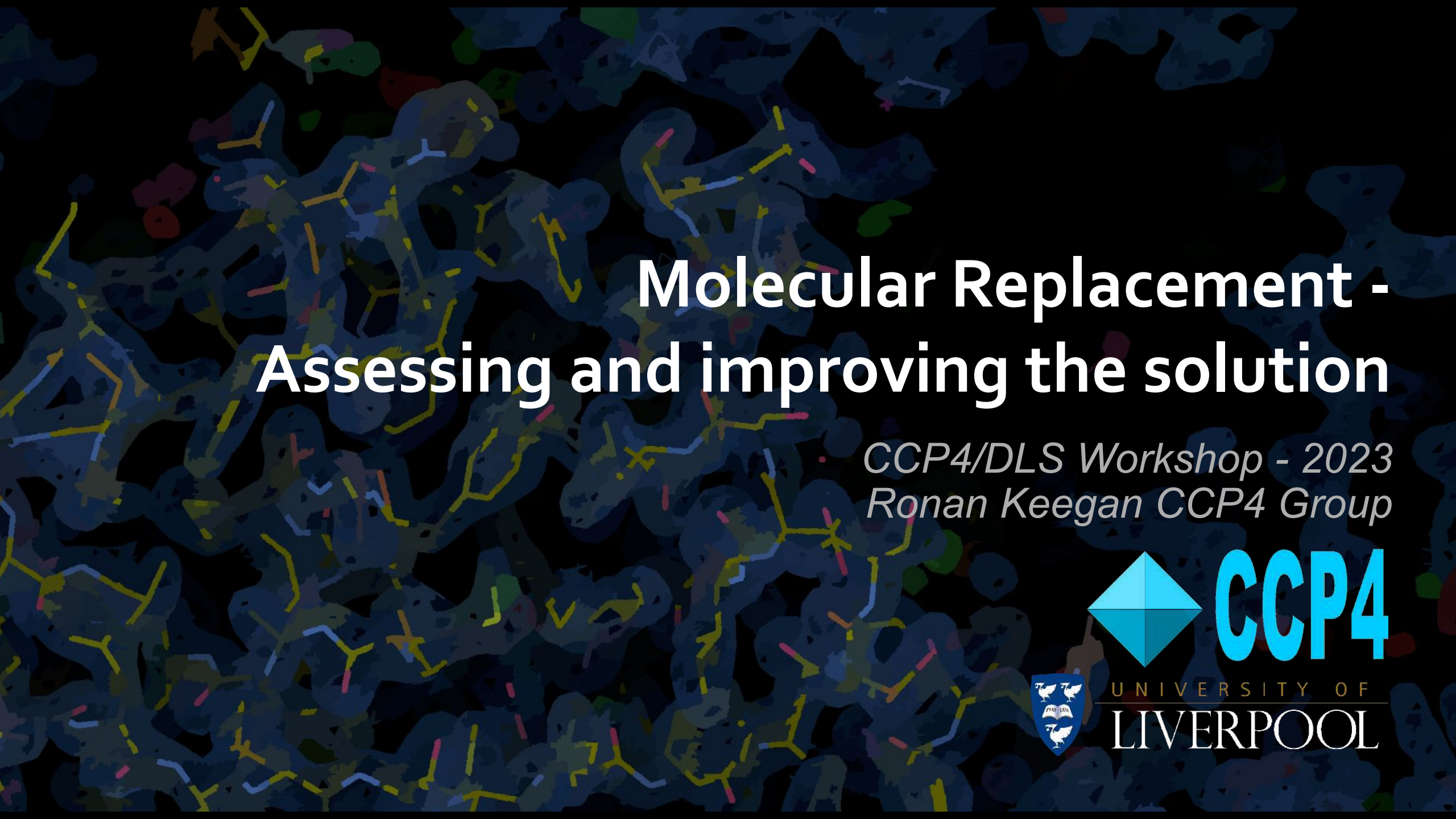


Molecular Replacement - Assessing and improving the solution

CCP4/DLS Workshop - 2023
Ronan Keegan CCP4 Group

CCP4

UNIVERSITY OF LIVERPOOL

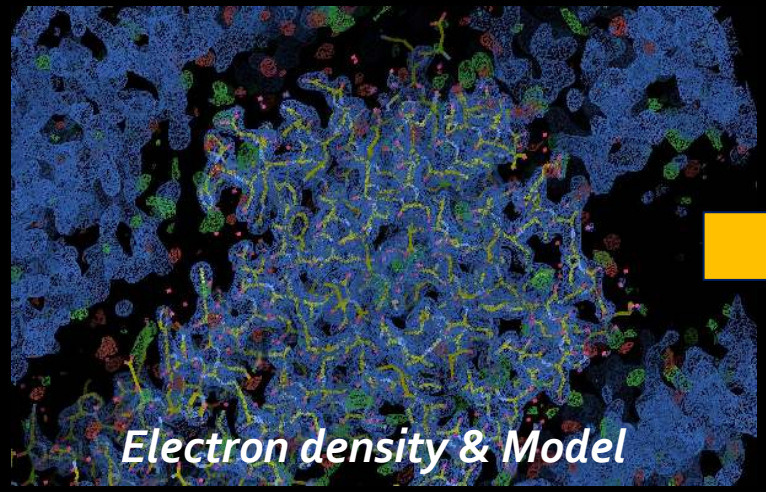
The Phase Problem

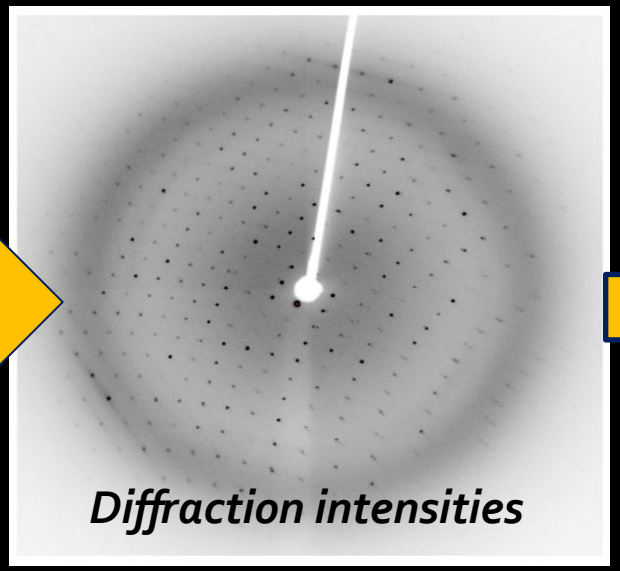
$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl}|F_o(hkl)|\cos[-2\pi(hx+ky+lz)+\varphi_{hkl}]$$

Electron density at (x,y,z)

Amplitudes

Phases





?

CCP4

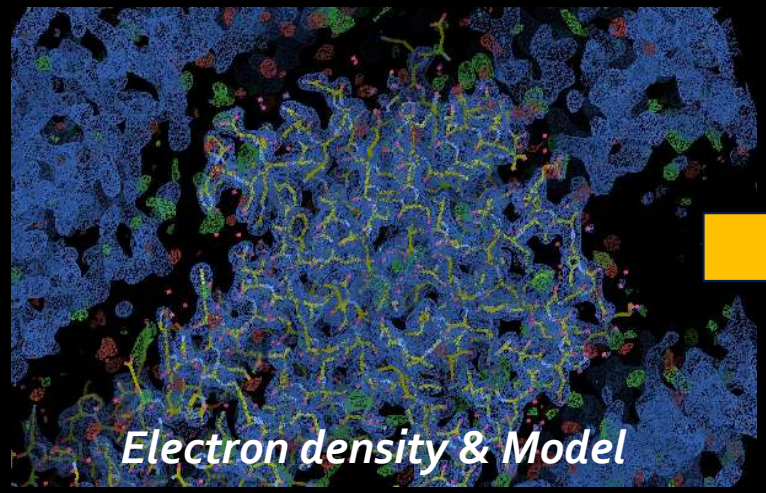
The Phase Problem: Molecular Replacement

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

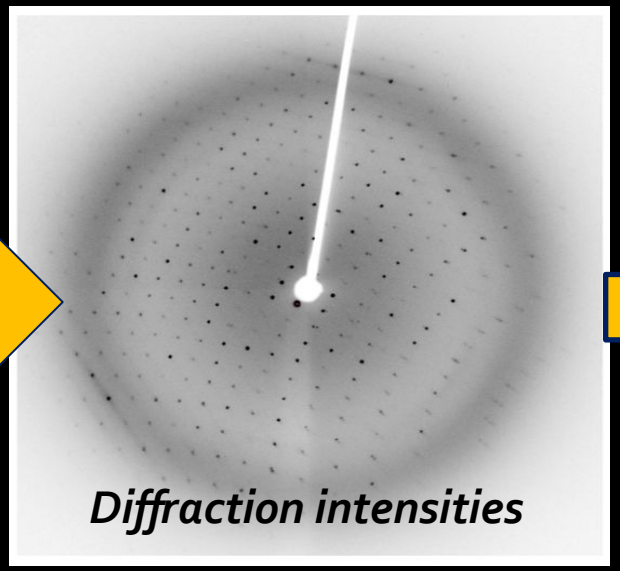
Electron density at (x,y,z)

Amplitudes

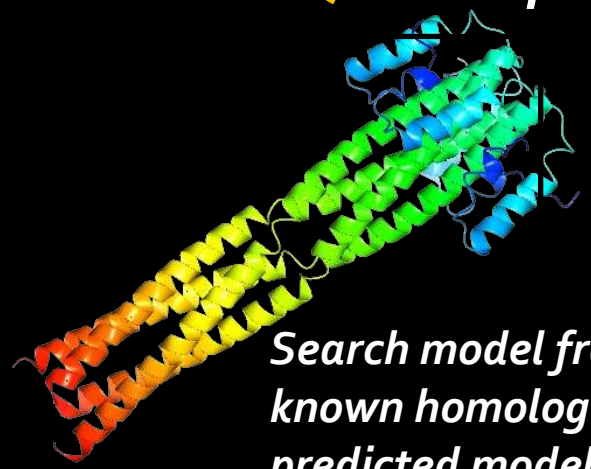
Molecular Replacement



Electron density & Model



Diffraction intensities



Search model from known homologue or predicted model

CCP4

Interpreting the results of Molecular Replacement

CCP4

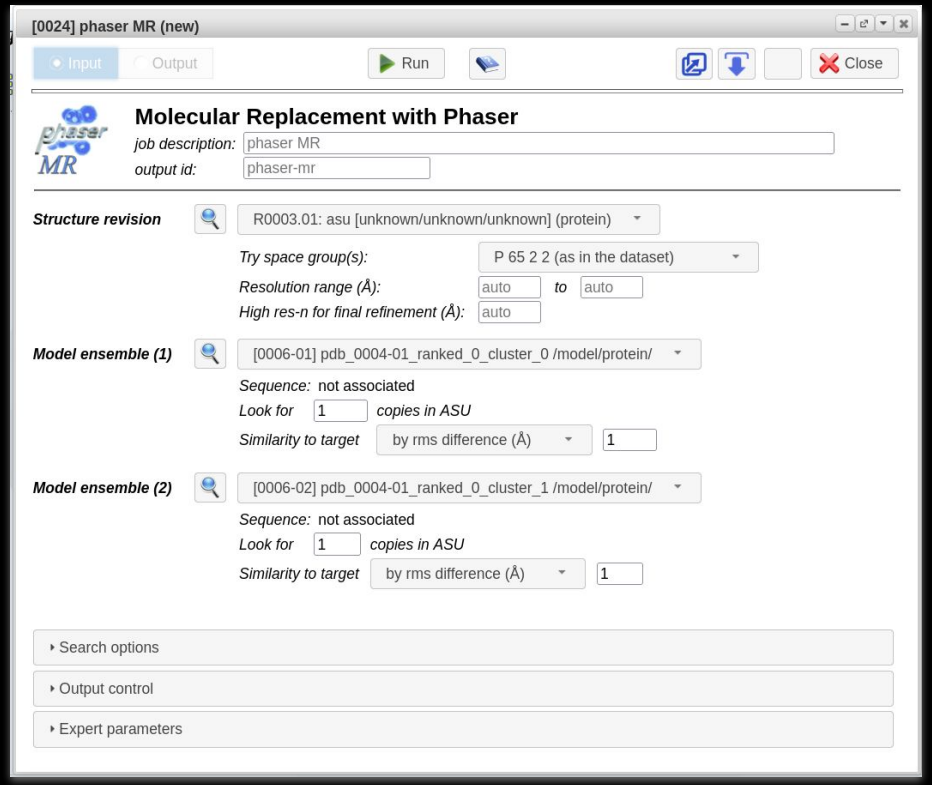
Molecular Replacement in CCP4

- 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
  - 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?

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

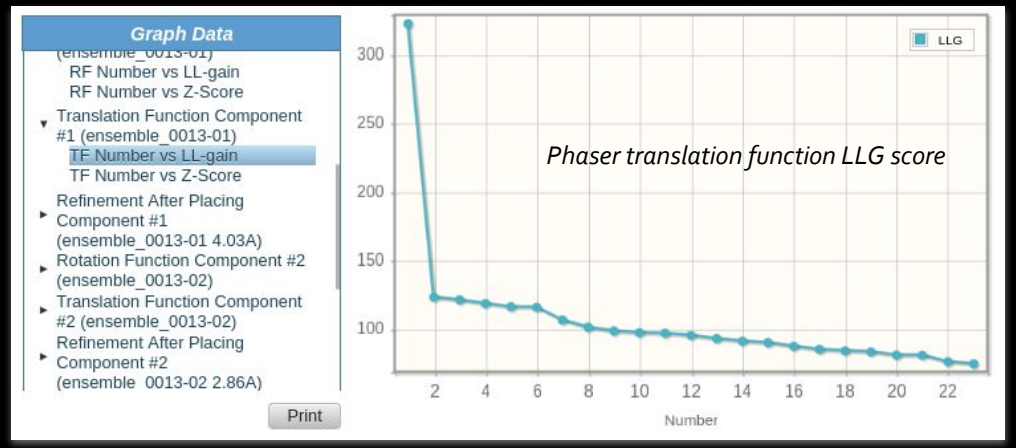
Assessing the MR Solution

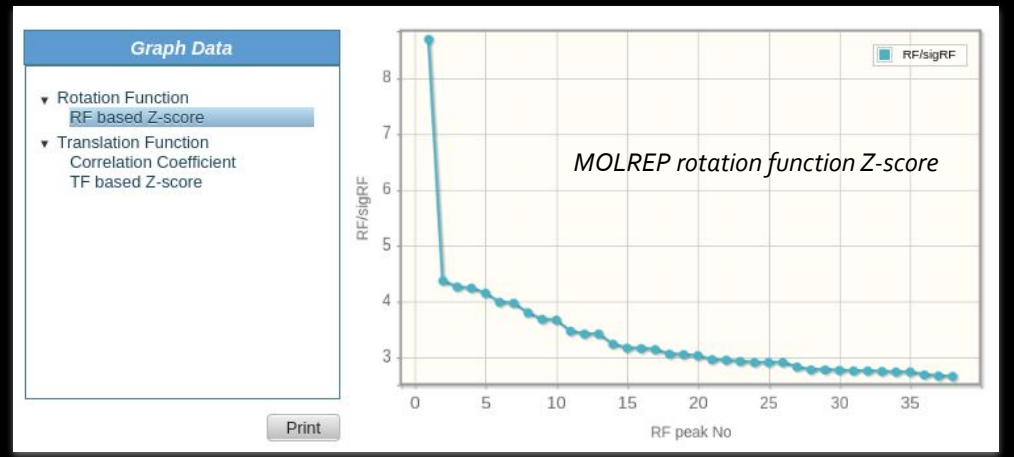
- In difficult cases the position may be correct but getting from the MR solution to a complete model may not be straight forward
- Assessment often involves performing additional structure solution steps such as refinement and density modification

Molecular Replacement Scoring

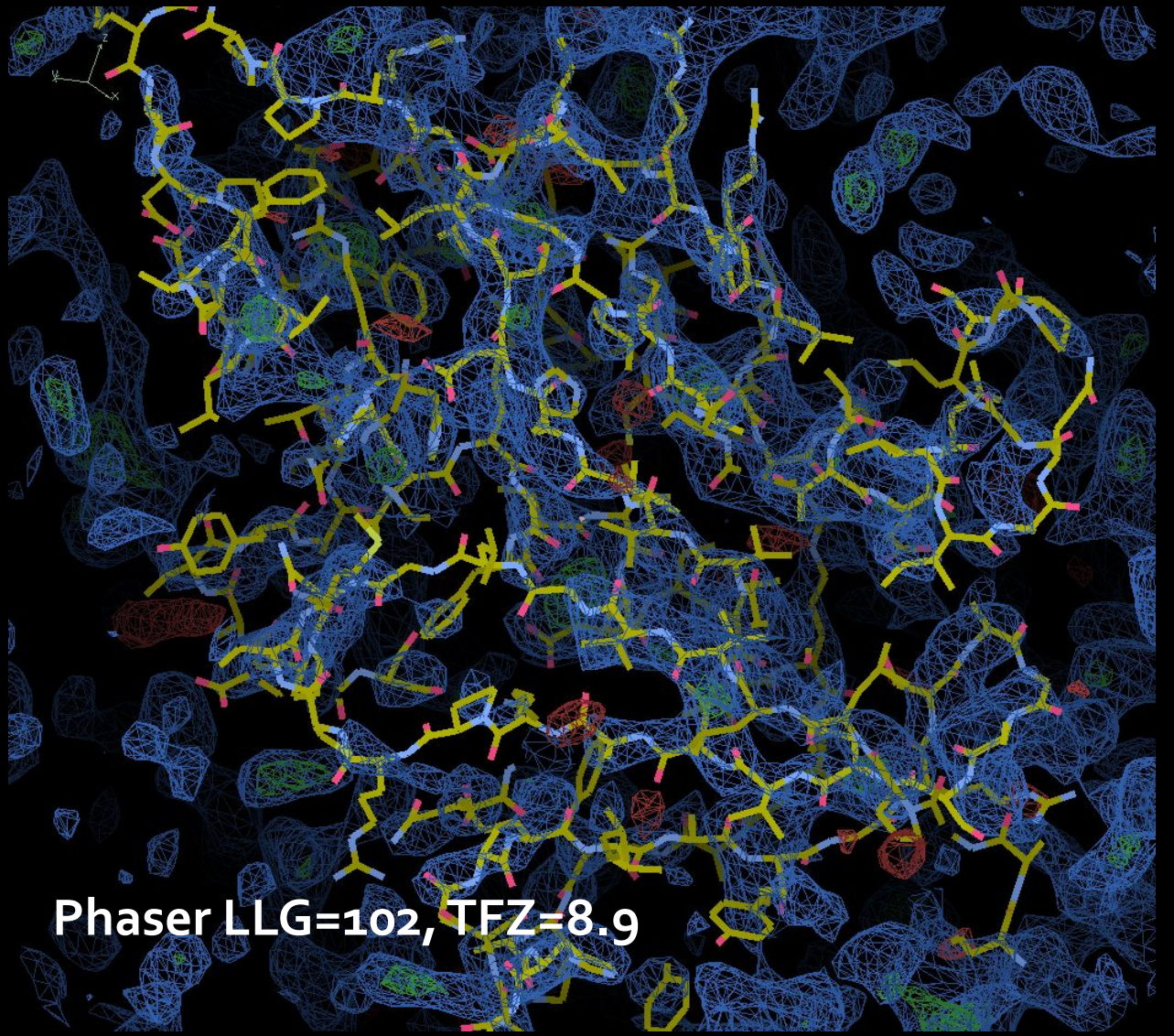
- Rough guide to MR program scoring
  - Phaser scores
    - LLG scores – has it increased by 60 or more after the placement of a new molecule?

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



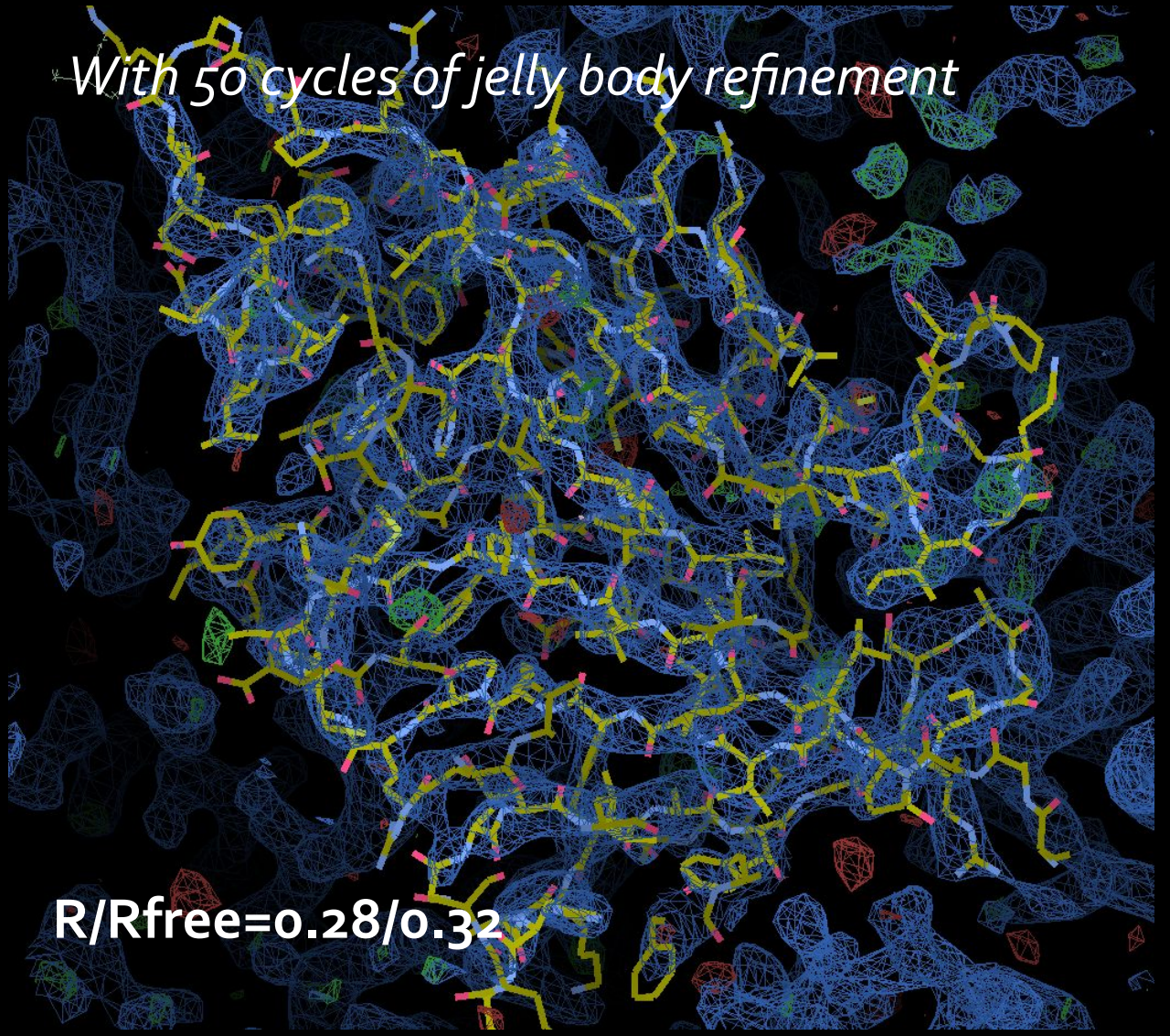


Refinement



CCP4

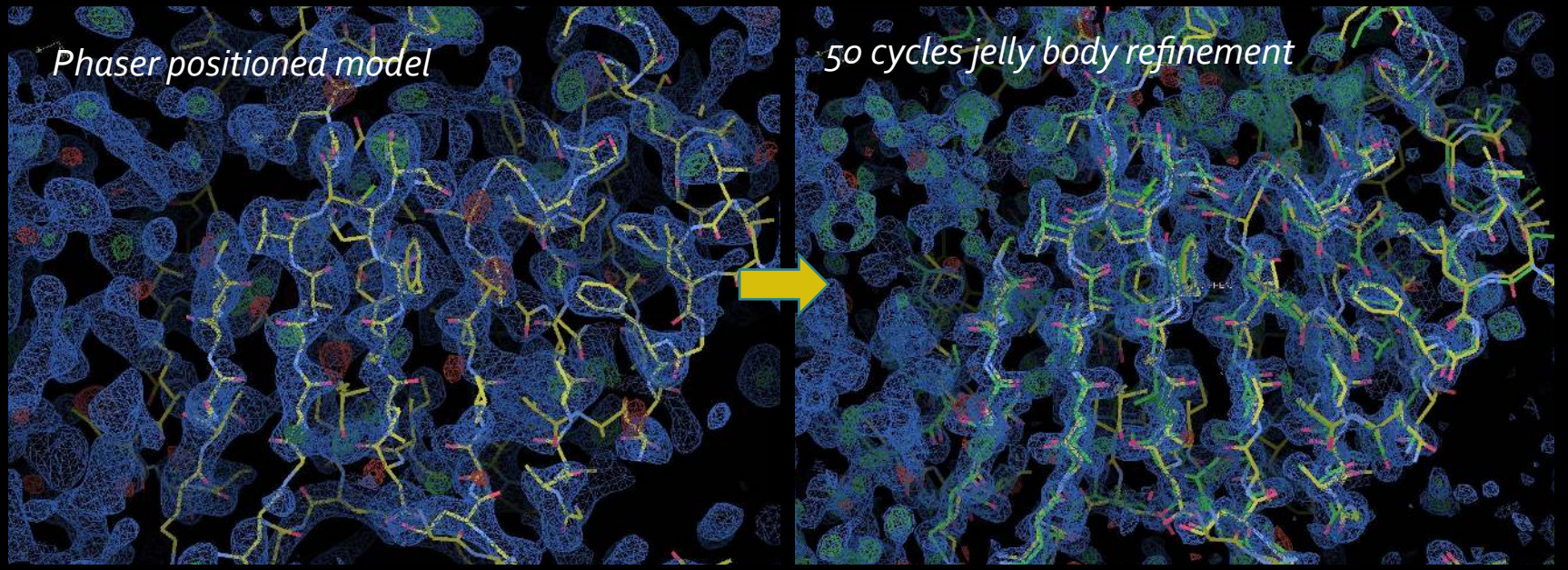
With 50 cycles of jelly body refinement



R/Rfree=0.28/0.32

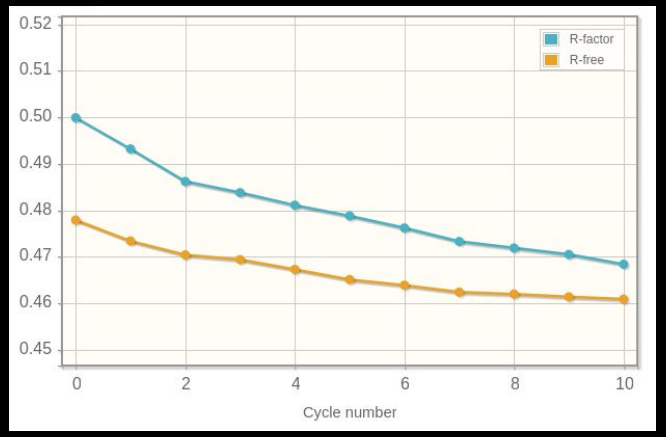
CCP4

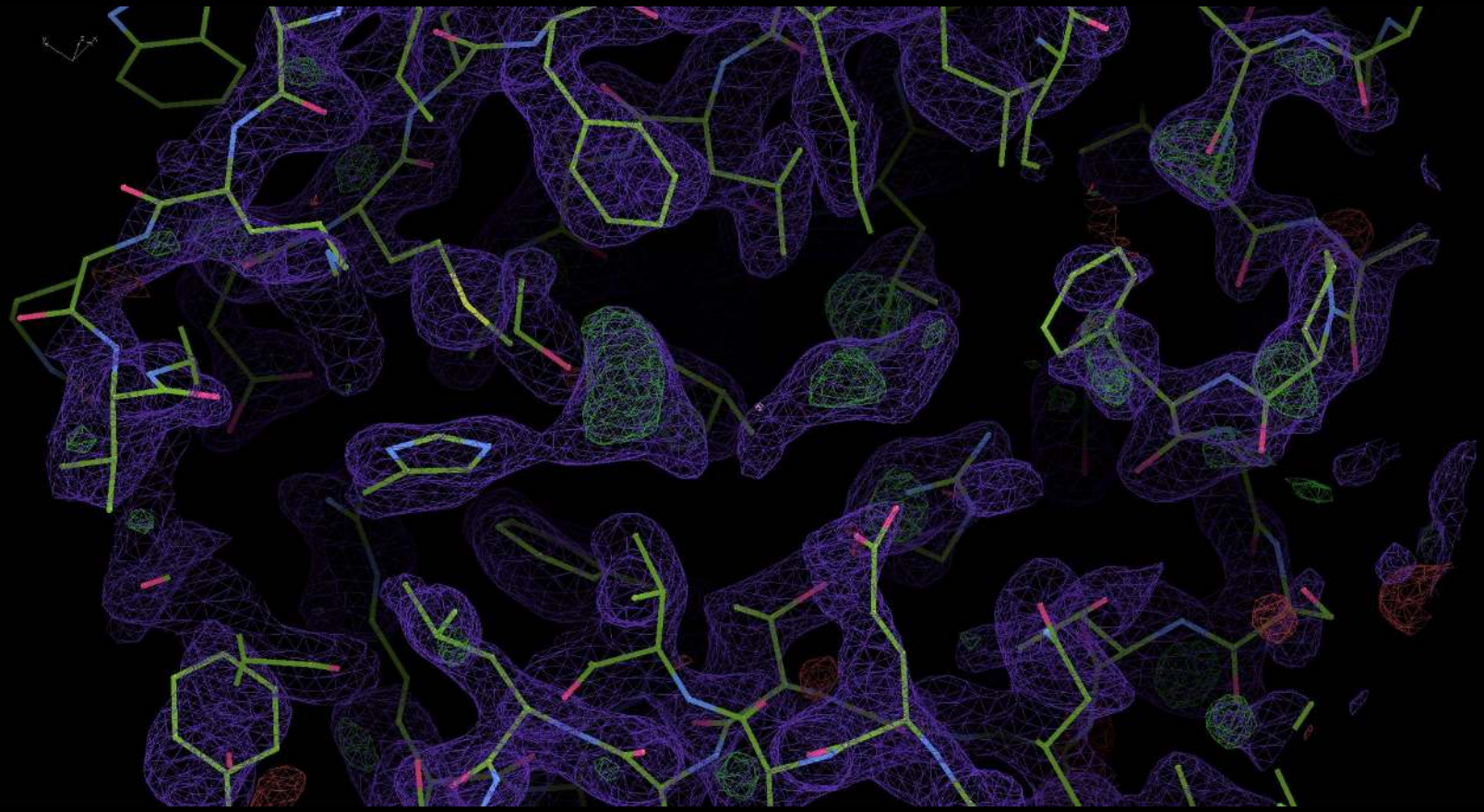
- Refinement
  - Look at Rfactor/Rfree
    - are they falling? Is Rfree below 0.5?
  - Use 50 cycles of jelly body refinement option in Refmac post MR



Refining predicted models used in Molecular Replacement

- Predicted models (*Alphafold2*, *Colabfold* etc.) are often significantly different in their main and side chain positioning to the crystal form despite making good MR search models
- Can require lots of cycles of jelly-body refinement in Refmac



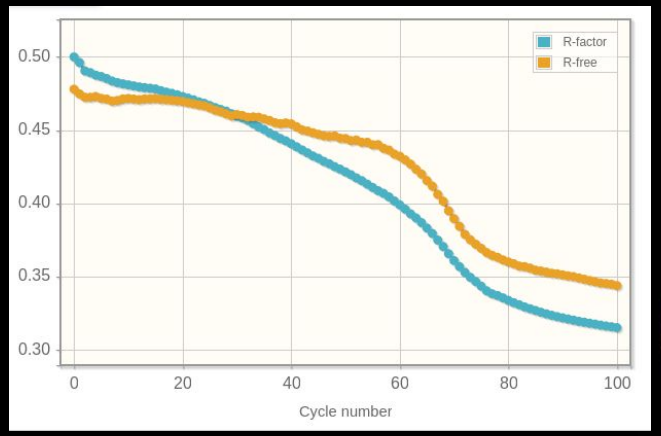


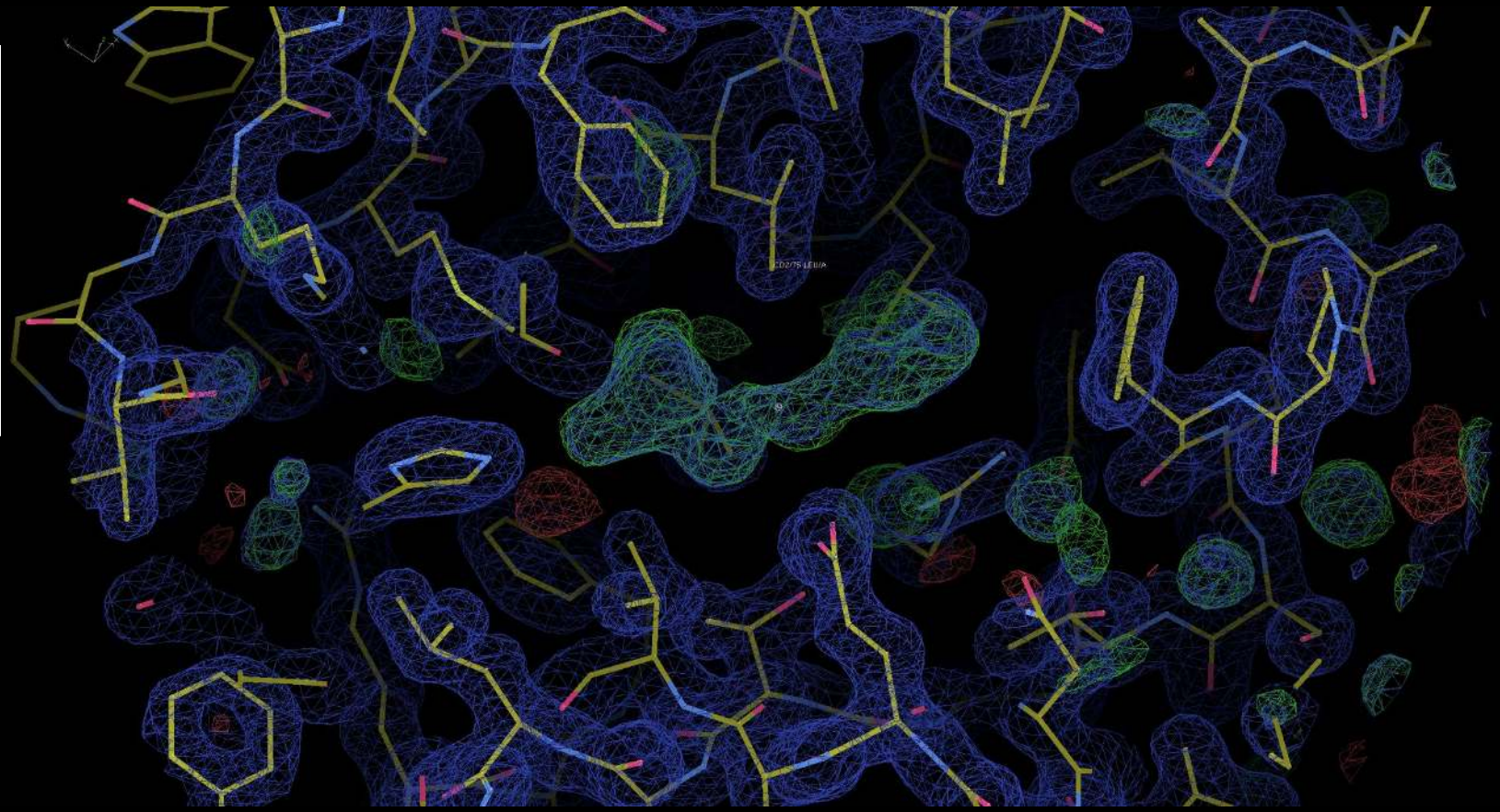
Example: 10 cycles of jelly-body refinement with Refmac
Rfactor = 0.46
Rfree = 0.47

CCP4

Refining predicted models used in Molecular Replacement

- Predicted models (*Alphafold2*, *Colabfold* etc.) are often significantly different in their main and side chain positioning to the crystal form despite making good MR search models
- Can require lots of cycles of jelly-body refinement in Refmac

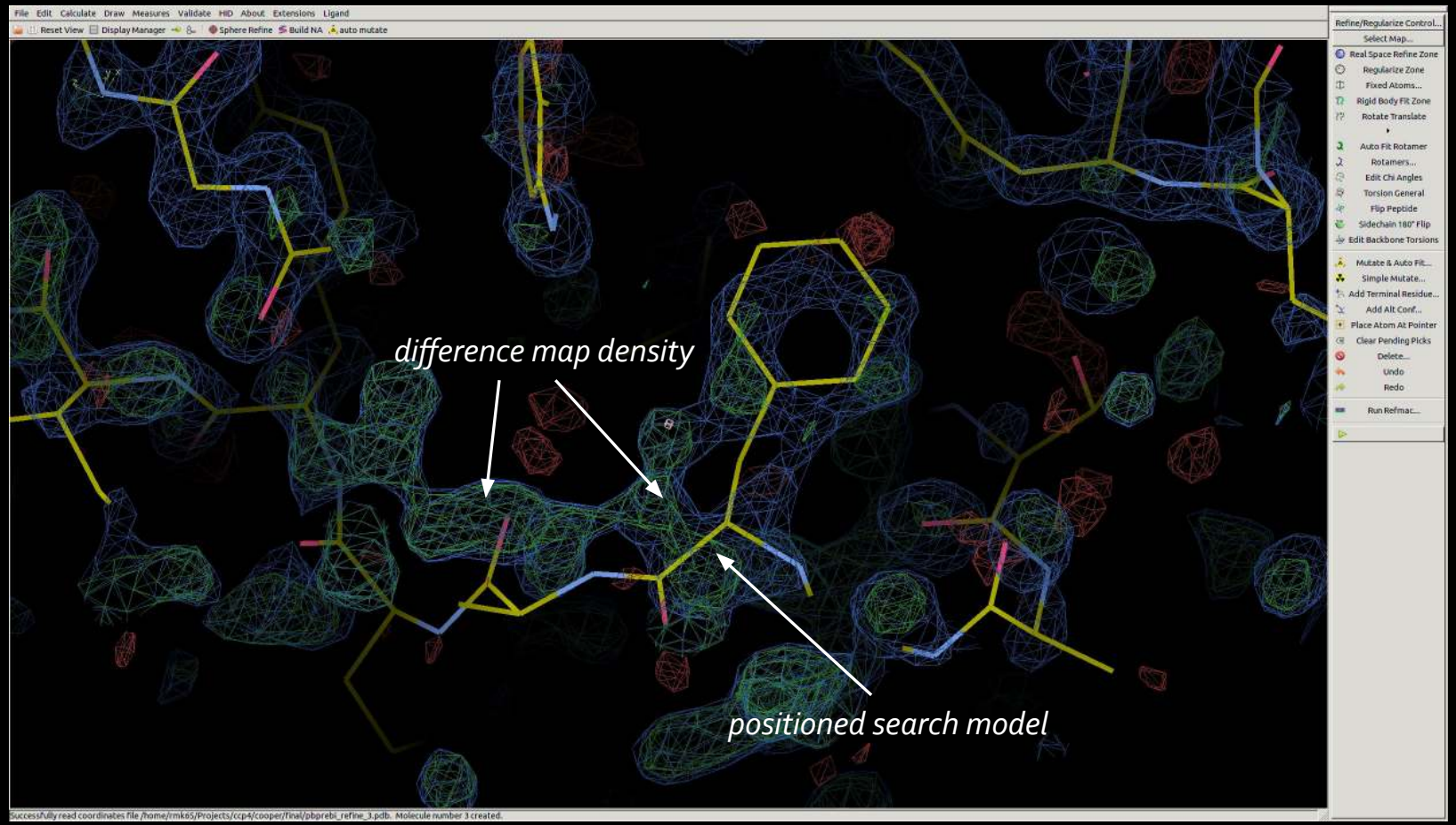




Example: 100 cycles of jelly-body refinement with Refmac
Rfactor = 0.35
Rfree = 0.32

CCP4

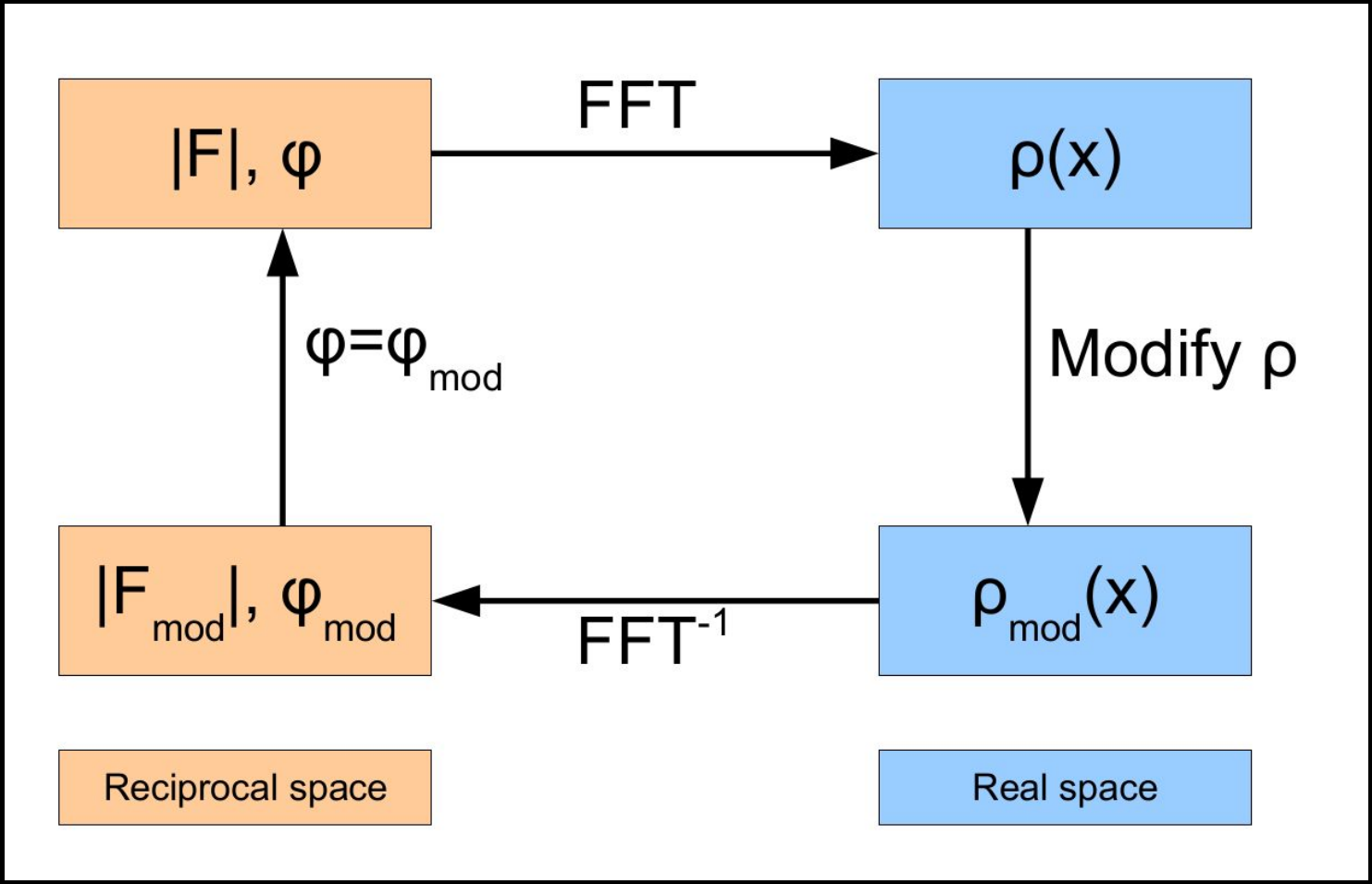
- Examine solution by eye
  - Use Coot to examine positioned models & maps



Density Modification and Model Building

CCP4

Density modification



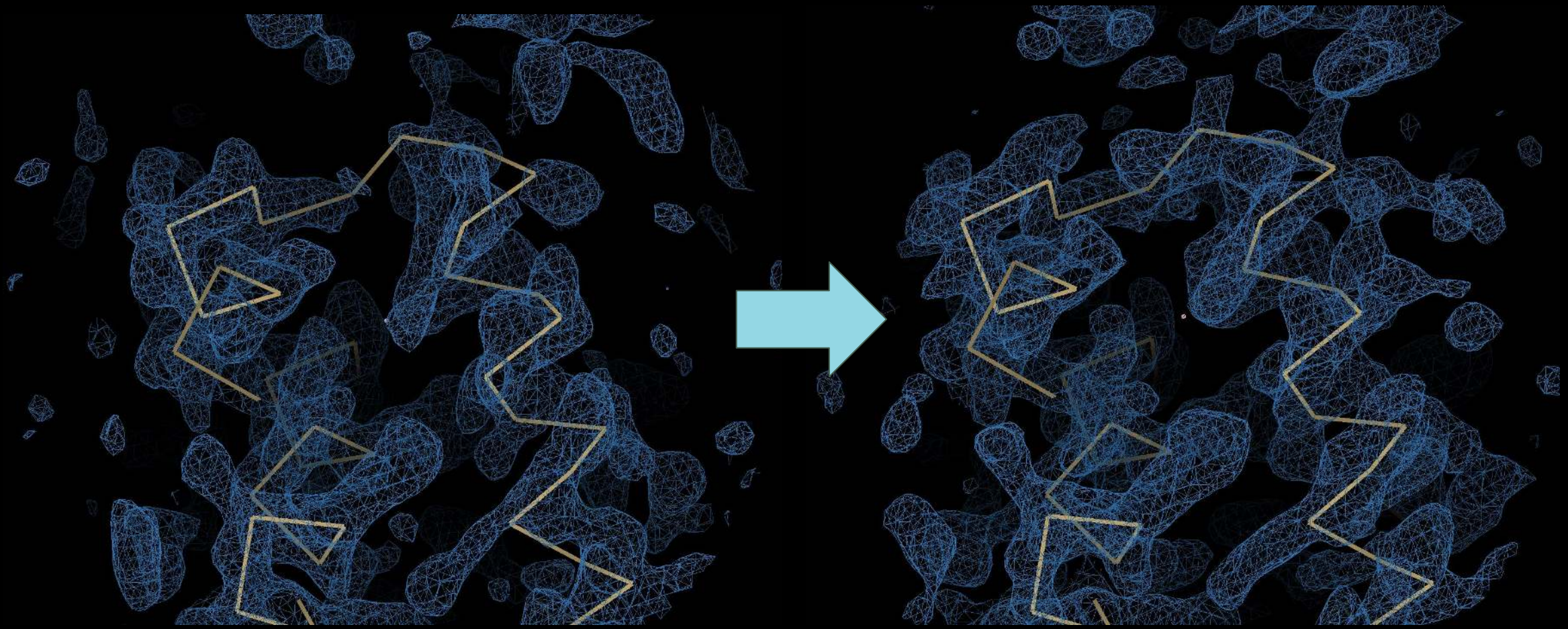
Main techniques:

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

(slide from Kevin Cowtan)

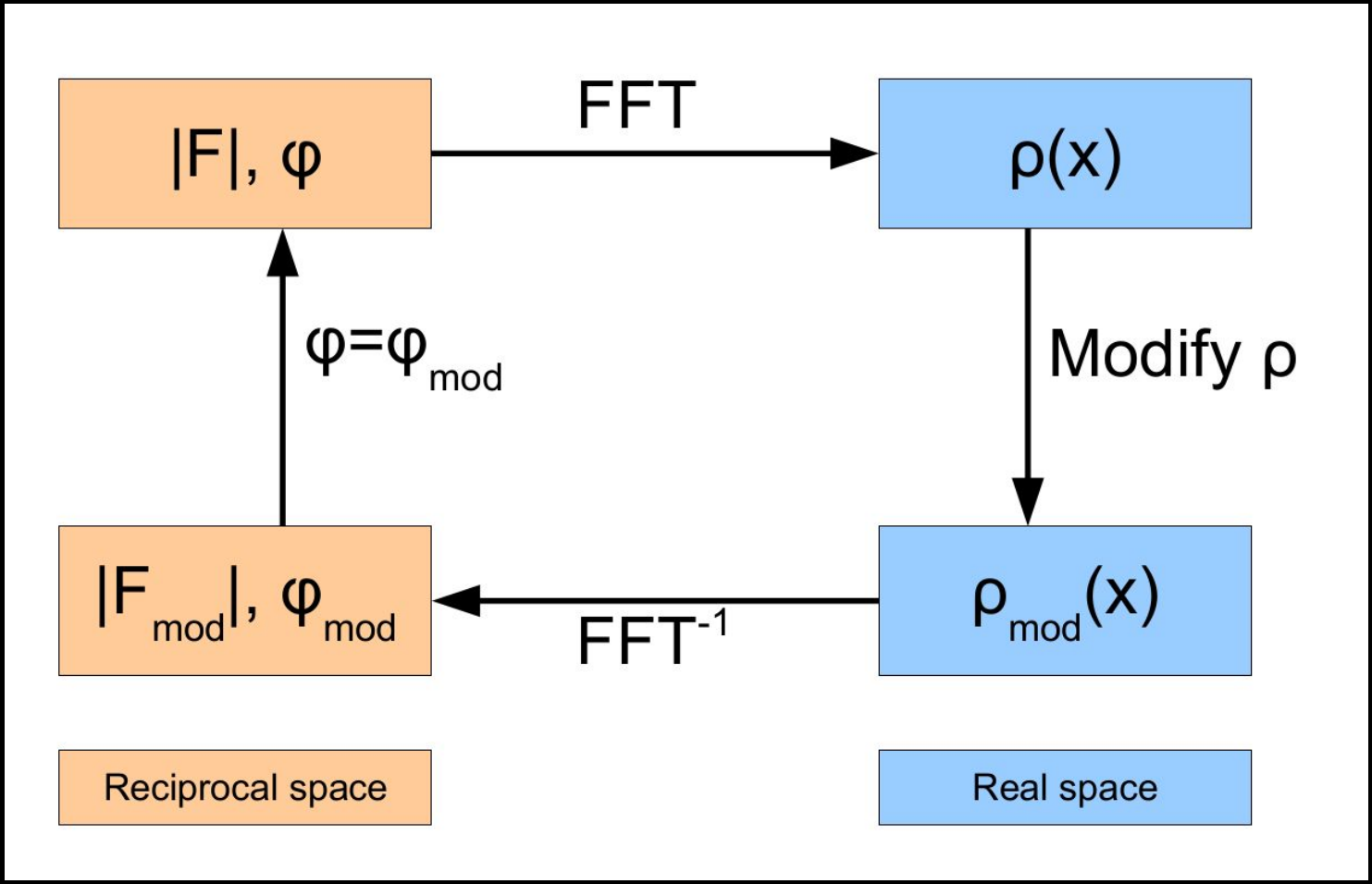
CCP4

Density modification



CCP4

Density modification



CCP4 Applications:

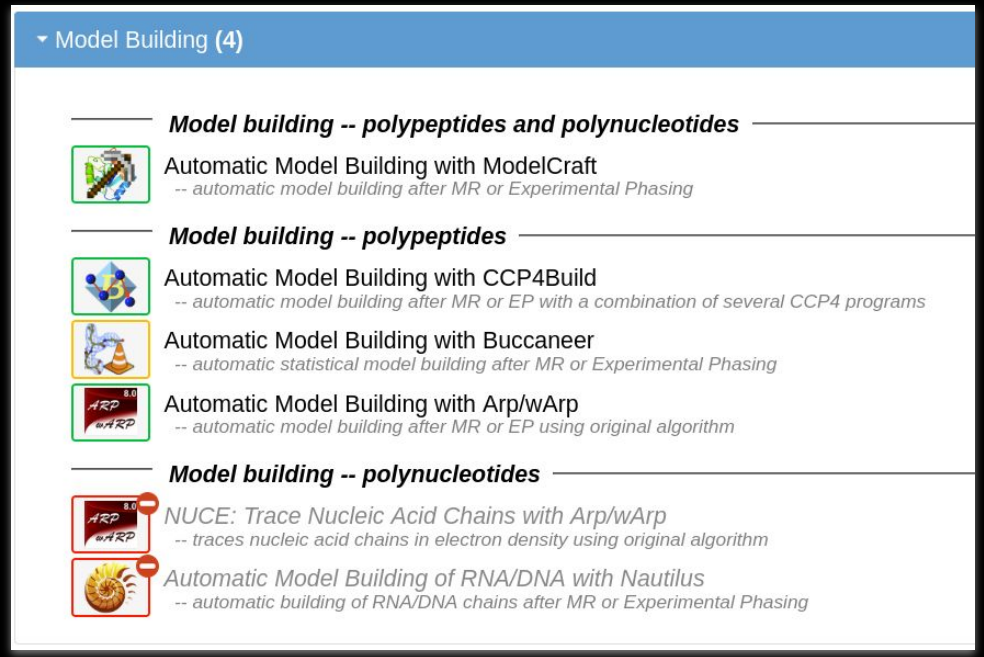
- Parrot
- SHELXE
- ACORN
- Pirate
- Solomon
- DM

(slide from Kevin Cowtan)

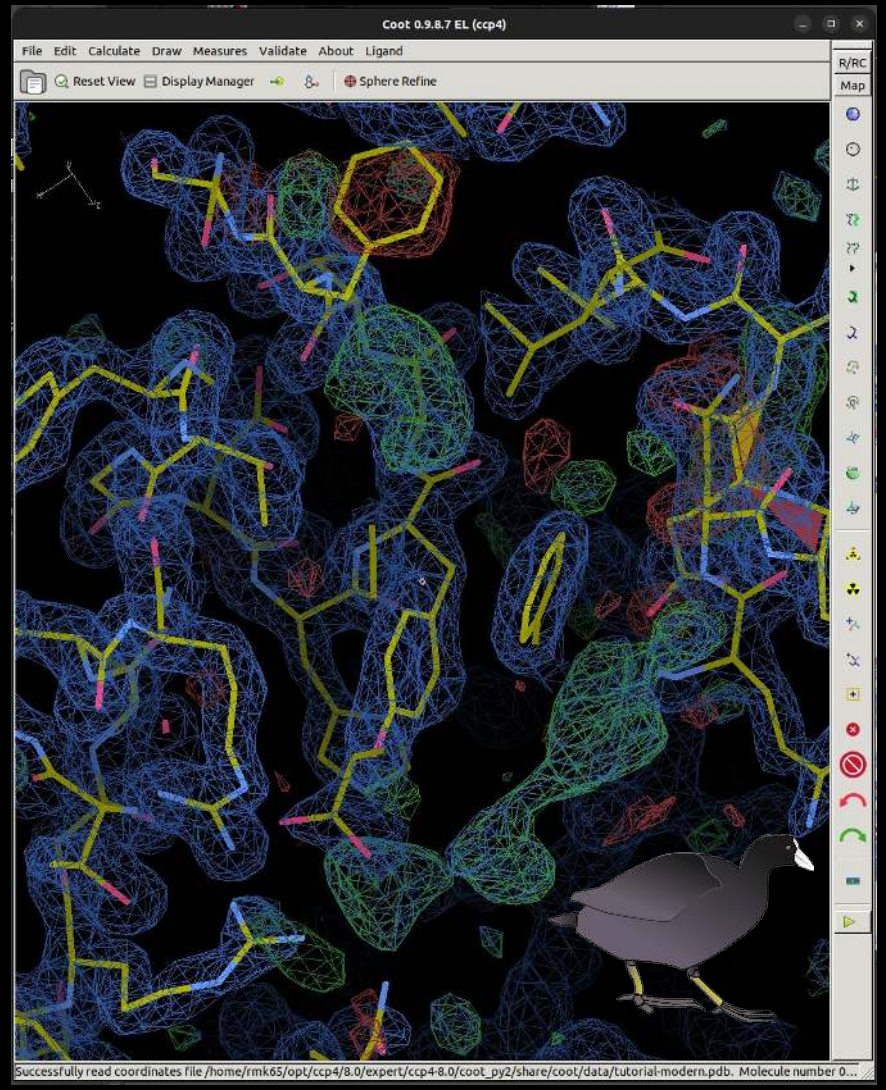
CCP4

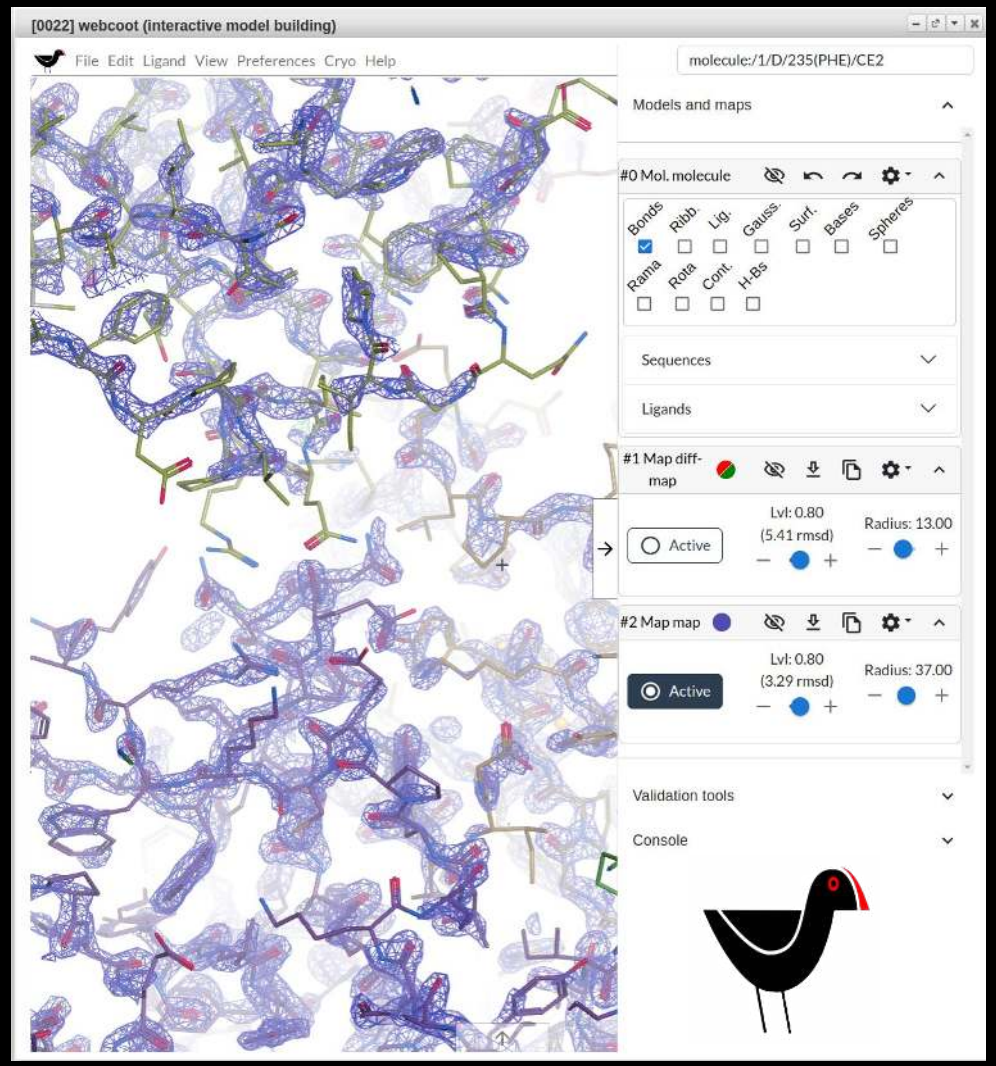
Model Building

- 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



Manual Model Building: Coot & Moorhen/WebCoot





CCP4

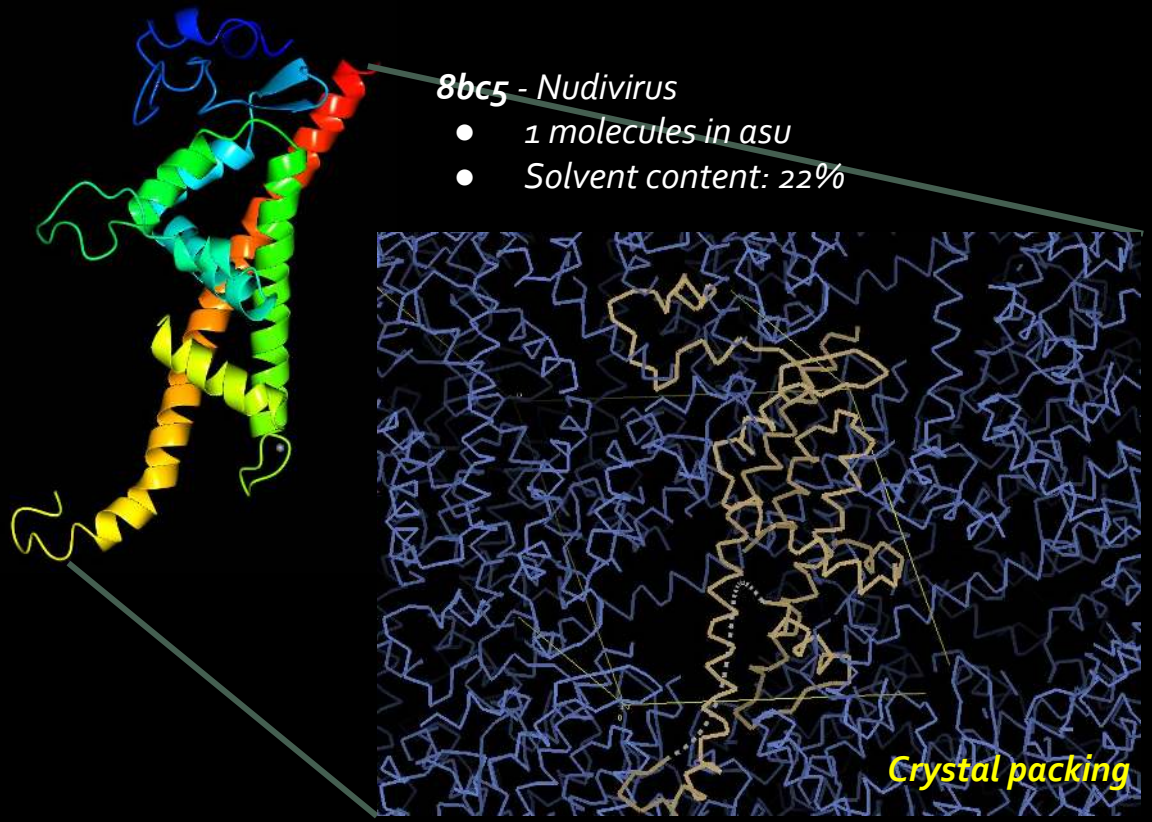
Common complications in Molecular Replacement

CCP4

Contents of the asymmetric unit

Contents of the asymmetric unit

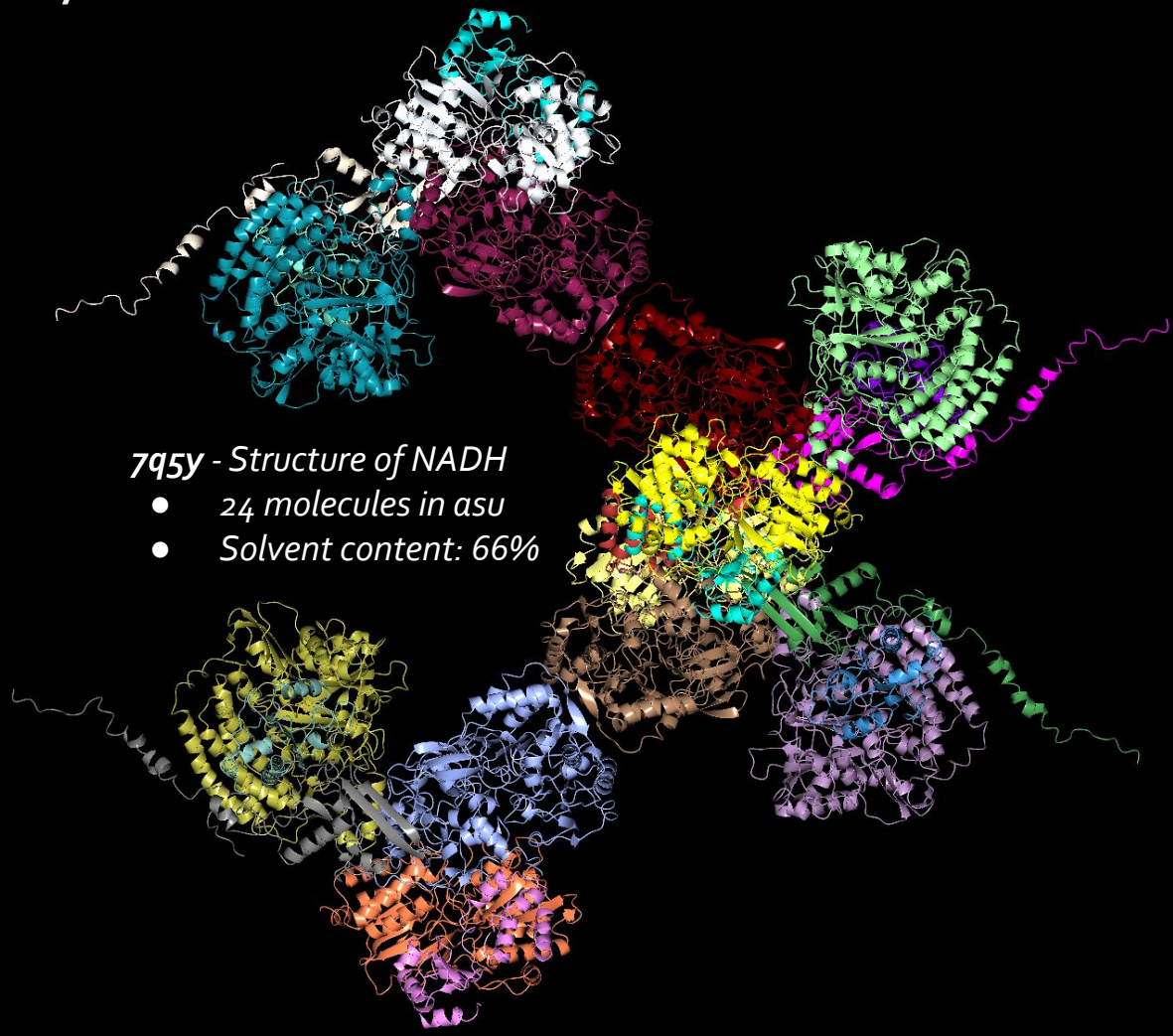
- How many copies/molecules in the asymmetric unit?



8bc5 *- Nudivirus*

- *1 molecules in asu*
- *Solvent content: 22%*

Crystal packing



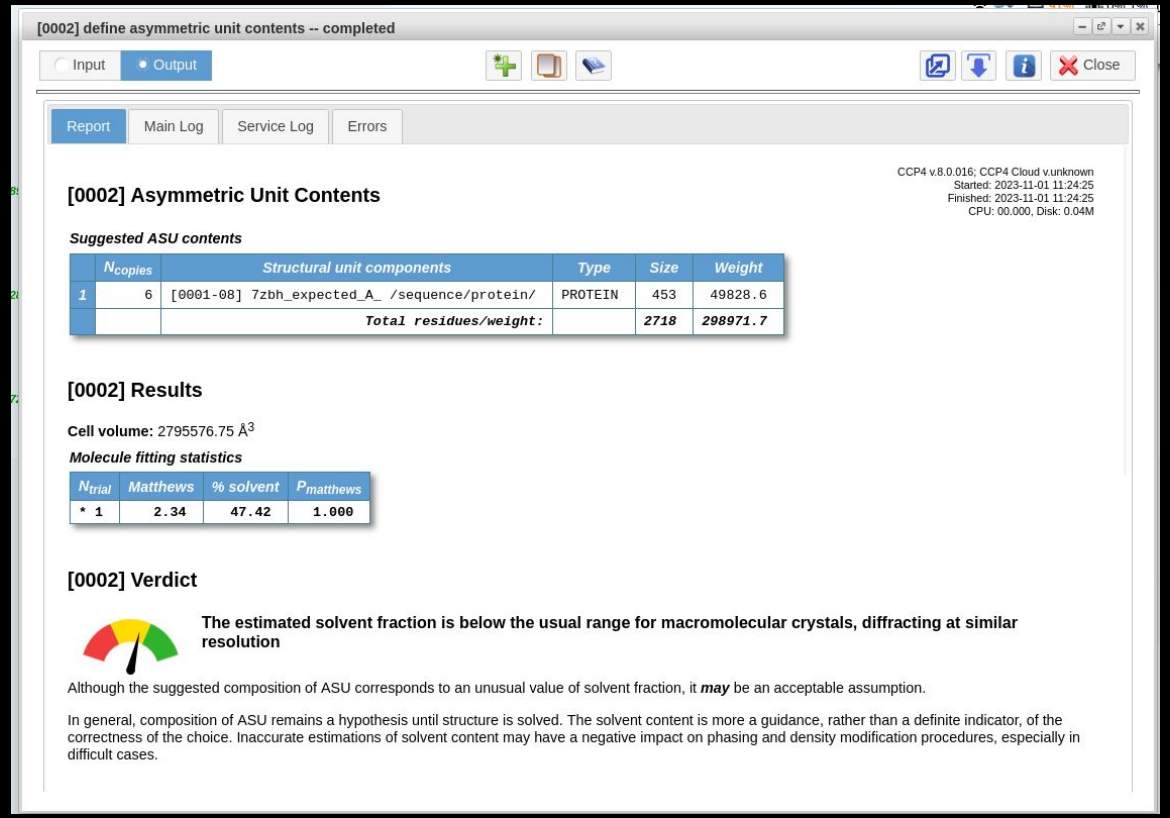
7q5y *- Structure of NADH*

- *24 molecules in asu*
- *Solvent content: 66%*

CCP4

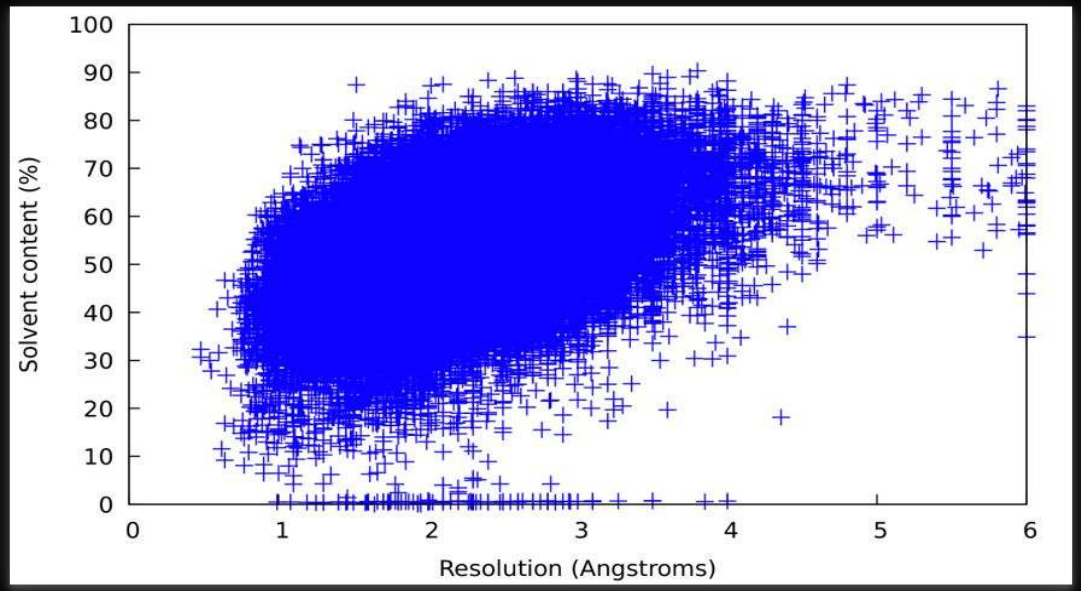
Contents of the asymmetric unit

- How many copies/molecules in the asymmetric unit?
- Estimate using Matthews Coefficient
  - Assumes roughly 50% of cell contents is solvent
  - Accounts for resolution



Contents of the asymmetric unit

- How many copies/molecules in the asymmetric unit?
- Estimate using Matthews Coefficient
  - Assumes roughly 50% of cell contents is solvent
  - Accounts for resolution

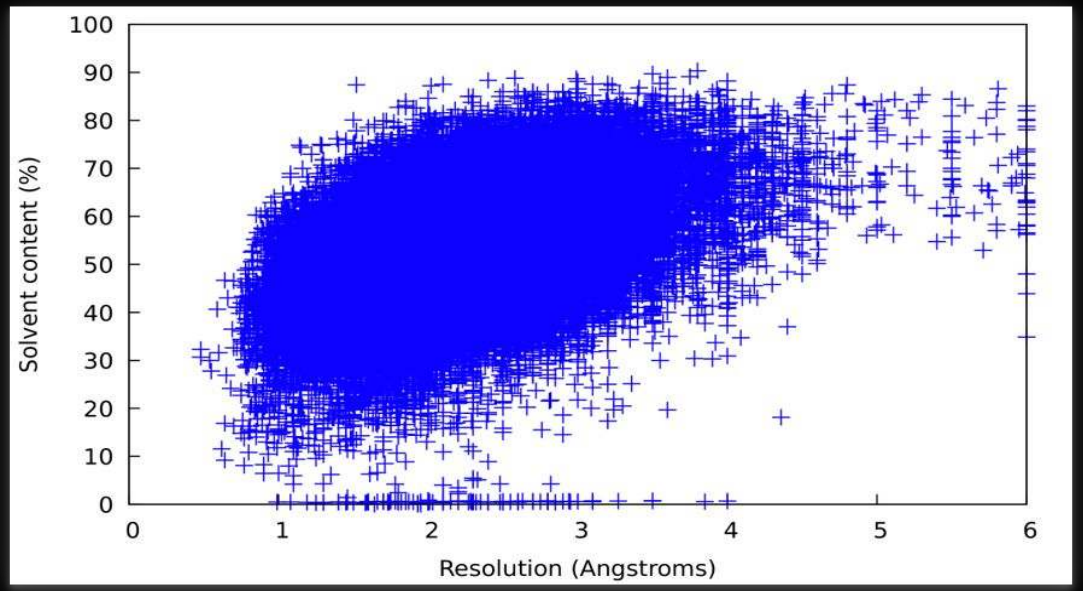


Distribution of solvent content across 156,000 X-ray datasets in the PDB

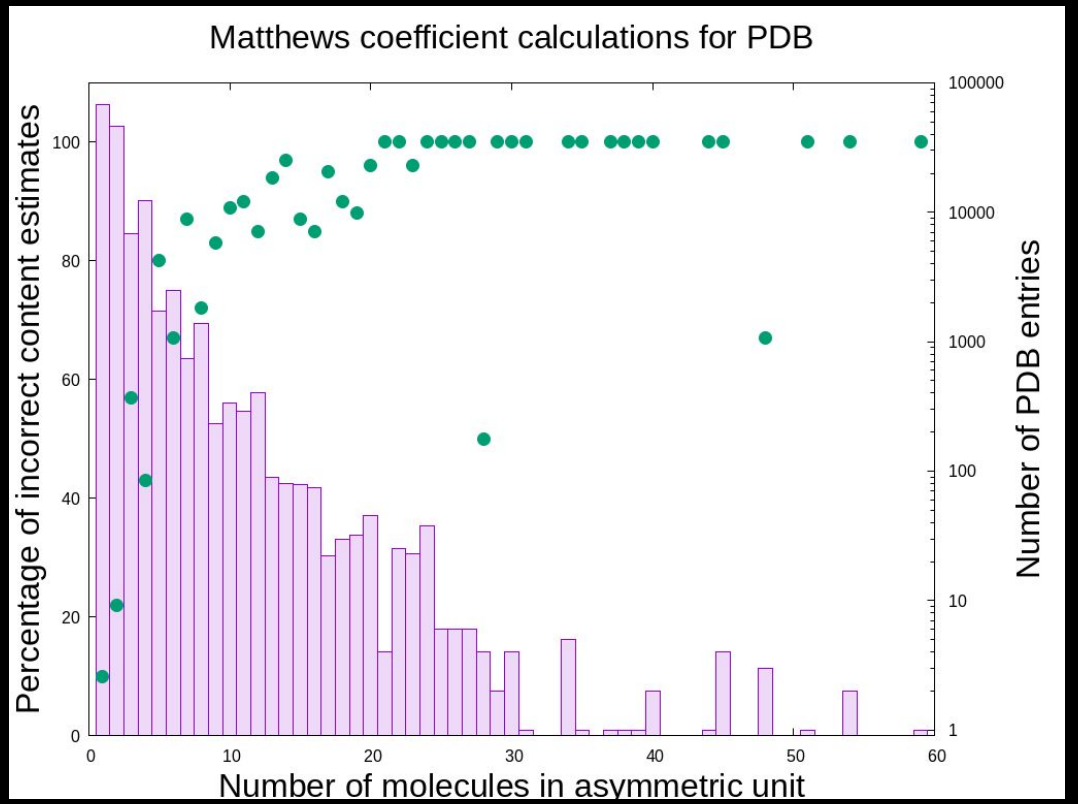
CCP4

Contents of the asymmetric unit

- How many copies/molecules in the asymmetric unit?
- Estimate using Matthews Coefficient
  - Assumes roughly 50% of cell contents is solvent
  - Accounts for resolution



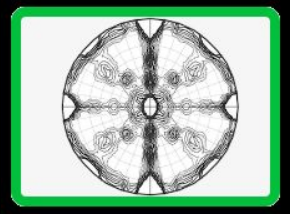
Distribution of solvent content across 156,000 X-ray datasets in the PDB

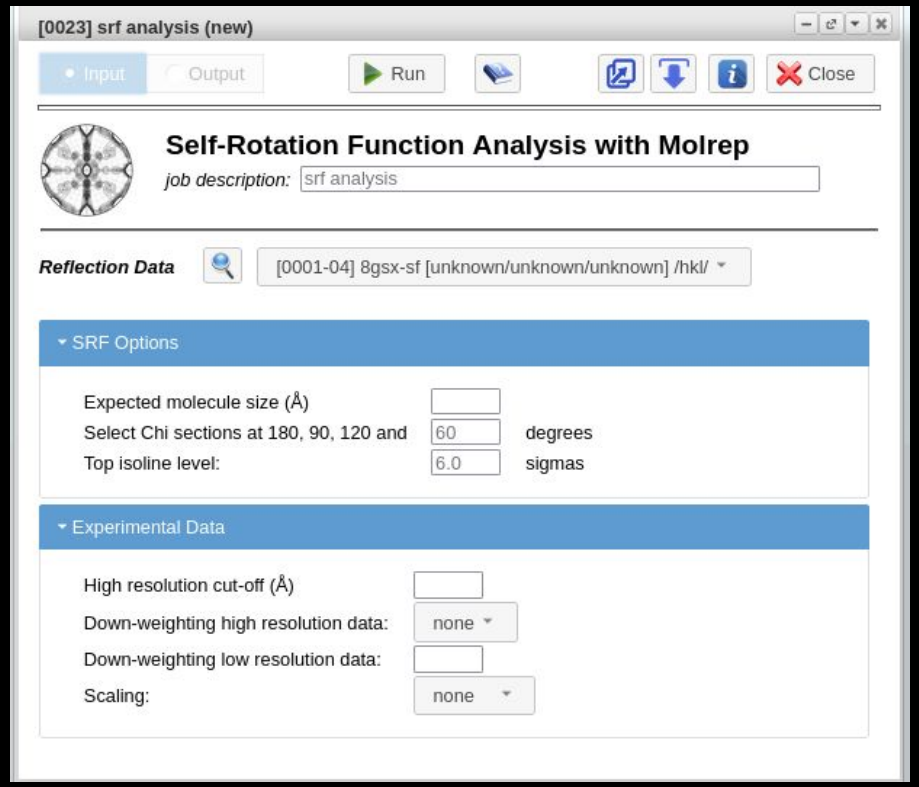


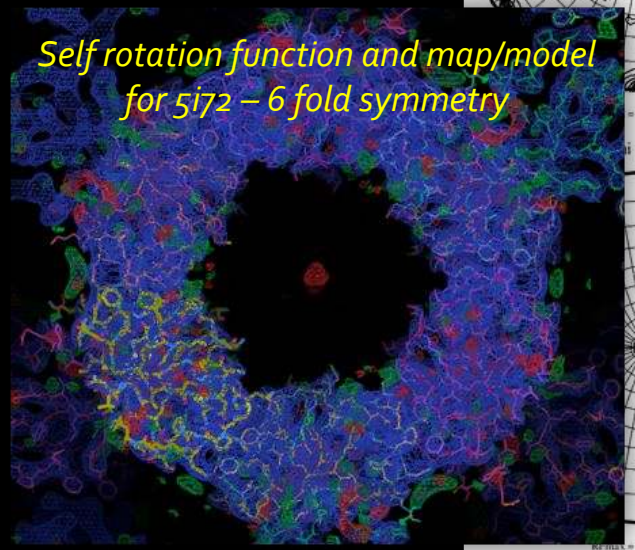
CCP4

Contents of the asymmetric unit

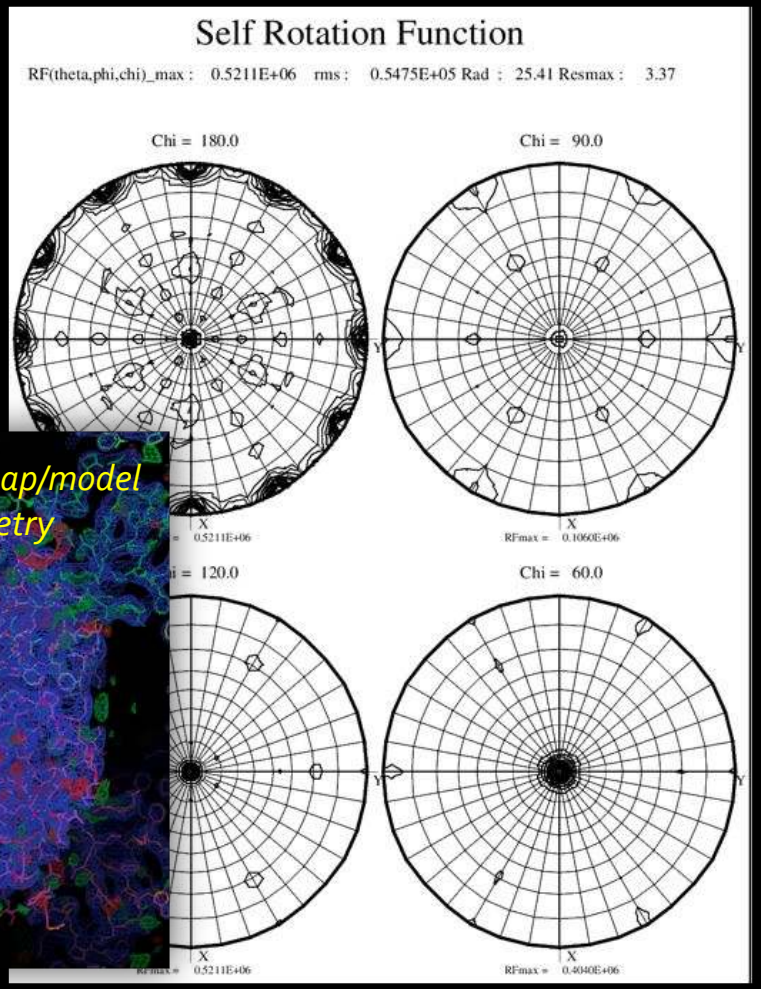
- Non-crystallographic symmetry (NCS)
- NCS copies can form dimer, trimer, etc..
- Evidence may be found in self rotation function







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



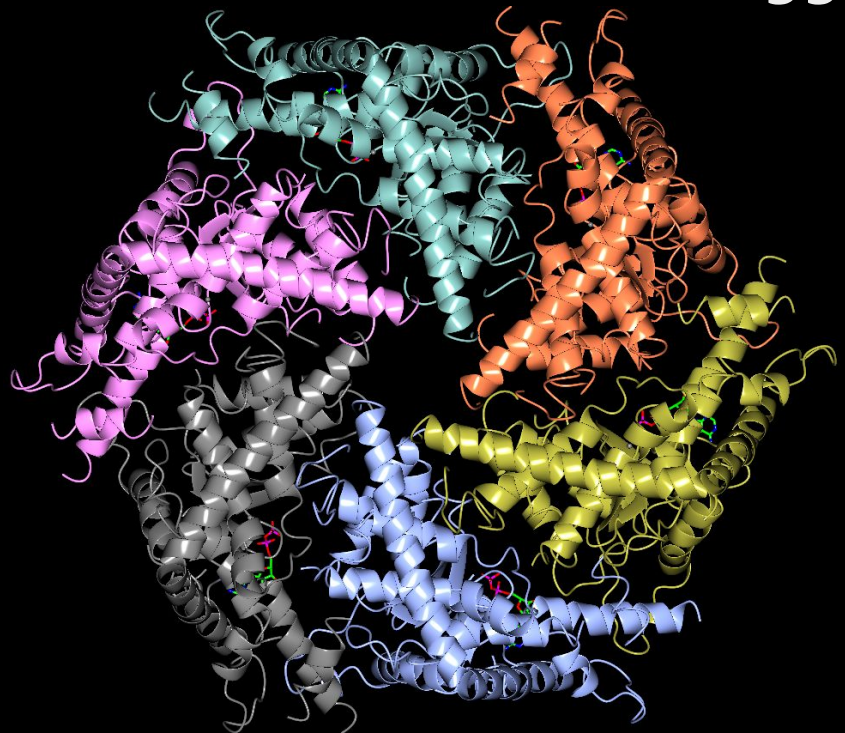
Contents of the asymmetric unit

- Complexes:
  - If there are “n” copies of the complex in the asu it is most likely that there will be “n” copies of each molecule in the complex
  - It is less common to have differing numbers of the the individual molecules
- Conformations:
  - Where there are multiple copies of a molecule in the asu it is most likely that each copy will have the same or a very similar conformational state to the other copies
  - In rare cases there may be different conformations present e.g. the open and closed form of a molecule
- Disorder in the data:
  - Where molecules protrude into solvent channels or with low resolution data, some molecules, or parts thereof, may be disordered

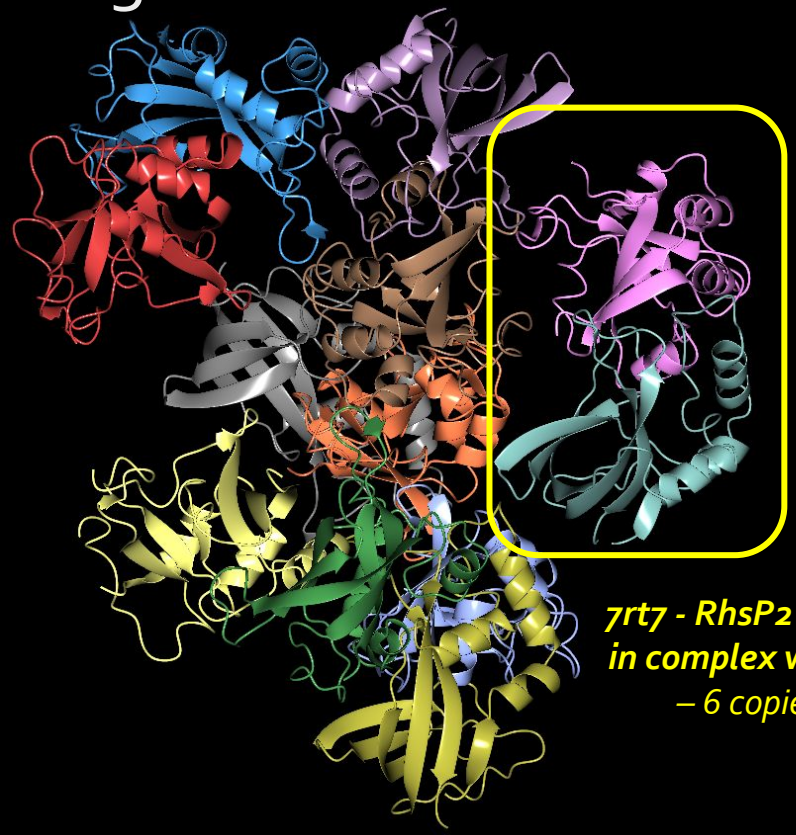
Multimers and Complexes

Multimers and Complexes

- Can present challenges for Molecular Replacement
- Molecular replacement is a signal-to-noise problem, the bigger the search model the bigger the signal



7zbh - ATP-dependent zinc metalloprotease
– 6 fold symmetry (hexamer)



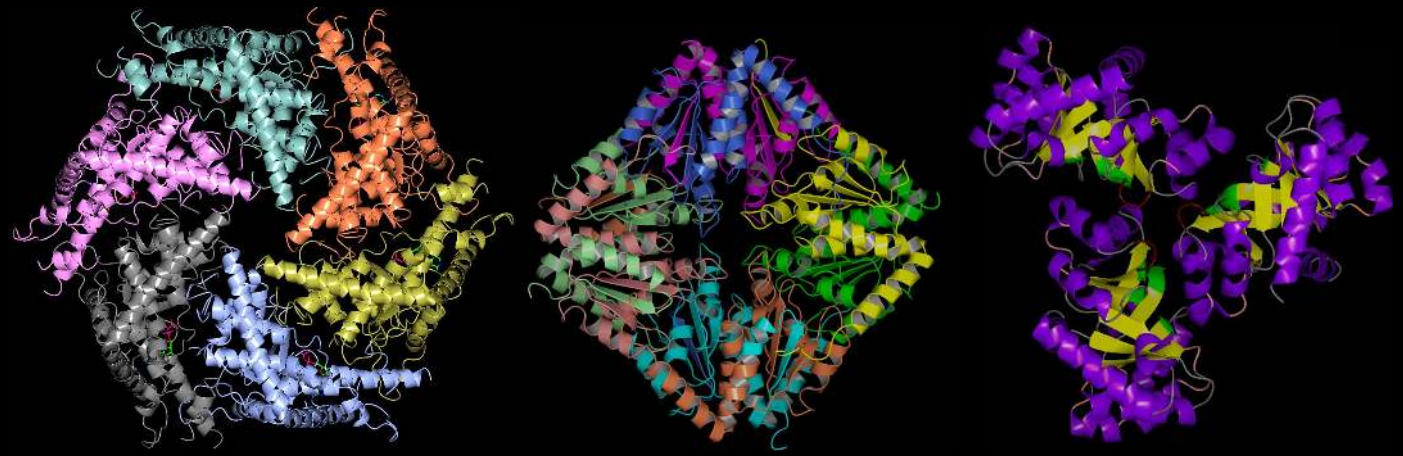
7rt7 - RhsP2 C-terminal toxin domain in complex with its immunity protein
– 6 copies of 2-chain complex

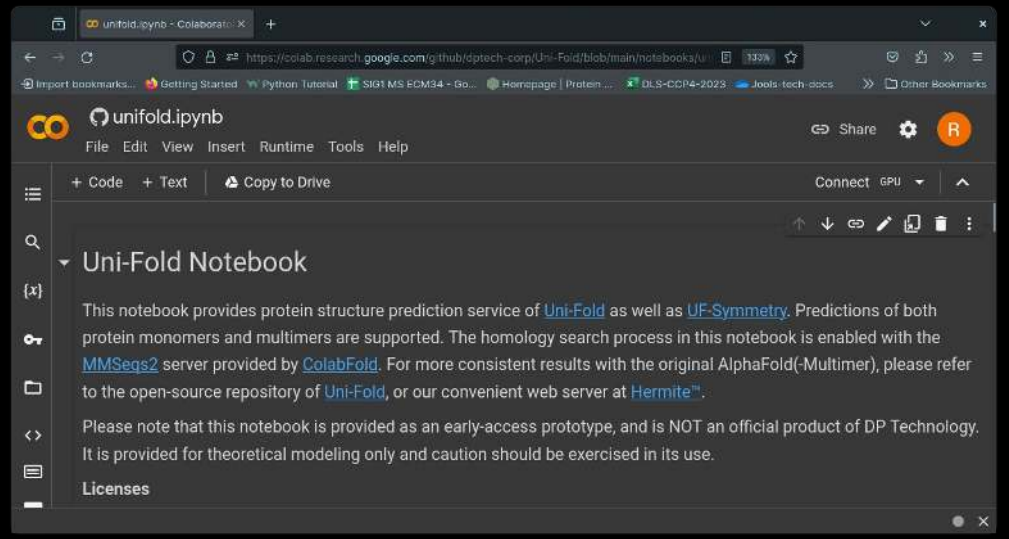
CCP4

Multimers and Complexes

Multimers:

- Multimeric forms of search models can improve the chances of success
  a. From related deposited structures which have multimeric forms
  b. Predicted multimer models from Colab servers:
     i. ***Alphafold2*** Deepmind Colab server
     ii. ***Unifold*** Colab server





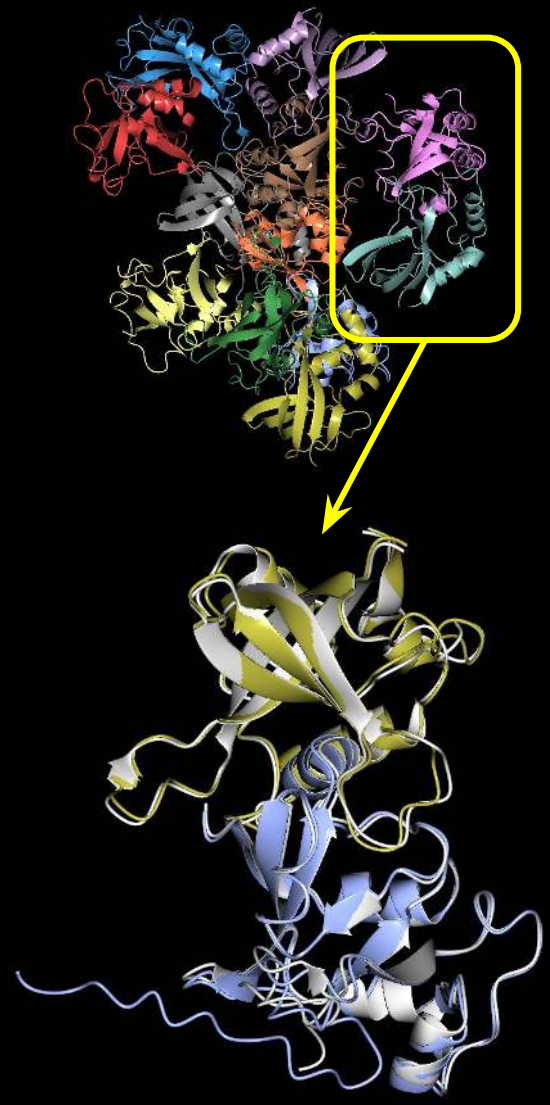
CCP4

Multimers and Complexes

Complexes:

- Complex forms of search models can also be generated by ***Alphafold2*** Deepmind Colab notebook

Larger search models allow for quicker and more clear-cut solutions in MR



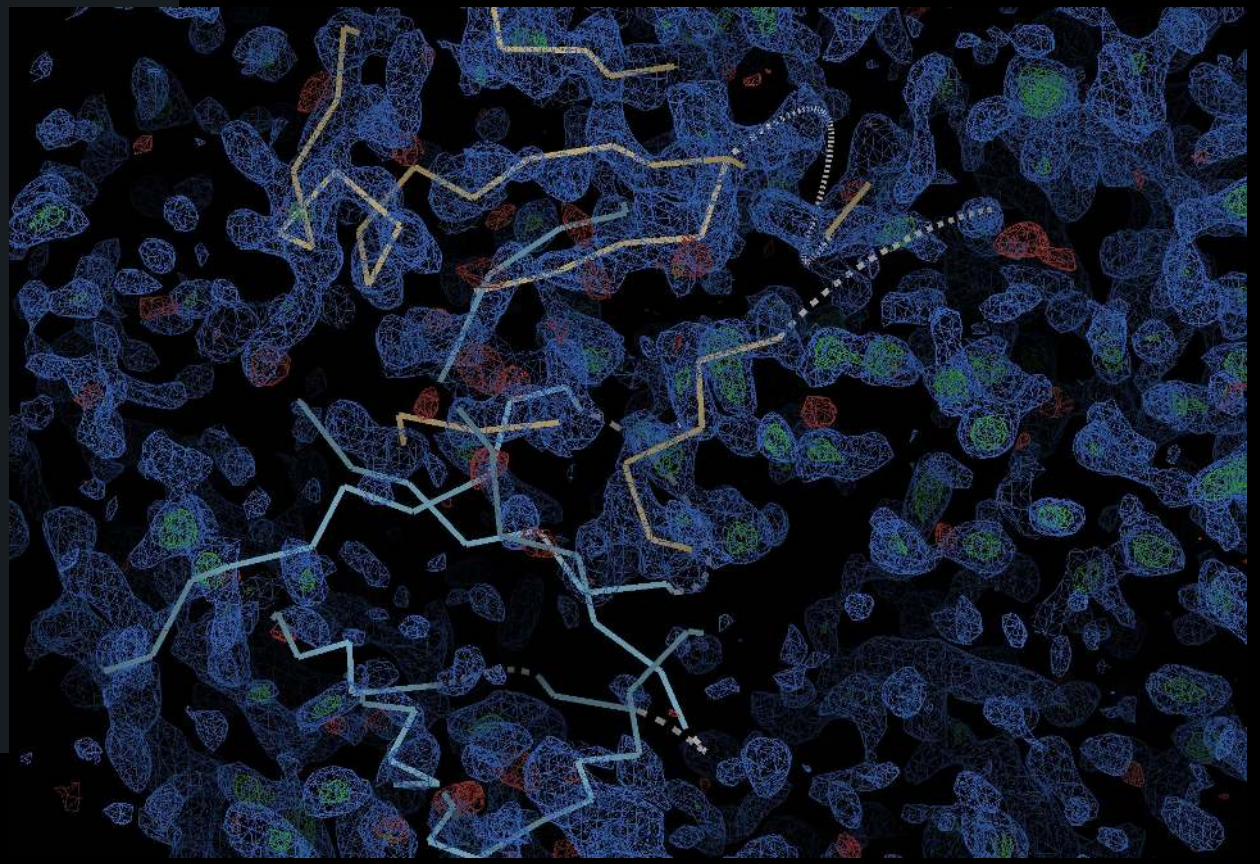
7rt7 - complex prediction from Alphafold2 aligned to experimental complex structure (white)

CCP4

Solving cases with many copies

- Check Phaser log file for low scoring placement
- Examine map and fit to density
- Remove poorly placed models, re-refine, and re-do MR with fixed model

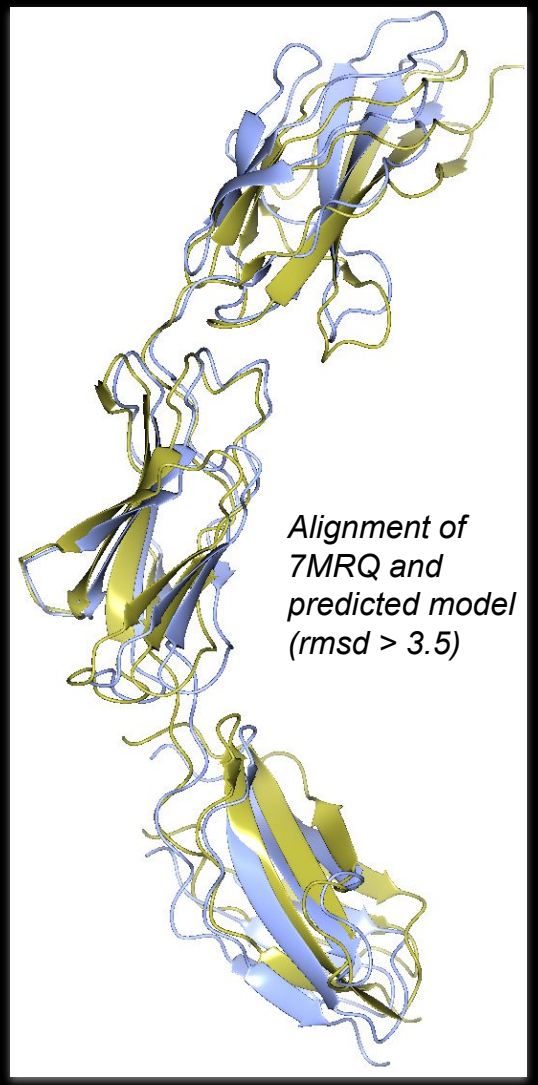
```
7115 ******************************************************************************
7116 *** Phaser Module: AUTOMATED MOLECULAR REPLACEMENT                    2.8.3 ***
7117 ******************************************************************************
7118
7119 ** SINGLE solution
7120
7121 ** Solution 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  236.5   31.8  133.9 FRAC  0.11 -0.28  0.74 BFAC 16.85
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
```

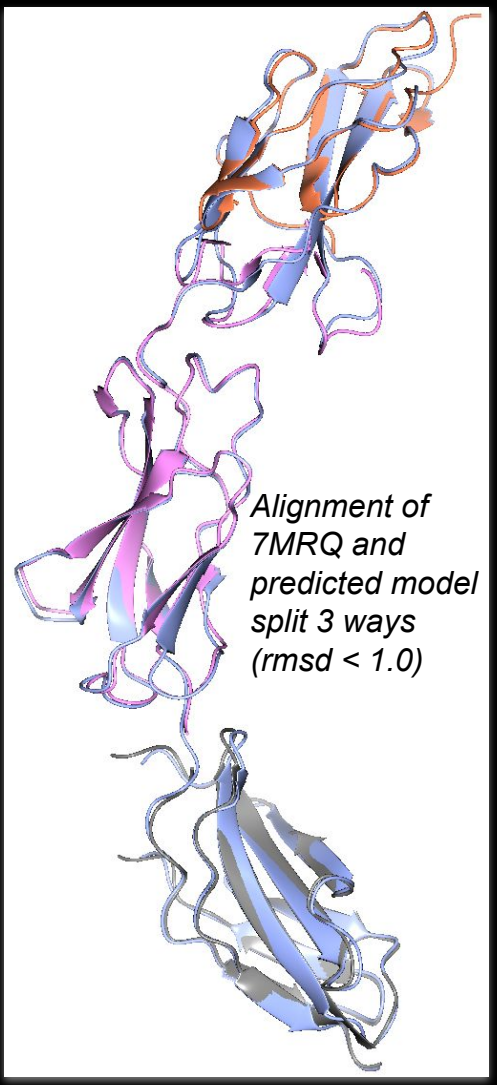


Multi-domain structures

Multi-domain structures

- The larger the target structure, the more likely there will be multiple domains
- A predicted model will often differ in the relative orientation of these domains when compared to the crystal structure
- Splitting a predicted search model into these domains can be a good strategy in MR



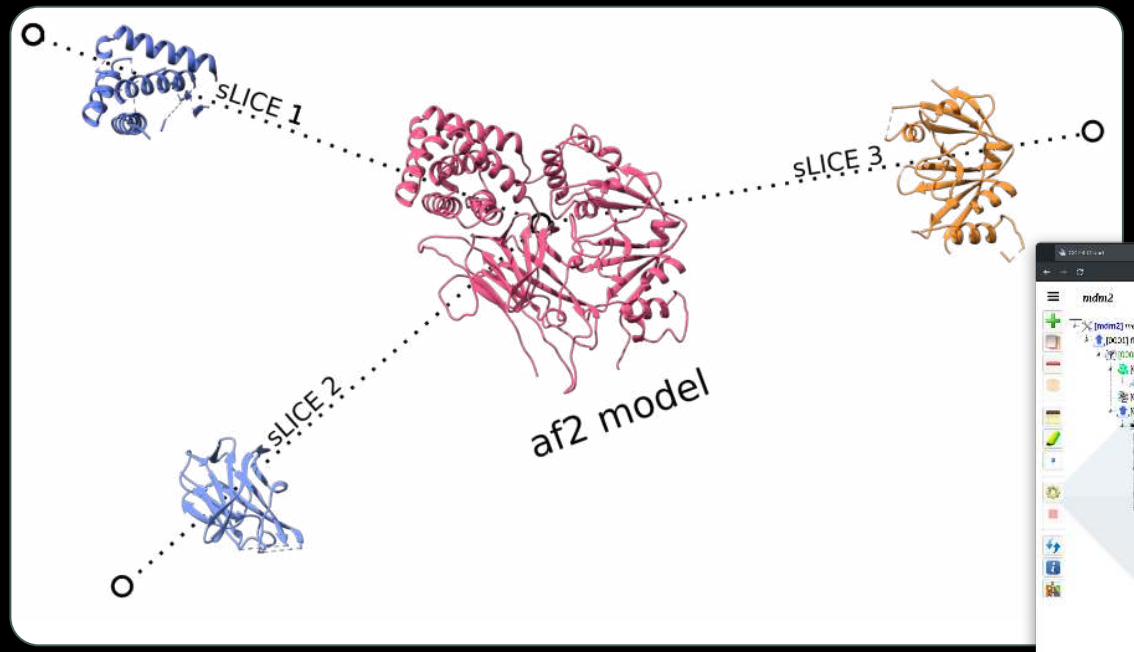


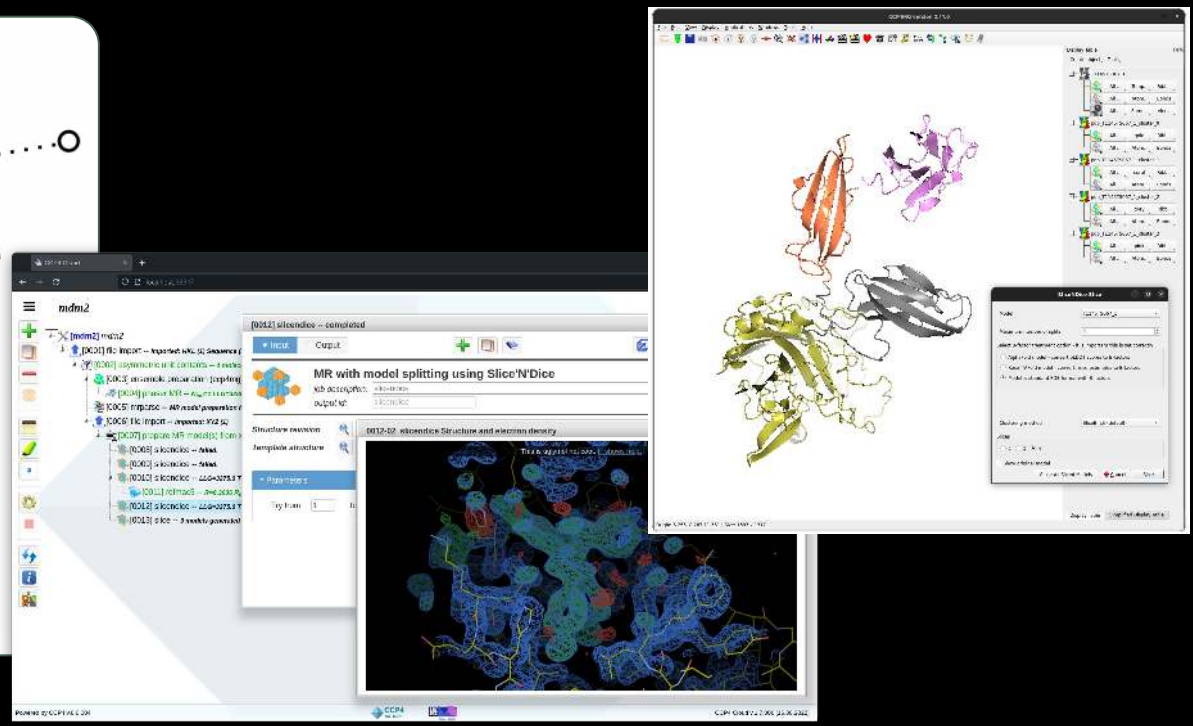
7mrq - Chicken CNTN4 FN1-FN3 domains with T751A, V752A, Y781A, E786A mutations

CCP4

Multi-domain structures

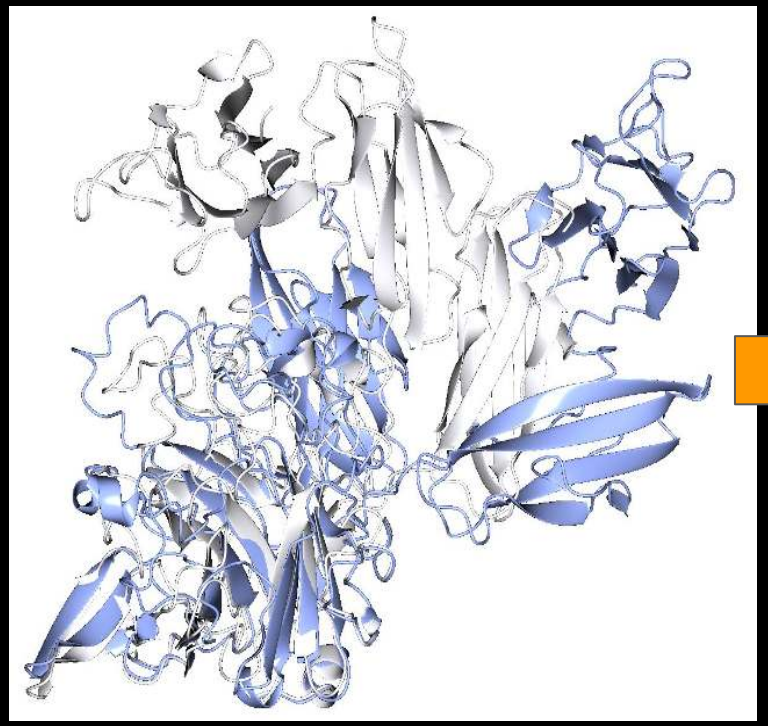
- Splitting search models can be done with the “slice” task in CCP4Cloud or using CCP4mg
- There is also a “slicendice” task which uses the split models in Molecular Replacement
- Uses atom clustering to decide where to split the domains





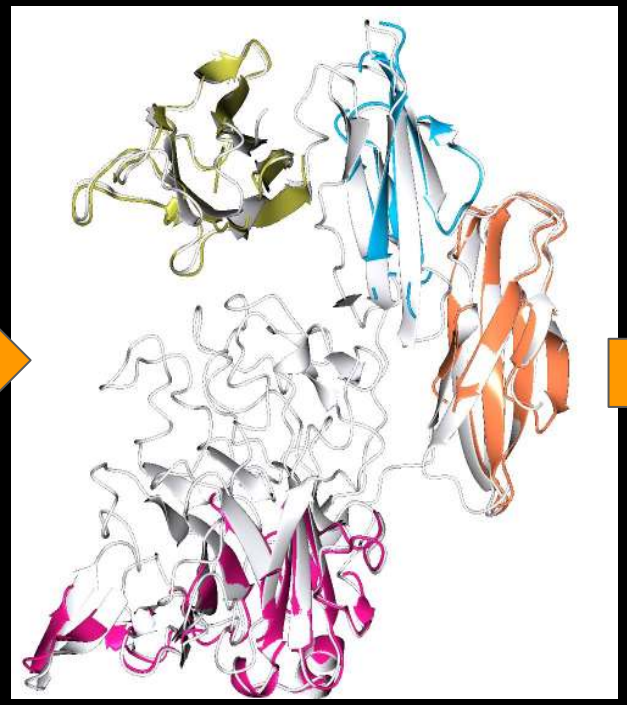
ESM-Fold Prediction

- Prediction (blue) aligned to target structure (white) on largest domain



Processing with *Slice'N'Dice*

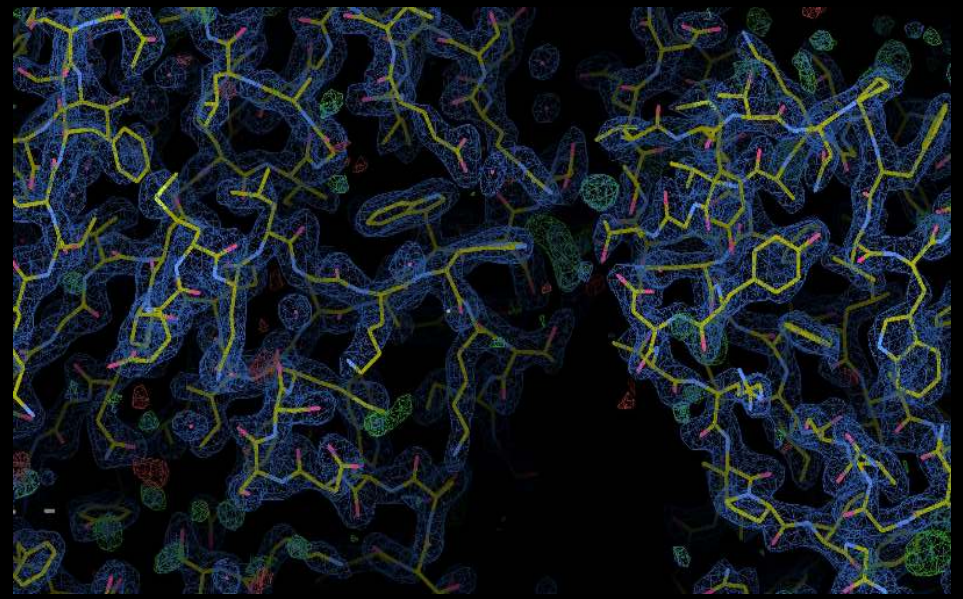
- 4-way Birch split of prediction
- Truncation to residues with a pLDDT > 70
- pLDDT converted to B-factor



(Domains aligned for comparison with target)

Automated X-ray structure solution

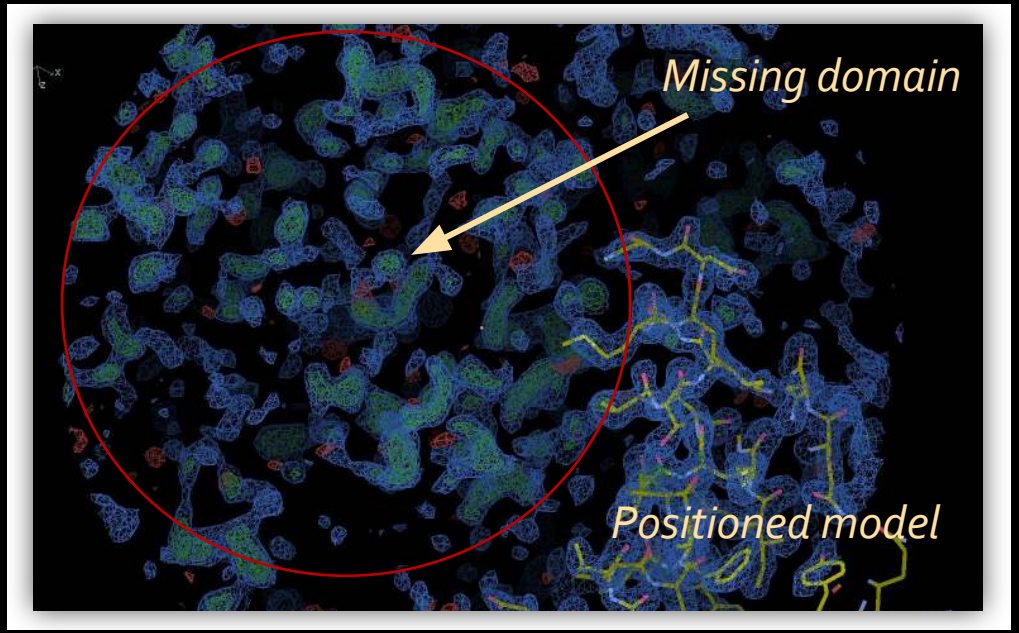
- *Slice'N'Dice* places 7 of the 8 domain models (2 copies of target in asymmetric unit of crystal) using *Phaser* (LLG=755 TFZ=16.1)
- Automated model building with *Modelcraft* brings model close to completion (R/Rfree 0.26/0.3)



2 copies (P212121, 2.2 Angstroms)

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

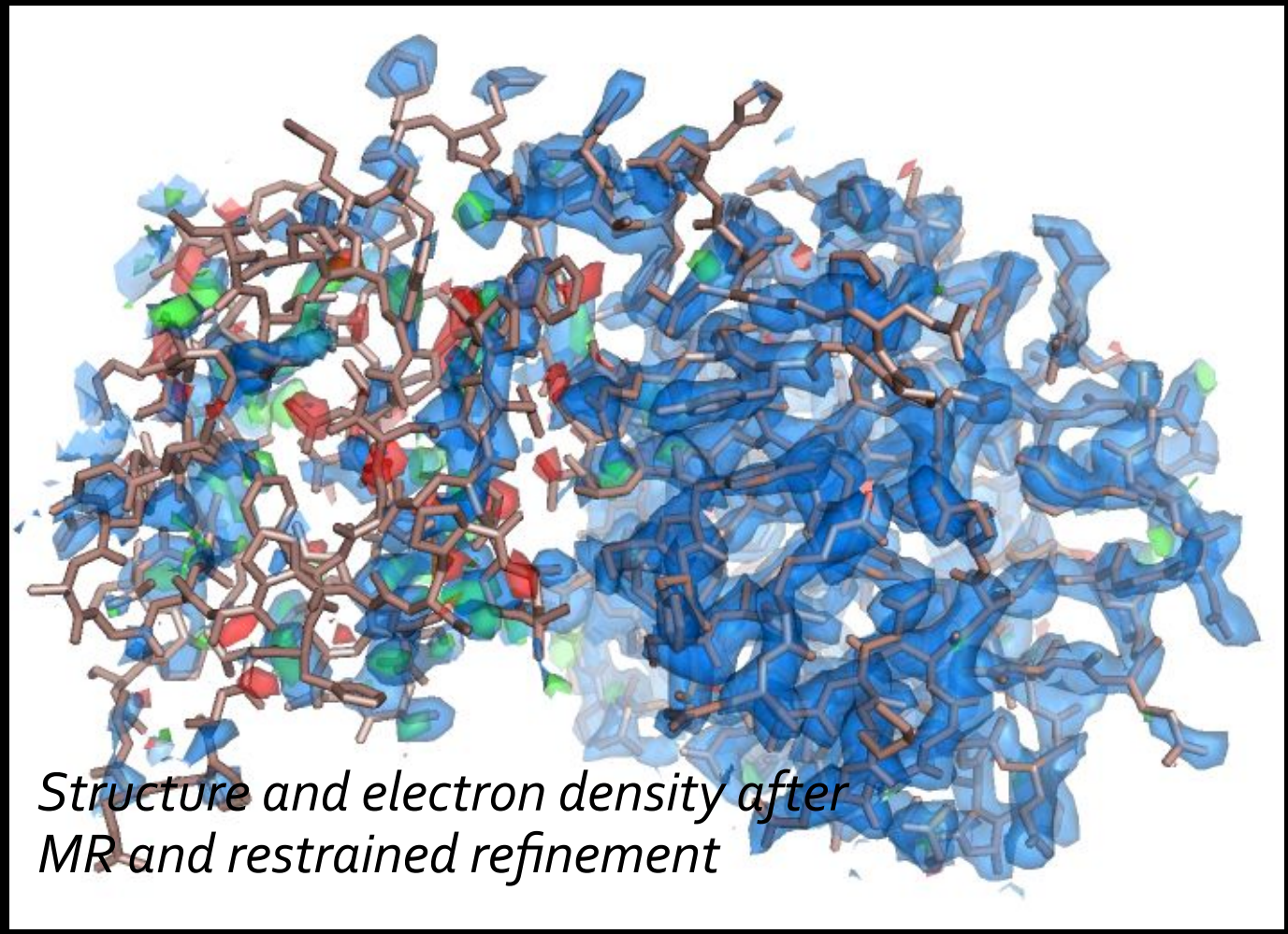
CCP4

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

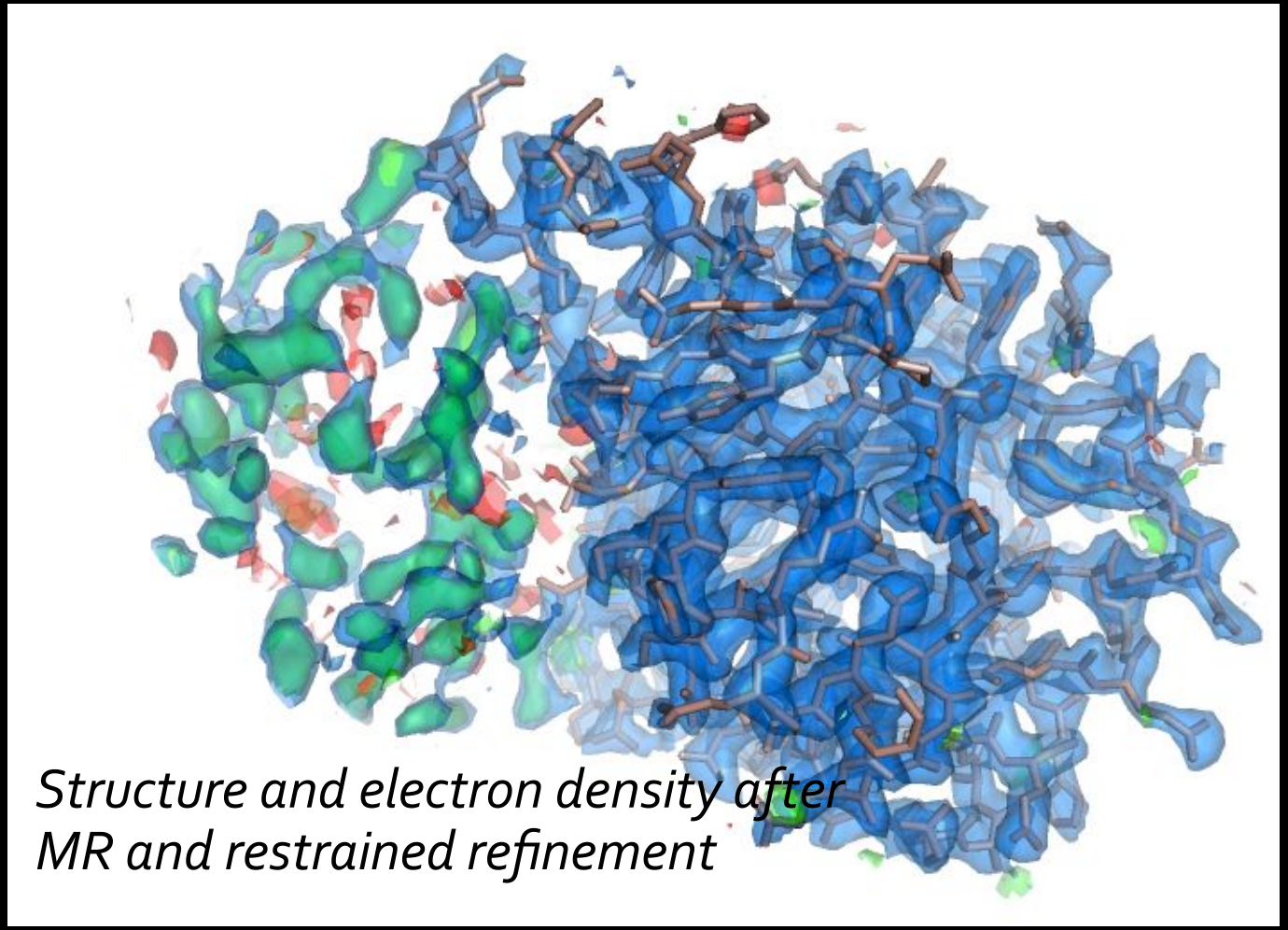
Phased Translation search and model Splitting

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***

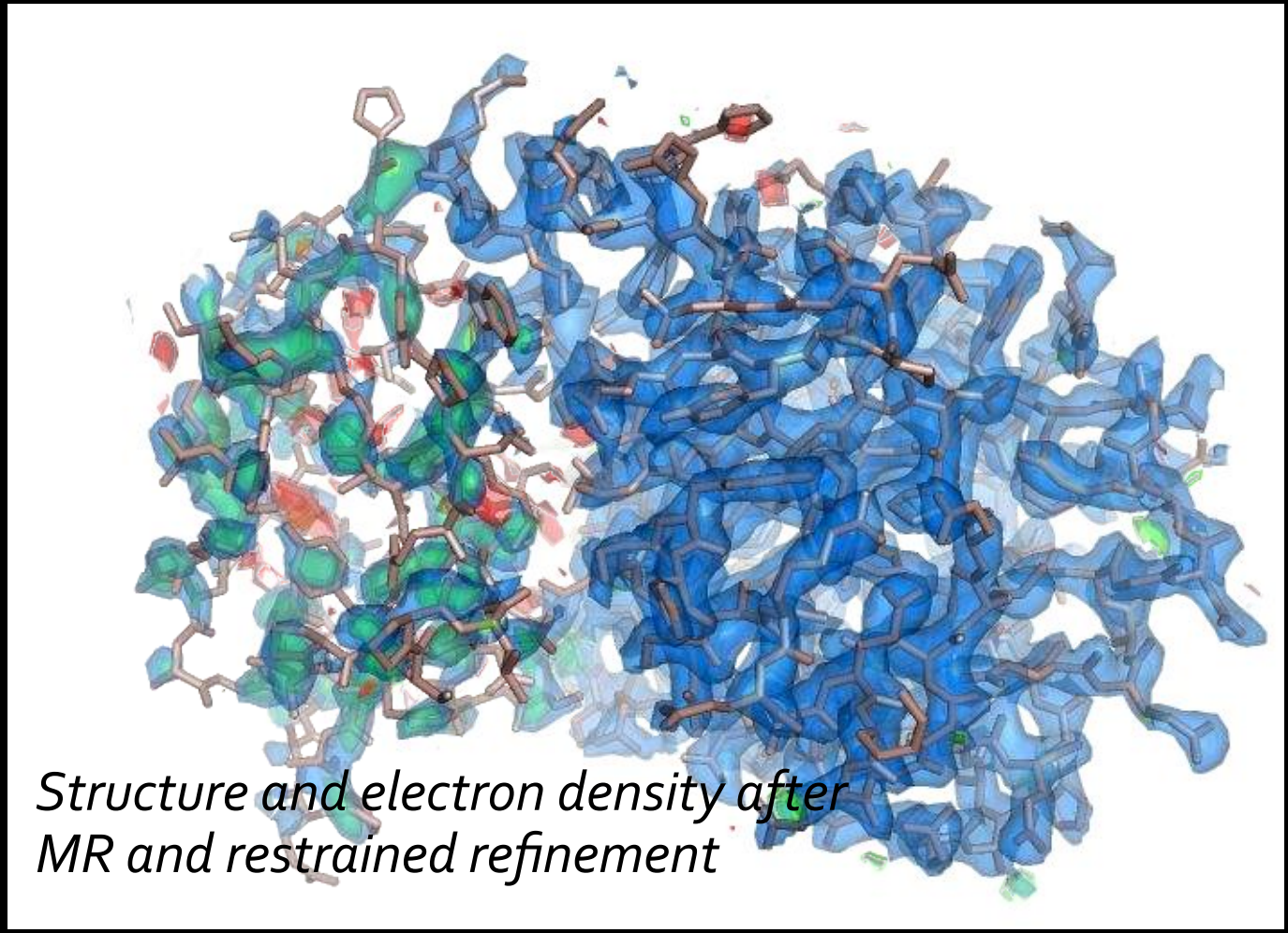
Search model: 1s20, chain A

Large Domain fixed

Small Domain search

Fit into density

(Molrep & Phaser)



Structure and electron density after MR and restrained refinement

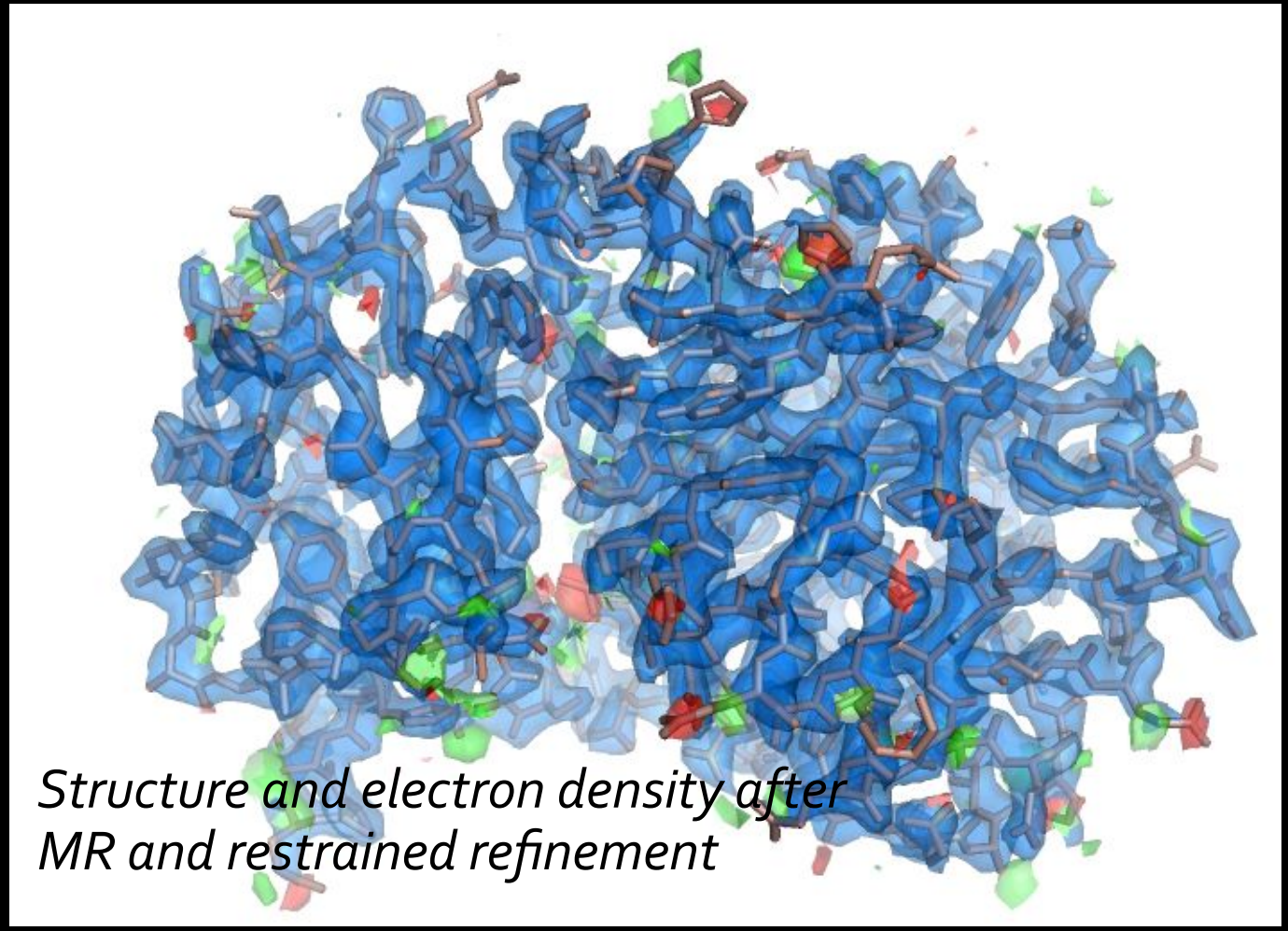
CCP4

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

Low confidence or inaccurate predicted models

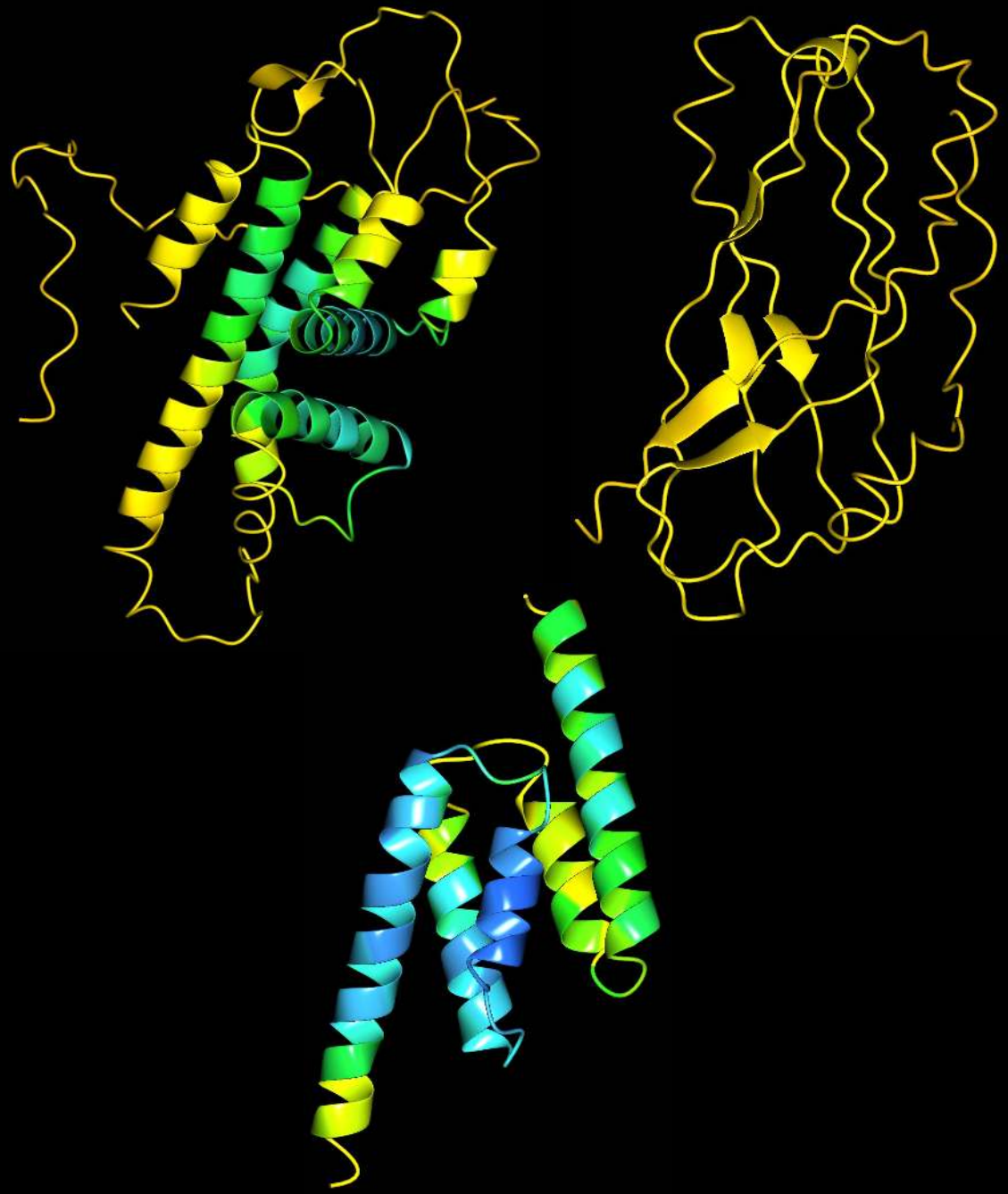
Predicted model scoring:

- Alphafold2 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

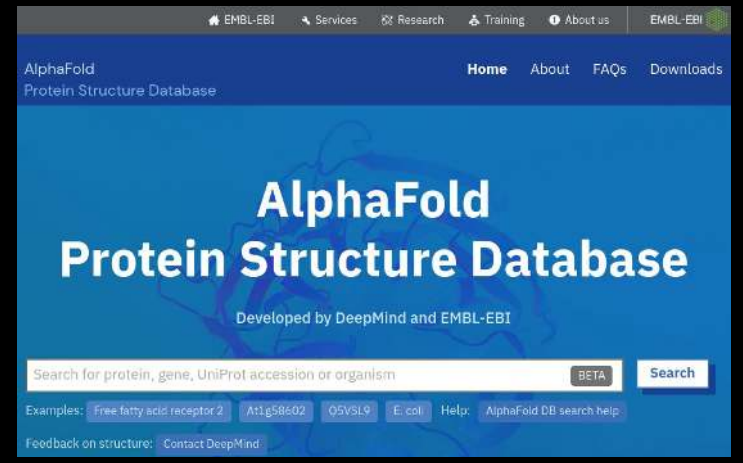
- Poor predictions will have low confidence scores (pLDDT<50) across much of the structure
- Low confidence predictions are nearly always poor MR search models
- High confidence predictions can sometimes differ significantly from the crystallised structure

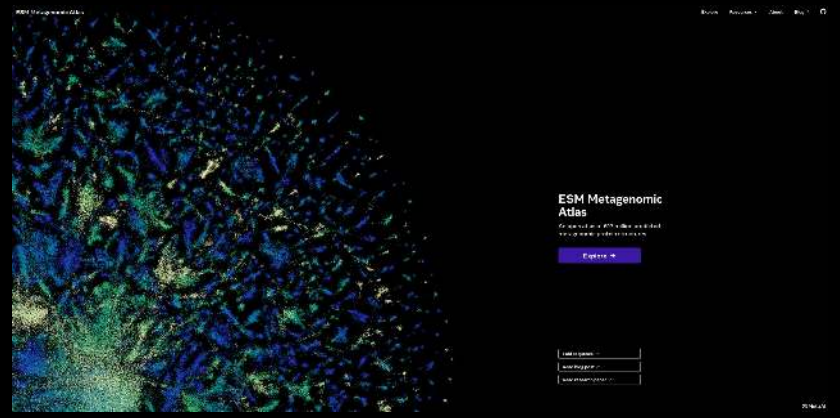


CCP4

What to do in these cases?

- Experimental phasing
- Alternative prediction methods:
  - *RoseTTAFold*
  - *ESMFold*
  - Give a template model to the prediction server (*Colabfold*)
- Look for possible homologues in the PDB or AFDB
- Use alternative Molecular Replacement tools



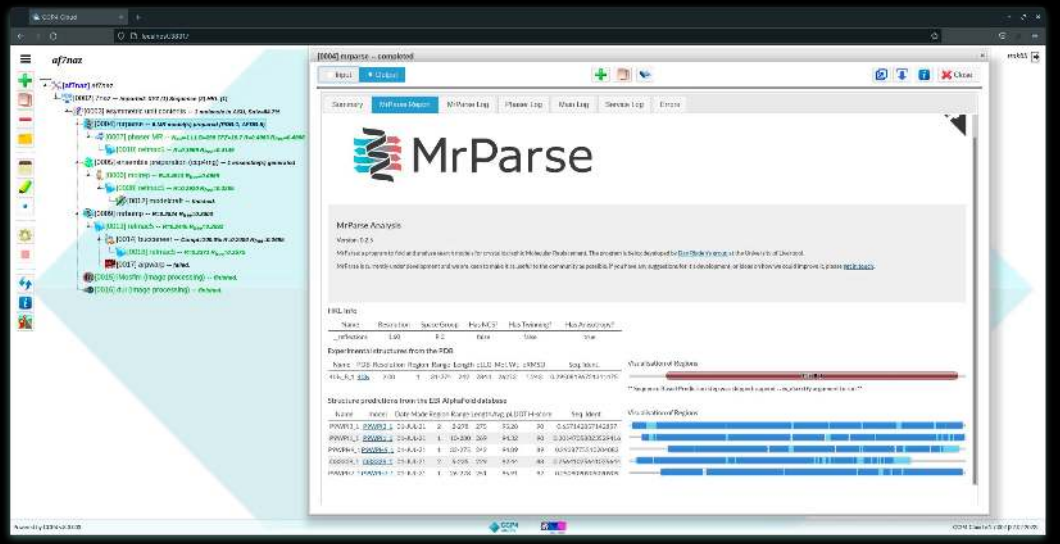


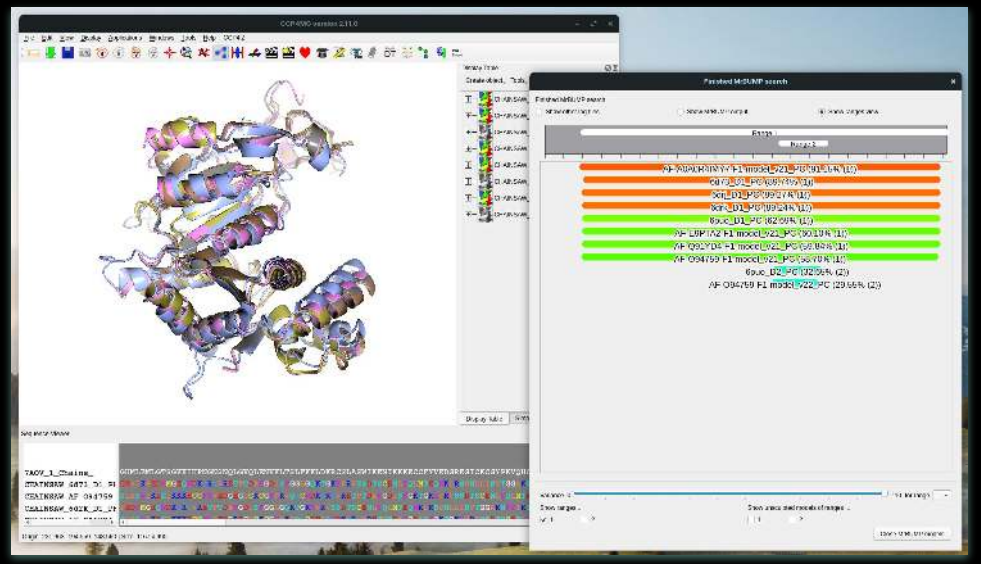
CCP4

Searching databases for search models

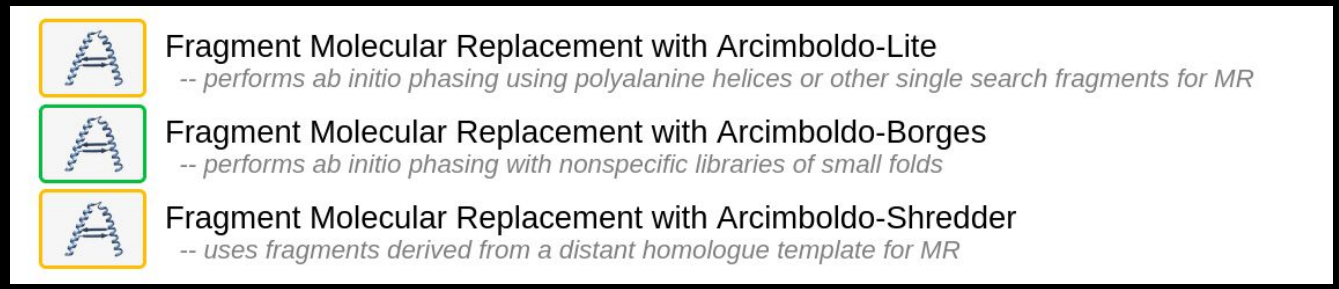
- *MrParse* can search the PDB (~200K sequences) and the Alphafold database (215 million)

- Search can also be performed in *CCP4mg* with hits aligned and displayed graphically





ARCIMBOLDO

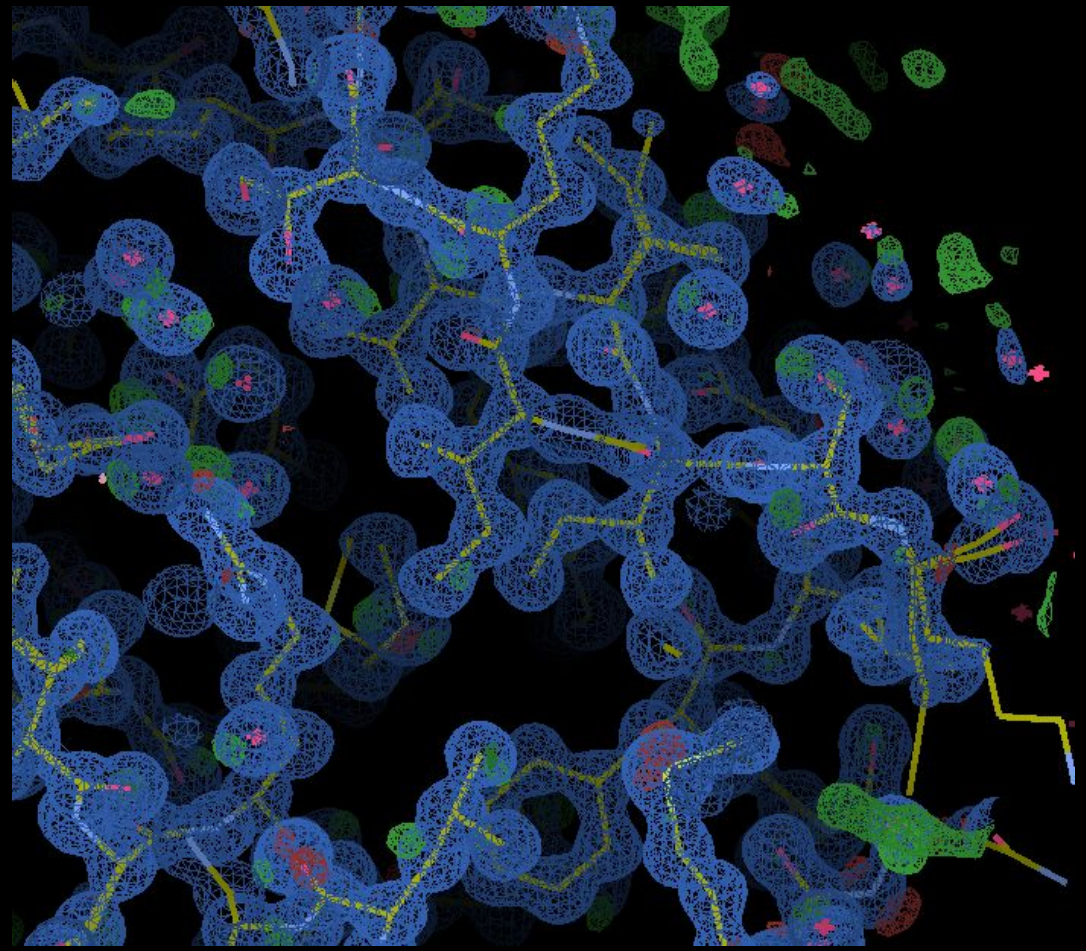


- Lite
  - Makes use of simple secondary structure elements such as helices in MR
  - Attempts to position fragments and build up the rest of the c-alpha backbone using SHELXE
- Borges
  - Similar to Lite but draws on a library of common motifs from the PDB as search models e.g. Zinc fingers or common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

Data resolution

Resolution and Molecular Replacement

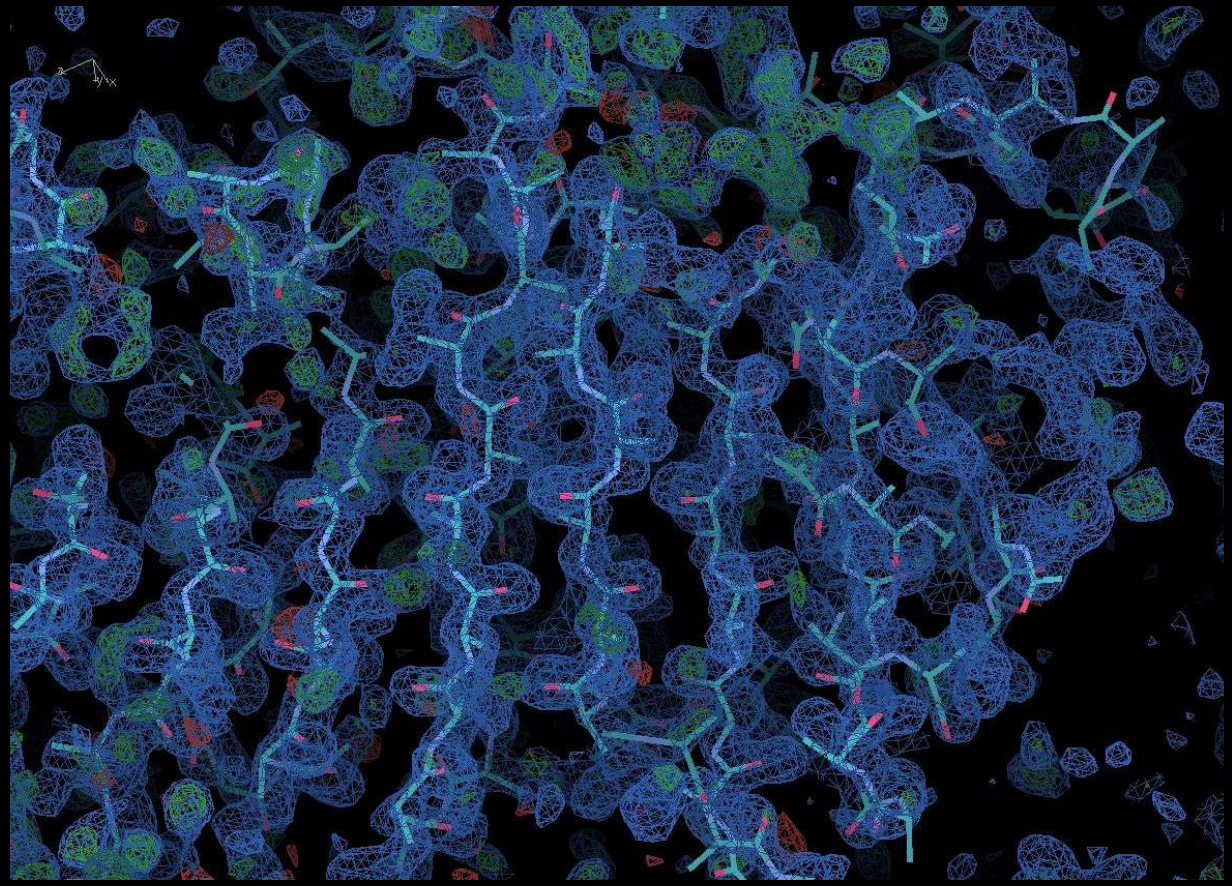
- Better than 1.0 Å – search model can be a small fragment or even a single atom in *Phaser*
  (*McCoy et al. 2017, PNAS*)



CCP4

Resolution and Molecular Replacement

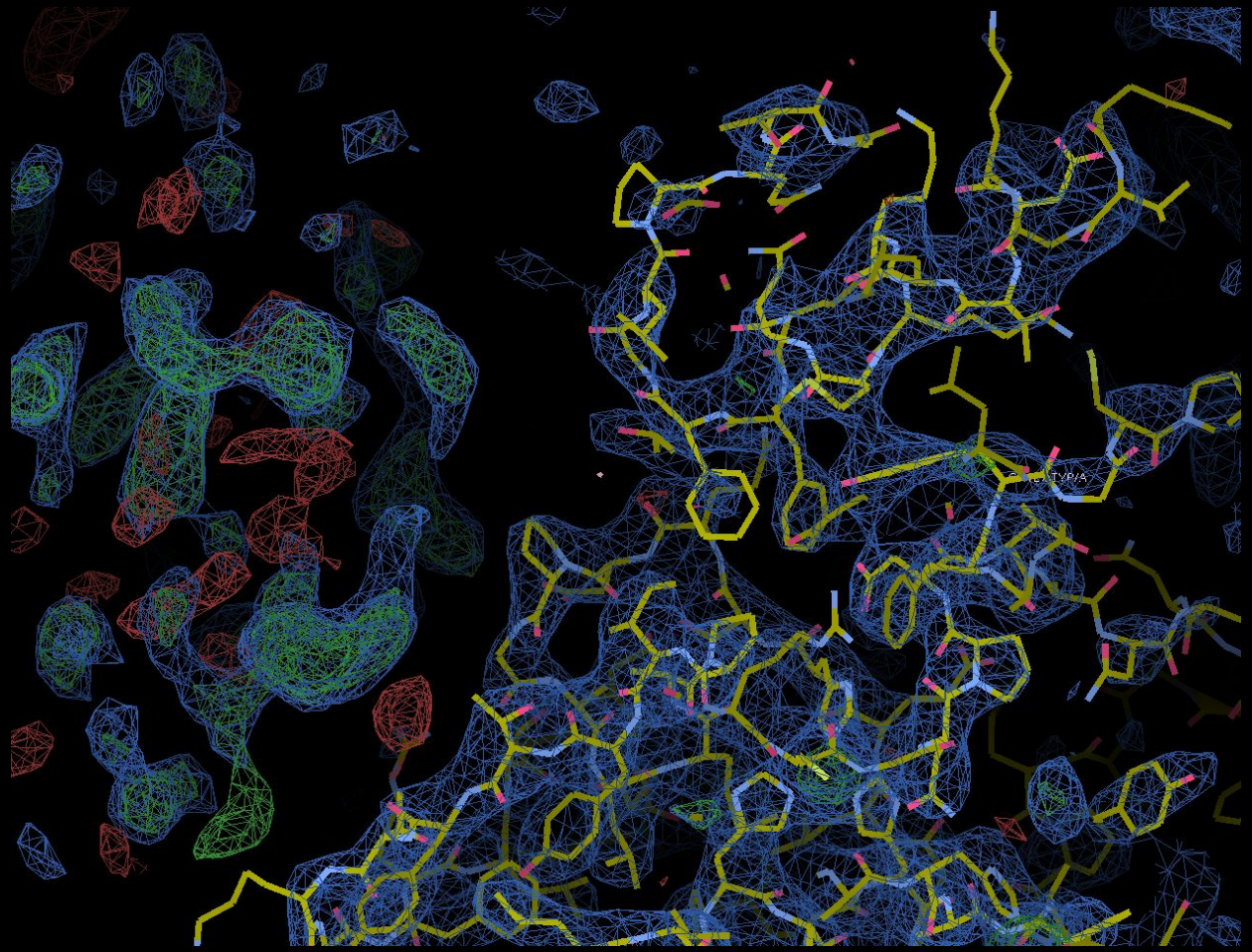
- At resolutions above ~ 2.5 Å
  - *Phaser* can place small fragments of total scattering
  - Applications like *SHELXE* and *Acorn* can improve through density modification (DM) and model building to a correct solution



CCP4

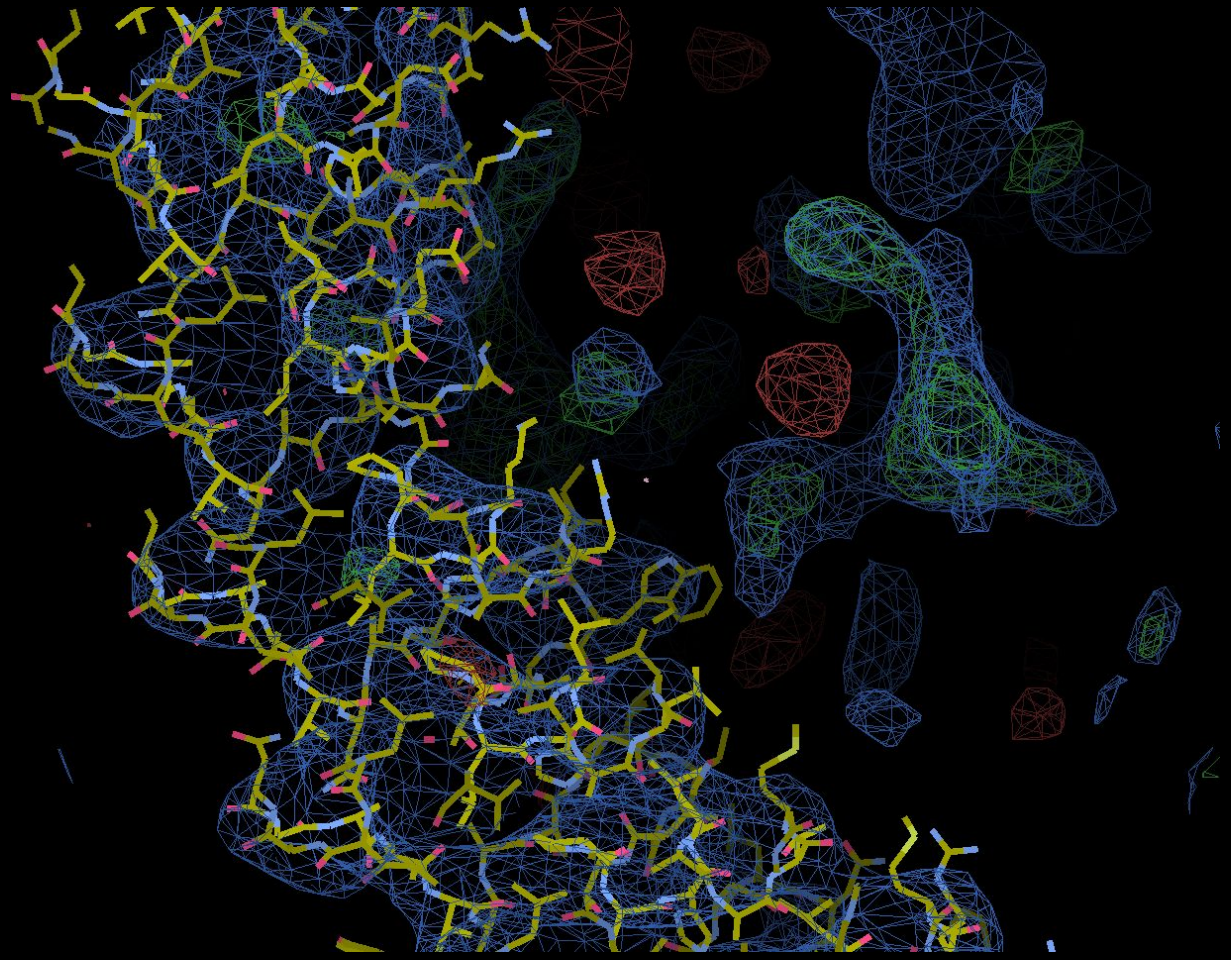
Resolution and Molecular Replacement

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



Resolution and Molecular Replacement

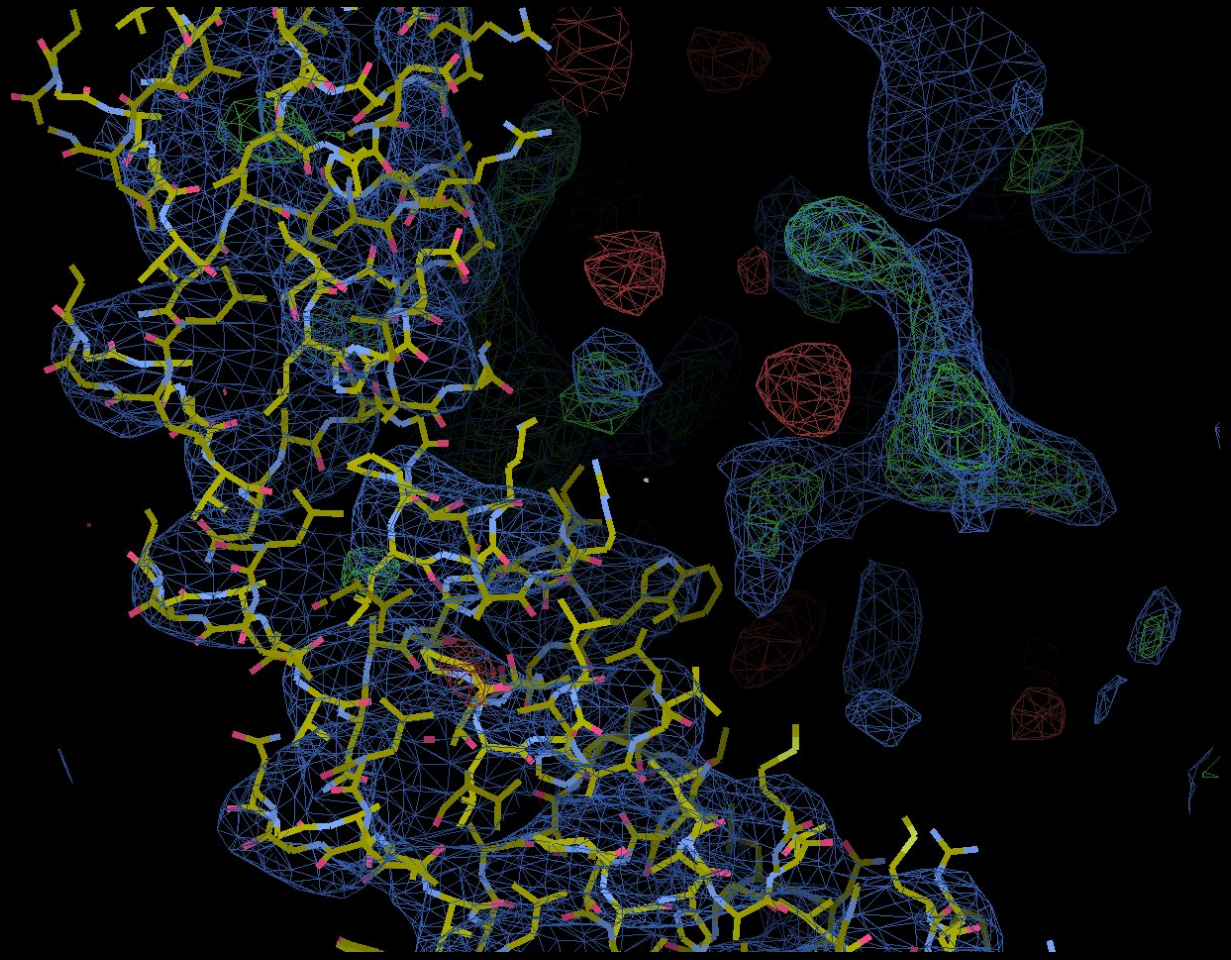
- Below 4Å automated model building becomes difficult



CCP4

Resolution and Molecular Replacement

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

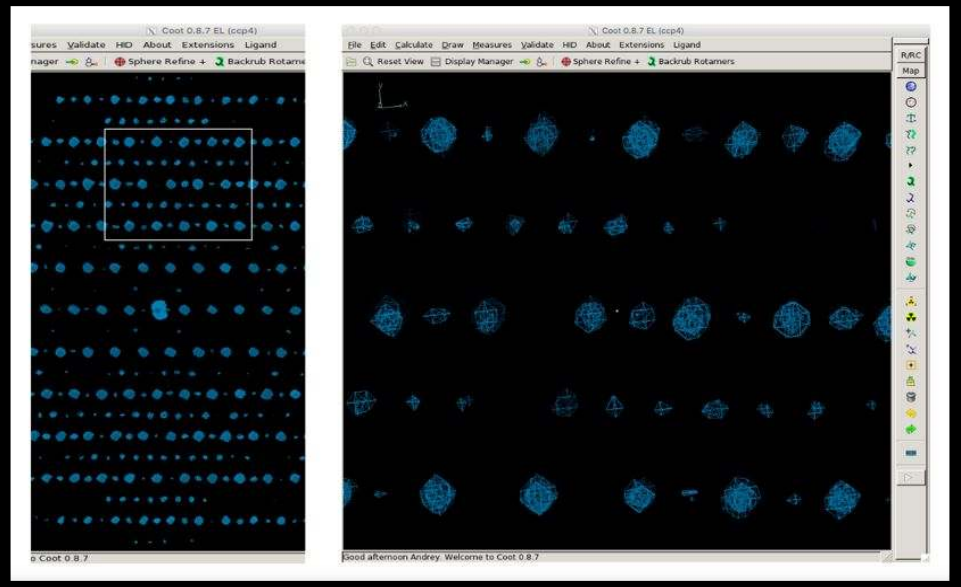


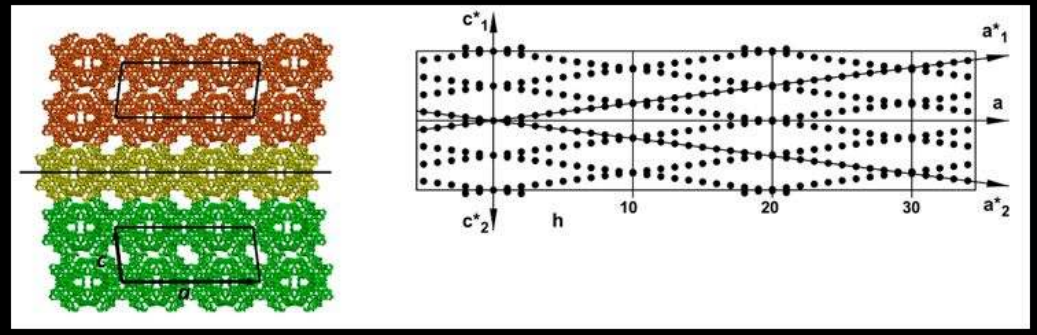
CCP4

Pathologies

Data Pathologies and Molecular Replacement

- Potential Problems:
  - Pseudo translation
  - Twinned data
- Both can complicate the determination of space group
- Use options in MR to try all possible space groups

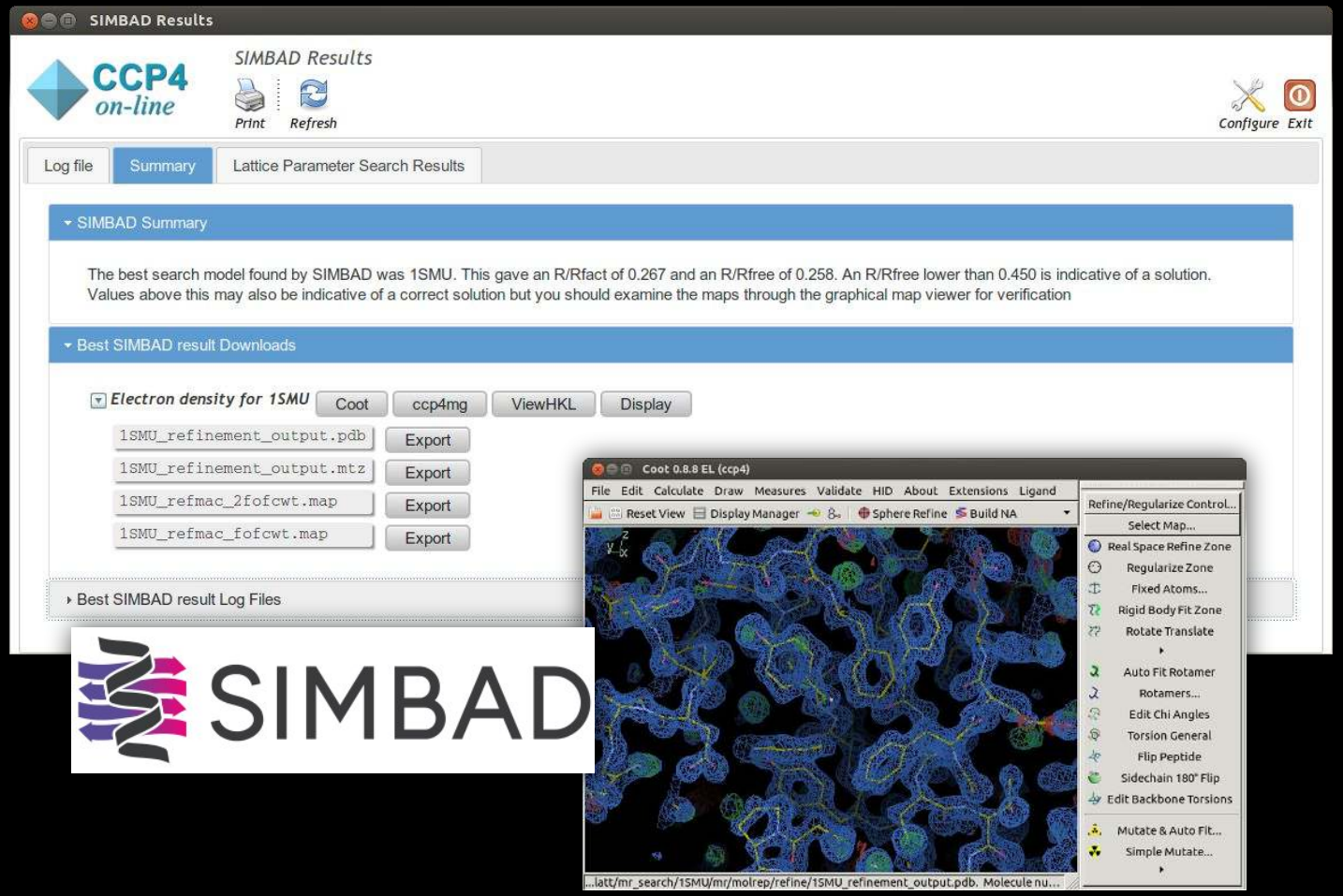




CCP4

Crystal contamination

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



CCP4