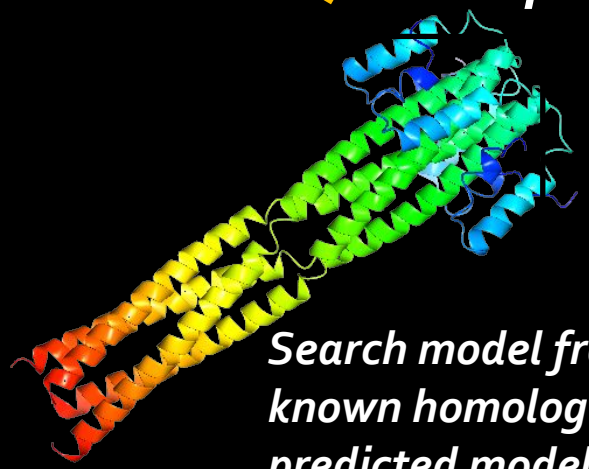
Molecular Replacement



Electron density & Model



Diffraction intensities



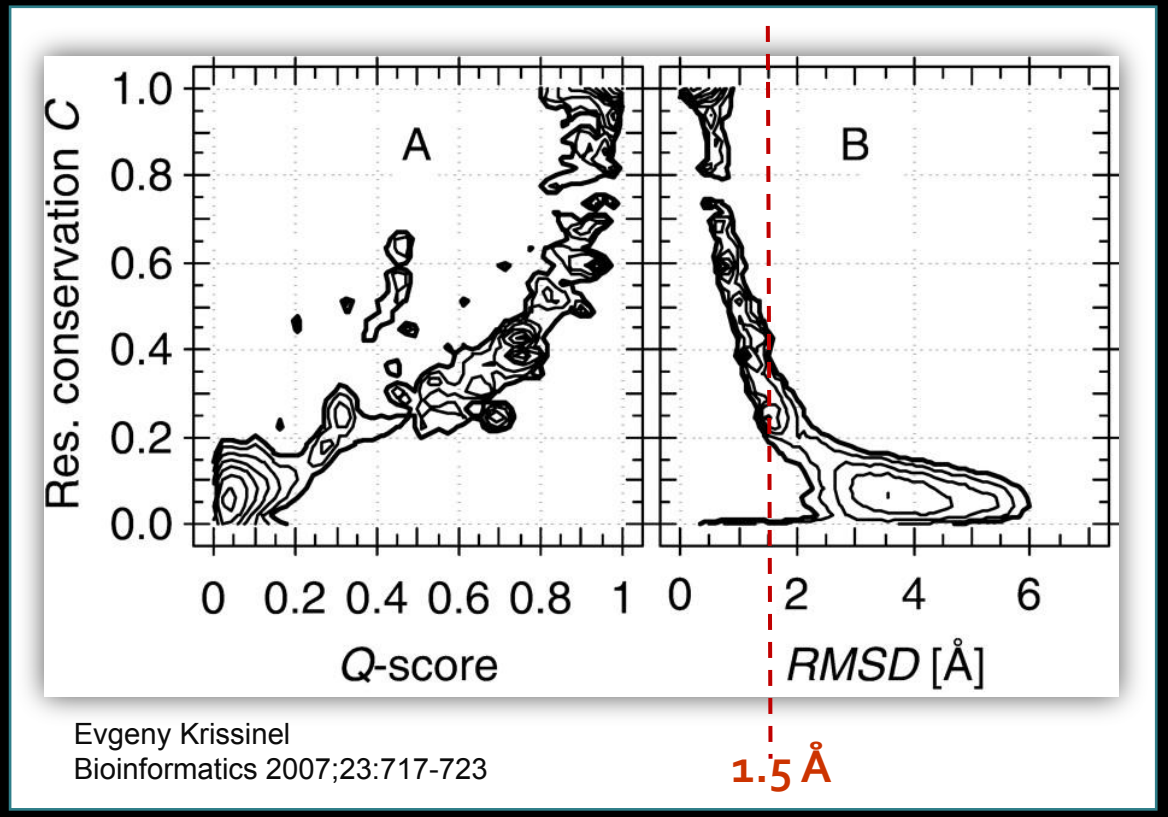
Search model from known homologue or predicted model

CCP4

Search Models

Before CASP14 (2021): *Finding a search model*

- How to find a search model?
  - Amino acid sequence similarity often correlates well with structural similarity



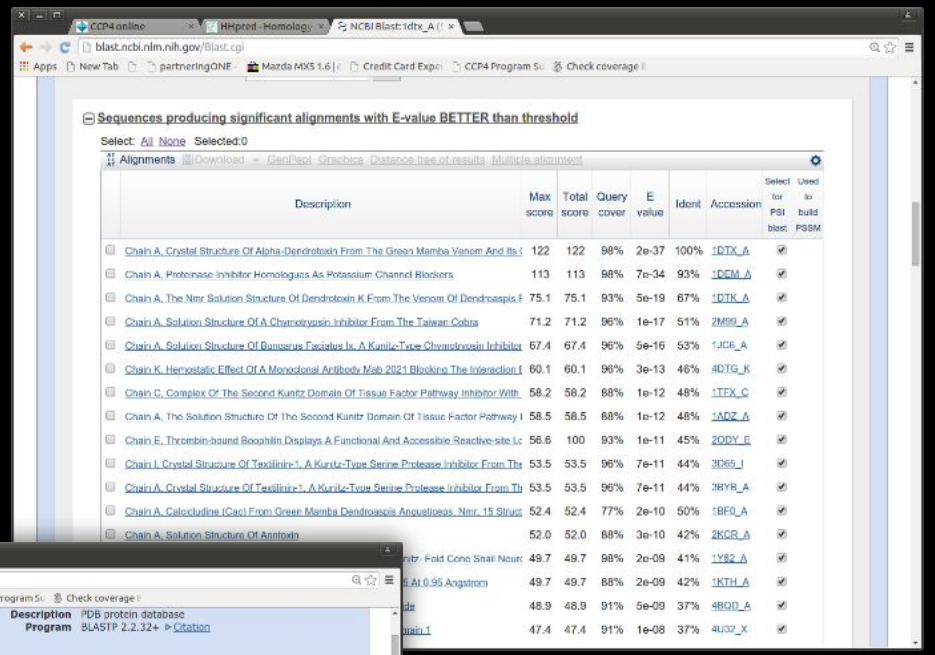
Evgeny Krissinel
Bioinformatics 2007;23:717-723

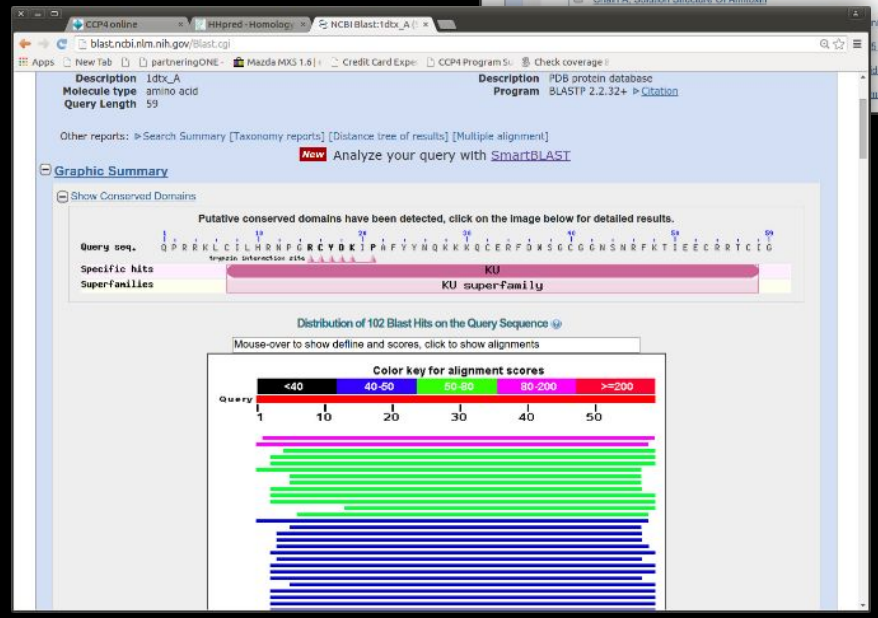
Krissinel et al. looked at structure-sequence relationship across entire PDB

CCP4

Before CASP14 (2021): *Finding a search model*

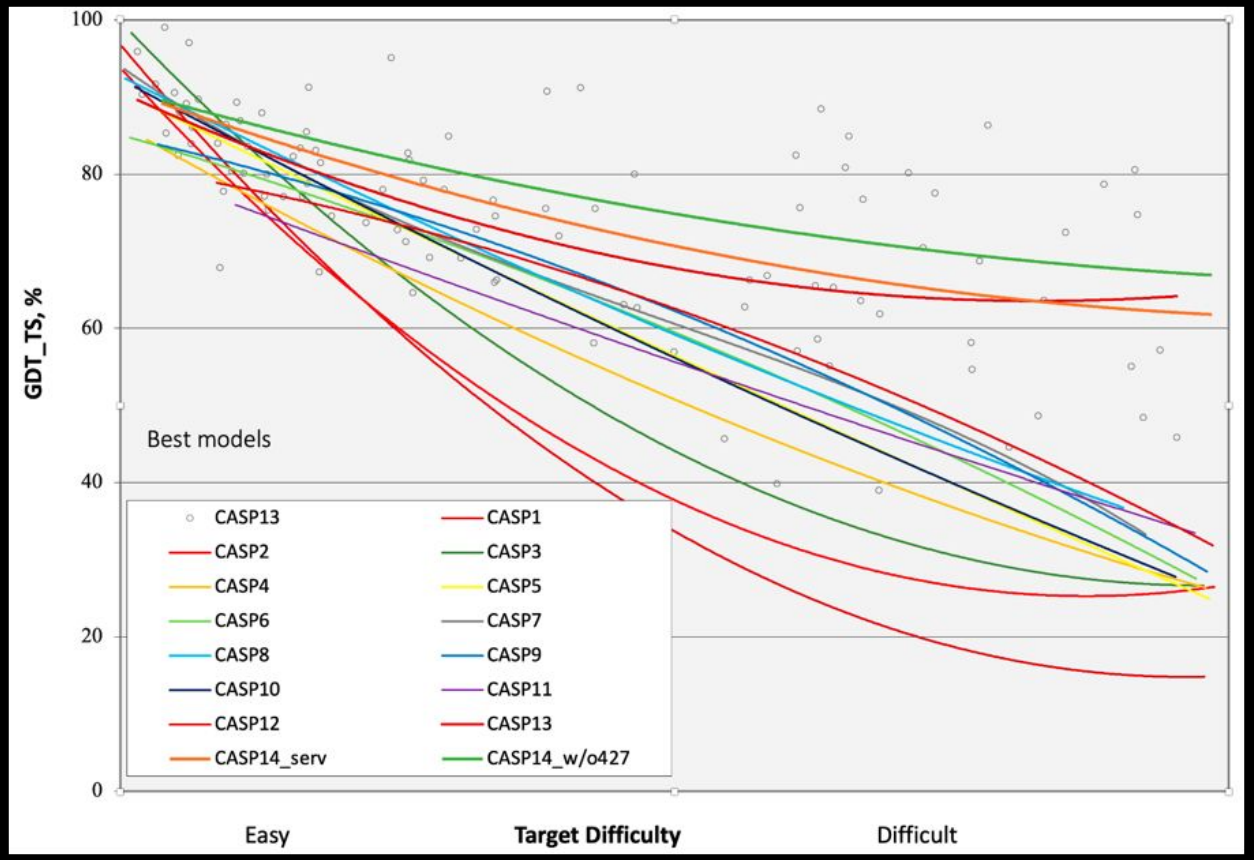
- Sequence searching e.g. PSI-Blast
  - Profile-based searching
  - Online server, fast
  - Works well at finding suitable homologues down to sequence identities of 30%





After CASP14 (2021): *Finding a search model*

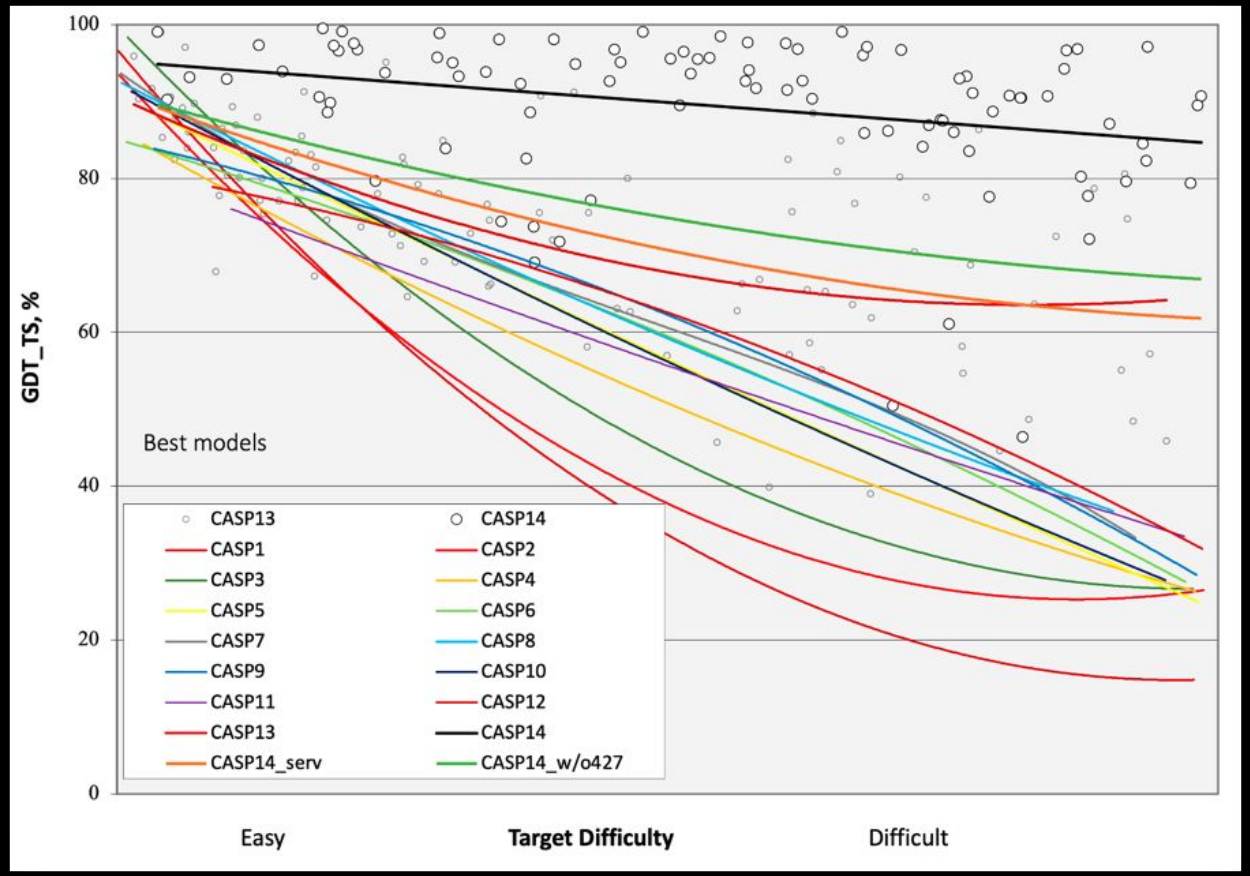
CASP: Critical Assessment of Techniques for Protein Structure Prediction



CCP4

After CASP14 (2021): *Finding a search model*

CASP: Critical Assessment of Techniques for Protein Structure Prediction

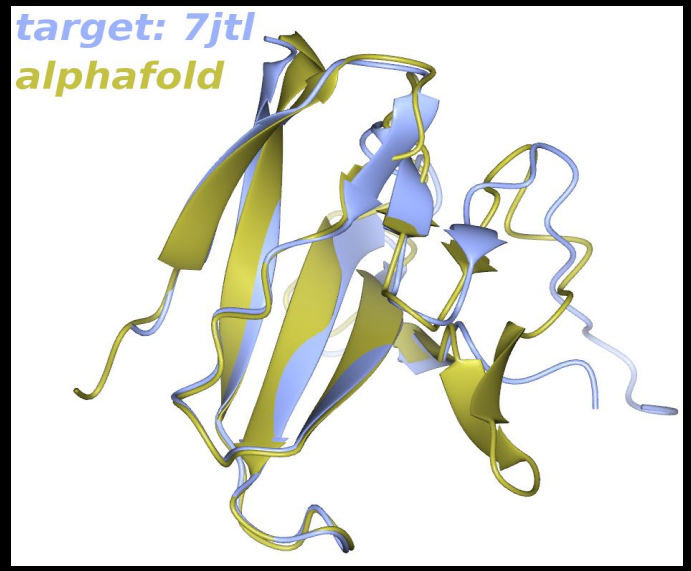


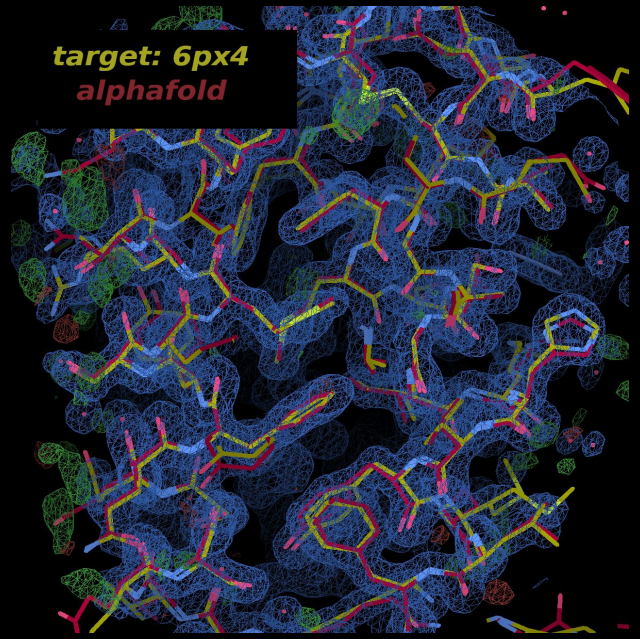
DeepMind

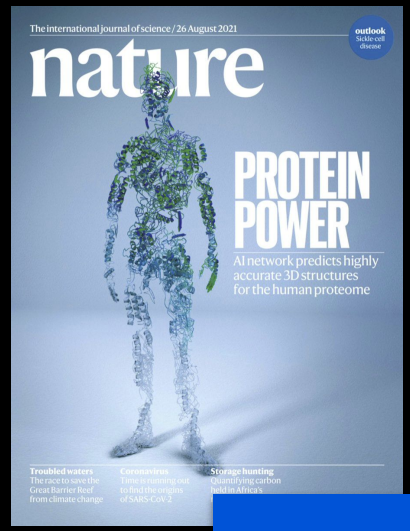
CCP4

After CASP14 (2021): *Finding a search model*

CASP: Critical Assessment of Techniques for Protein Structure Prediction







DeepMind

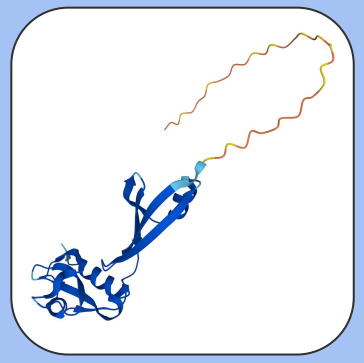
CCP4

Molecular Replacement using predicted models

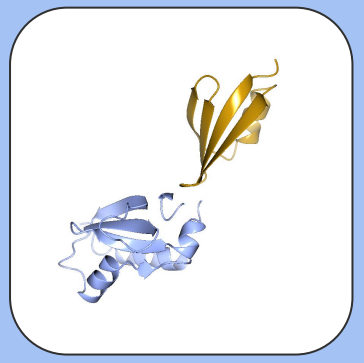
CCP4

3 Step Process

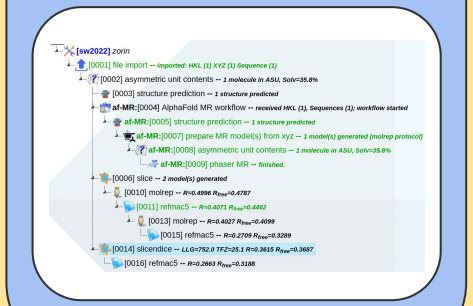
Generate or obtain predicted model

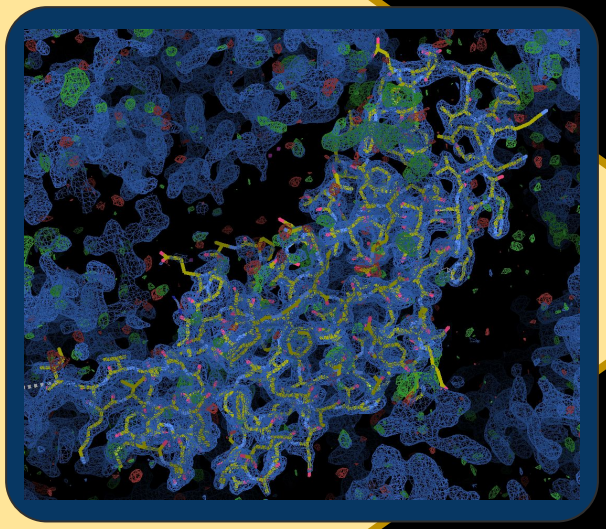


Process predicted model



Manual or automated MR

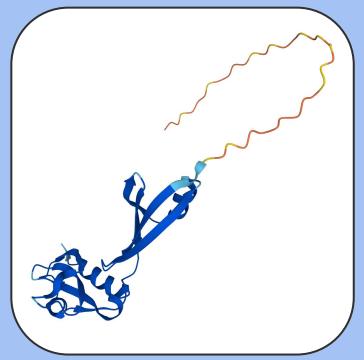




CCP4

Generate or obtain predicted model

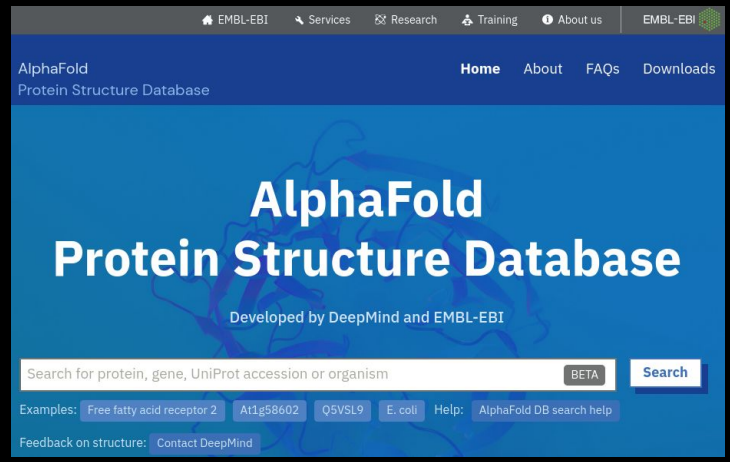
Generate or obtain predicted model

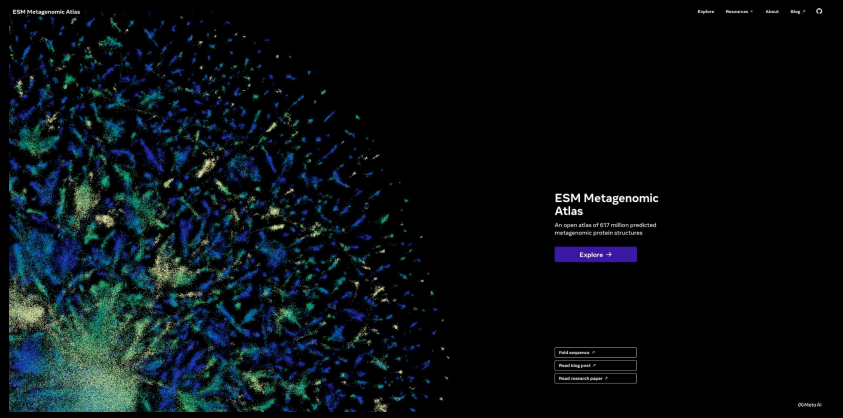


CCP4

Generate or obtain predicted model

- Four routes:
  a. Searching the ***EBI-AlphaFold*** (200million) or ***ESMAtlas*** (600 million) ***Databases***
  b. Using online ***Google Colab***, ***RoseTTAFold or ESMFold*** servers to make predictions
  c. ***CCP4Cloud*** prediction task
  d. ***User's local installation*** - supported by local CCP4Cloud





AlphaFold2.ipynb

RoseTTAFold

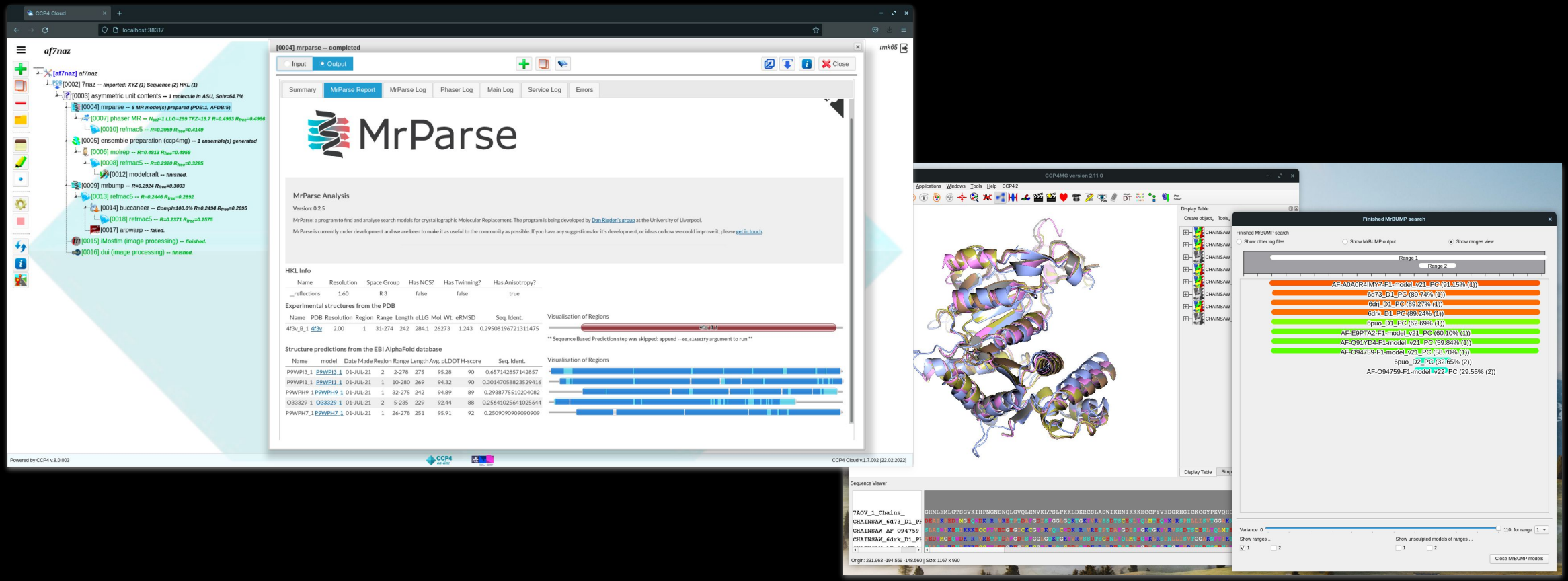
Science and Technology Facilities Council

iris

CCP4

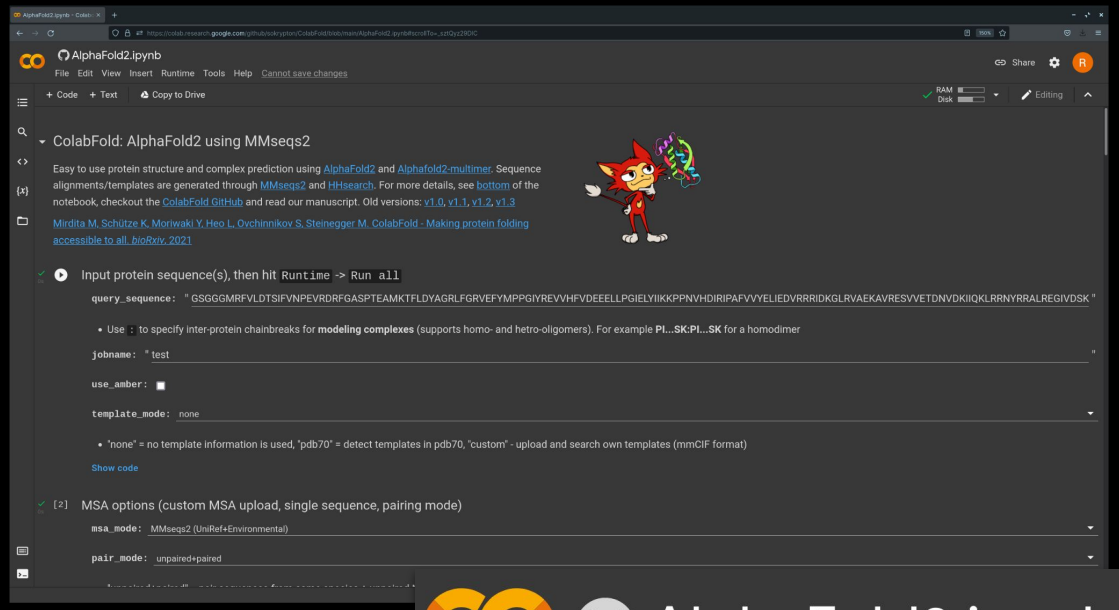
Generate or obtain predicted model

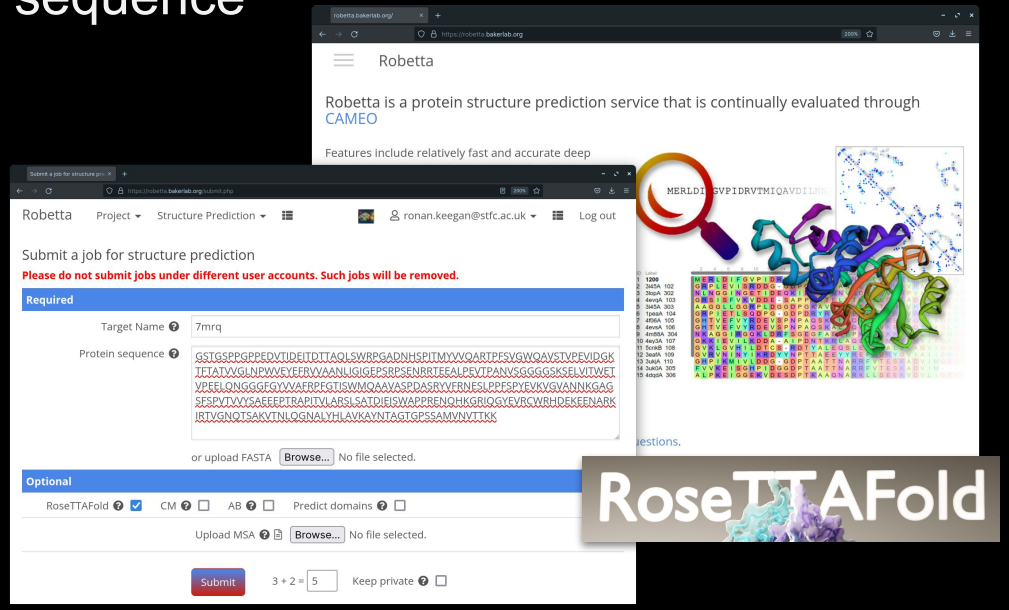
- Searching the ***EBI-Alphafold and ESMAtlas Database***:

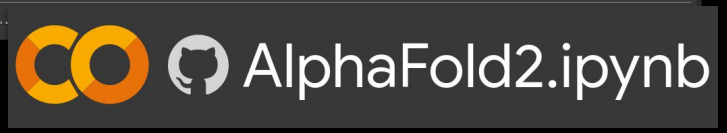


Generate or obtain predicted model

- Predicted model generation online:
  - ***Google Colab*** - provide sequence and generate AlphaFold prediction in the Google Cloud
  - Robetta server - ***RoseTTAFold*** - requires sequence



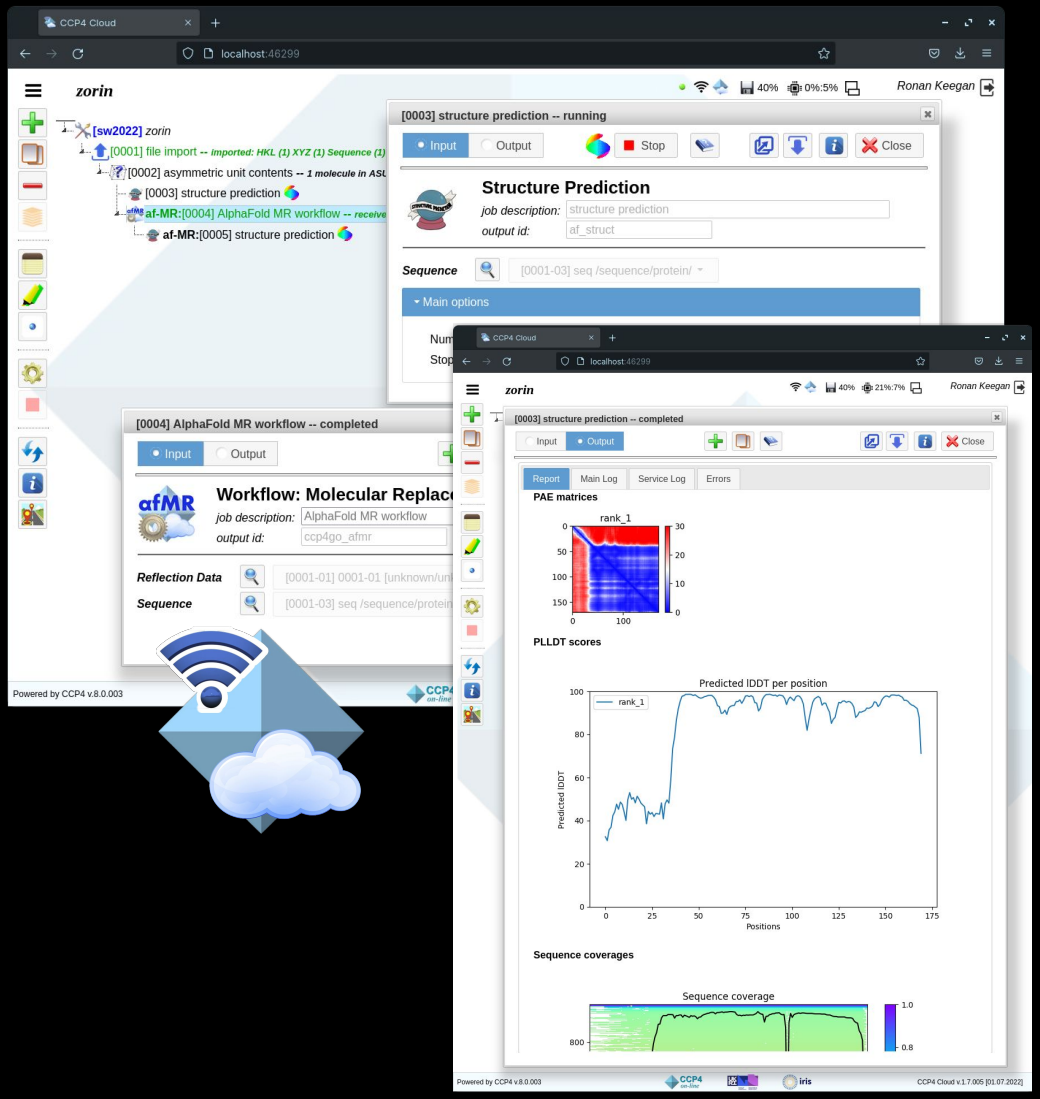




CCP4

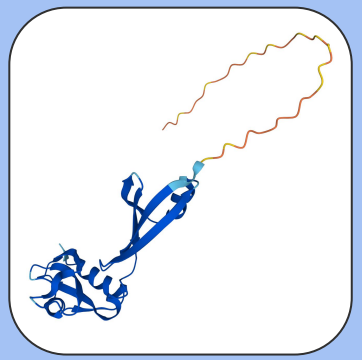
Generate or obtain predicted model

- Structure prediction through ***CCP4Cloud***:
  - If AlphaFold is installed locally, CCP4Cloud “desktop” can be linked to this
  - CCP4Cloud “remote”, hosted at Rutherford Lab has servers running AlphaFold
  - CCP4Cloud “afMR” workflow - automated structure determination including structure prediction task



Process predicted model

Generate or obtain predicted model

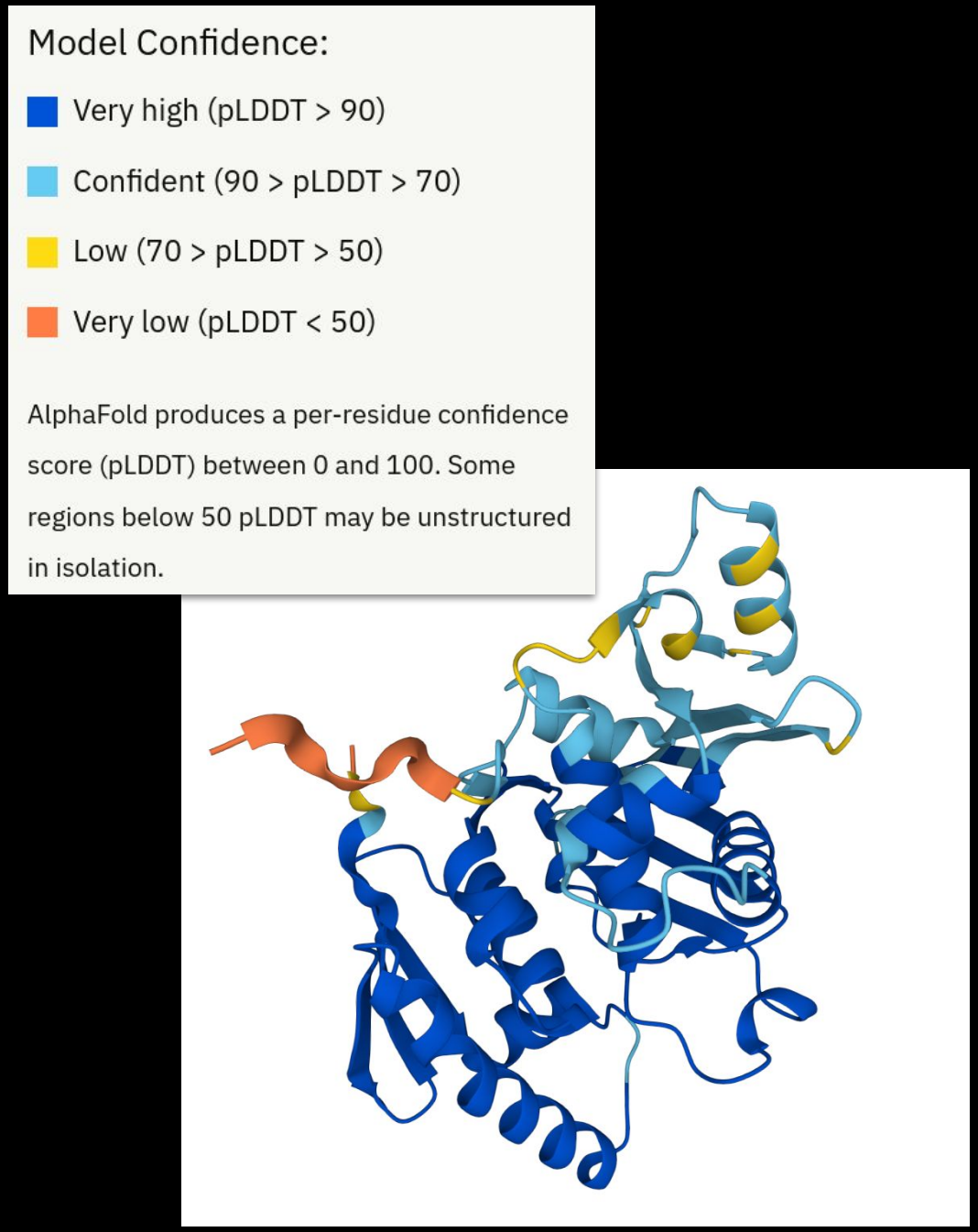


Process predicted model

CCP4

Process predicted model

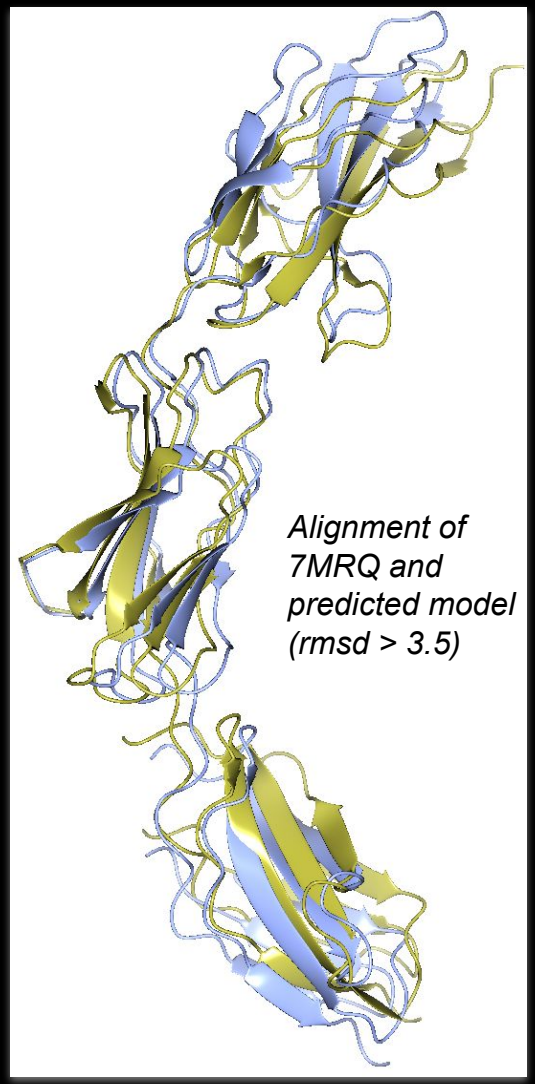
- Predicted model scoring:
  - ***pLDDT*** - confidence indicator for residues
  - Low confidence residues should be removed before use of model in MR
  - B-factor column in PDB file used to store these values - should be converted to B-factor before use in MR - important for ***Phaser***
  - ***RoseTTAFold*** uses rmsd estimate - treatment also required

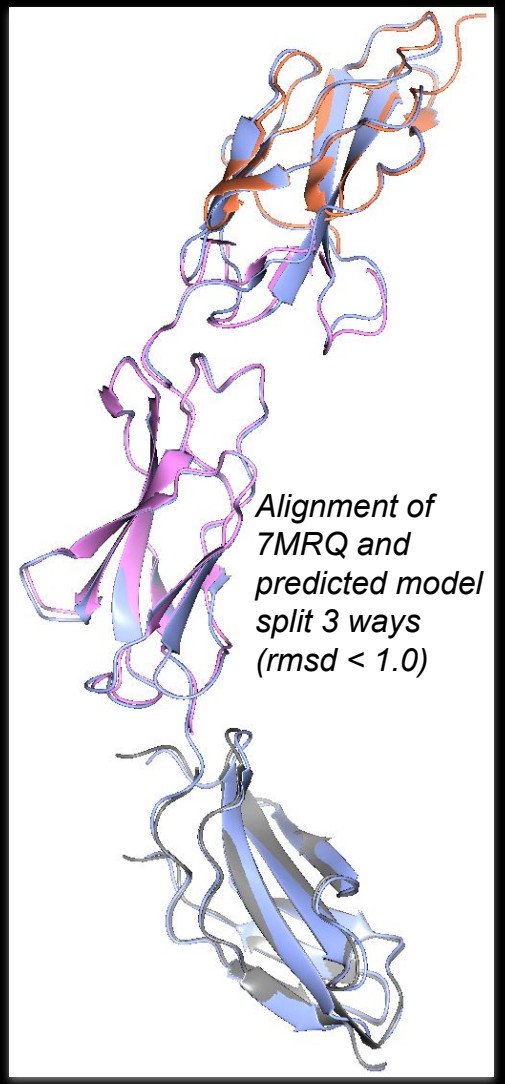


CCP4

Process predicted model

- Identifying domains or rigid units for MR:
  - Conformation of predictions can vary from crystal structures, particularly for larger molecules (>250 residues)
  - Splitting of models into domains can make placement of models through MR easier

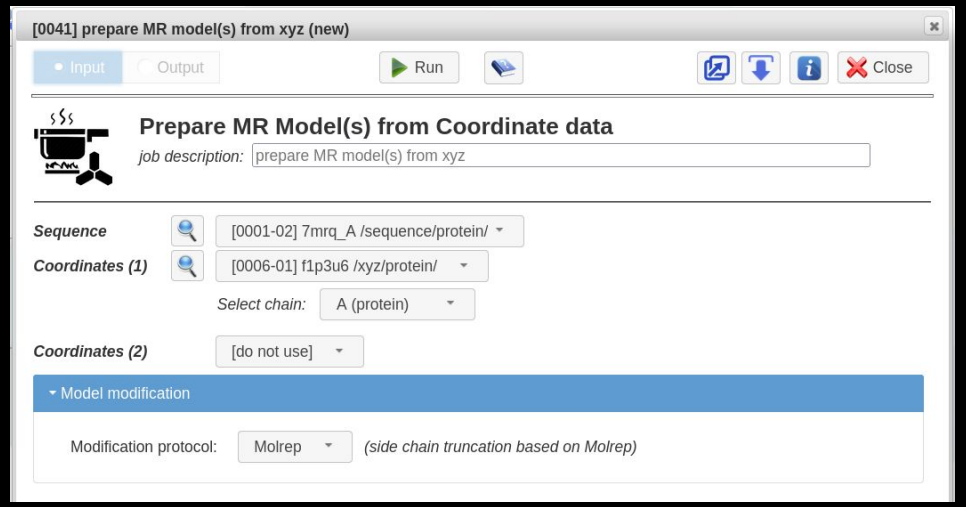


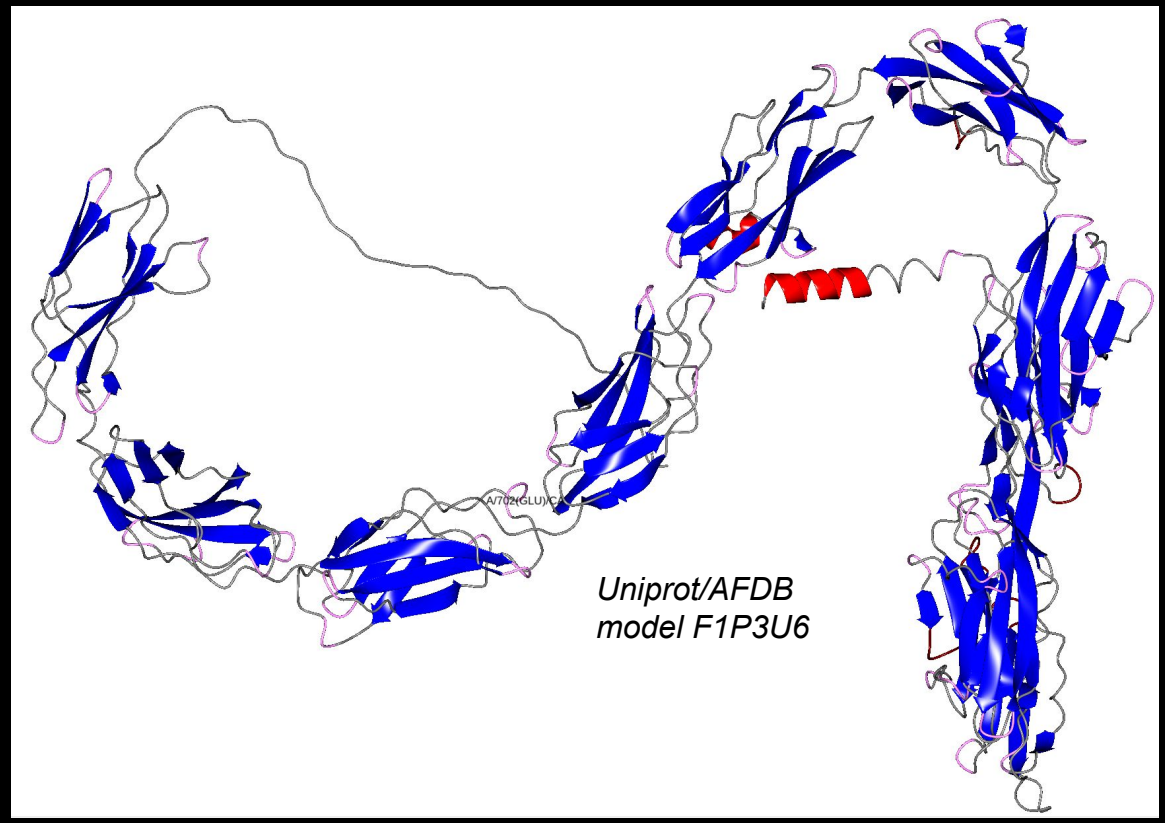


CCP4

Process predicted model

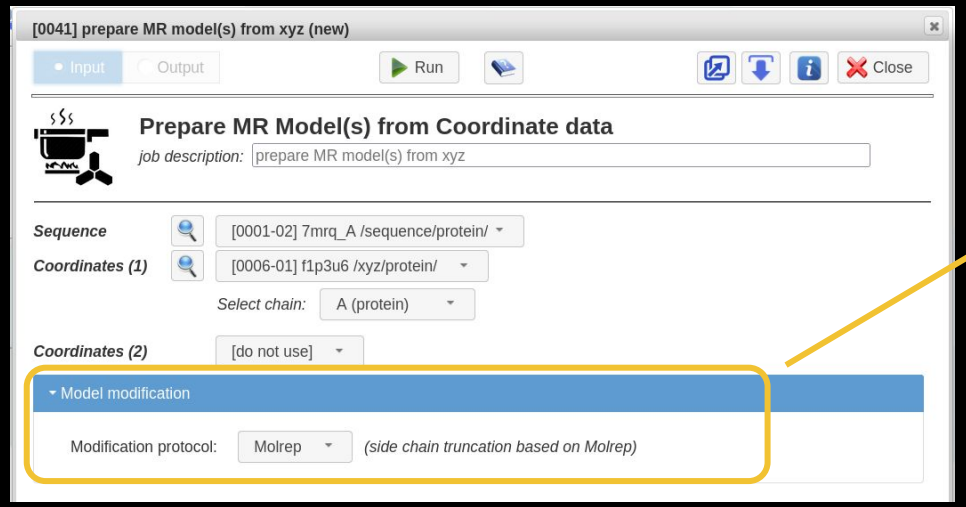
- Importing from Uniprot:
  - Truncate to region of interest
  - Can use MR model processing tasks to remove parts not needed
  - Use “Molrep” modification protocol in “Prepare MR Model(s) from Coordinate data task

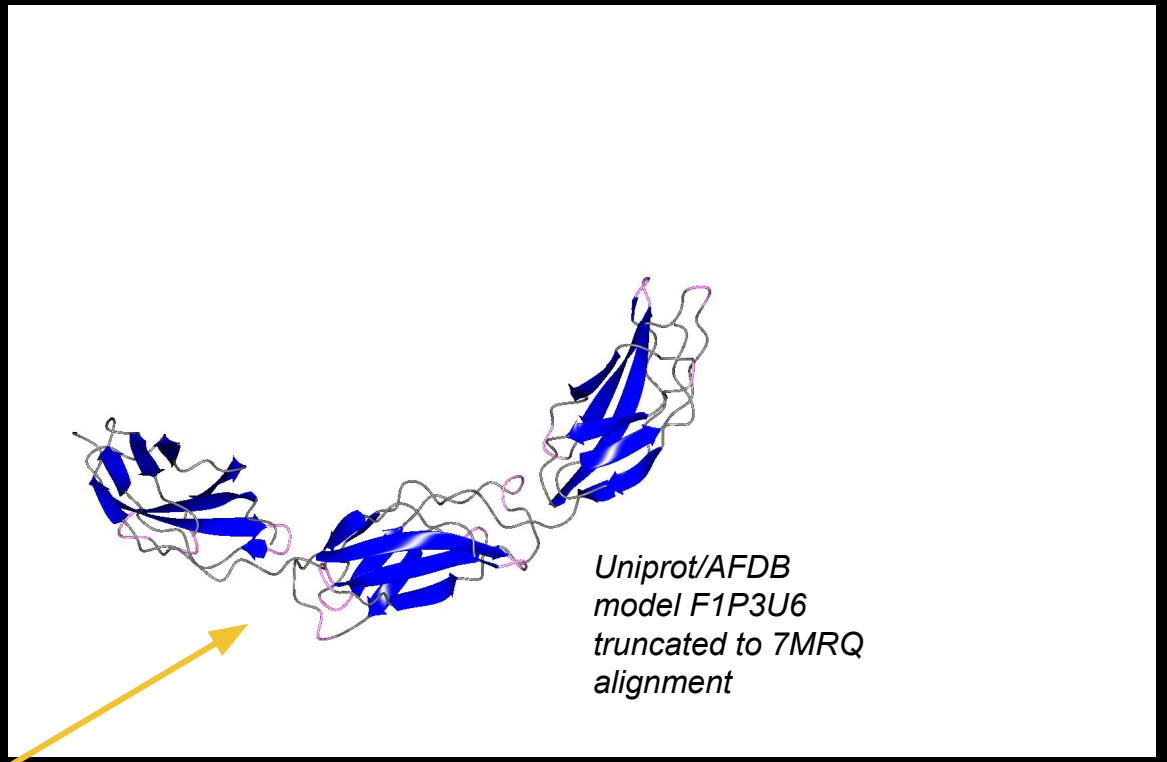




Process predicted model

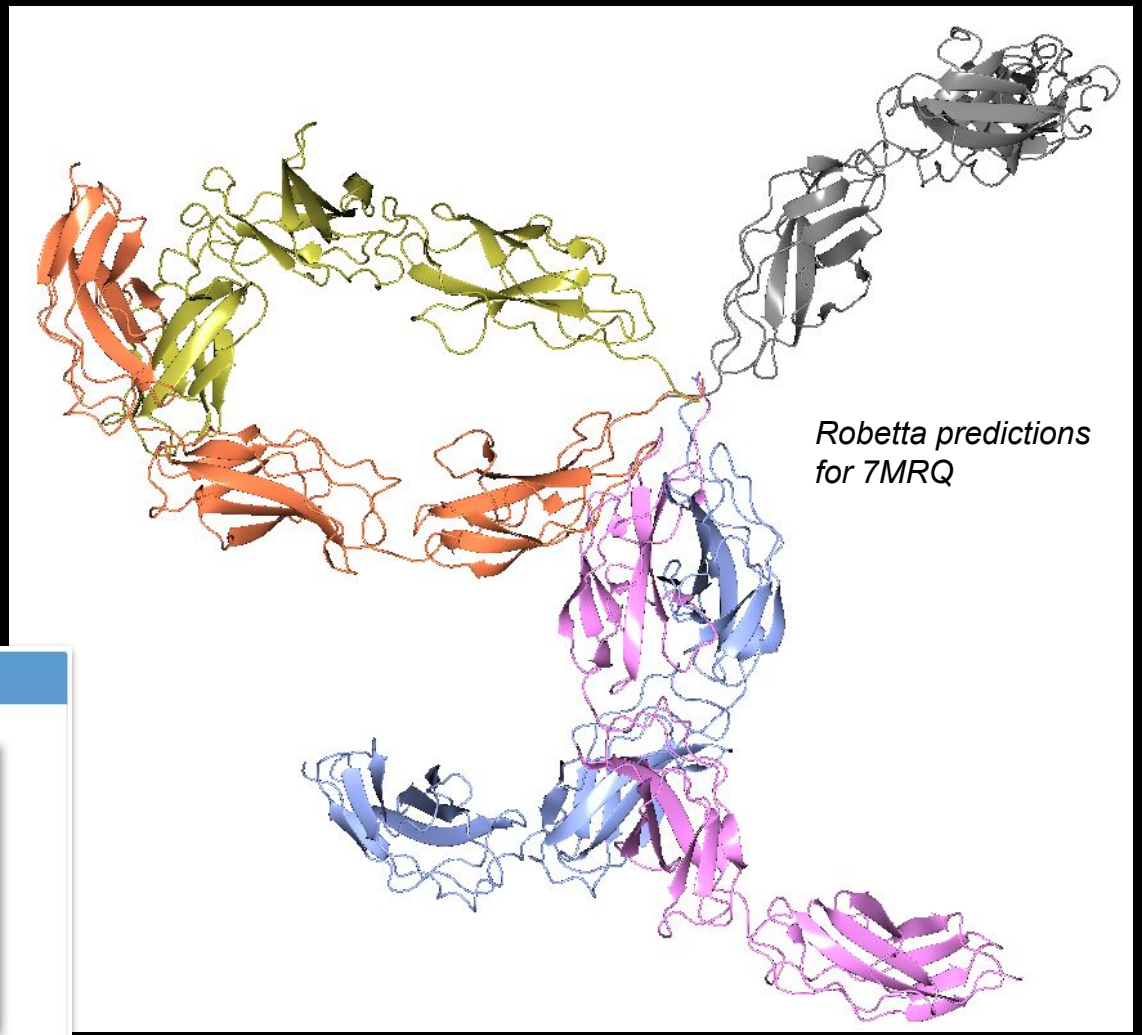
- Importing from Uniprot:
  - Truncate to region of interest
  - Can use MR model processing tasks to remove parts not needed
  - Use “Molrep” modification protocol in “Prepare MR Model(s) from Coordinate data task

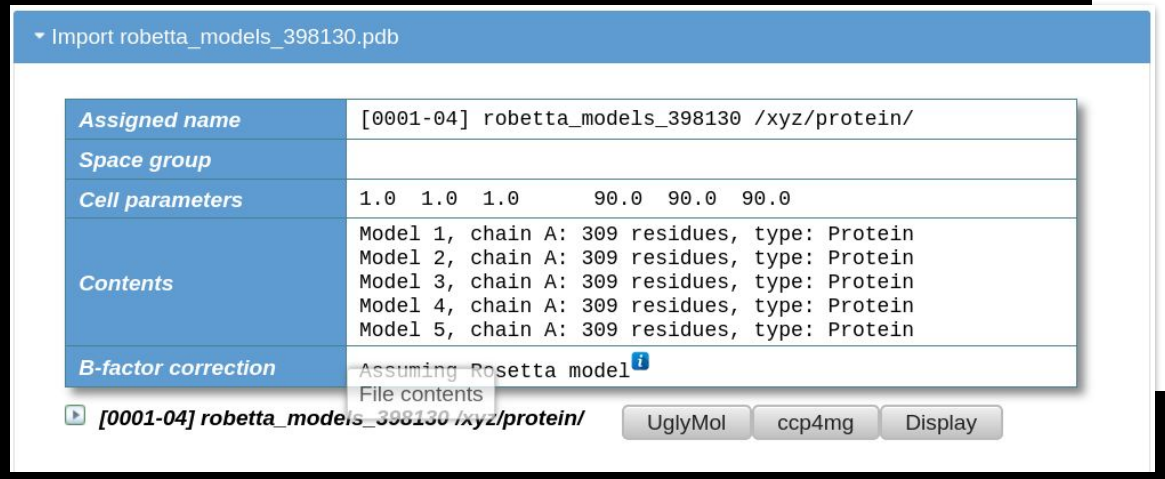




Process predicted model

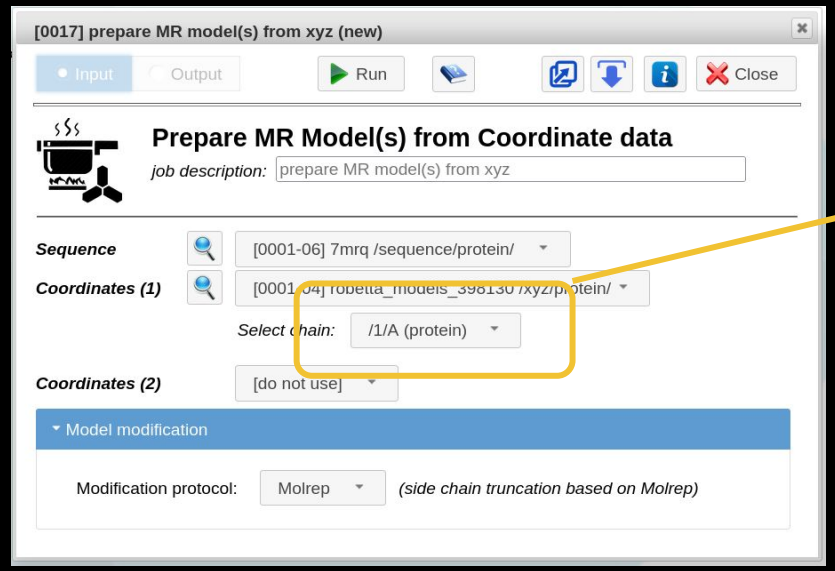
- RoseTTAFold predictions
  - RoseTTAFold provides 5 predictions by default
  - Models are contained in single PDB file
  - Select out individual models using prepare model from coordinates task

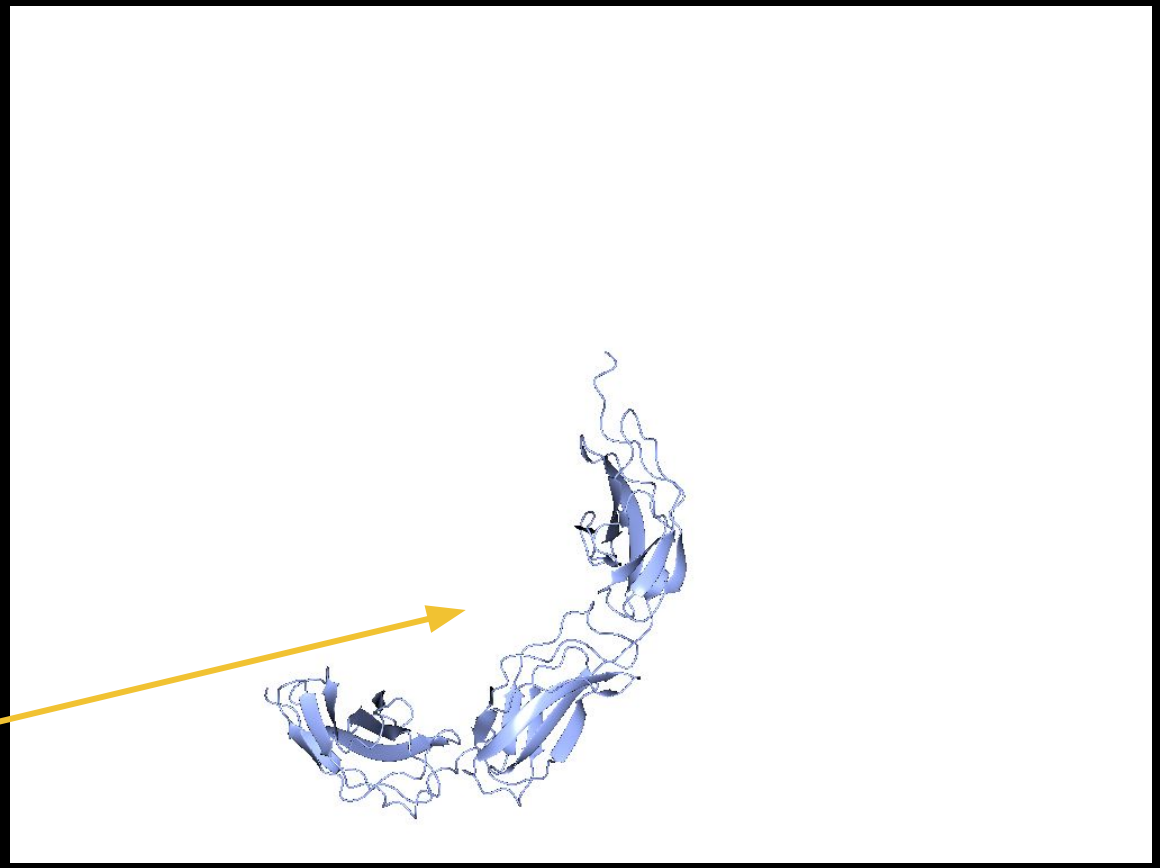




Process predicted model

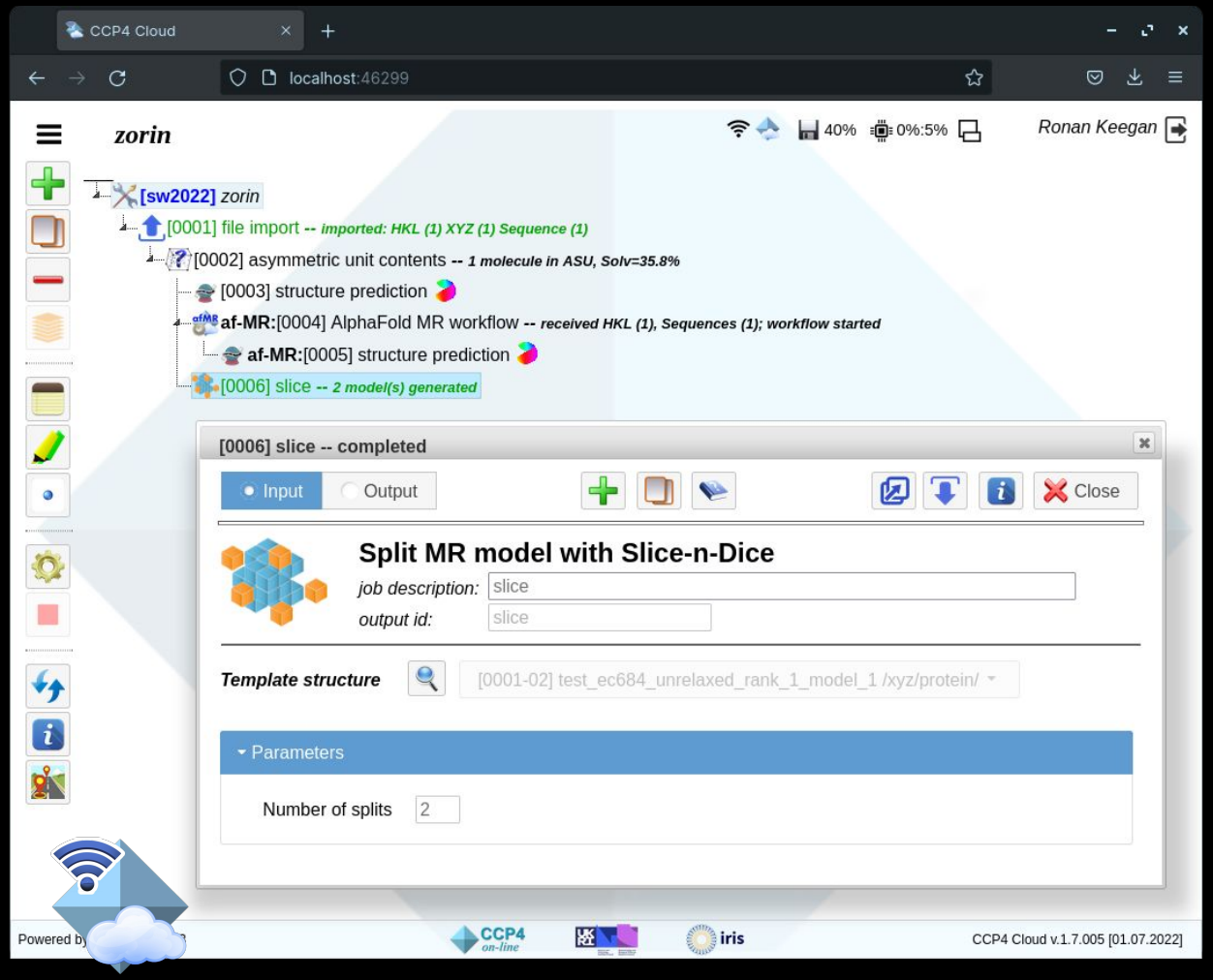
- Multi-model prediction pdb files
  - *RoseTTAFold* provides 5 predictions by default. Some AF servers do similar.
  - Models are contained in single PDB file
  - Select out individual models using prepare model from coordinates task





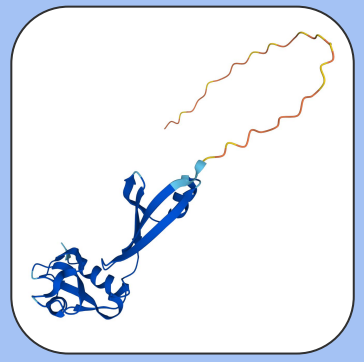
Process predicted model

- Processing in ***CCP4Cloud***:
  - *Importing models:*
    - automatically corrects confidence scores on model import (e.g. pLDDT->B-factor)
  - *Slice task:*
    - remove low confidence residues and split model into domains
    - Makes use of machine learning cluster analysis to divide input model

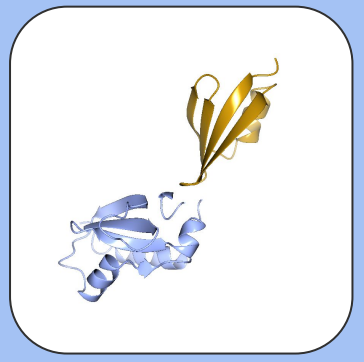


Manual or automated MR

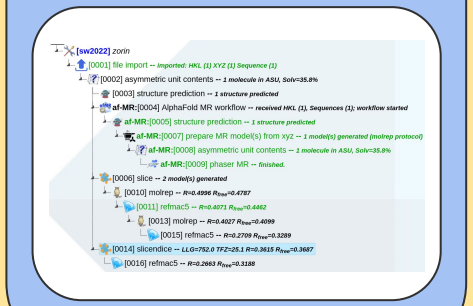
Generate or obtain predicted model

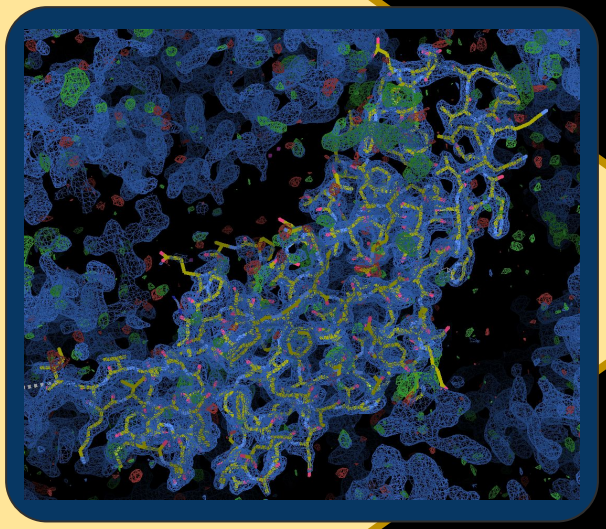


Process predicted model



Manual or automated MR

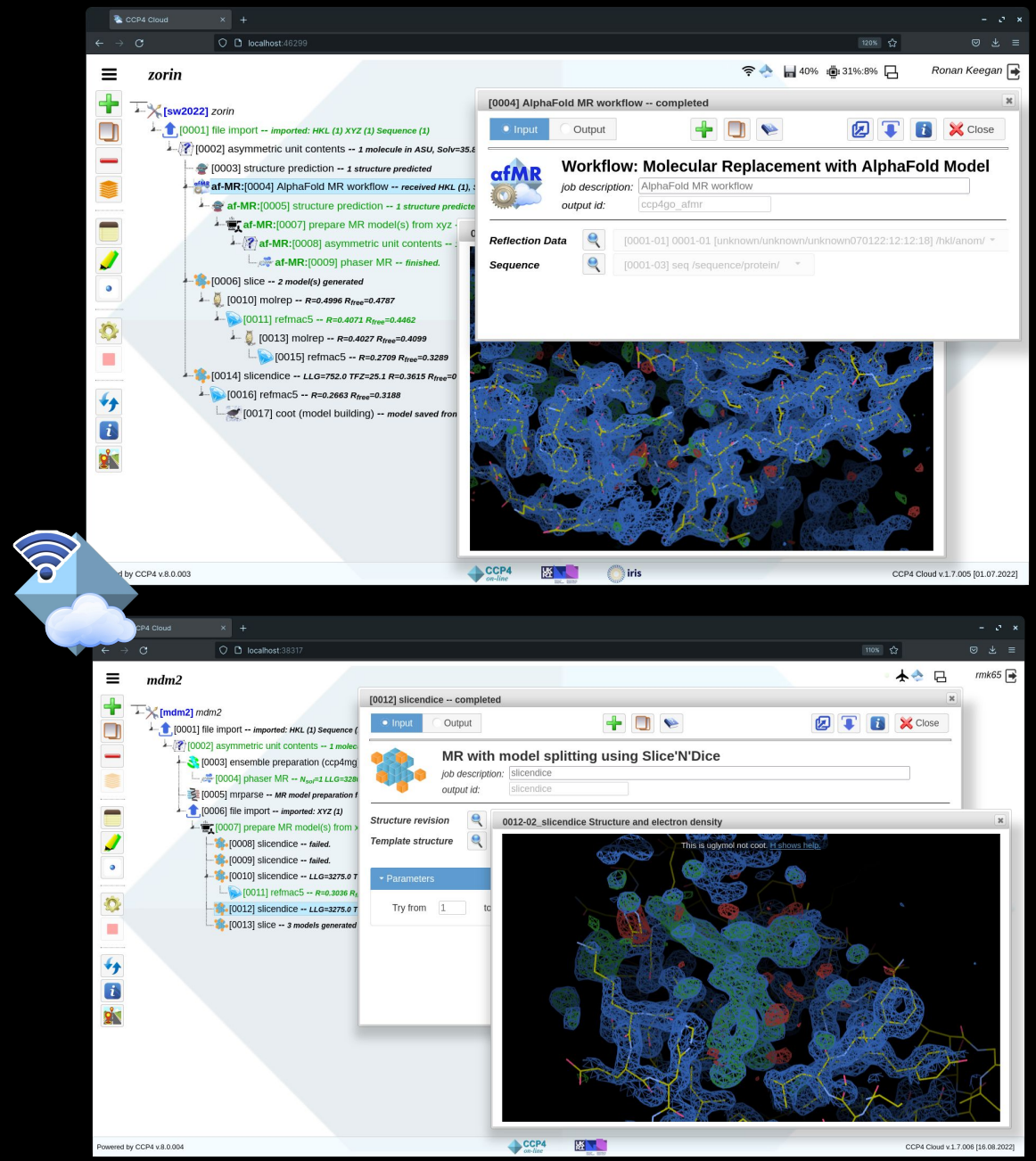




CCP4

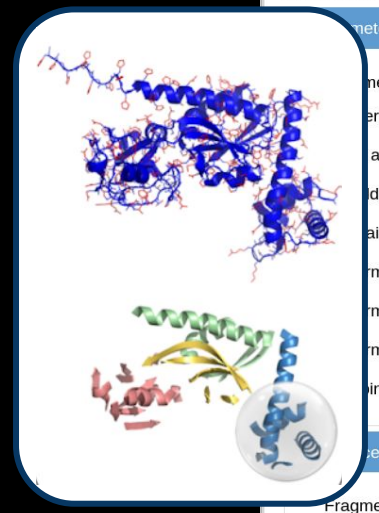
Manual or automated MR

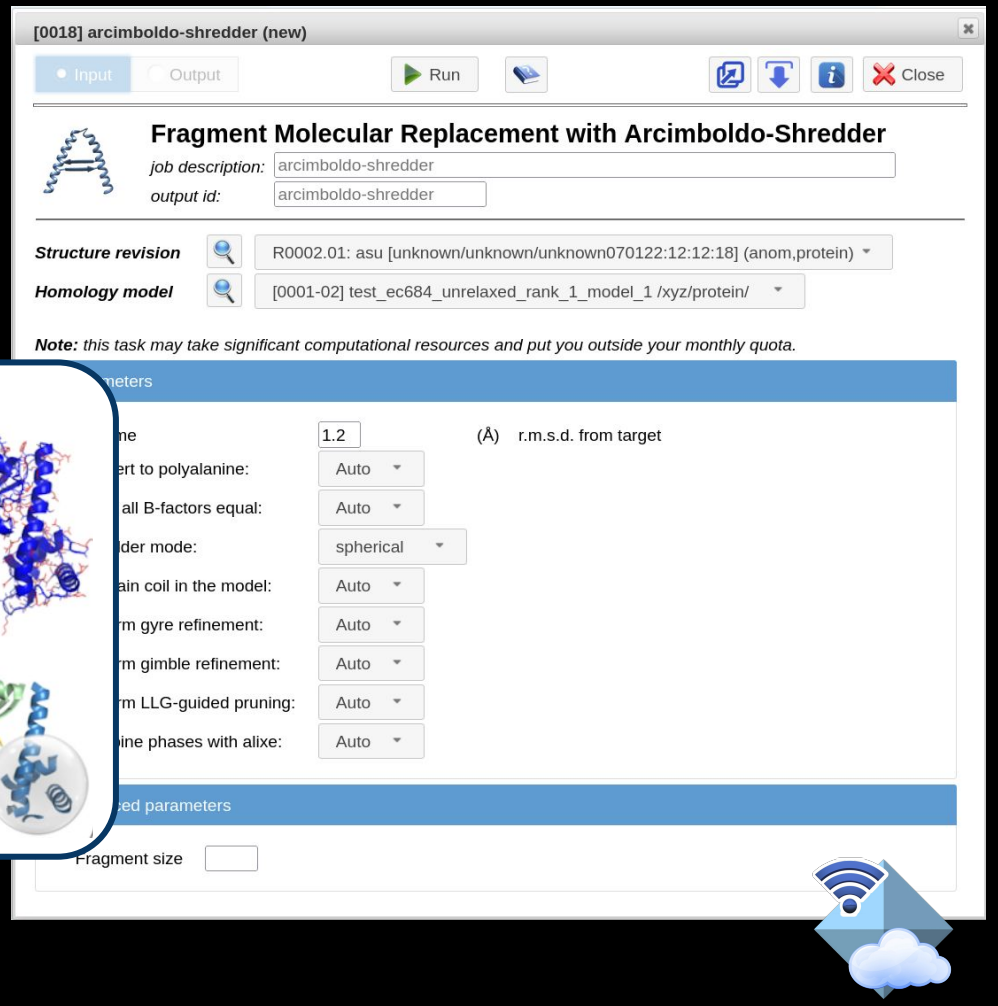
- Once processed, models are suitable for use in *Phaser* and *Molrep*
- New automated MR tools:
  - *CCP4Cloud* MR workflow
    - All steps from generation of predicted model to MR and refinement
  - *SliceNDice*
    - Automated processing and splitting of predicted models followed by MR using *Phaser* and refinement in *Refmac*
    - Trials several splits of model (default 1,2 and 3 splits)



Manual or automated MR

- ARCIMBOLDO_SHREDDER
  - Verify MR solution
  - Phase a structure - automatically pre-process an *Alphafold* or *RoseTTAFold* model by eliminating unstructured and disconnected areas
  - Eliminate bias from models

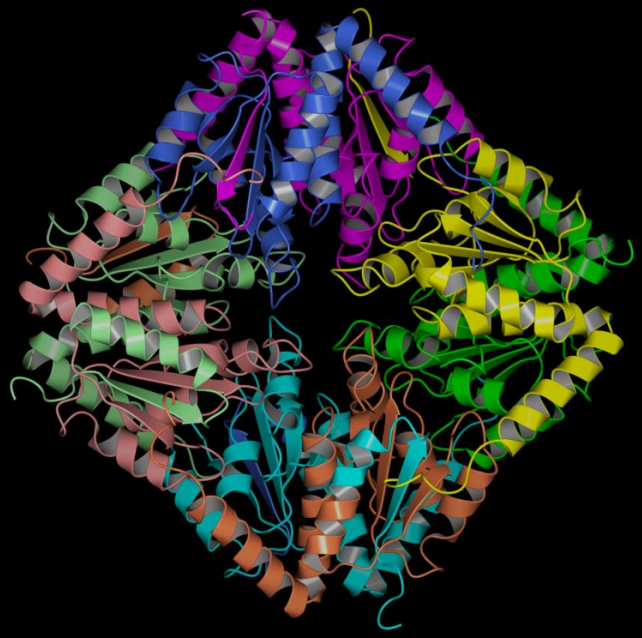


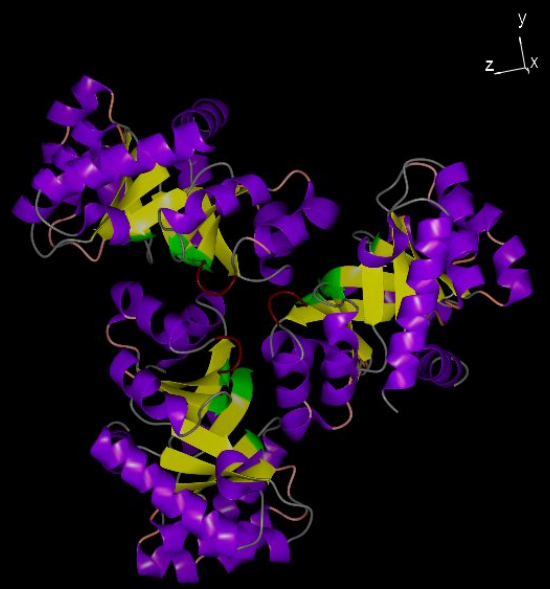


CCP4

Search Models

- ***Multimers** as search models*
  - A single chain search model can be too small if the target has crystallised in multimeric form
  - The signal for the correct position is too weak against the background noise of incorrect positions
  - Particularly a problem at lower resolutions and crystals with high symmetry





CCP4

Experimental Data and Molecular Replacement

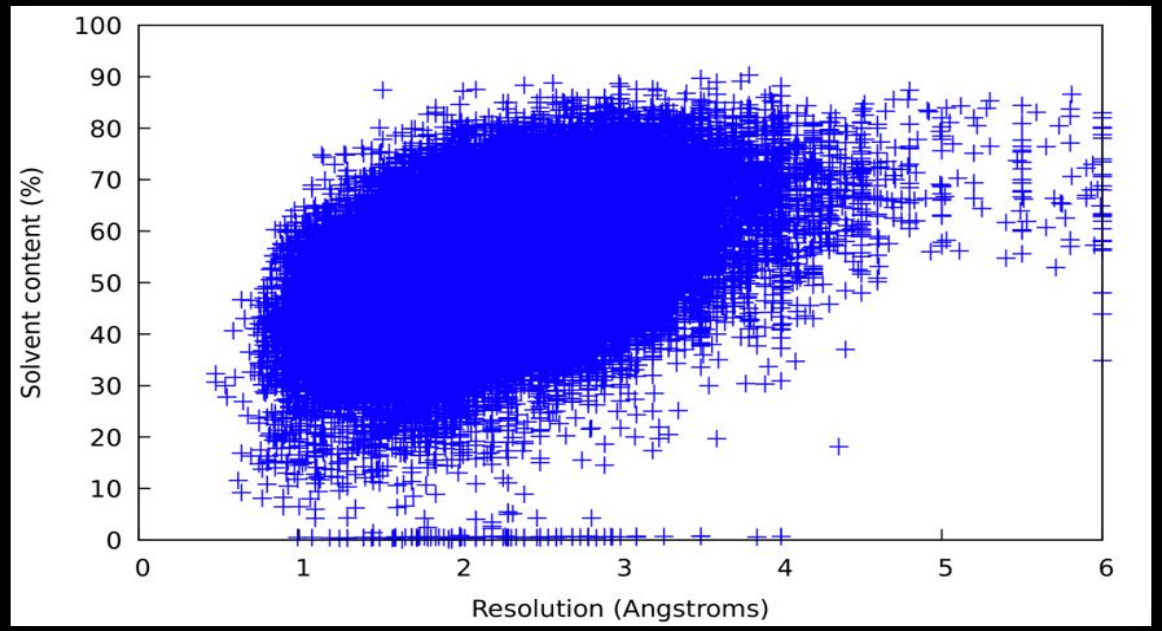
CCP4

Experimental data

- Data issues can have an impact on how well MR will work

Experimental data

- Data issues can have an impact on how well MR will work
- Things to think about:
  - How many copies in the asymmetric unit?
    - Estimate through Matthews Coefficient

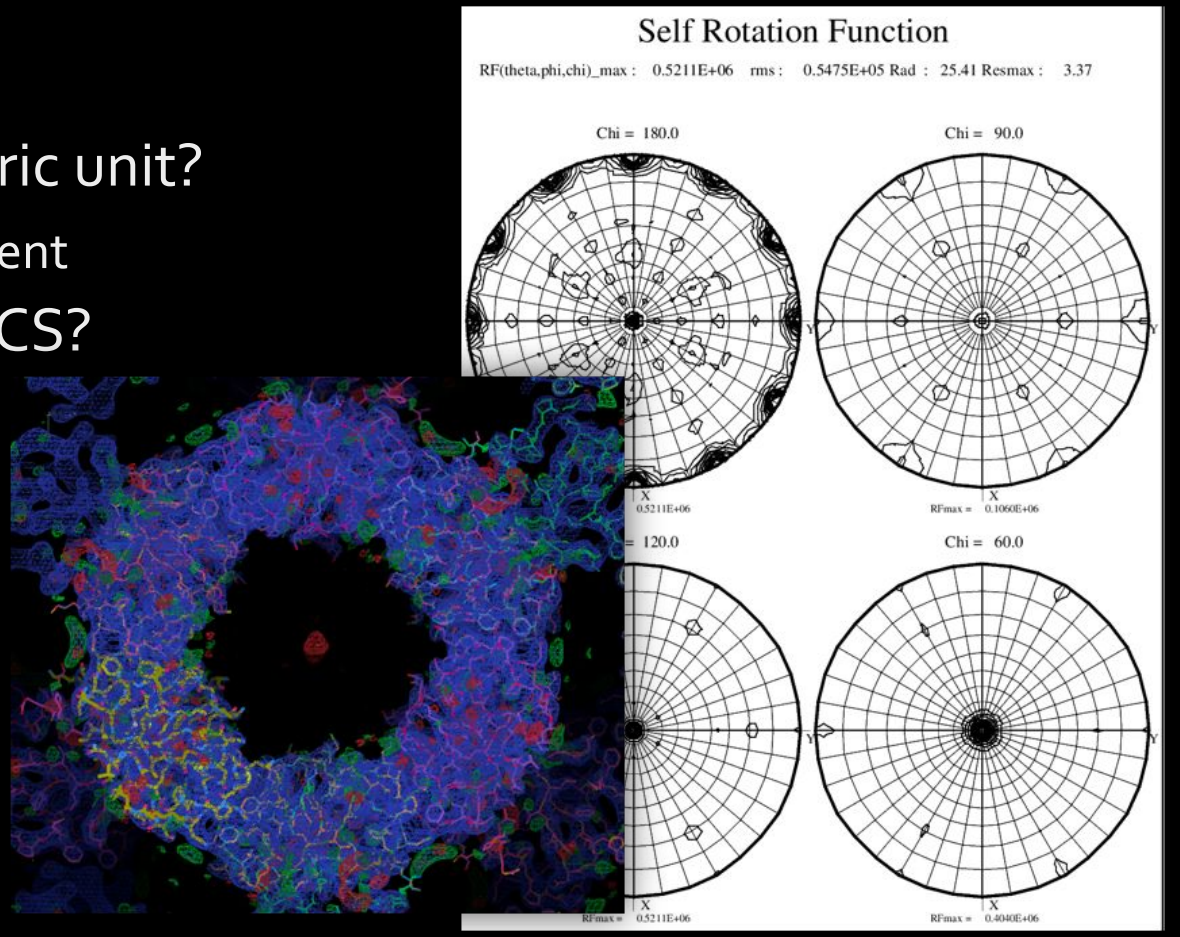


Distribution of solvent content across 156,000 X-ray datasets in the PDB

CCP4

Experimental data

- Data issues can have an impact on how well MR will work
- Things to think about:
  - How many copies in the asymmetric unit?
    - Estimate through Matthews Coefficient
  - Self-rotation function – signs of NCS?



Self rotation function and map/model for 5i72 – 6 fold symmetry

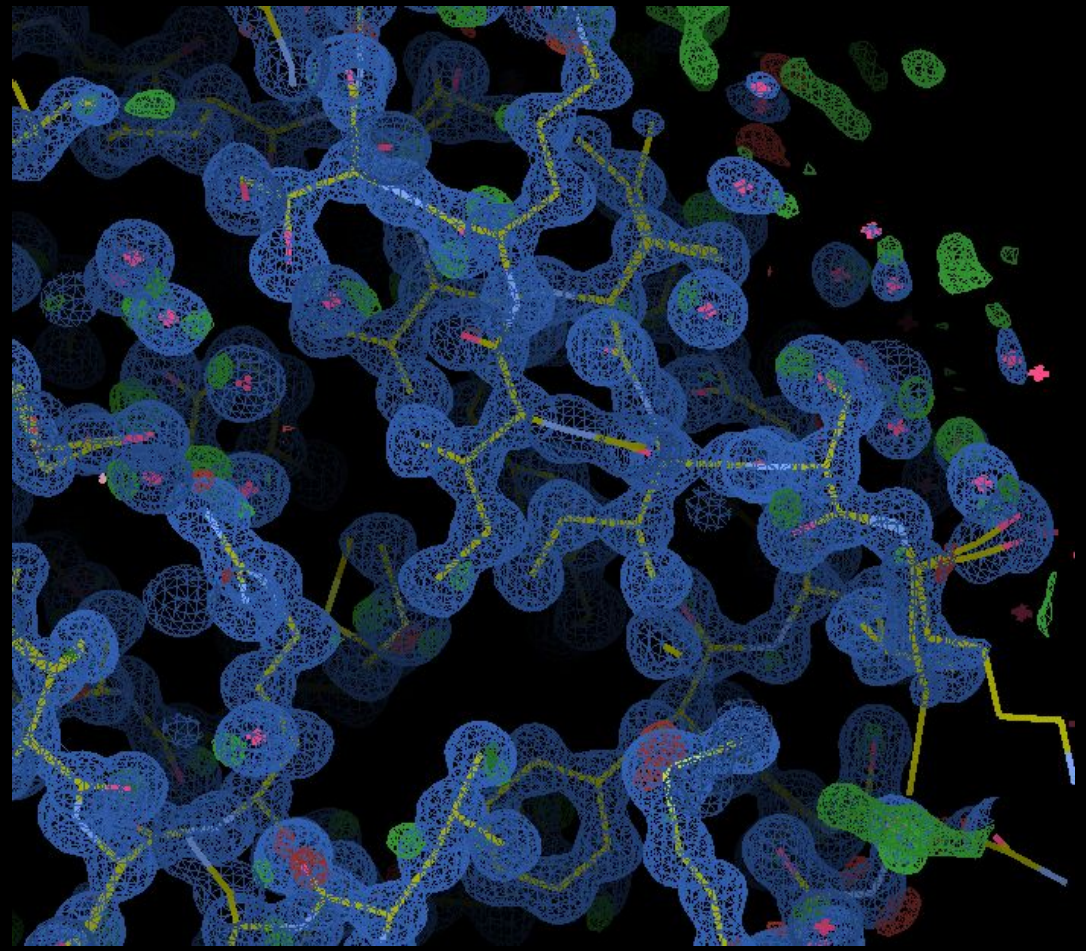
CCP4

Experimental data

- Data issues can have an impact on how well MR will work
- Things to think about:
  - How many copies in the asymmetric unit?
    - Estimate through Matthews Coefficient
  - Self-rotation function – signs of NCS?
  - Resolution of the data?

Experimental data: resolution and MR

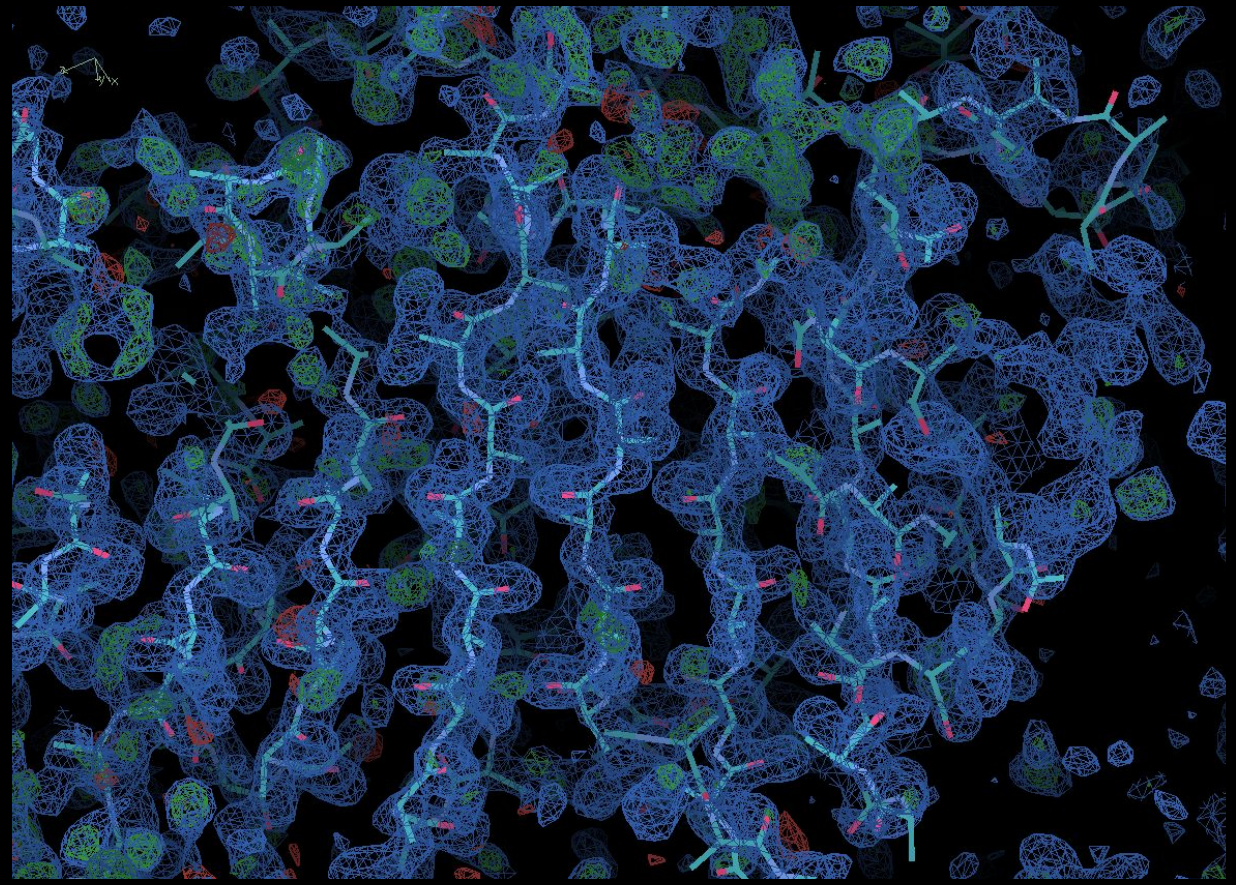
- Better than 1.0 Å – search model can be a small fragment or even a single atom in *Phaser*
  (*McCoy et al. 2017, PNAS*)



CCP4

Experimental data: resolution and MR

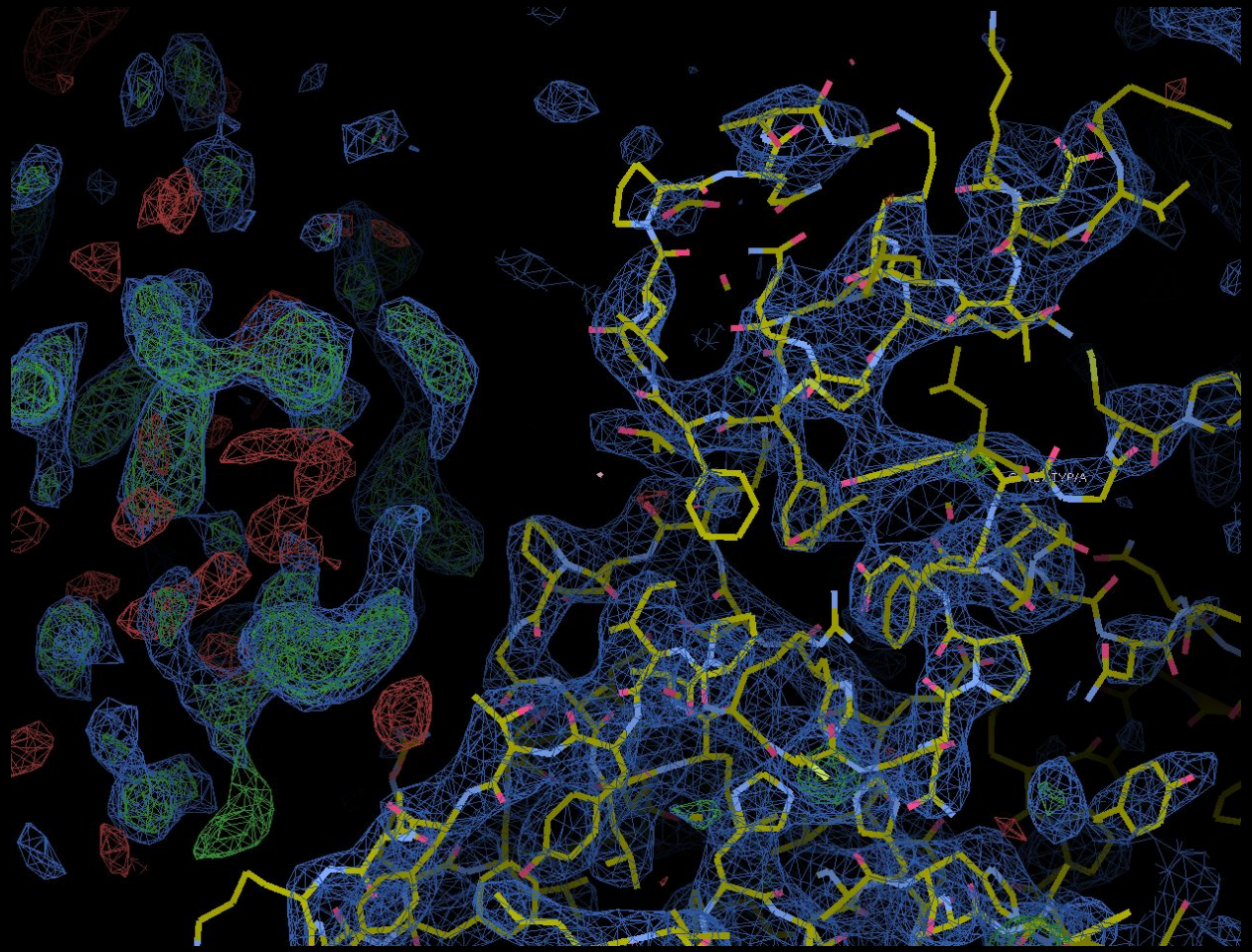
- At resolutions above ~ 2.5 Å
  - *Phaser* can place small fragments of total scattering
  - Applications like *SHELXE* and *Acorn* can improve through density modification (DM) and model building to a correct solution



CCP4

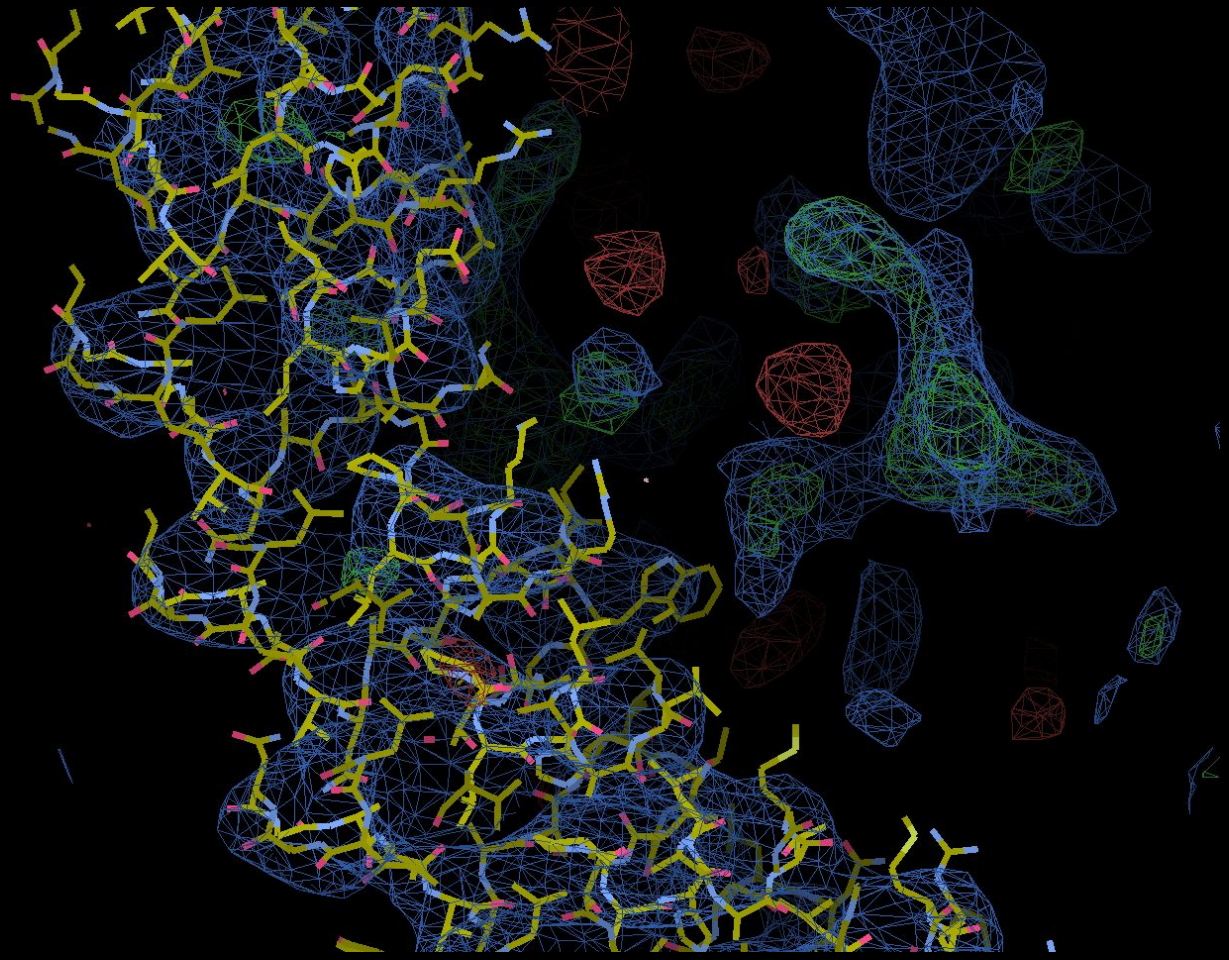
Experimental data: resolution and MR

- 2.5Å or worse
  - DM techniques and fragment placing is less effective
  - Larger search models are required



Experimental data: resolution and MR

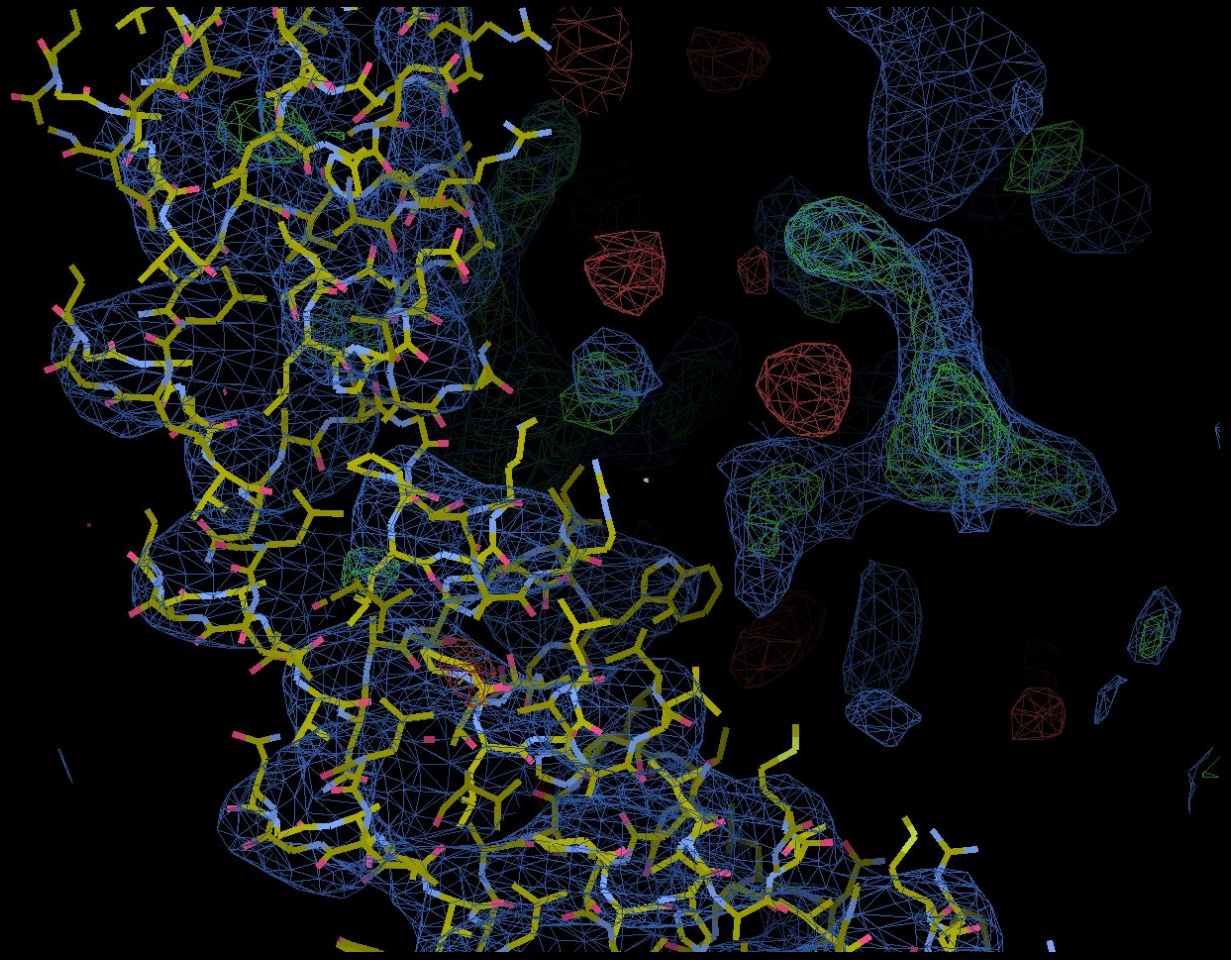
- Below 4Å automated model building becomes difficult



CCP4

Experimental data: resolution and MR

- Below 4Å automated model building becomes difficult
- However, given suitable search models, MR can be used to build up a model structure



CCP4

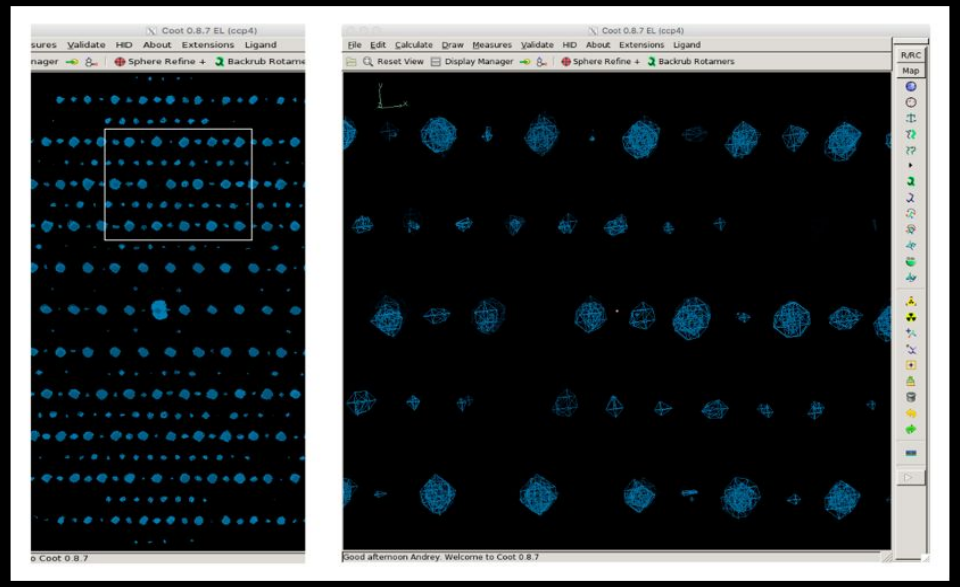
Experimental data

- Data issues can have an impact on how well MR will work
- Things to think about:
  - How many copies in the asymmetric unit?
    - Estimate through Matthews Coefficient
  - Self-rotation function – signs of NCS?
  - Resolution of the data?
- Potential Problems:

CCP4

Experimental data

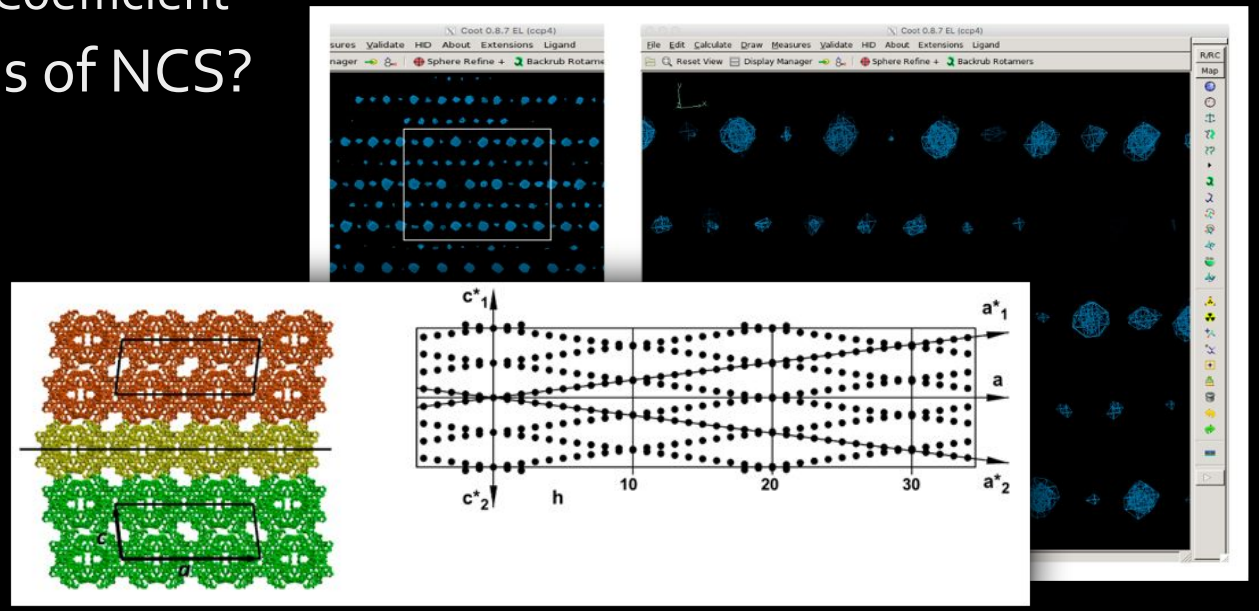
- Data issues can have an impact on how well MR will work
- Things to think about:
  - How many copies in the asymmetric unit?
    - Estimate through Matthews Coefficient
  - Self-rotation function – signs of NCS?
  - Resolution of the data?
- Potential Problems:
  - Pseudo translation



CCP4

Experimental data

- Data issues can have an impact on how well MR will work
- Things to think about:
  - How many copies in the asymmetric unit?
    - Estimate through Matthews Coefficient
  - Self-rotation function – signs of NCS?
  - Resolution of the data?
- Potential Problems:
  - Pseudo translation
  - Twinned data

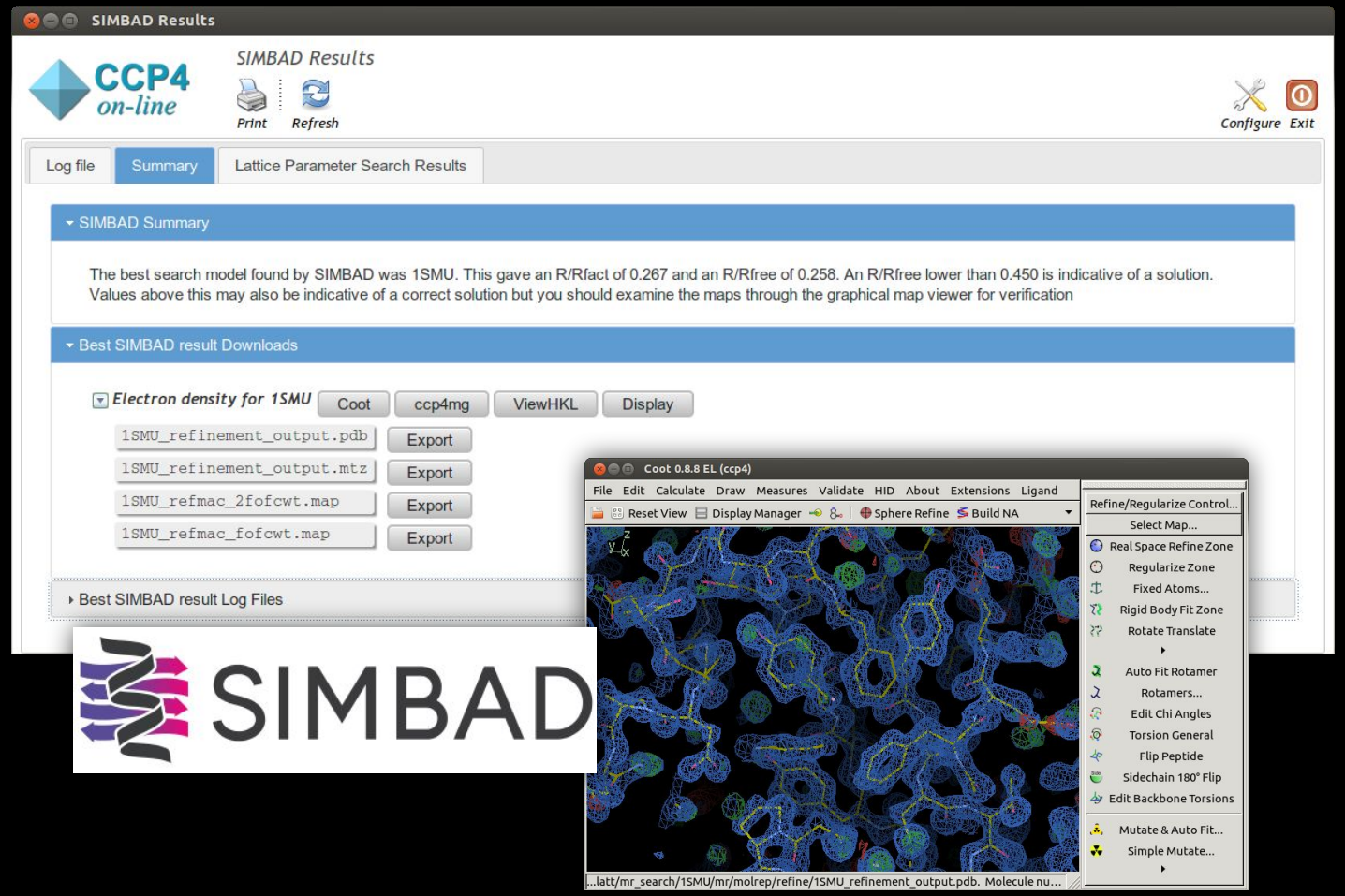


CCP4

Molecular Replacement in CCP4

CCP4

- **SIMBAD**: *Sequence-less MR*
  1. nearest cell (lattice)
  2. contaminants
  3. entire non-redundant domain database (90000 models)



CCP4

Molecular Replacement in CCP4

- CCP4 has several programs for doing Molecular Replacement
  - Amore
    - Manual steps but very fast
  - 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

Molecular Replacement: Phaser

- Important points on using Phaser
  - Phaser accounts for errors in:

Molecular Replacement: Phaser

- Important points on using Phaser
  - Phaser accounts for errors in:
    1. Model
       - Provide accurate details of AU composition
       - Use RMS value rather than sequence identity and try different values if first attempt doesn't work

Molecular Replacement: Phaser

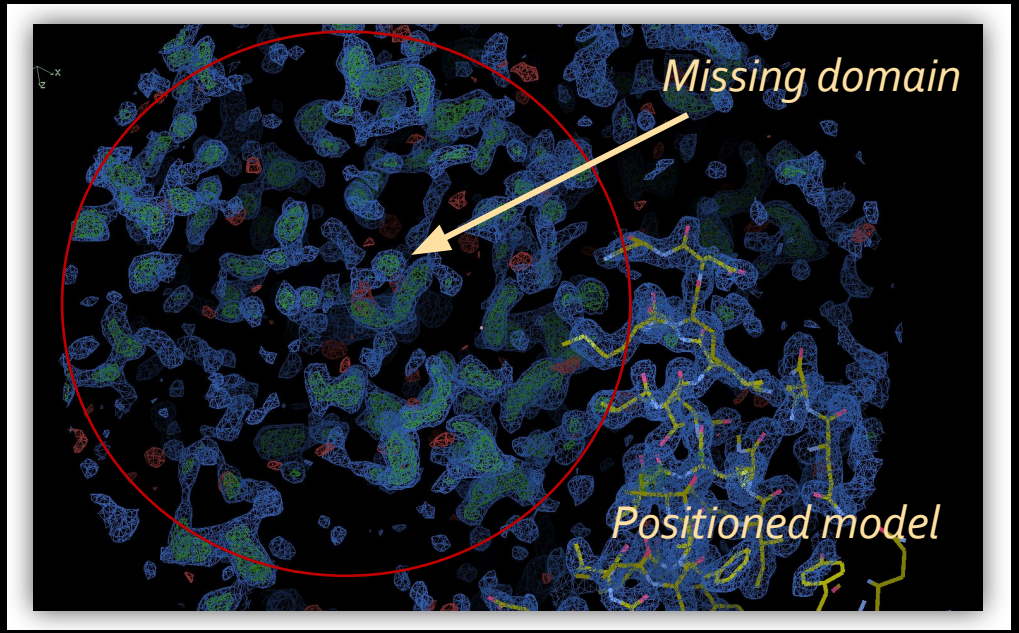
- Important points on using Phaser
  - Phaser accounts for errors in:
    1. Model
       - Provide accurate details of AU composition
       - Use RMS value rather than sequence identity and try different values if first attempt doesn't work
    2. Data
       - Provide intensities – internally works out amplitudes accounting for experimental errors

Molecular Replacement: Phaser

- Important points on using Phaser
  - Phaser accounts for errors in:
    1. Model
       - Provide accurate details of AU composition
       - Use RMS value rather than sequence identity and try different values if first attempt doesn't work
    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

CCP4

Phased Translation search



- Used when searching for several copies or dealing with a complex
- Available through MOLREP (3 protocols) and Phaser
- Often more successful than standard MR search approach particularly when looking for small domains

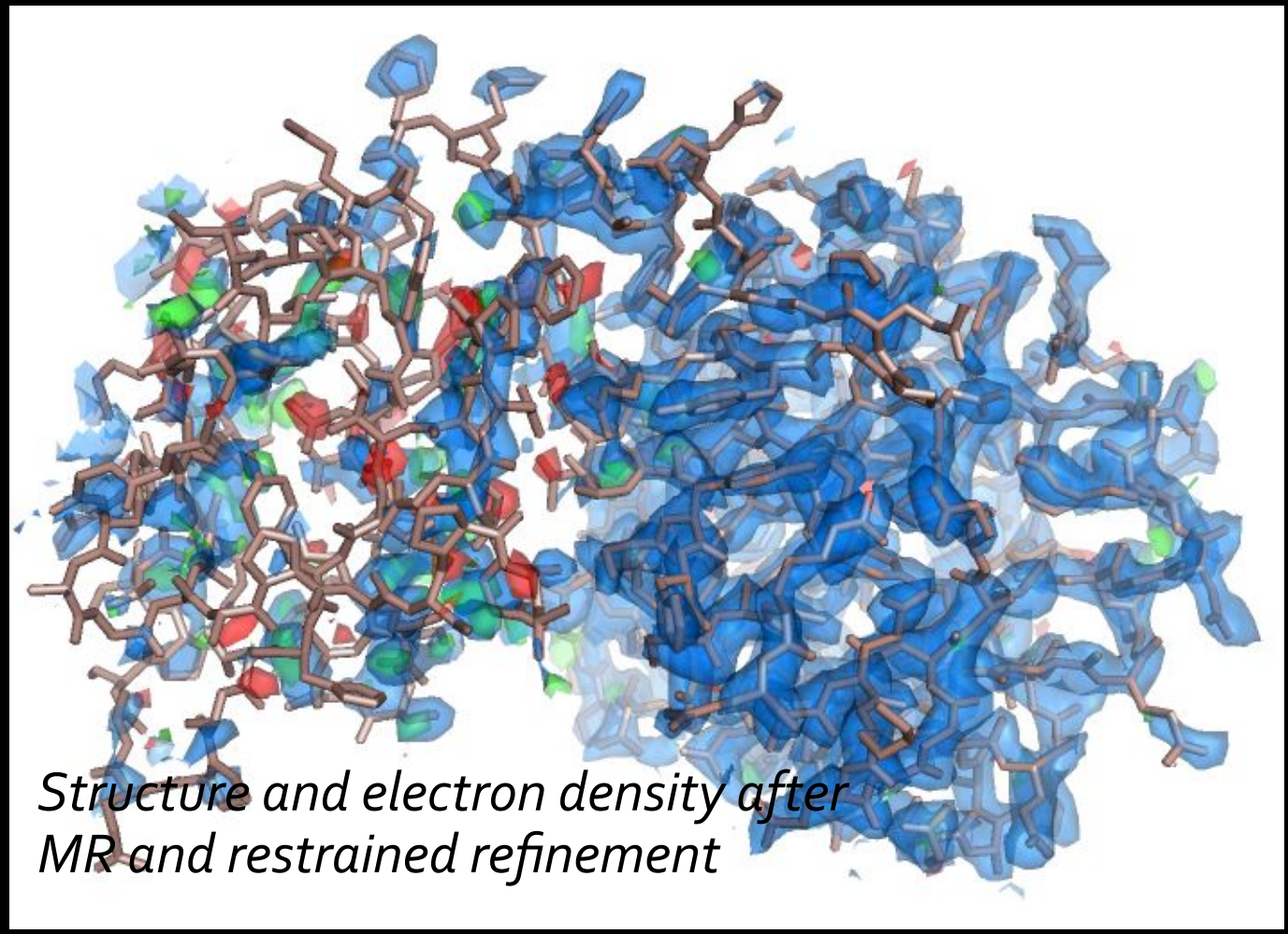
CCP4

Phased Translation search and model Splitting

Phased translation search:

Example: ***1tj3***

Search model: 1s20, chain A



Structure and electron density after MR and restrained refinement

CCP4

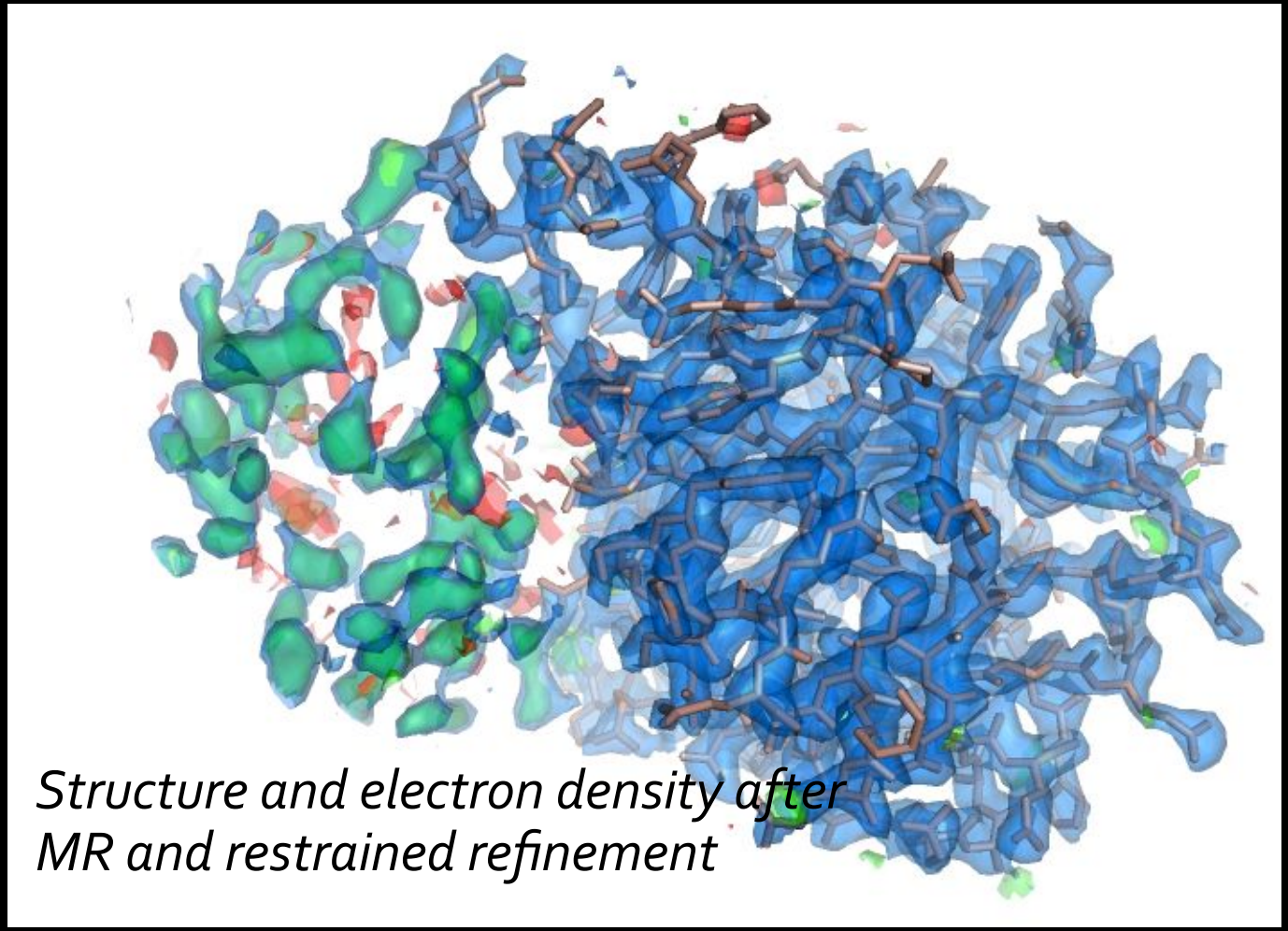
Phased Translation search and model Splitting

Phased translation search:

Example: ***1tj3***

Search model: 1s20, chain A

Large Domain only



Structure and electron density after MR and restrained refinement

CCP4

Phased Translation search and model Splitting

Phased translation search:

Example: ***1tj3***

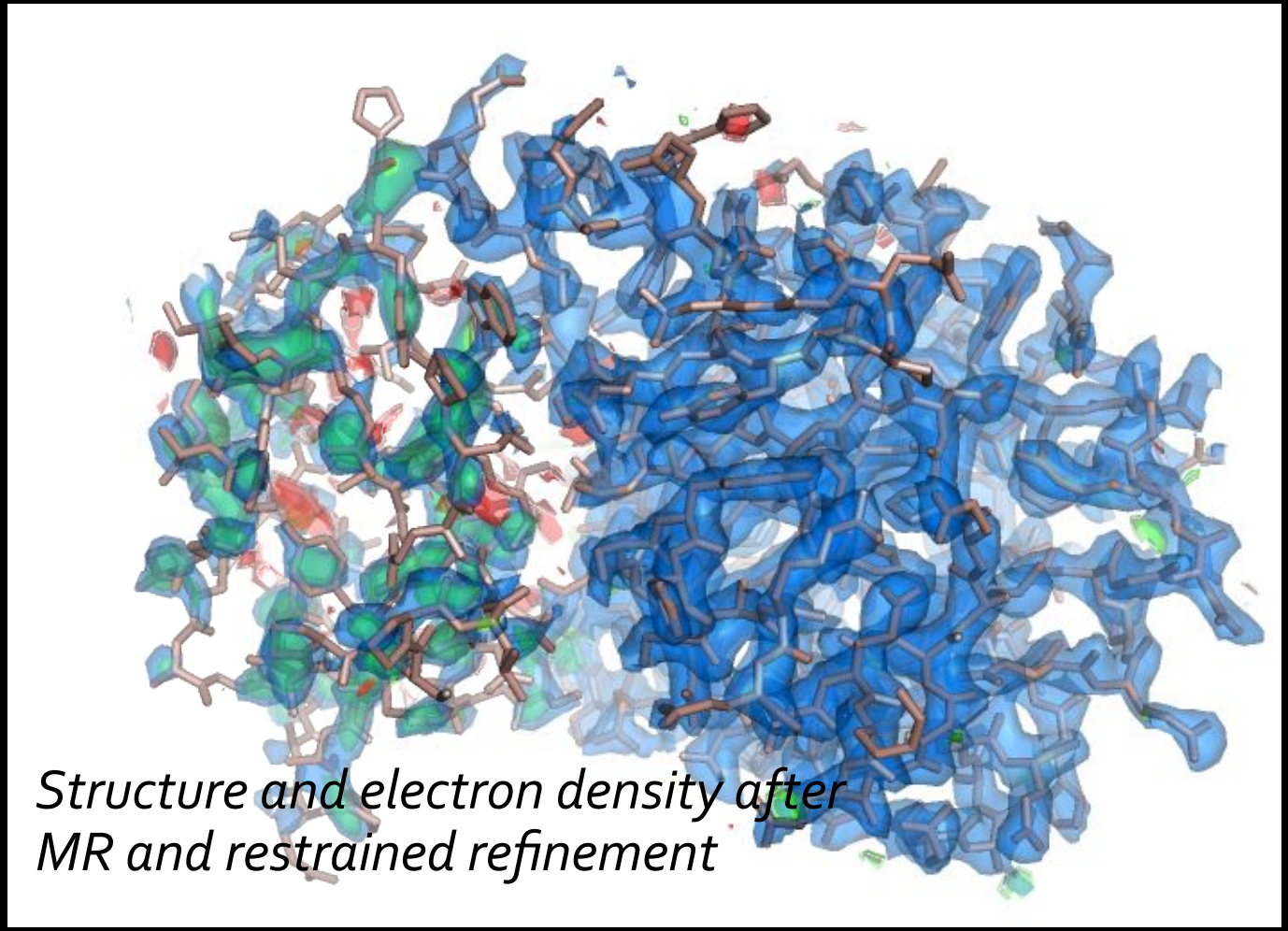
Search model: 1s20, chain A

Large Domain fixed

Small Domain search

Fit into density

(Molrep & Phaser)



Structure and electron density after MR and restrained refinement

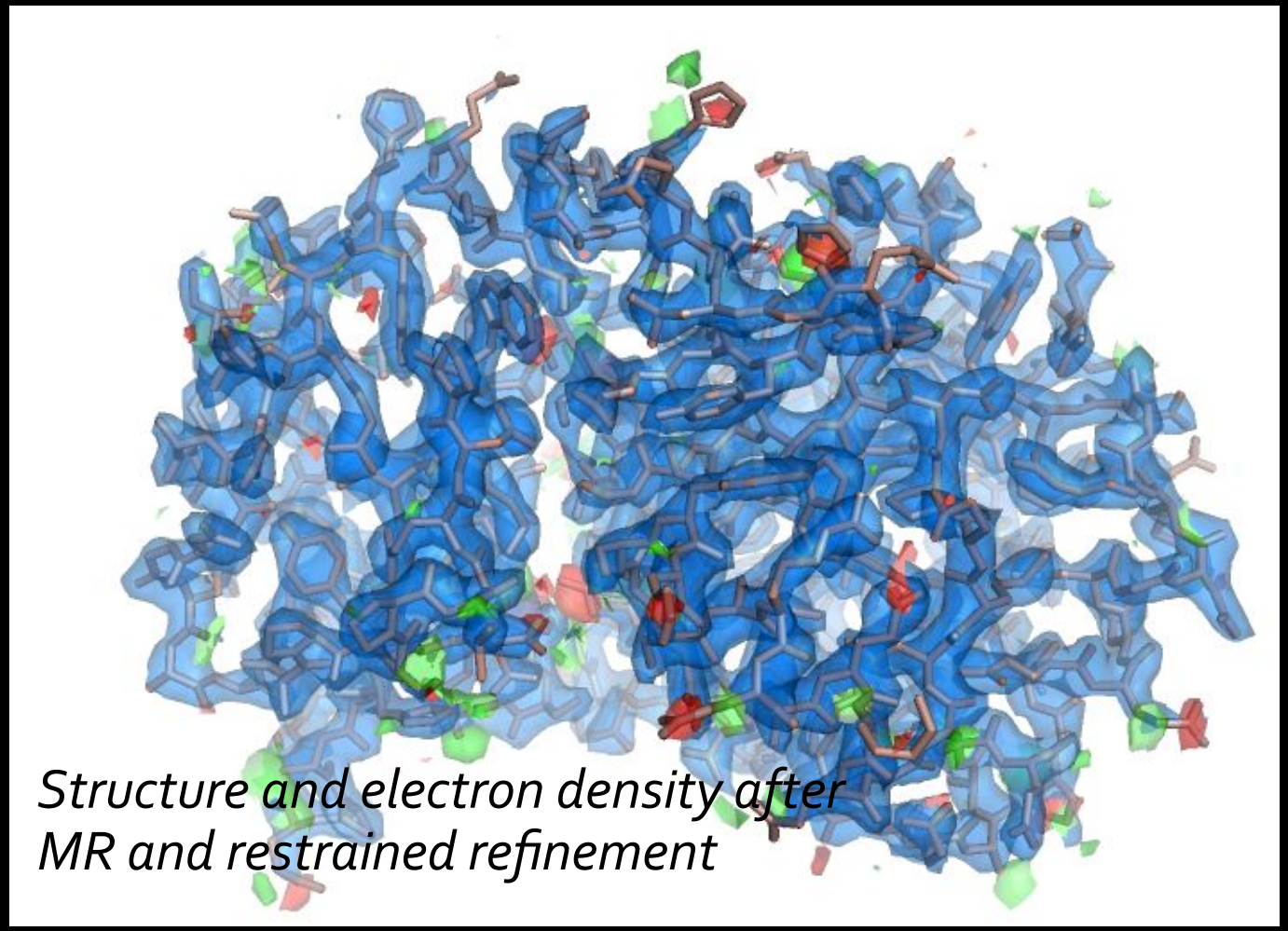
CCP4

Phased Translation search and model Splitting

Phased translation search:

Example: ***1tj3***

Complete structure and electron density after restrained refinement



Structure and electron density after MR and restrained refinement

CCP4

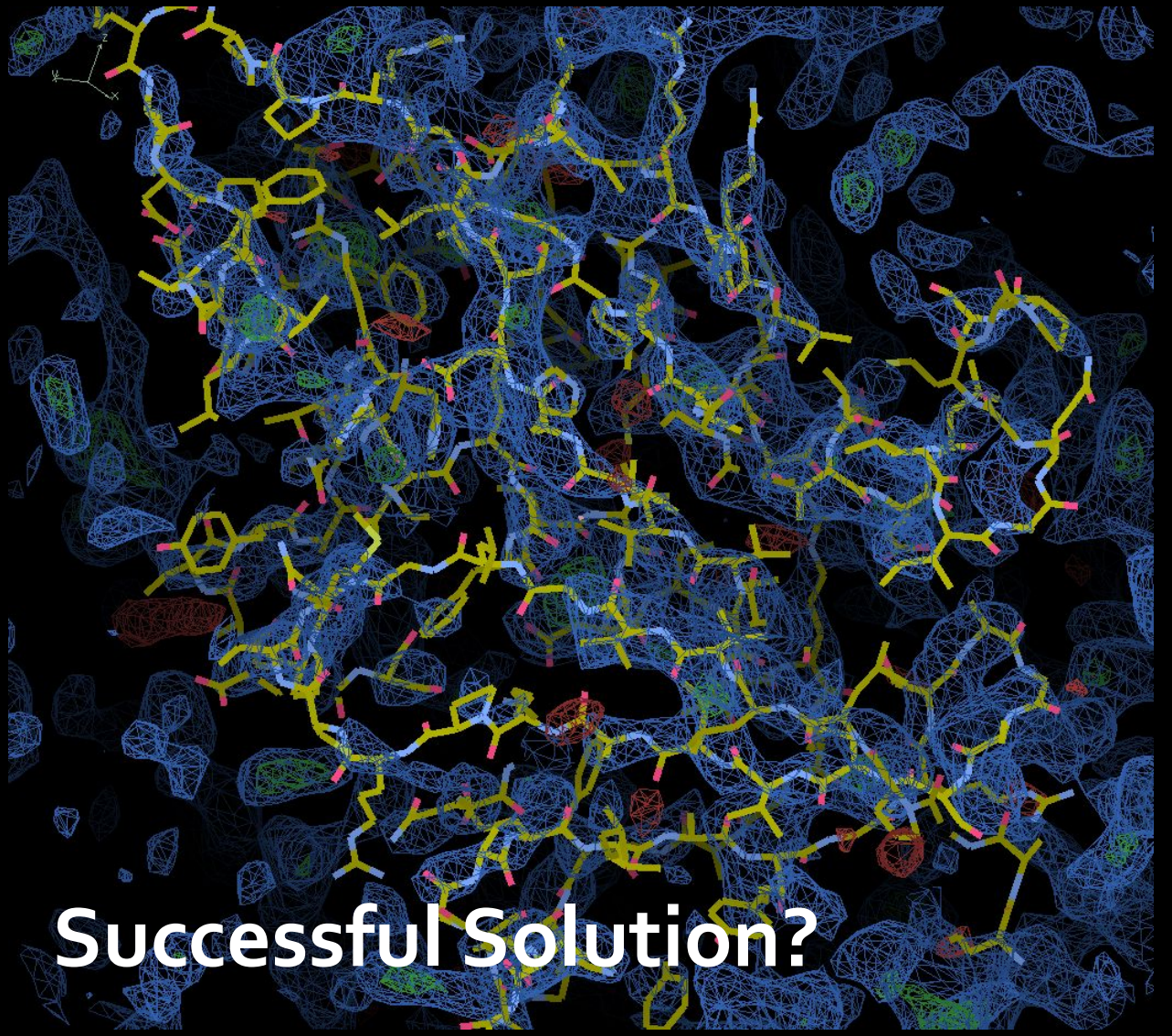
How do I know my MR solution is successful?

Assessing the MR Solution

- Terminology
  - MR solution
  - Successful MR solution
- What is a successful MR solution?
  - A search model placed in the target unit cell such that it will provide us with sufficiently accurate phase estimates for the target
  - Enables us to complete the model through model building and refinement

Assessing the MR Solution

- In difficult cases the position may be correct but getting from the MR solution to a complete model may not be straight forward
- Assessment often involves performing additional structure solutions steps such as refinement and density modification



Successful Solution?

CCP4

With 50 cycles of jelly body refinement

Successful Solution? Yes

CCP4

Assessing the solution:
Molecular Replacement Scoring

- Rough guide to MR program scoring

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

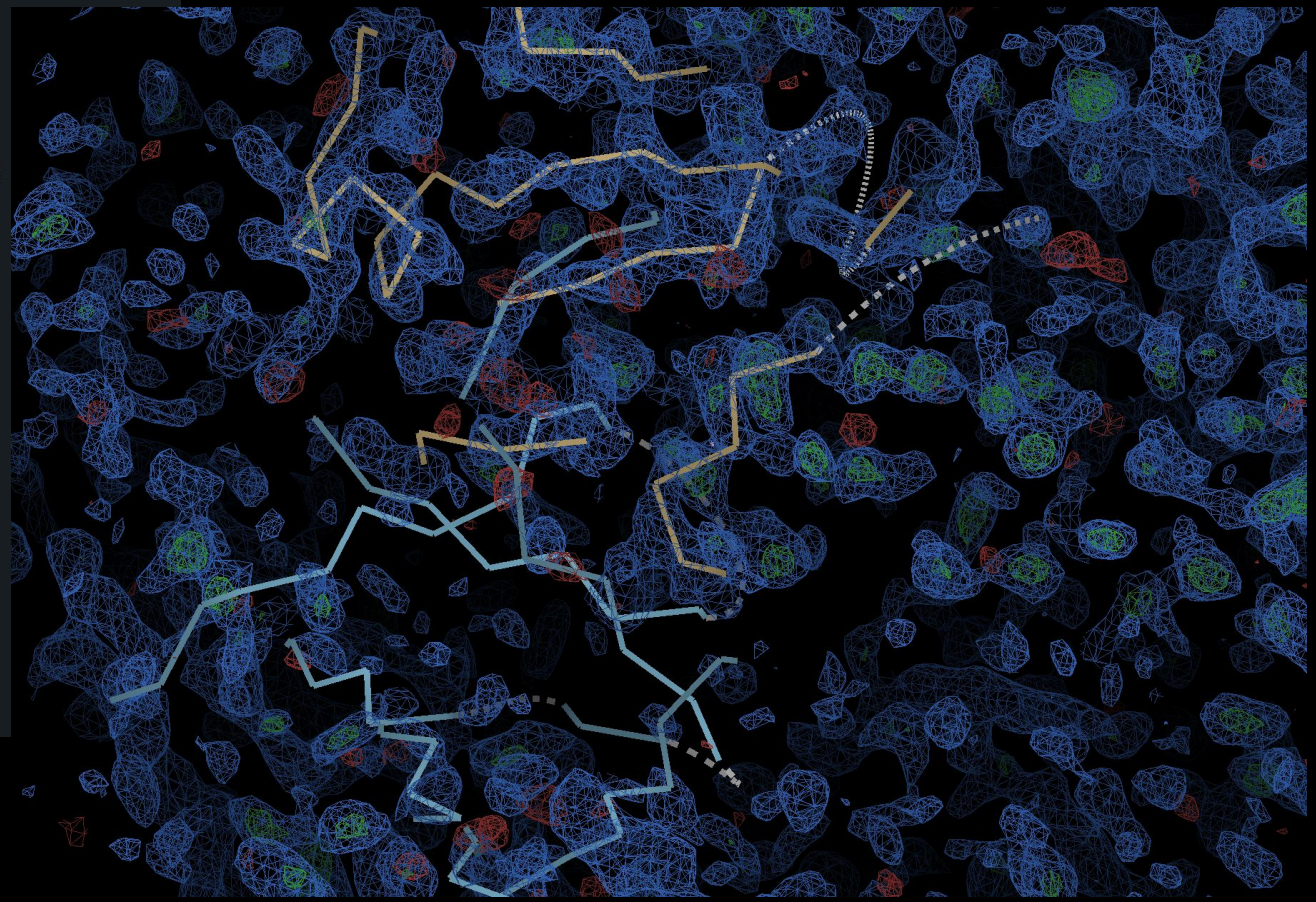
      (resolution and space group dependent)
    - TFZ – greater than 8?
    - Few or single solution almost always indicative of success

- 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?

Solving cases with many copies

```
7115 ***************************************************************************************
7116 *** Phaser Module: AUTOMATED MOLECULAR REPLACEMENT                          2.8.3 ***
7117 ***************************************************************************************
7118
7119 ** SINGLE solution
7120
7121 ** Solution written to PDB file:  phaser_mr_output.1.pdb
7122 ** Solution written to MTZ file:  phaser_mr_output.1.mtz
7123    Solution annotation (history):
7124    SOLU SET  RFZ=6.7 TFZ=7.5 PAK=0 LLG=45 TFZ==8.1 RFZ=4.4 TFZ=12.9 PAK=0 LLG=137 TFZ==13.6 RFZ=3.4 TFZ=4.8 PAK=32
7125     LLG=164 TFZ==6.9 RFZ=2.5 TFZ=12.8 PAK=32 LLG=262 TFZ==14.8 RFZ=2.3 TFZ=5.7 PAK=32 LLG=297 TFZ==10.4 RFZ=1.7 TFZ=5.9
7126     PAK=32 LLG=329 TFZ==10.1 RFZ=2.3 TFZ=10.3 PAK=32 LLG=360 TFZ==11.7 RFZ=1.8 TFZ=11.0 PAK=32 LLG=398 TFZ==13.8
7127     LLG=755 TFZ==16.1 PAK=32 LLG=755 TFZ==16.1
7128    SOLU SPAC P 21 21 21
7129    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_0 EULER  113.3   55.3  188.8 FRAC  0.31  0.02  0.08 BFAC  0.30
7130     #TFZ==8.1
7131    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_0 EULER  117.2  142.7   12.8 FRAC -0.01  0.34  0.30 BFAC -3.30
7132     #TFZ==13.6
7133    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_1 EULER  234.1   83.3  143.6 FRAC  0.12 -0.67  0.24 BFAC 16.96
7134     #TFZ==6.9
7135    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_1 EULER  206.5   31.0  188.9 FRAC  0.11 -0.28  0.74 BFAC -16.35
7136     #TFZ==14.8
7137    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_2 EULER  121.6  108.6   37.2 FRAC  0.02  0.21  0.30 BFAC -1.54
7138     #TFZ==10.4
7139    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_2 EULER  283.6   75.3  214.8 FRAC  0.07  0.12  0.57 BFAC  0.93
7140     #TFZ==10.1
7141    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_3 EULER   70.2  115.8  249.6 FRAC -0.06  0.34  0.36 BFAC -1.70
7142     #TFZ==11.7
7143    SOLU 6DIM ENSE pdb_0012-01_T1145TS067_1_cluster_3 EULER  335.3   67.9   70.7 FRAC  0.12 -0.02  0.52 BFAC  0.97
7144     #TFZ==16.1
7145    SOLU ENSEMBLE pdb_0012-01_T1145TS067_1_cluster_0 VRMS DELTA -1.1862 #RMSD  1.20 #VRMS  0.50
7146    SOLU ENSEMBLE pdb_0012-01_T1145TS067_1_cluster_1 VRMS DELTA -0.6488 #RMSD  1.20 #VRMS  0.89
7147    SOLU ENSEMBLE pdb_0012-01_T1145TS067_1_cluster_2 VRMS DELTA -0.8625 #RMSD  1.20 #VRMS  0.76
7148    SOLU ENSEMBLE pdb_0012-01_T1145TS067_1_cluster_3 VRMS DELTA -1.0630 #RMSD  1.20 #VRMS  0.61
7149
7150 CPU Time: 0 days 6 hrs 26 mins 9.75 secs (  23169.75 secs)
7151 Finished: Fri Nov 25 23:52:13 2022
7152
```



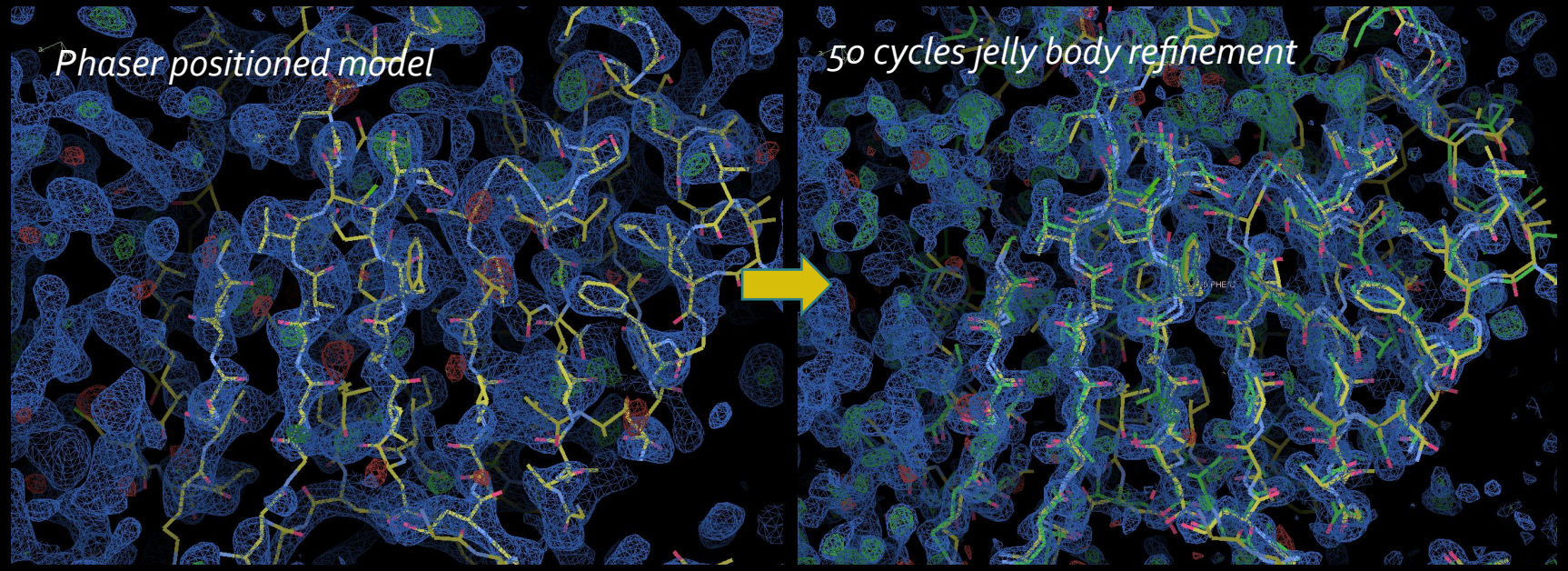
CCP4

Assessing and improving the solution: Refinement

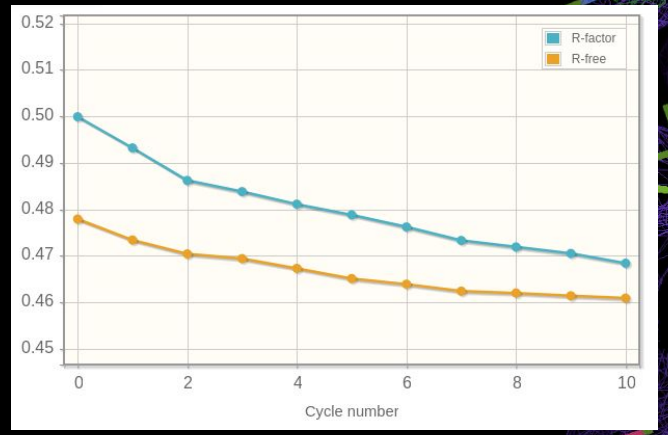
- Refinement
  - Look at Rfactor/Rfree
    - are they falling? Is Rfree below 0.5?

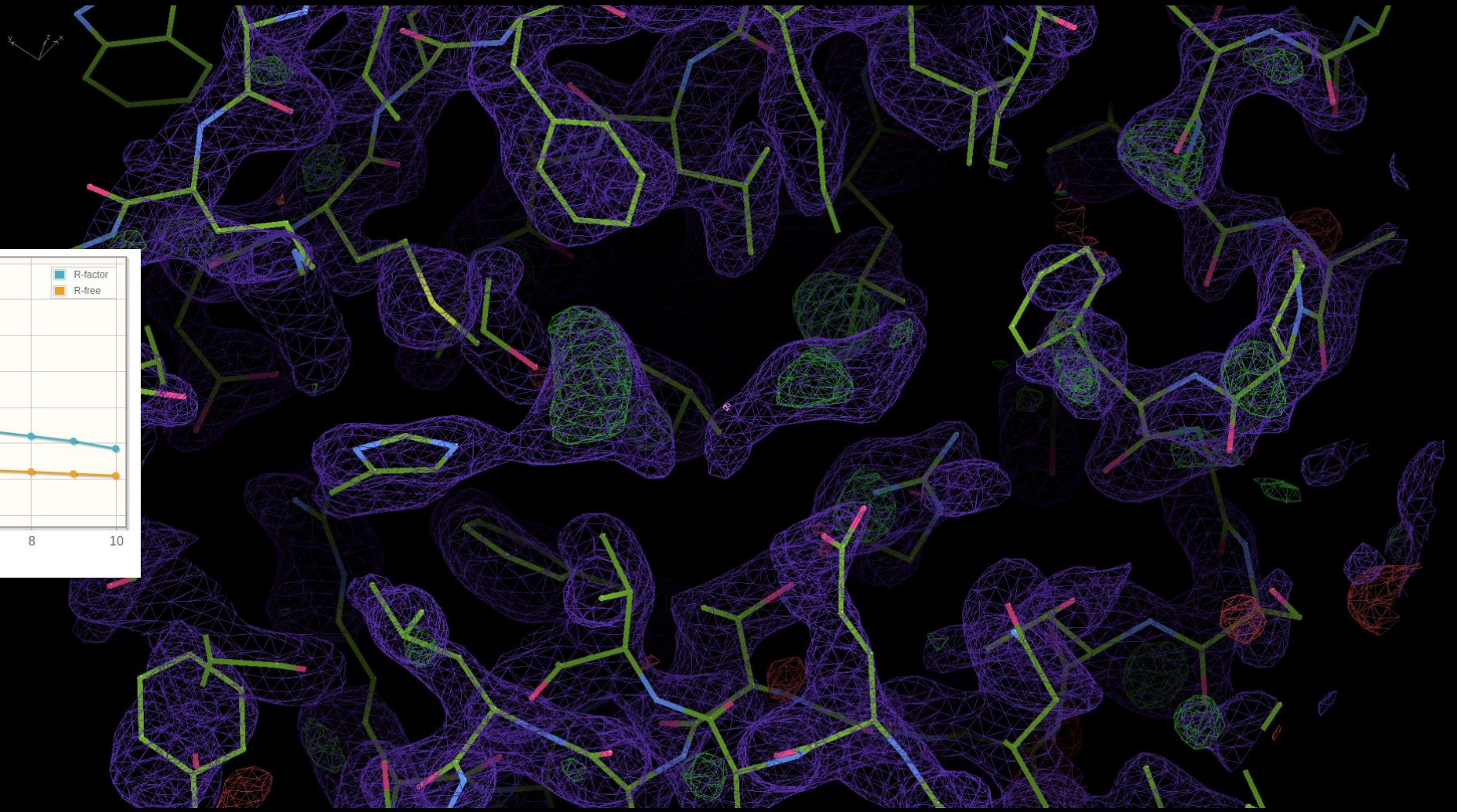
- Refinement
  - Look at Rfactor/Rfree
    - are they falling? Is Rfree below 0.5?
  - Use 50 cycles of jelly body refinement option in Refmac post MR

- Refinement
  - Look at Rfactor/Rfree
    - are they falling? Is Rfree below 0.5?
  - Use 50 cycles of jelly body refinement option in Refmac post MR



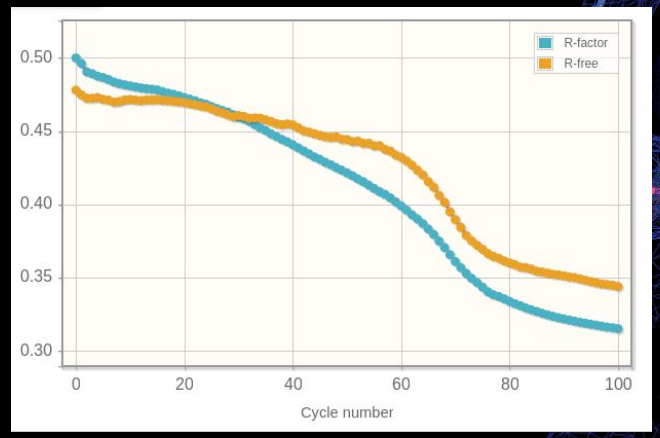
Refining predicted models

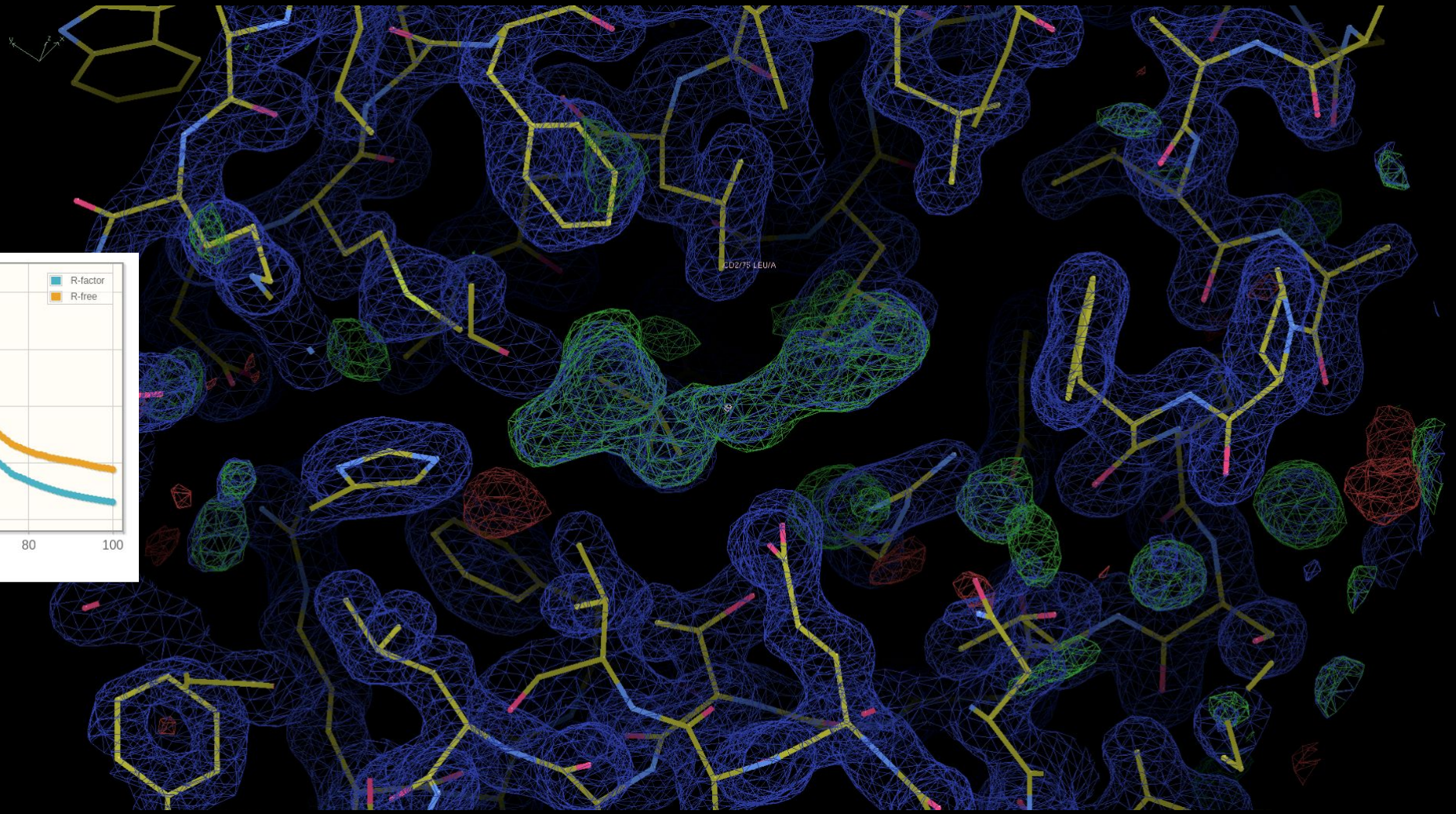




CCP4

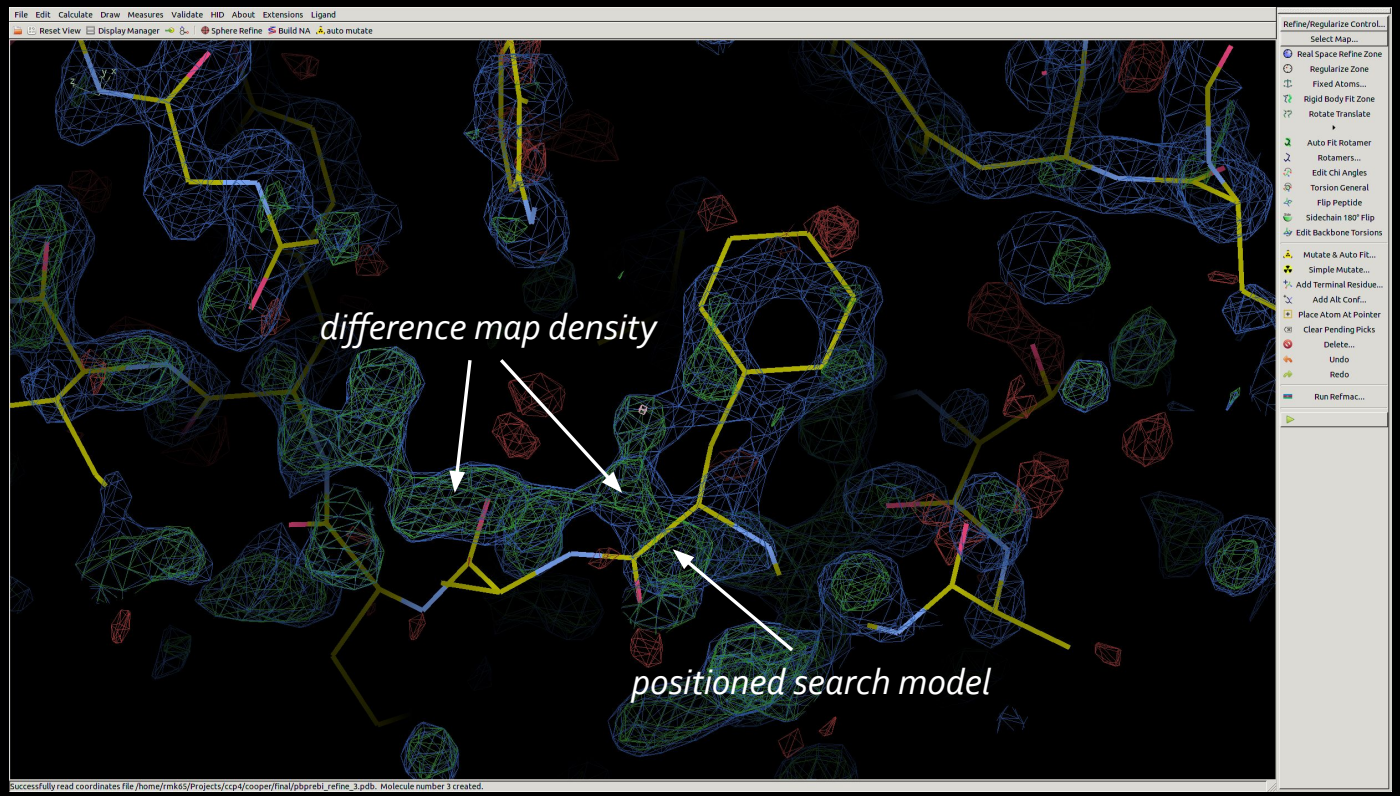
Refining predicted models





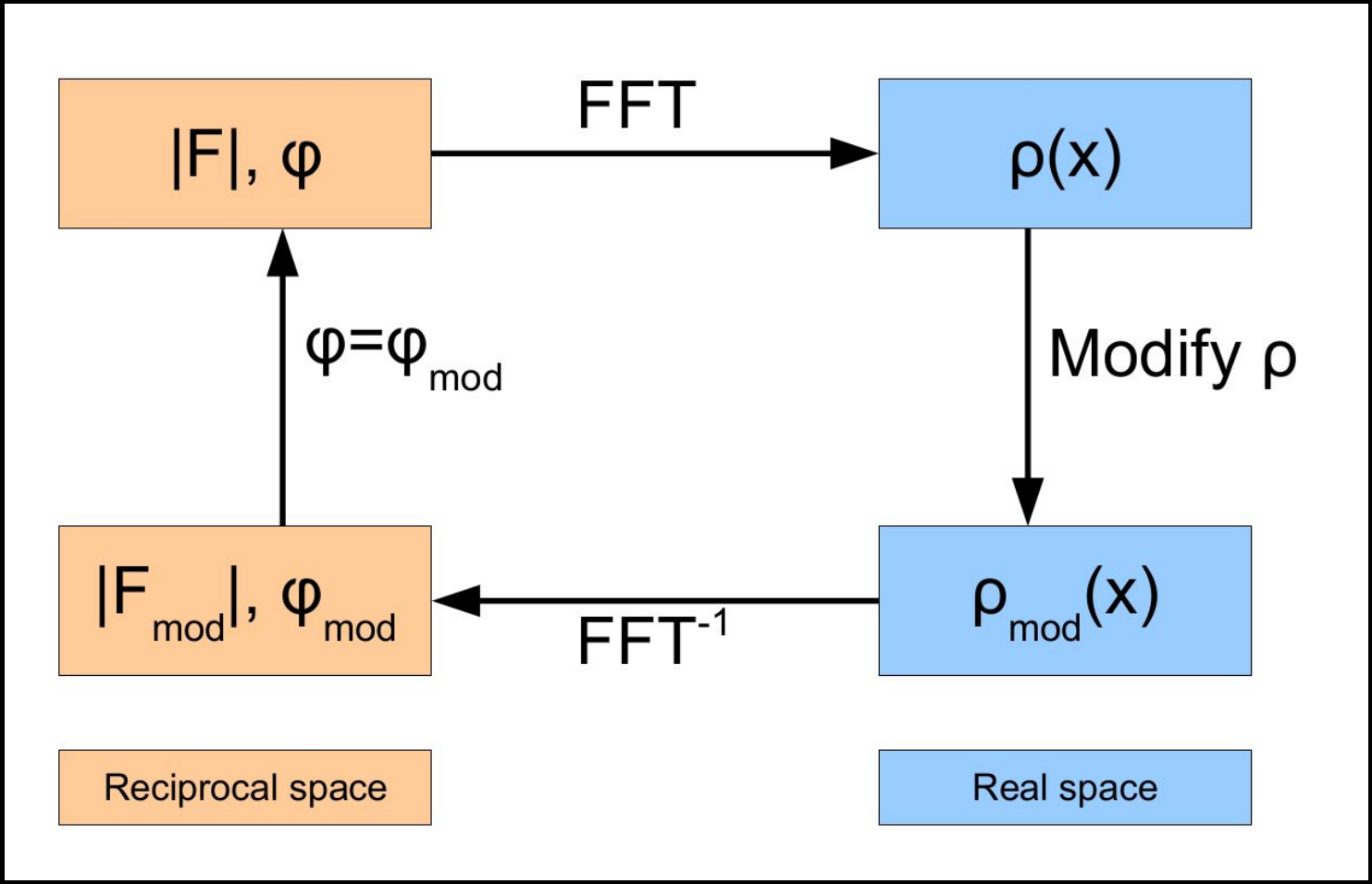
CCP4

- Examine solution by eye
  - Use Coot to examine positioned models & maps



Assessing and improving the solution: Density Modification

Density modification



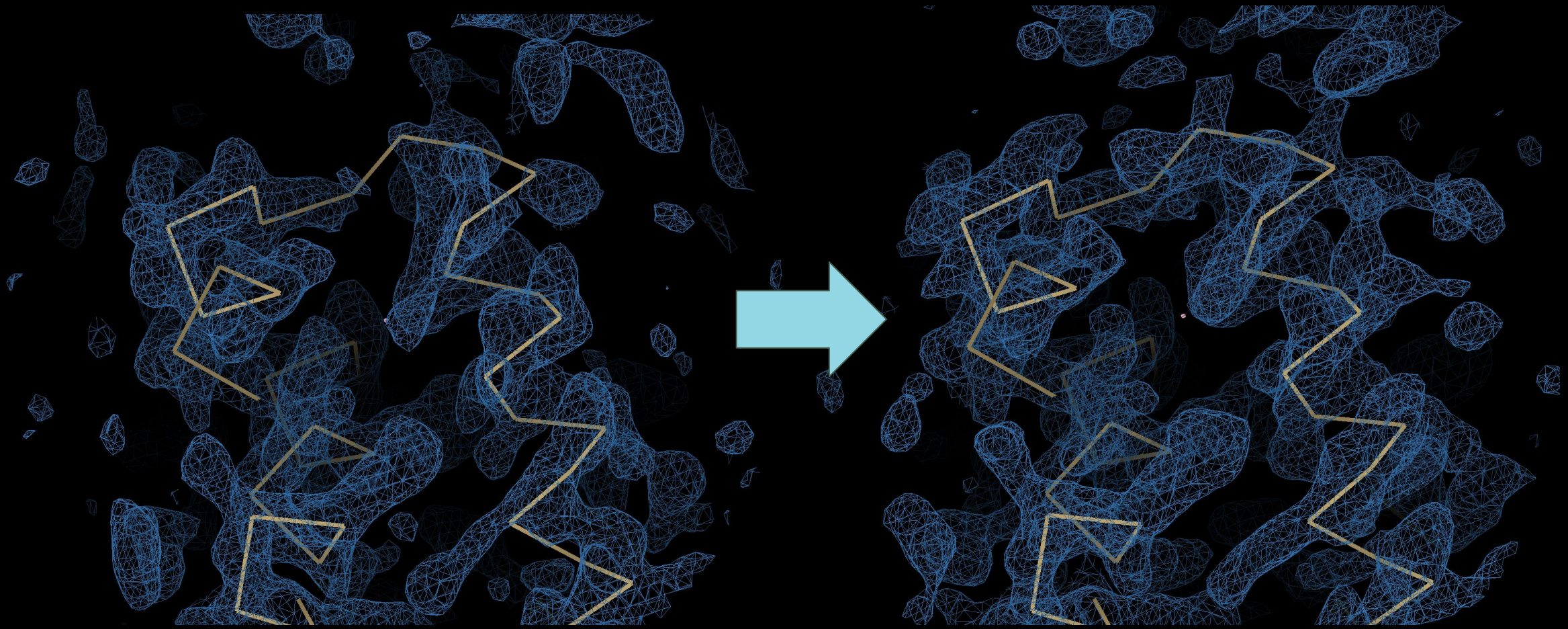
Main techniques:

1. Solvent flattening
2. Histogram matching
3. NCS averaging
4. C-alpha tracing

(slide from Kevin Cowtan)

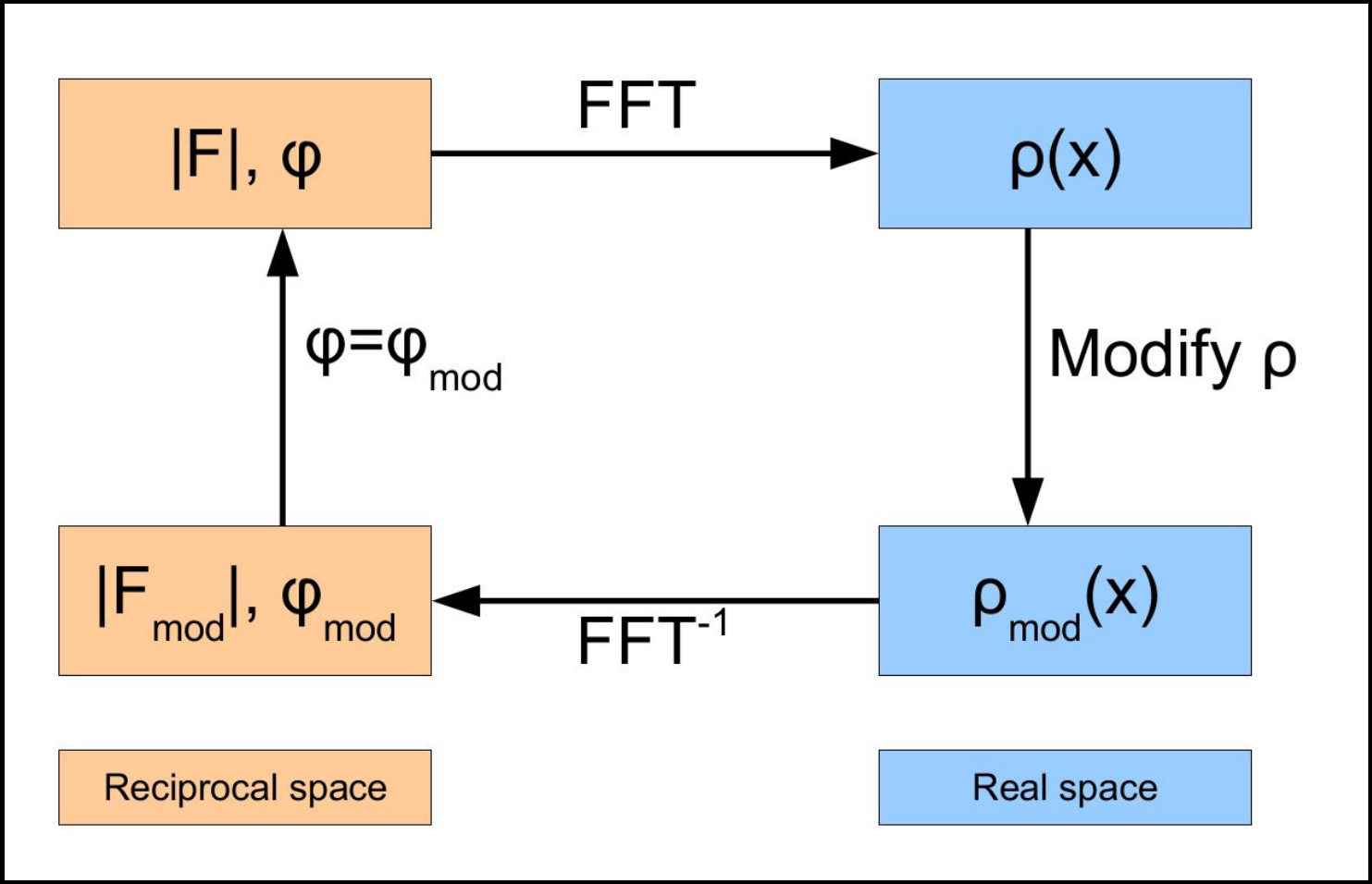
CCP4

Density modification



CCP4

Density modification



CCP4 Applications:

- Parrot
- SHELXE
- ACORN
- Pirate
- Solomon
- DM

(slide from Kevin Cowtan)

CCP4

Assessing the MR Solution: Model Building

- Automatic model building: ***Buccaneer, ModelCraft & ARP/wARP***
  - Can be used post-MR for generation of better model and phases for the target
  - Rebuilding parts that may not be present in search model
  - Useful for assessing whether or not your positioned MR model is true – eliminates bias

CCP4

Automated Molecular Replacement in CCP4

- Several automation pipelines for MR in CCP4:
  - ***MrBUMP*** – model search, preparation, MR and refinement
  - ***BALBES*** – model search in custom version of PDB database
  - ***MoRDa*** – similar to BALBES

ARCIMBOLDO

- Lite
  - Makes use of simple secondary structure elements such as helices in MR
  - Attempts to position fragments and build up the rest of the c-alpha backbone using SHELXE
- Borges
  - Similar to Lite but draws on a library of common motifs from the PDB as search models e.g. Zinc fingers or common beta sheet fragments
- Shredder
  - Cuts distant homologues into fragments and will use them as search models in ARCIMBOLDO
- **http://chango.ibmb.csic.es/ARCIMBOLDO/**

Acknowledgements

- Alexey Vagin & Andrey Lebedev – ***MoRDa and Molrep***
- Daniel Rigden, Jens Thomas, Felix Simkovic & Adam Simpkin - ***University of Liverpool***
- Stuart McNicholas, Keith Wilson & Christian Roth - ***University of York***
- Randy Read, Airlie McCoy & Gabor Bunkozci - ***Phaser***
- Martyn Winn, Charles Ballard, Eugene Krissinel, Ville Uski & Kyle Stevenson - ***CCP4***
- Marcin Wojdyr - ***Global Phasing***
- Isabel Uson, Andrea Thorn, Tim Gruene & George Sheldrick – ***SHELXE***
- Garib Murshudov, Rob Nicholls & Fei Long – ***Refmac and BALBES***
- Kevin Cowtan & Paul Bond – ***Buccaneer & DM***
- Victor Lamzin & Grzegorz Chojnowski - ***ARP/wARP***
- Thanks to authors of all other underlying programs