

Molecular Replacement - Assessing and improving the solution

CCP4/DLS Workshop - 2023
Ronan Keegan CCP4 Group



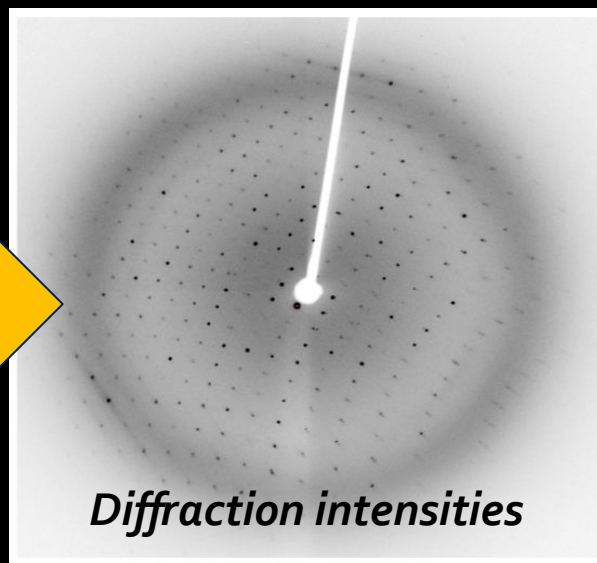
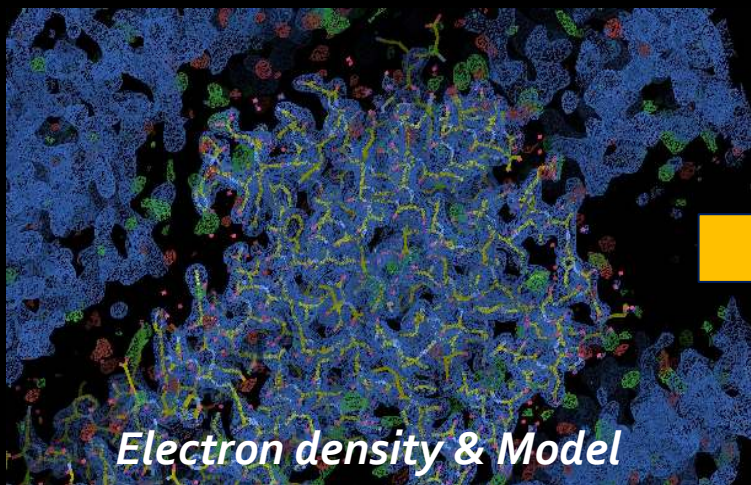
The Phase Problem

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

Electron density
at (x,y,z)

Amplitudes

Phases



?

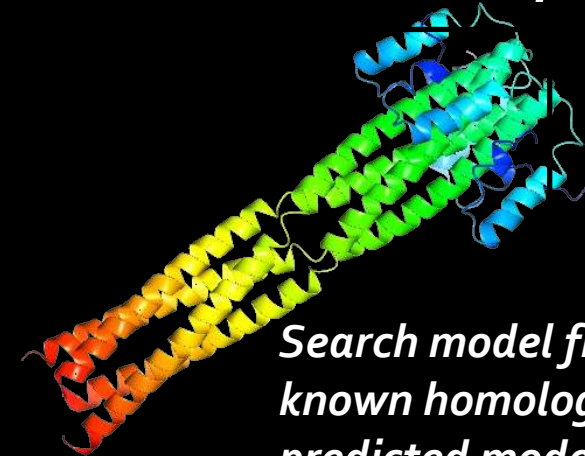
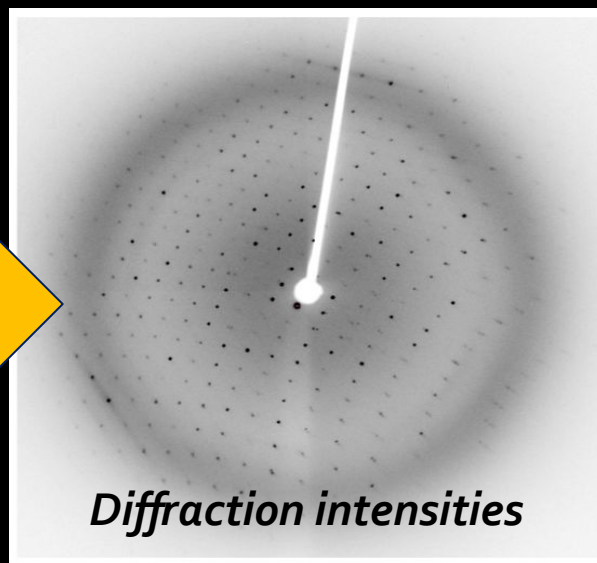
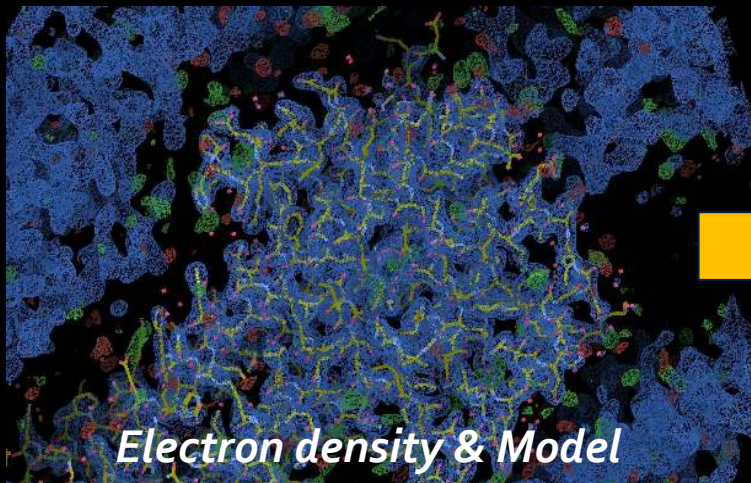
The Phase Problem: Molecular Replacement

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

Electron density
at (x,y,z)

Amplitudes

Molecular
Replacement



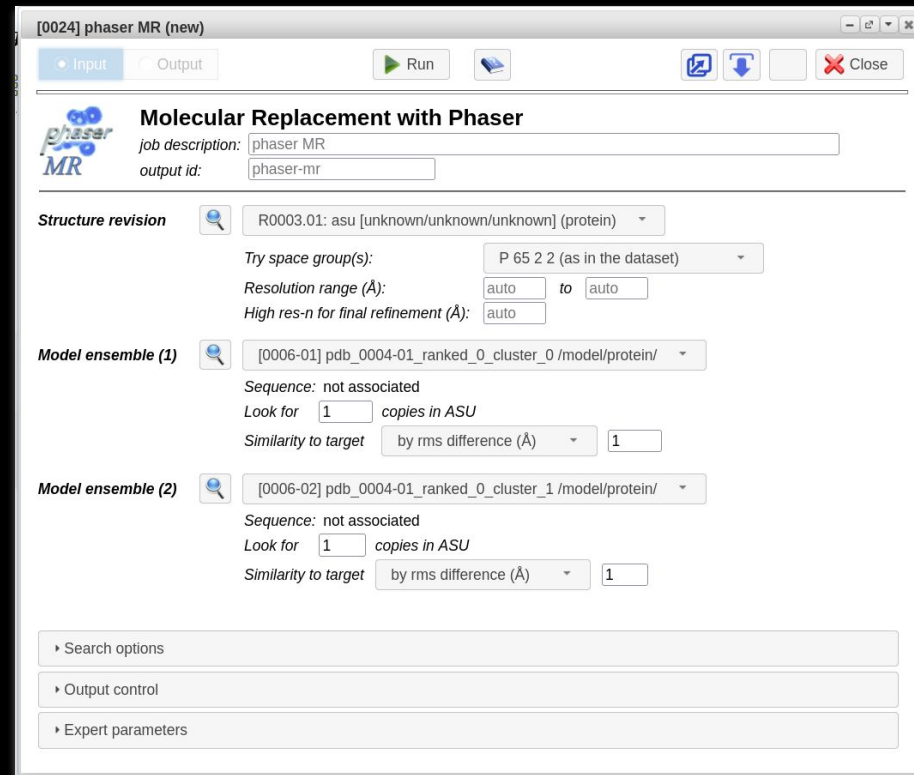
Interpreting the results of Molecular Replacement

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

Molecular Replacement: Phaser

- Important points on using Phaser
 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



The screenshot shows the 'Molecular Replacement with Phaser' window. It includes fields for 'job description' (phaser MR) and 'output id' (phaser-mr). Under 'Structure revision', it shows 'R0003.01: asu [unknown/unknown/unknown] (protein)' and 'Try space group(s): P 65 2 2 (as in the dataset)'. The 'Resolution range (Å)' is set to 'auto' to 'auto', and 'High res-n for final refinement (Å)' is 'auto'. Under 'Model ensemble (1)', it shows '[0006-01] pdb_0004-01_ranked_0_cluster_0 /model/protein/' with 'Sequence: not associated', 'Look for 1 copies in ASU', and 'Similarity to target by rms difference (Å) 1'. A second model ensemble (2) is also shown with similar settings. At the bottom, there are expandable sections for 'Search options', 'Output control', and 'Expert parameters'.

How do I know my MR solution is successful?

Assessing the MR Solution

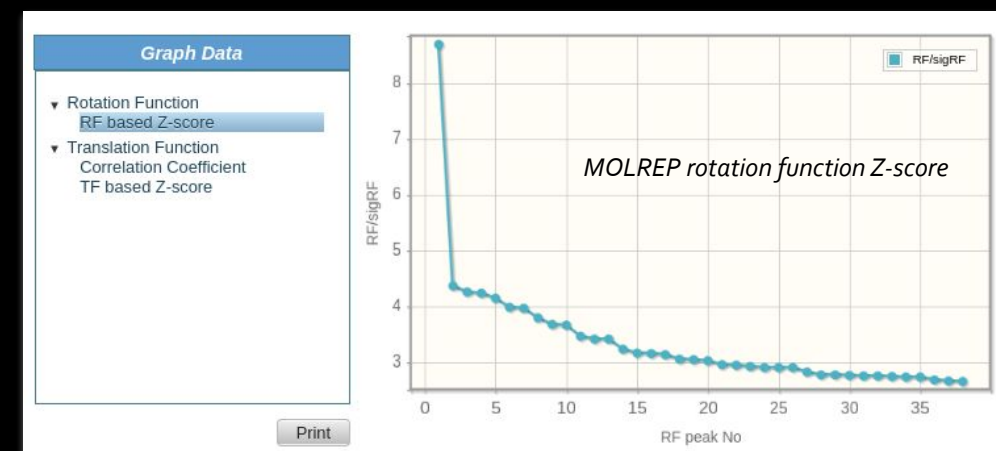
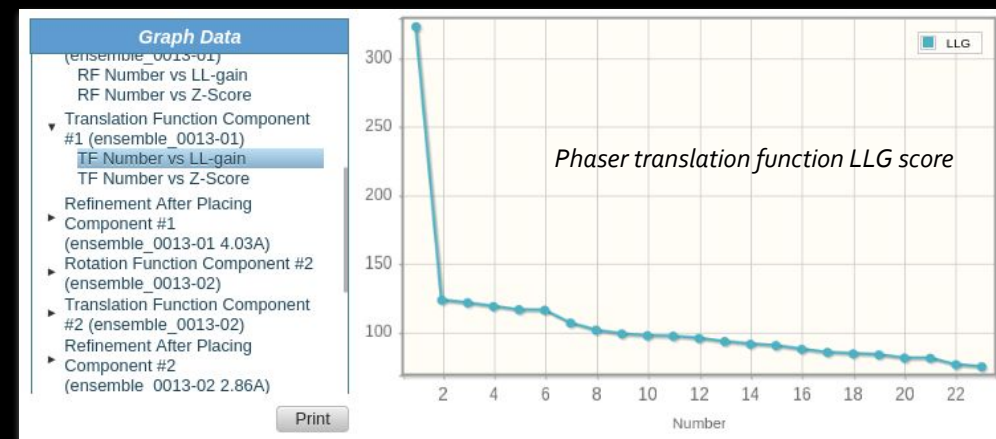
- Terminology
 - MR solution
 - Successful MR solution
- What is a successful MR solution?
 - A search model placed in the target unit cell such that it will provide us with sufficiently accurate phase estimates for the target
 - Enables us to complete the model through model building and refinement

Assessing the MR Solution

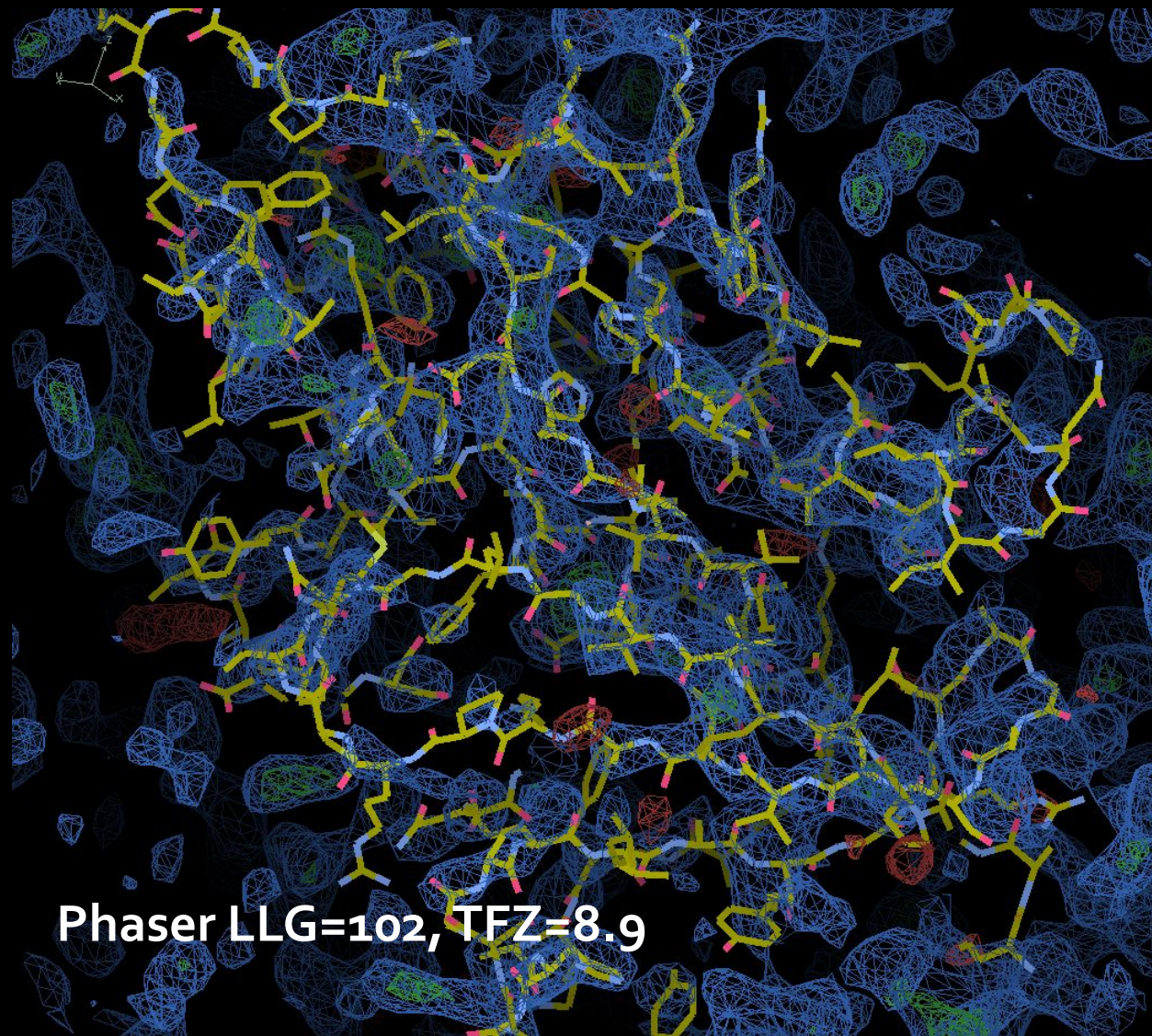
- In difficult cases the position may be correct but getting from the MR solution to a complete model may not be straight forward
- Assessment often involves performing additional structure solution steps such as refinement and density modification

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



With 50 cycles of jelly body refinement

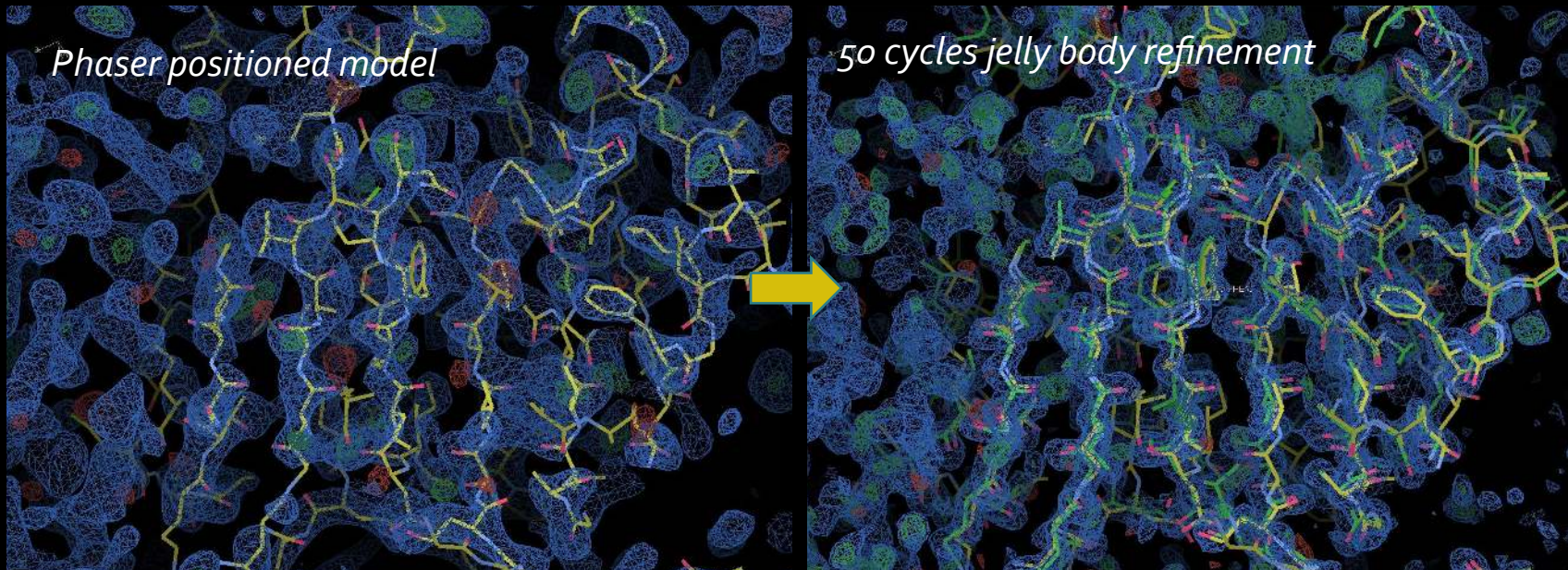
$R/R_{\text{free}}=0.28/0.32$

- Refinement

- Look at Rfactor/Rfree

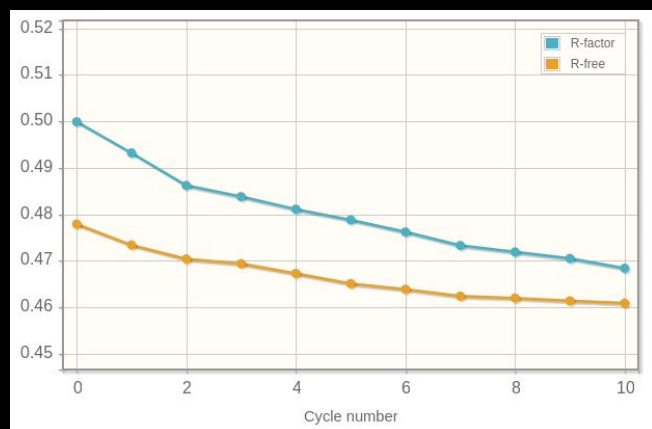
- are they falling? Is Rfree below 0.5?

- Use 50 cycles of jelly body refinement option in Refmac post MR

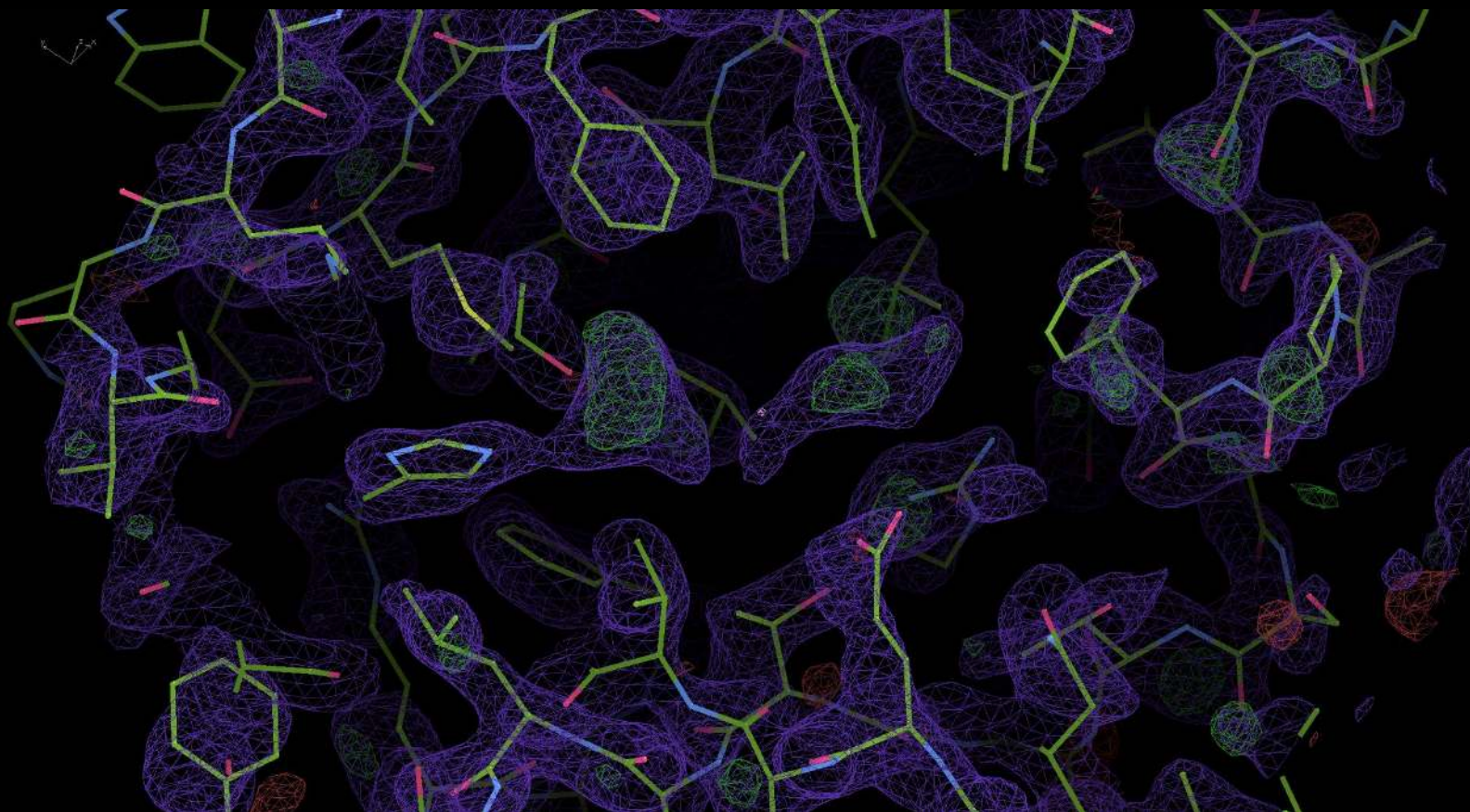


Refining predicted models used in Molecular Replacement

- 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

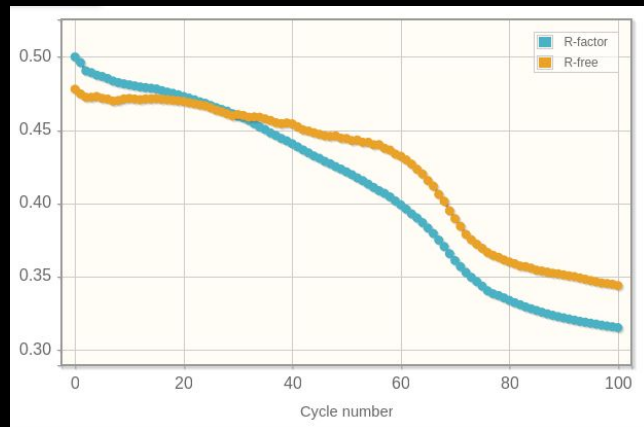


*Example: 10 cycles of
jelly-body refinement with
Refmac*
Rfactor = 0.46
Rfree = 0.47

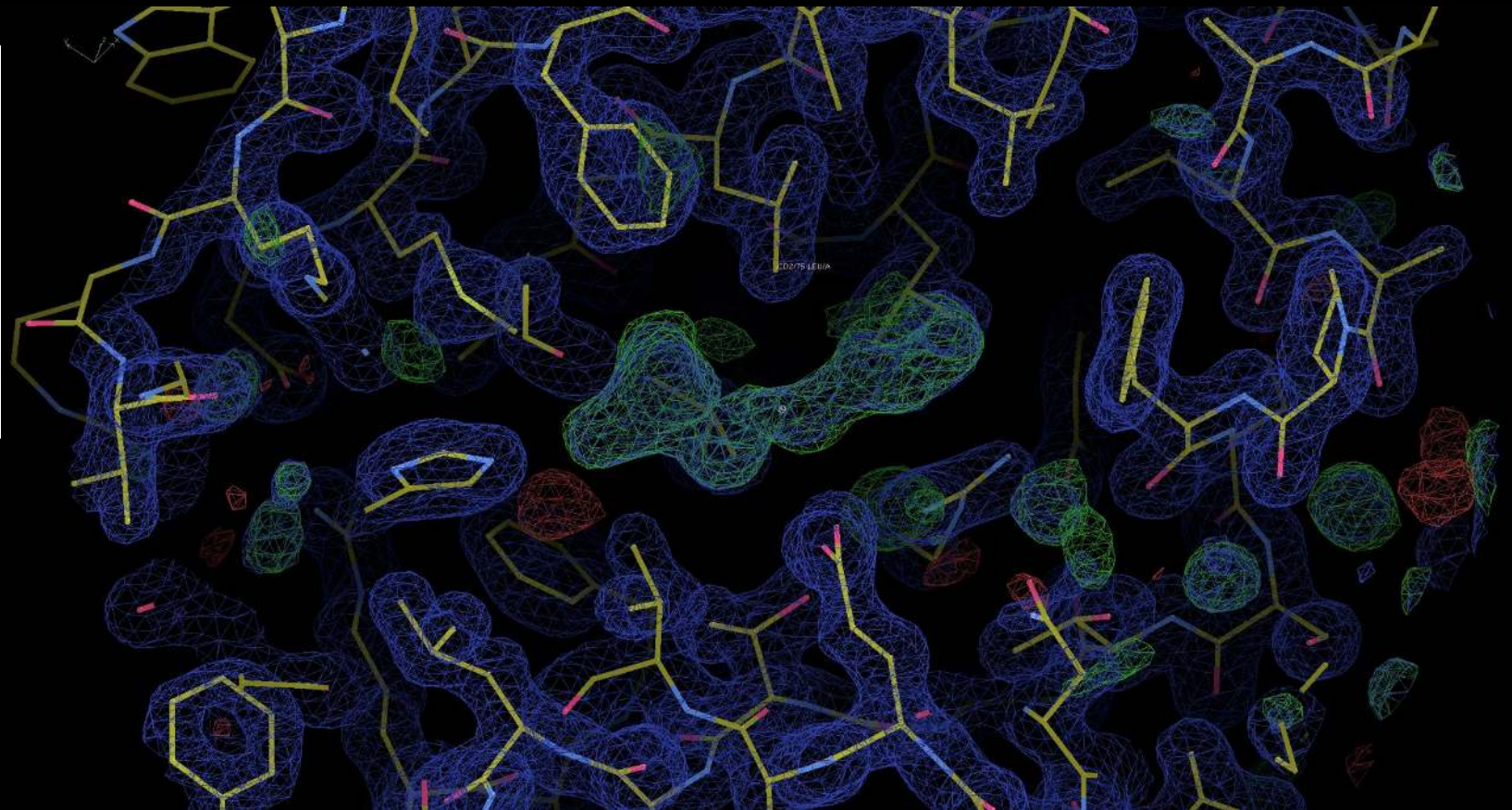


Refining predicted models used in Molecular Replacement

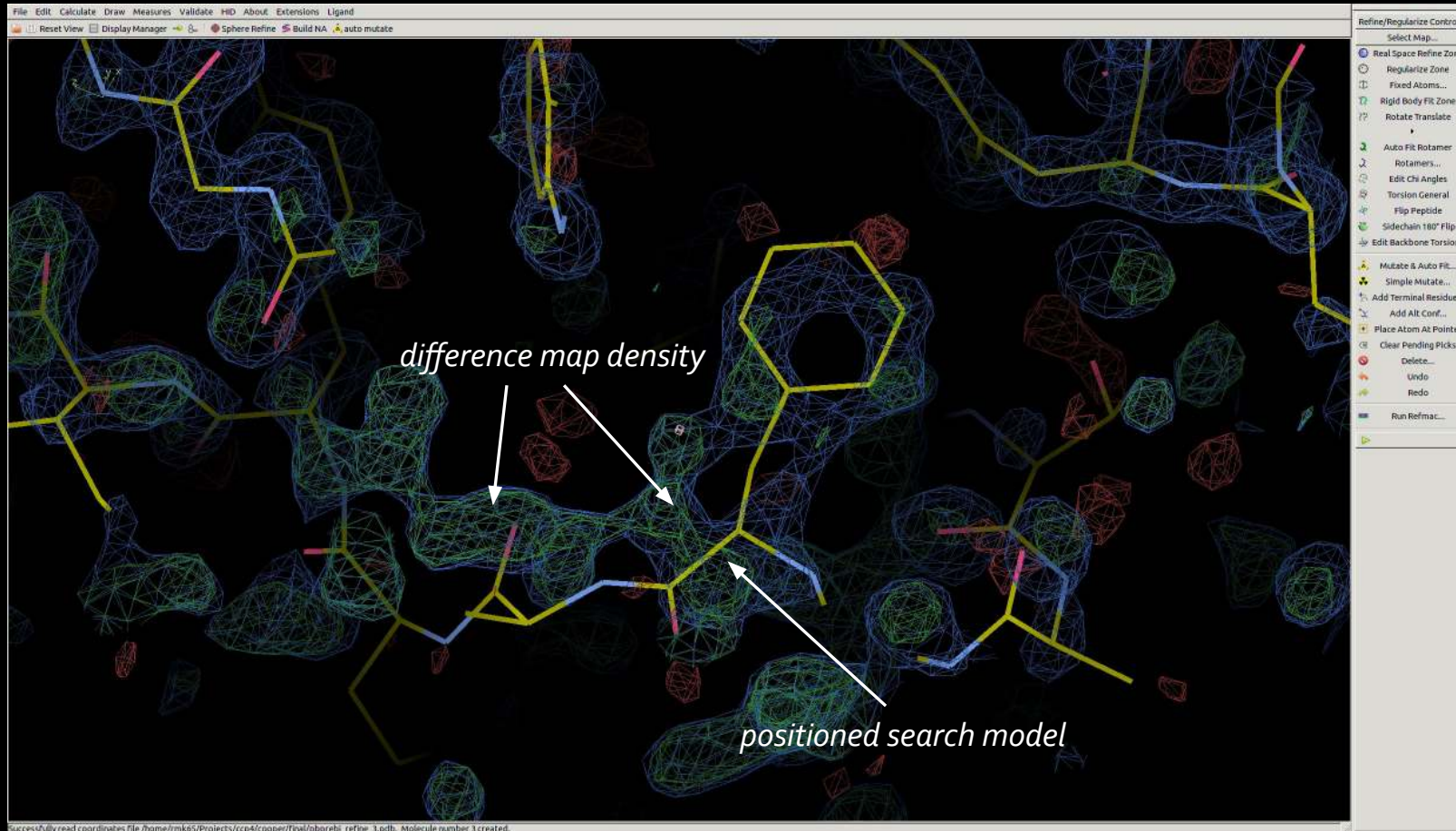
- Predicted models (*AlphaFold2*, *Colabfold* etc.) are often significantly different in their main and side chain positioning to the crystal form despite making good MR search models
- Can require lots of cycles of jelly-body refinement in Refmac



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

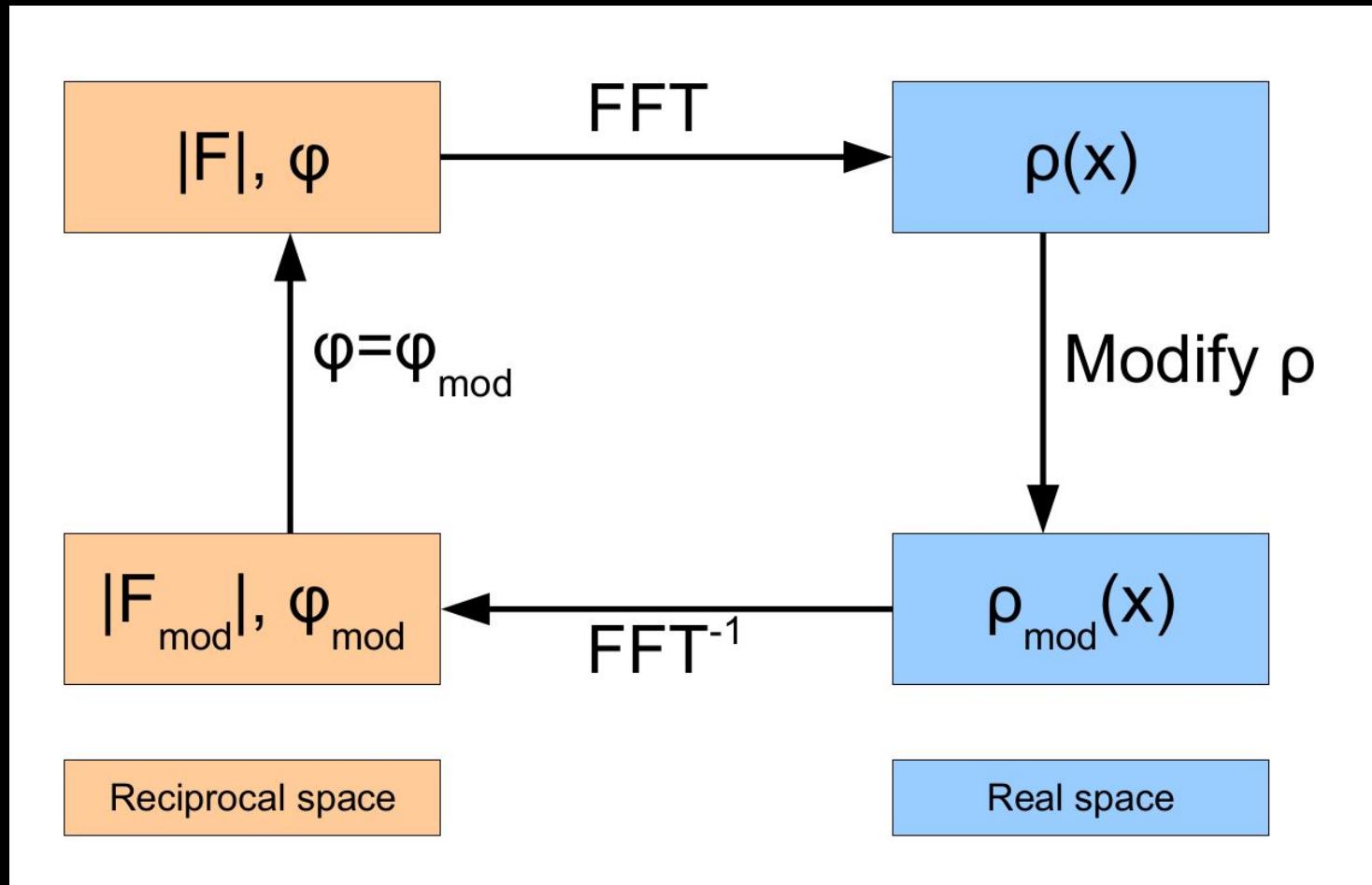


- Examine solution by eye
 - Use Coot to examine positioned models & maps



Density Modification and Model Building

Density modification

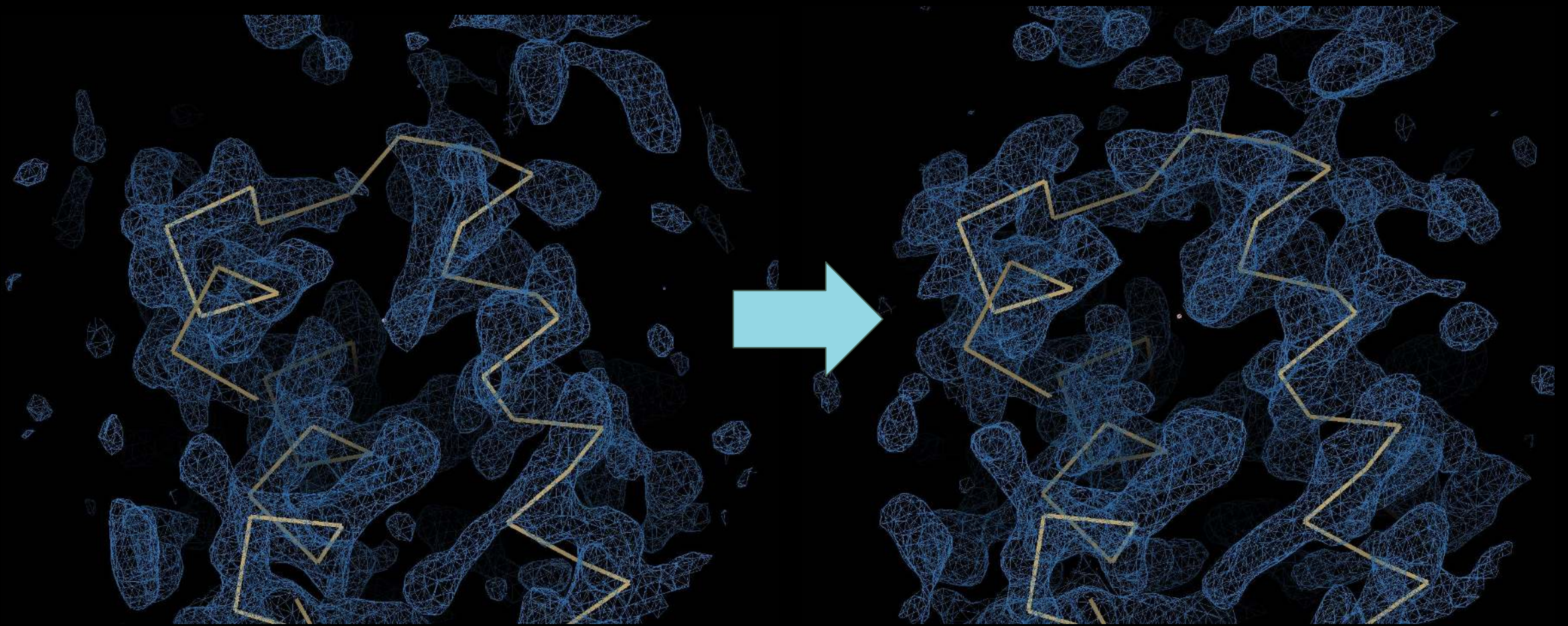


Main techniques:

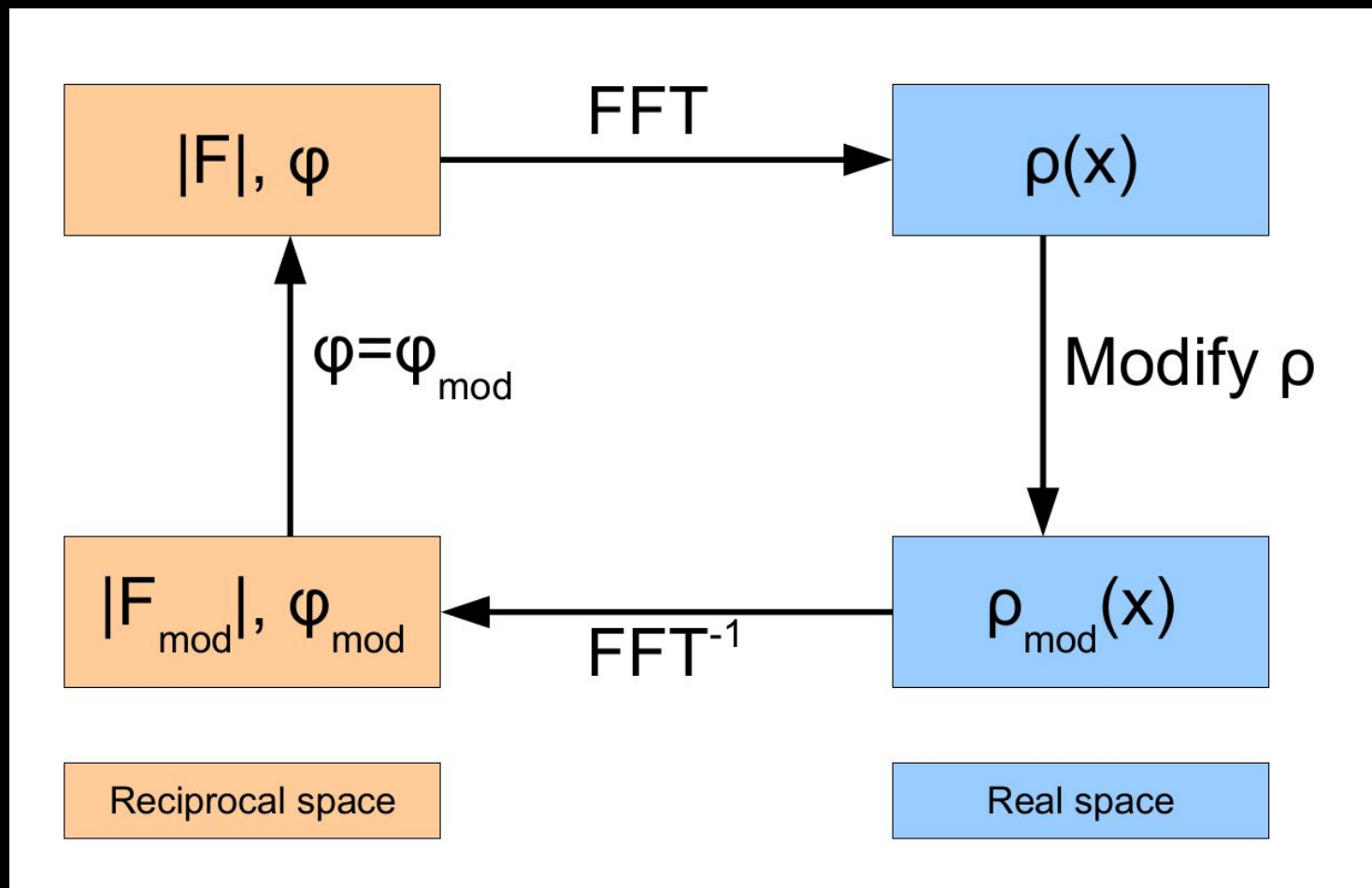
1. Solvent flattening
2. Histogram matching
3. NCS averaging
4. C-alpha tracing

(slide from Kevin Cowtan)

Density modification



Density modification



CCP4 Applications:

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


(slide from Kevin Cowtan)


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


▼ Model Building (4)


 **Model building -- polypeptides and polynucleotides**
Automatic Model Building with ModelCraft
-- automatic model building after MR or Experimental Phasing

 **Model building -- polypeptides**
Automatic Model Building with CCP4Build
-- automatic model building after MR or EP with a combination of several CCP4 programs

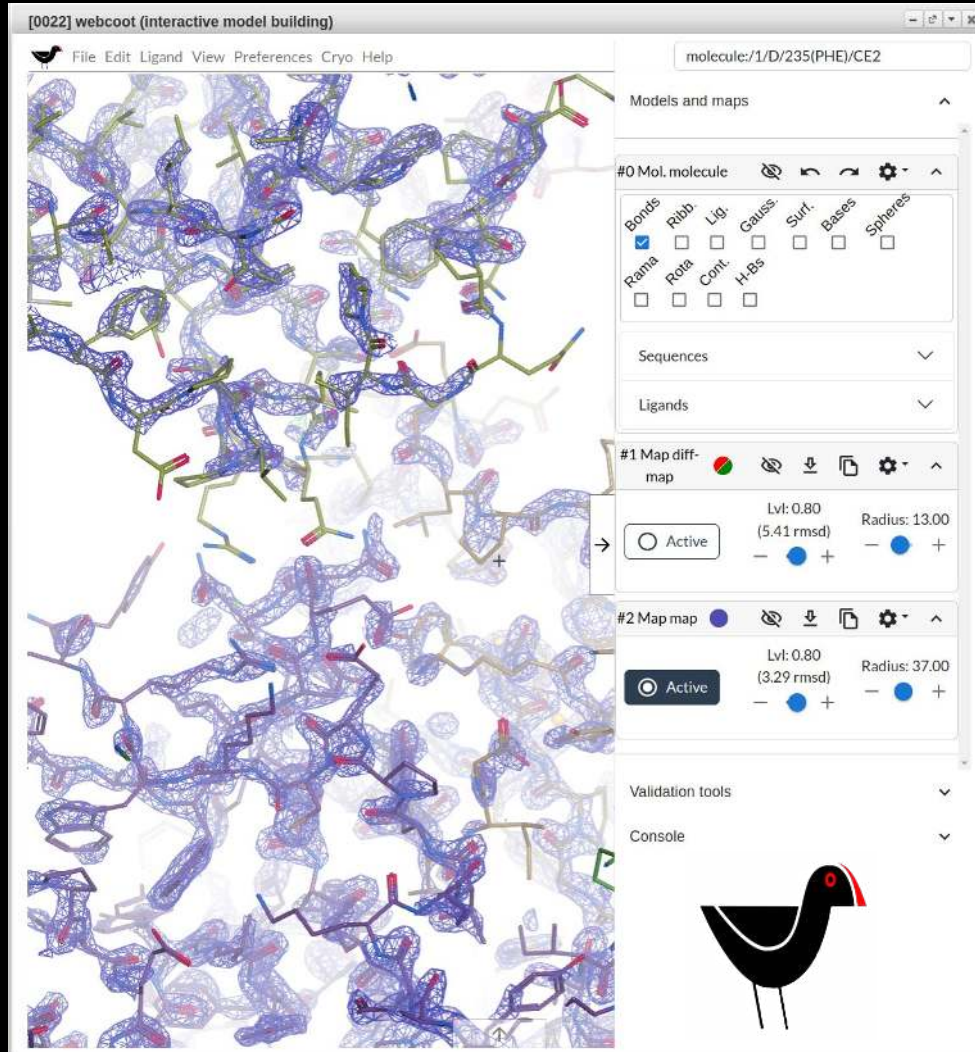
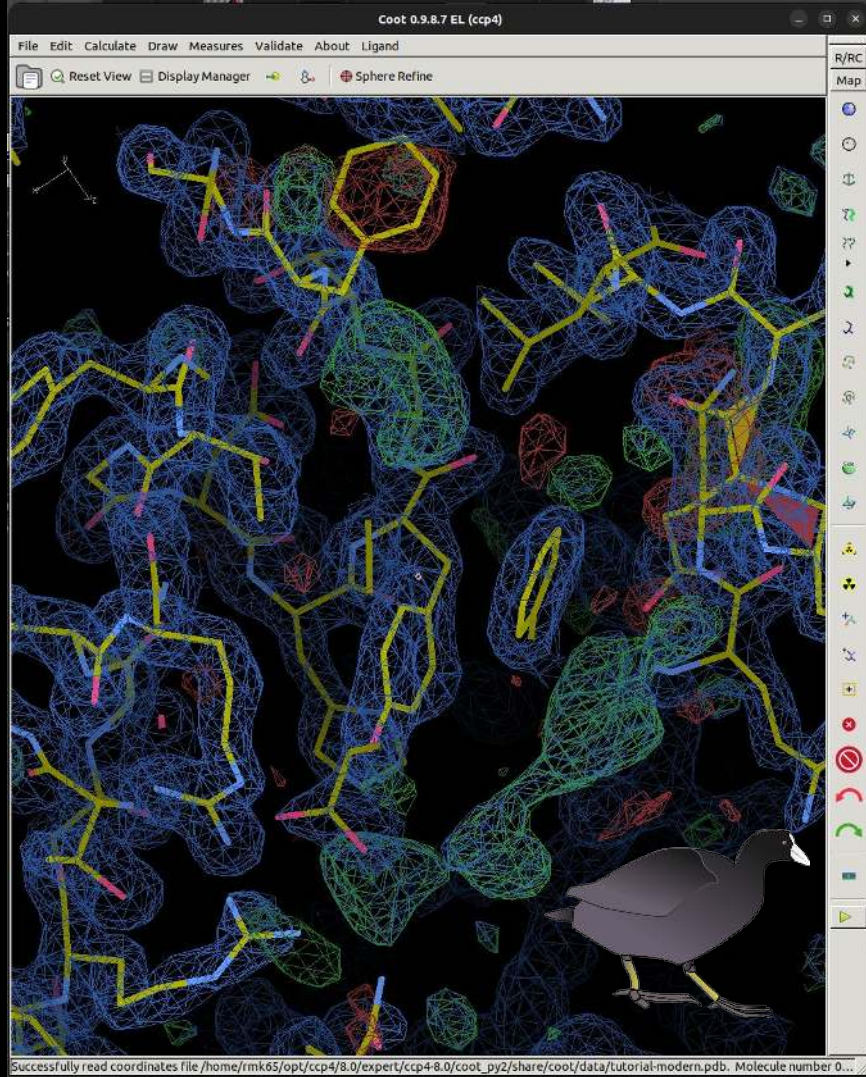
 Automatic Model Building with Buccaneer
-- automatic statistical model building after MR or Experimental Phasing

 Automatic Model Building with Arp/wArp
-- automatic model building after MR or EP using original algorithm

 **Model building -- polynucleotides**
NUCE: Trace Nucleic Acid Chains with Arp/wArp
-- traces nucleic acid chains in electron density using original algorithm

 Automatic Model Building of RNA/DNA with Nautilus
-- automatic building of RNA/DNA chains after MR or Experimental Phasing

Manual Model Building: Coot & Moorhen/WebCoot

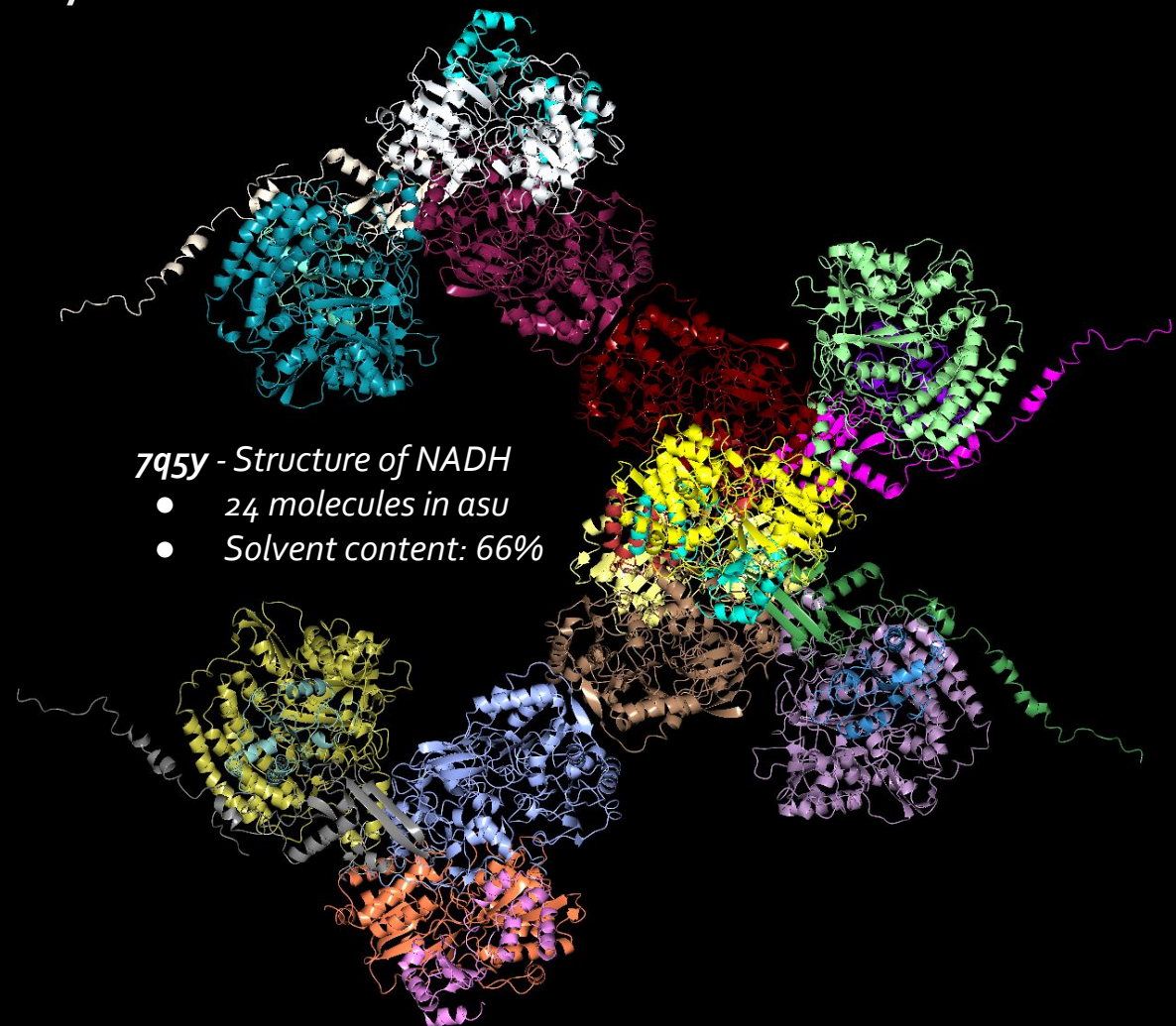
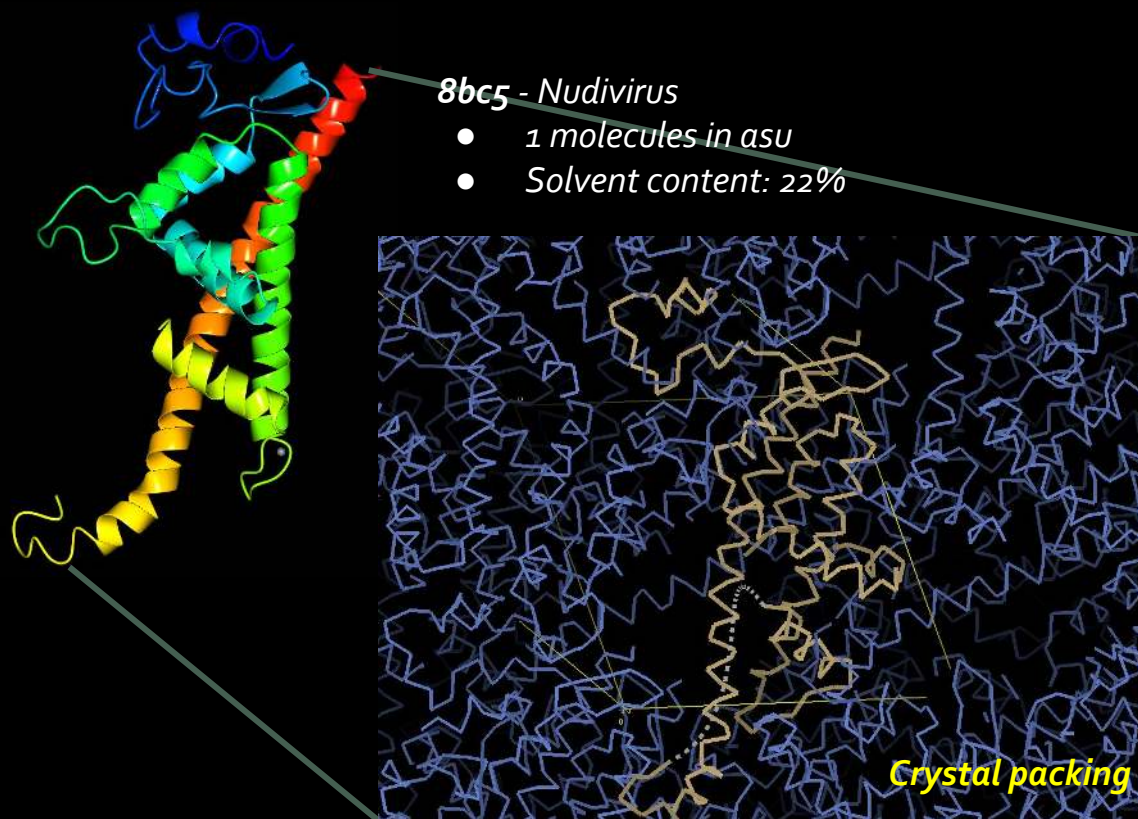


Common complications in Molecular Replacement

Contents of the asymmetric unit

Contents of the asymmetric unit

- How many copies/molecules in the asymmetric unit?



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

[0002] define asymmetric unit contents -- completed

Input Output

Report Main Log Service Log Errors

[0002] Asymmetric Unit Contents

CCP4 v8.0.016: CCP4 Cloud vunknown
Started: 2023-11-01 11:24:25
Finished: 2023-11-01 11:24:25
CPU: 00.000, Disk: 0.04M

Suggested ASU contents

	<i>N_{copies}</i>	<i>Structural unit components</i>	<i>Type</i>	<i>Size</i>	<i>Weight</i>
1	6	[0001-08] 7zbh_expected_A_ /sequence/protein/	PROTEIN	453	49828.6
Total residues/weight:				2718	298971.7


[0002] Results

Cell volume: 2795576.75 Å³

Molecule fitting statistics

<i>N_{trial}</i>	<i>Matthews</i>	<i>% solvent</i>	<i>P_{matthews}</i>
* 1	2.34	47.42	1.000

[0002] Verdict

 The estimated solvent fraction is below the usual range for macromolecular crystals, diffracting at similar resolution

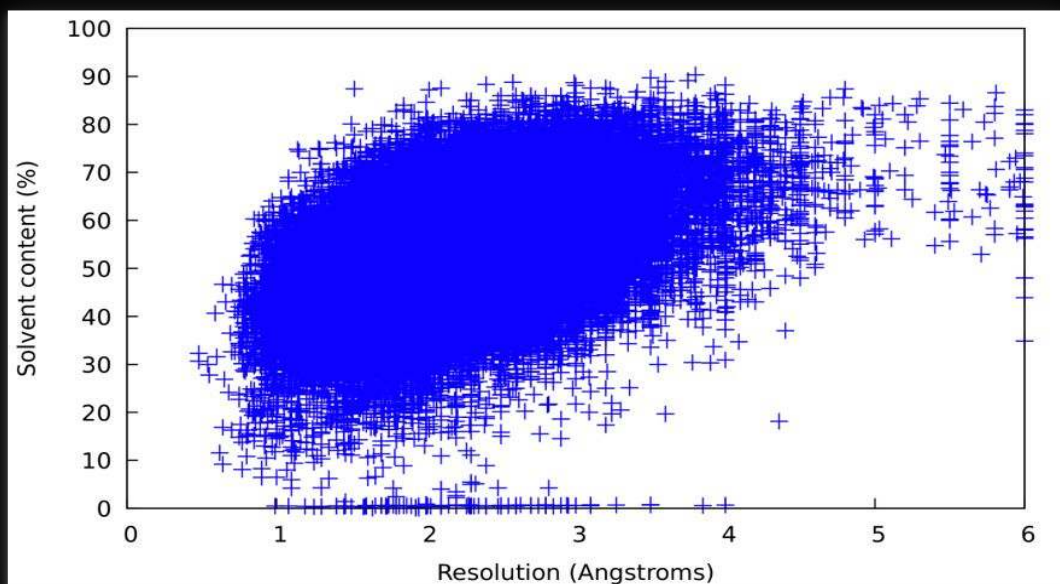
Although the suggested composition of ASU corresponds to an unusual value of solvent fraction, it *may* be an acceptable assumption.

In general, composition of ASU remains a hypothesis until structure is solved. The solvent content is more a guidance, rather than a definite indicator, of the correctness of the choice. Inaccurate estimations of solvent content may have a negative impact on phasing and density modification procedures, especially in difficult cases.



Contents of the asymmetric unit

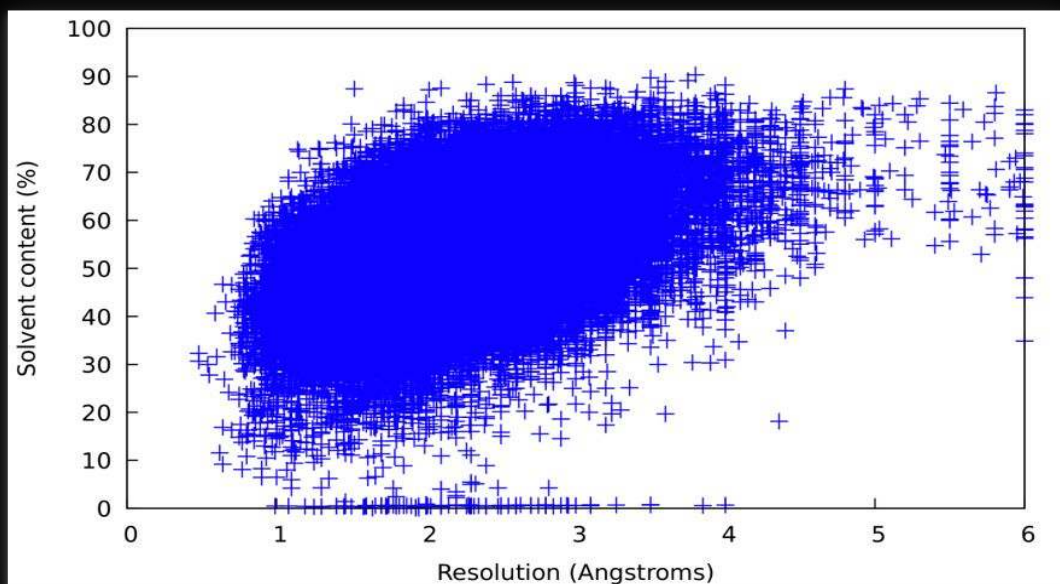
- How many copies/molecules in the asymmetric unit?
- Estimate using Matthews Coefficient
 - Assumes roughly 50% of cell contents is solvent
 - Accounts for resolution



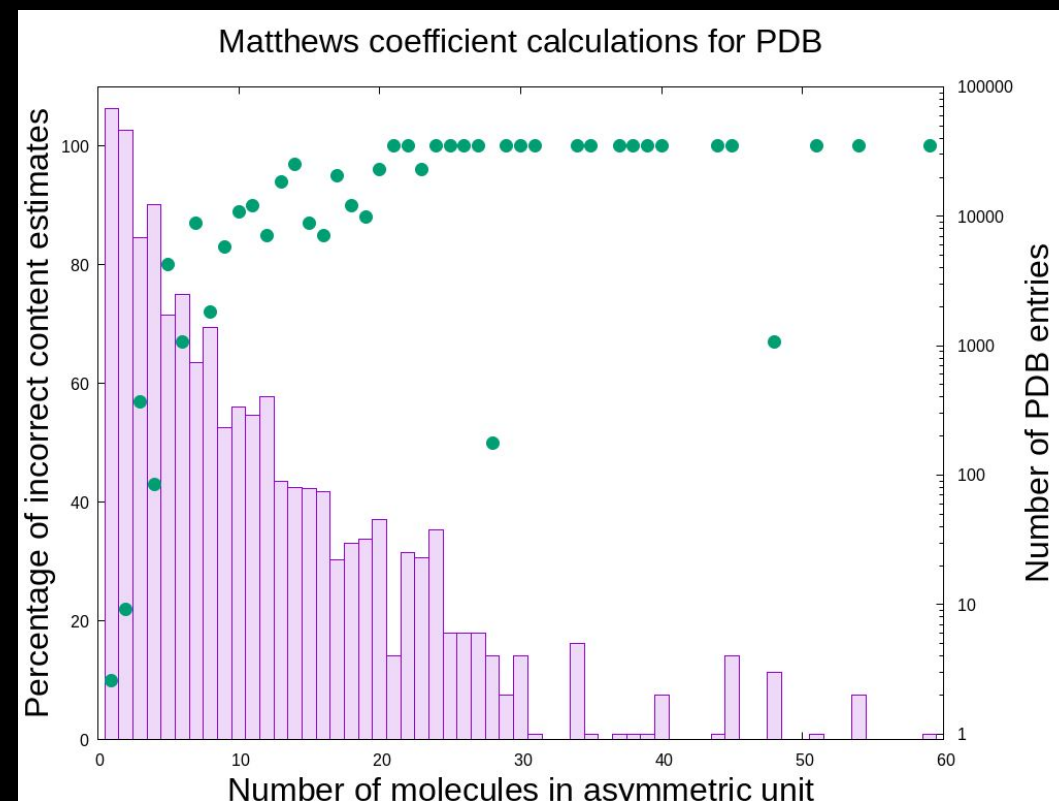
*Distribution of solvent content across
156,000 X-ray datasets in the PDB*

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

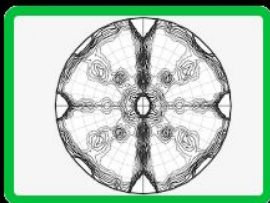


*Distribution of solvent content across
156,000 X-ray datasets in the PDB*



Contents of the asymmetric unit

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



[0023] srf analysis (new)

Input Output Run

Self-Rotation Function Analysis with Molrep
job description: srf analysis

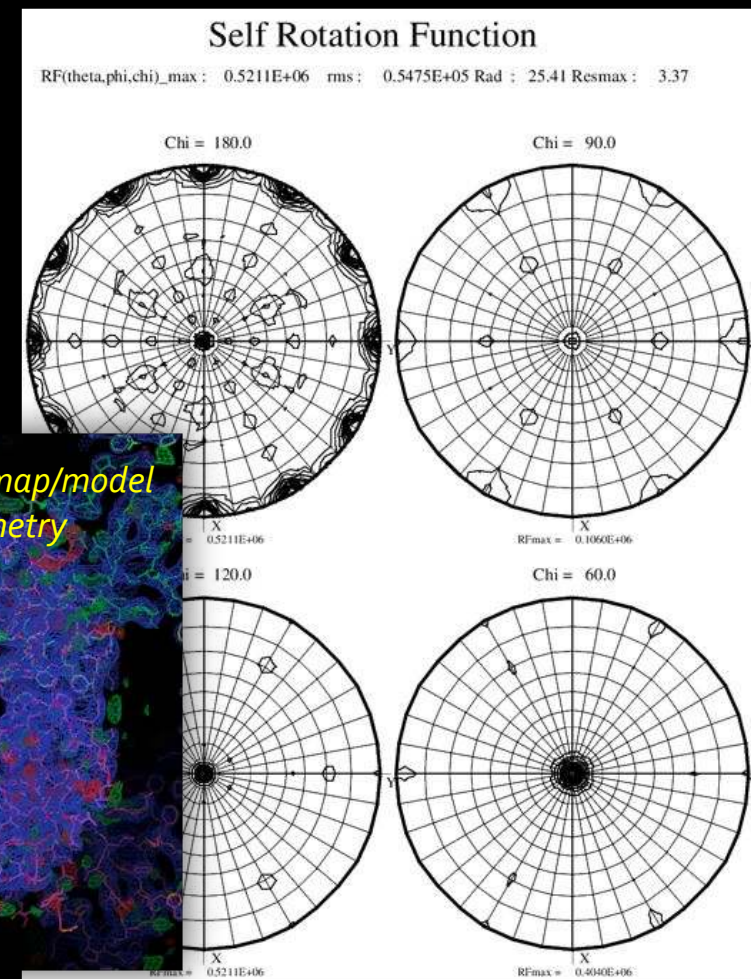
Reflection Data [0001-04] 8gsx-sf [unknown/unknown/unknown] /hkl/

▼ SRF Options

Expected molecule size (Å)
Select Chi sections at 180, 90, 120 and degrees
Top isoline level: sigmas

▼ Experimental Data

High resolution cut-off (Å)
Down-weighting high resolution data: none
Down-weighting low resolution data:
Scaling: none



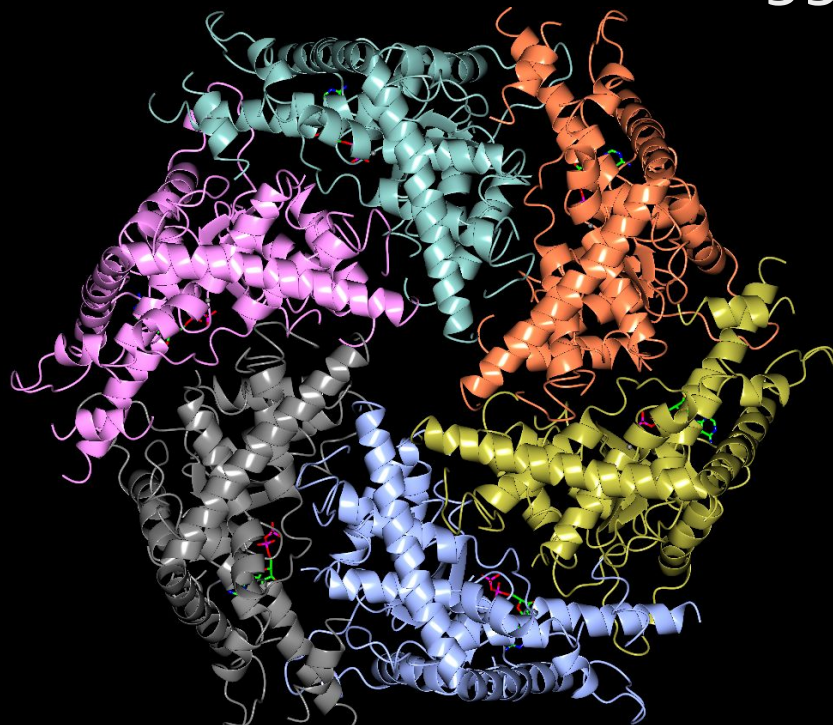
Contents of the asymmetric unit

- 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

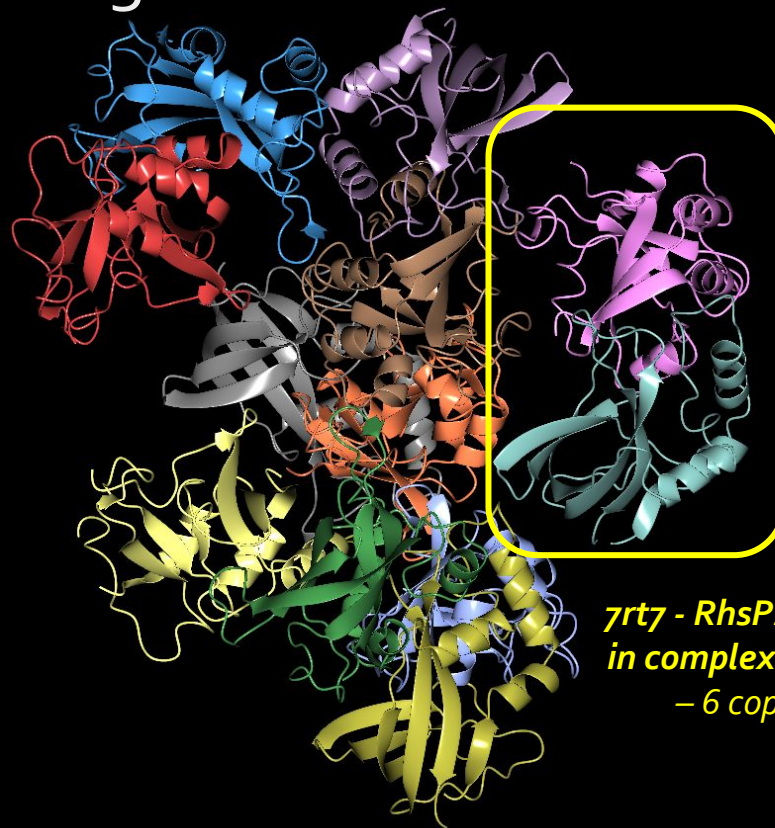
Multimers and Complexes

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)

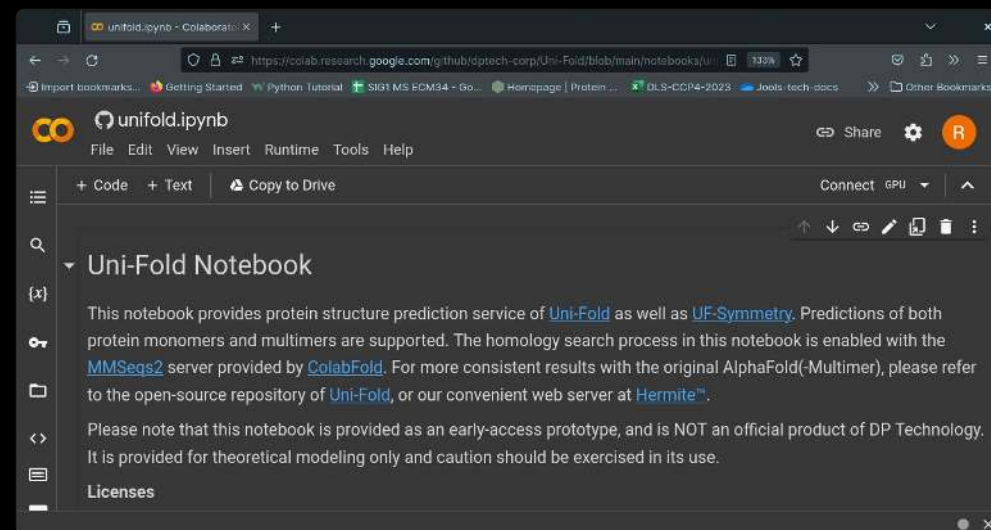
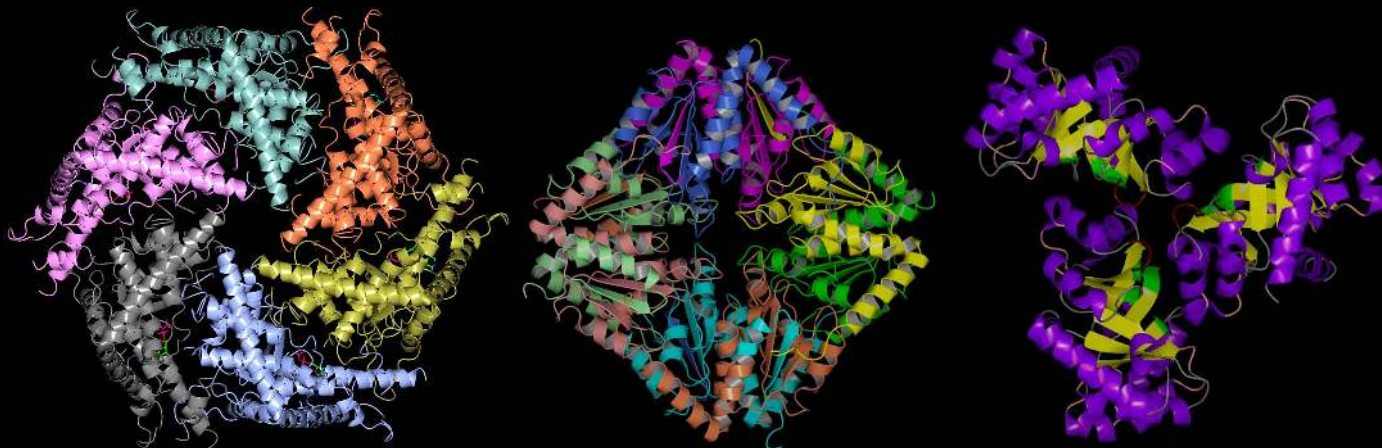


7rt7 - RhsP2 C-terminal toxin domain in complex with its immunity protein
– 6 copies of 2-chain complex

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

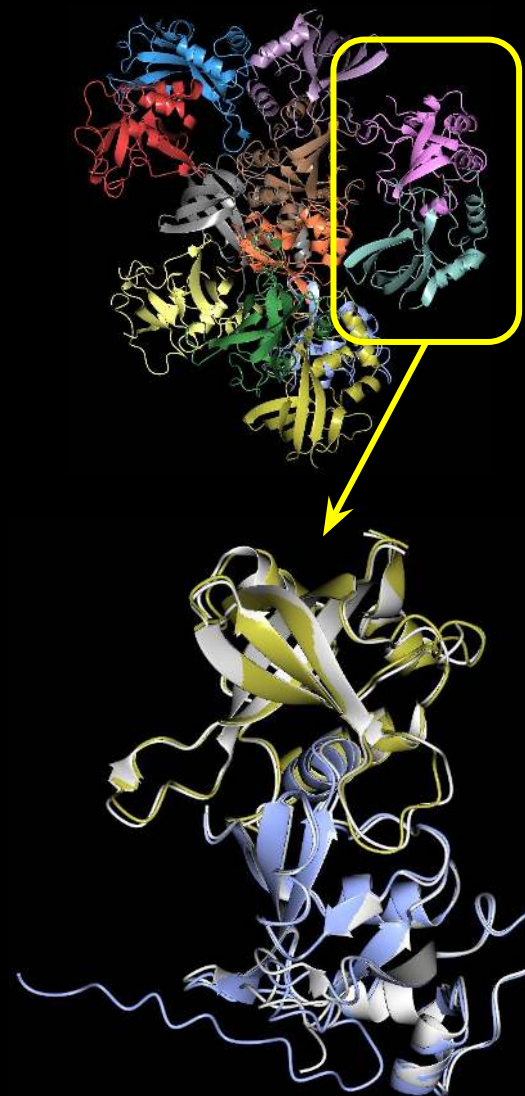


Multimers and Complexes

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

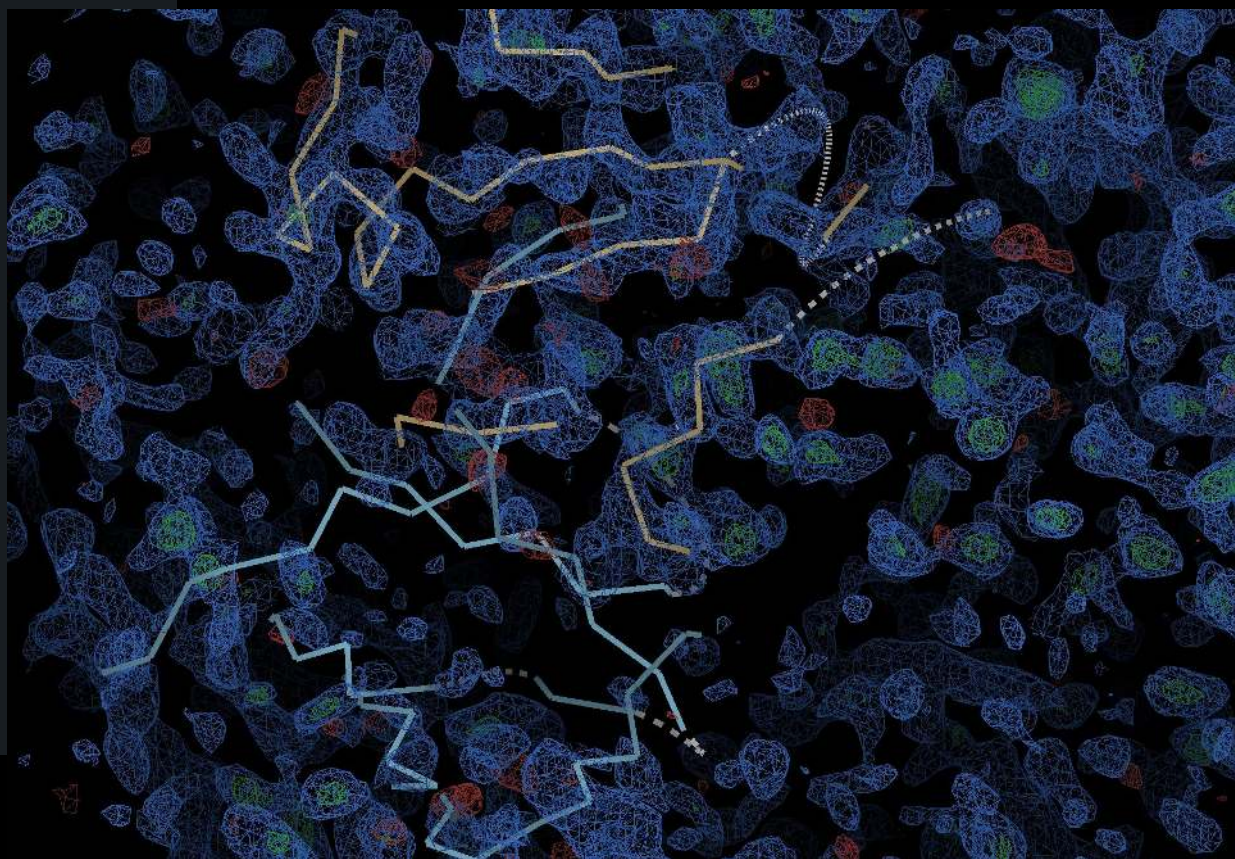


*7rt7 - complex prediction from
AlphaFold2 aligned to experimental
complex structure (white)*

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

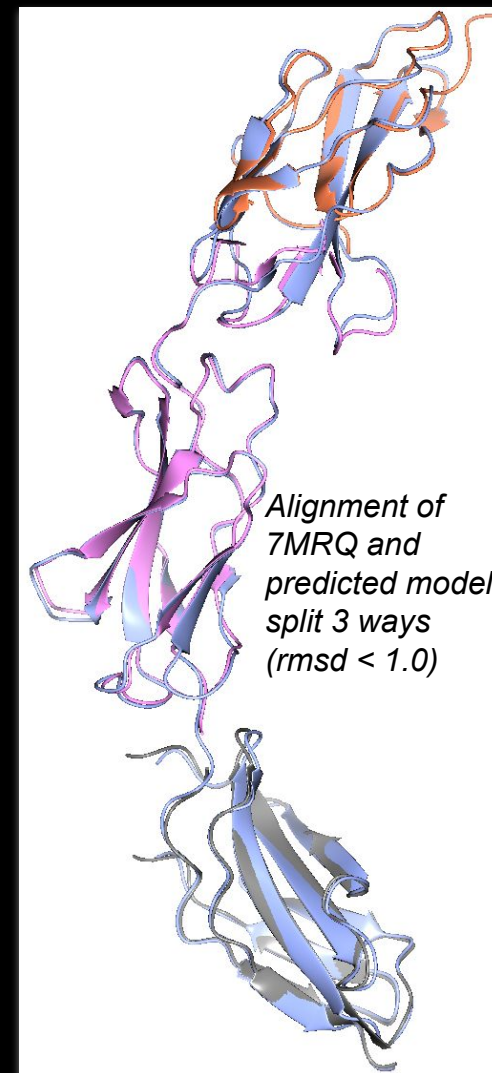
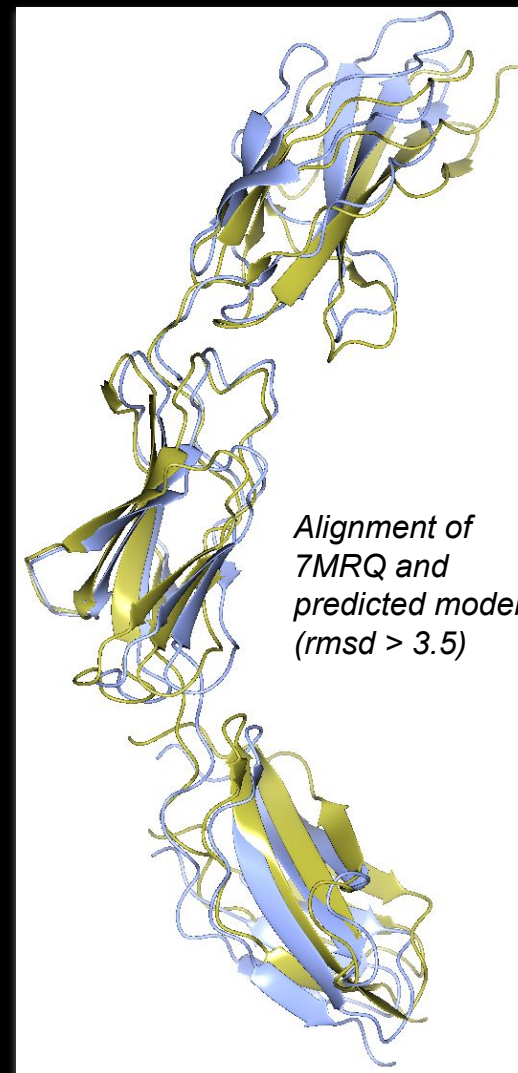
```
7115 *****
7116 *** Phaser Module: AUTOMATED MOLECULAR REPLACEMENT 2.8.3 ***
7117 *****
7118
7119 ** SINGLE solution
7120
7121 ** Solution written to PDB file: phaser_mr_output.1.pdb
7122 ** Solution written to MTZ file: phaser_mr_output.1.mtz
7123 Solution annotation (history):
7124 SOLU SET RFZ=6.7 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 200.5 31.0 100.9 FRAC -0.11 -0.20 -0.74 BFAC 16.00
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

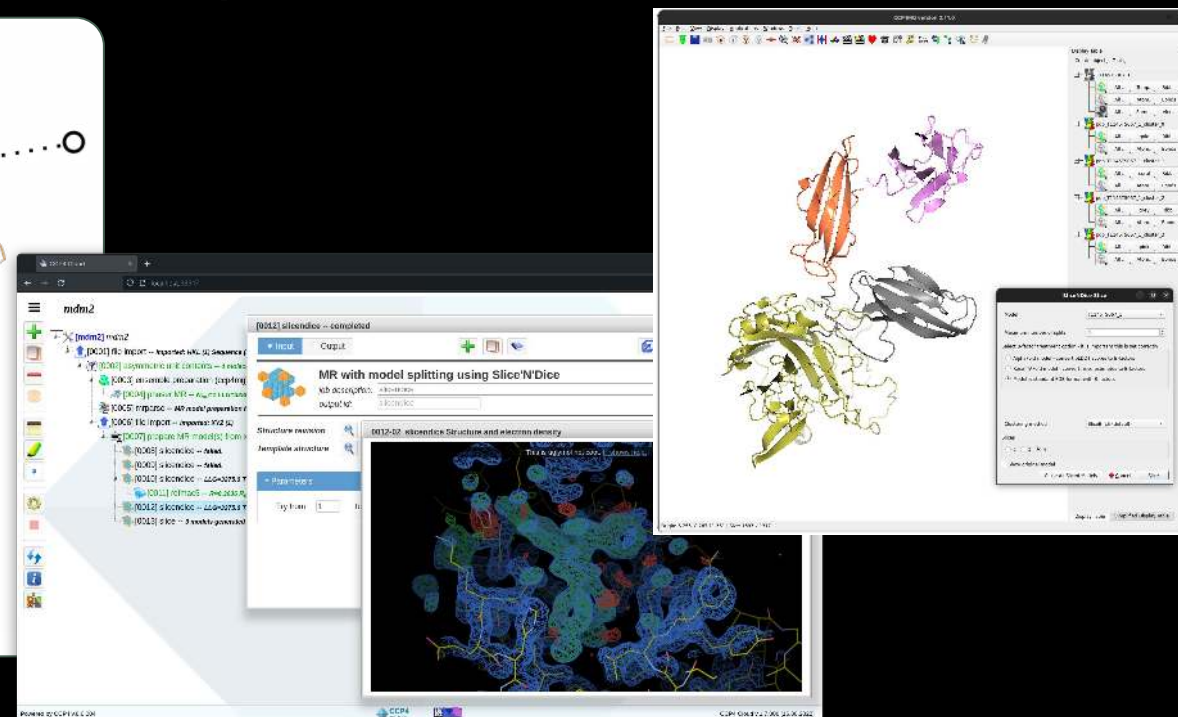
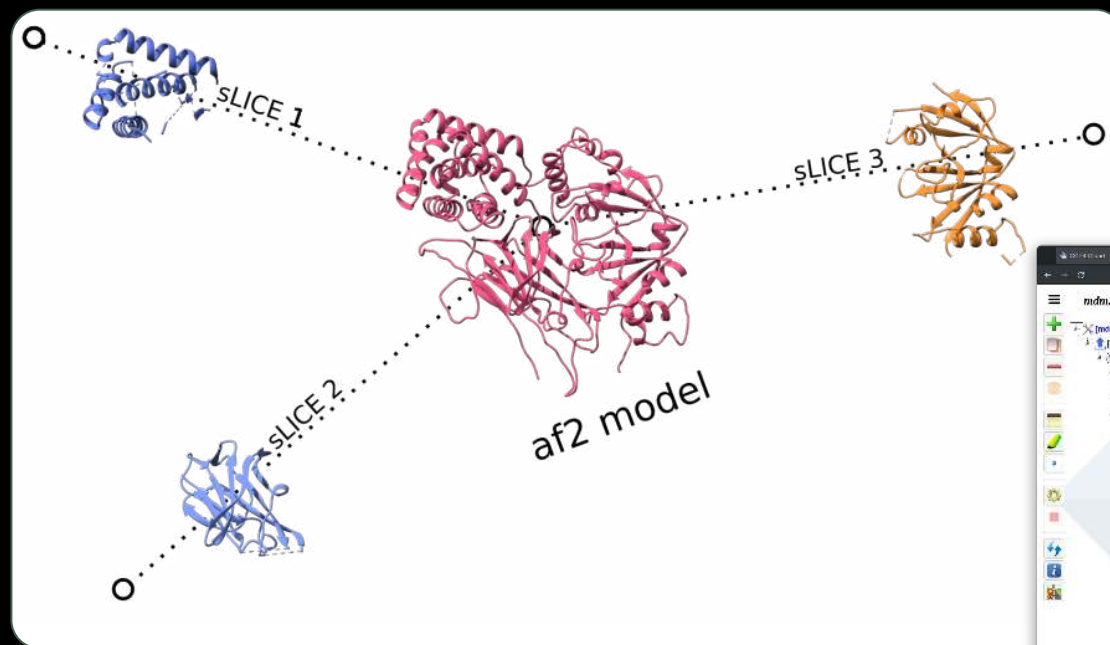
- The larger the target structure, the more likely there will be multiple domains
- A predicted model will often differ in the relative orientation of these domains when compared to the crystal structure
- Splitting a predicted search model into these domains can be a good strategy in MR



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

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

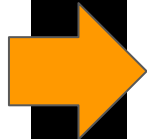
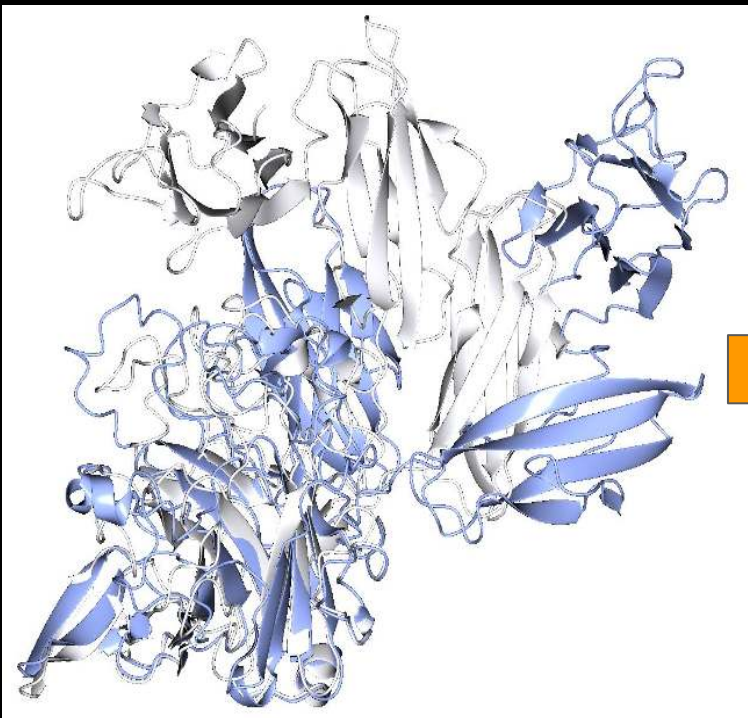
- Prediction (blue) aligned to target structure (white) on largest domain

Processing with *Slice'N'Dice*

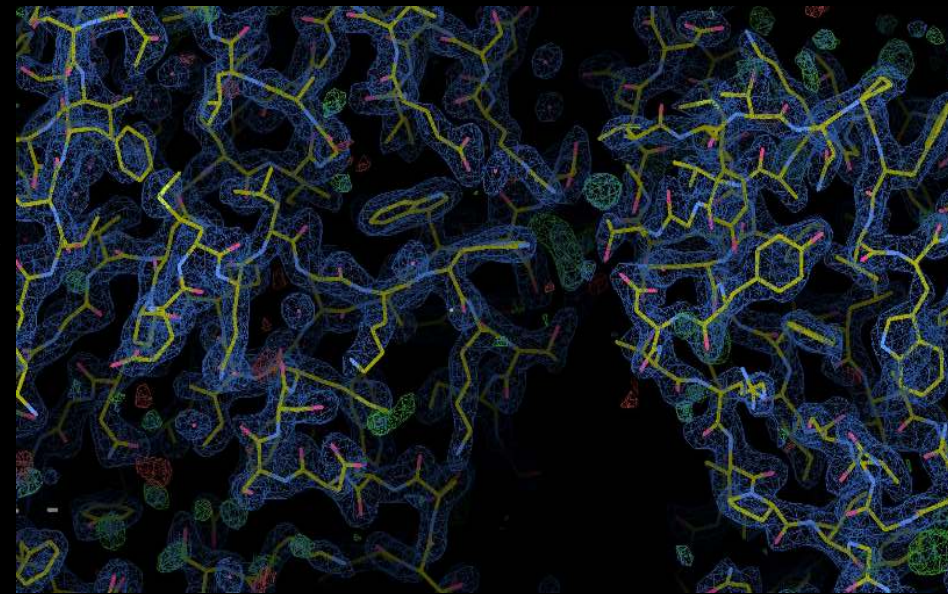
- 4-way Birch split of prediction
- Truncation to residues with a pLDDT > 70
- pLDDT converted to B-factor

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)

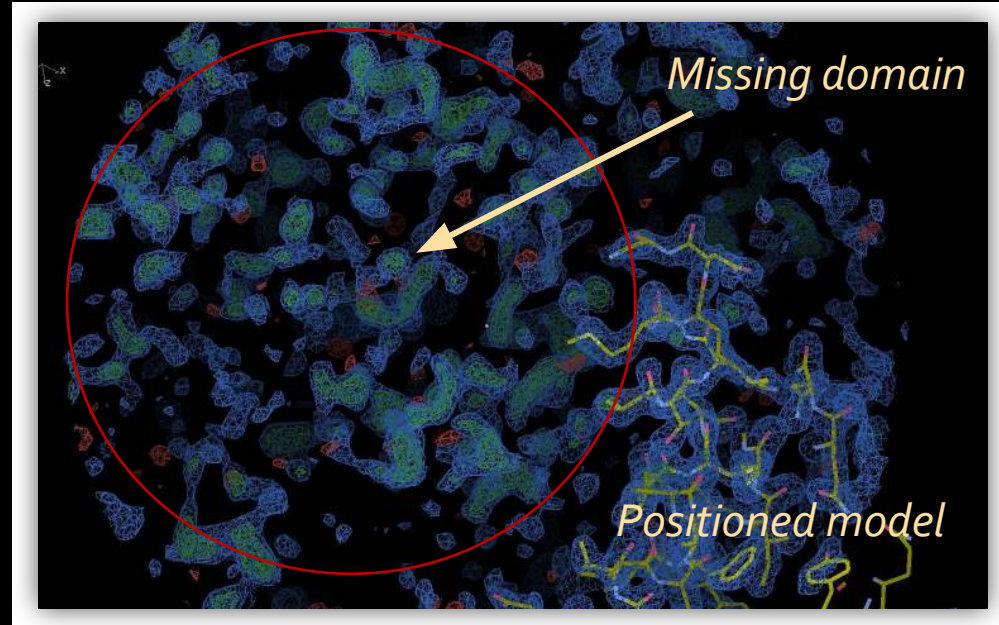


*(Domains aligned for
comparison with target)*



2 copies (P212121, 2.2 Angstroms)

Phased Translation search



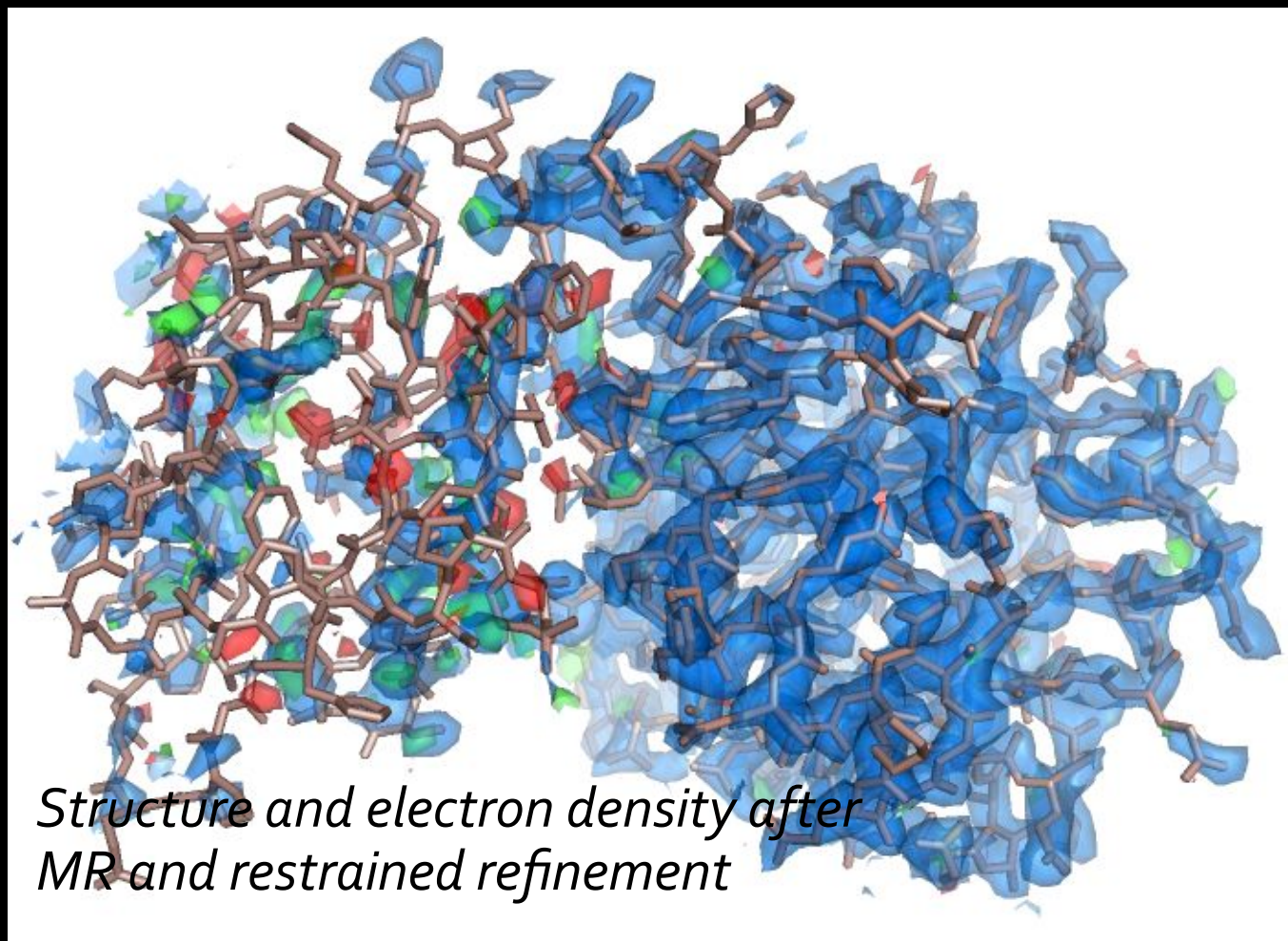
- Used when searching for several copies or dealing with a complex
- Available through MOLREP (3 protocols) and Phaser
- Often more successful than standard MR search approach particularly when looking for small domains

Phased Translation search and model Splitting

Phased translation search:

Example: 1tj3

Search model: 1s20, chain A



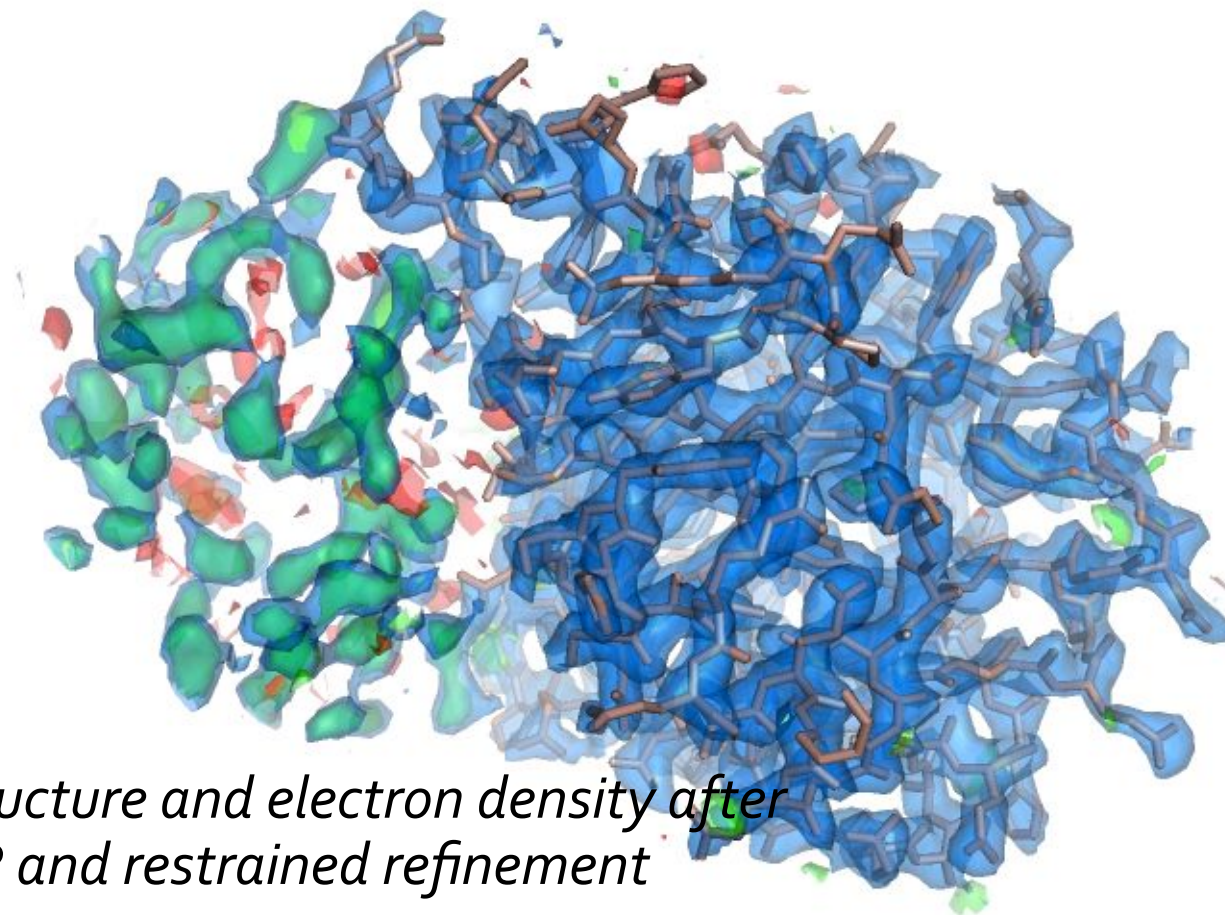
Phased Translation search and model Splitting

Phased translation search:

Example: 1tj3

Search model: 1s20, chain A

Large Domain only



*Structure and electron density after
MR and restrained refinement*

Phased Translation search and model Splitting

Phased translation search:

Example: 1tj3

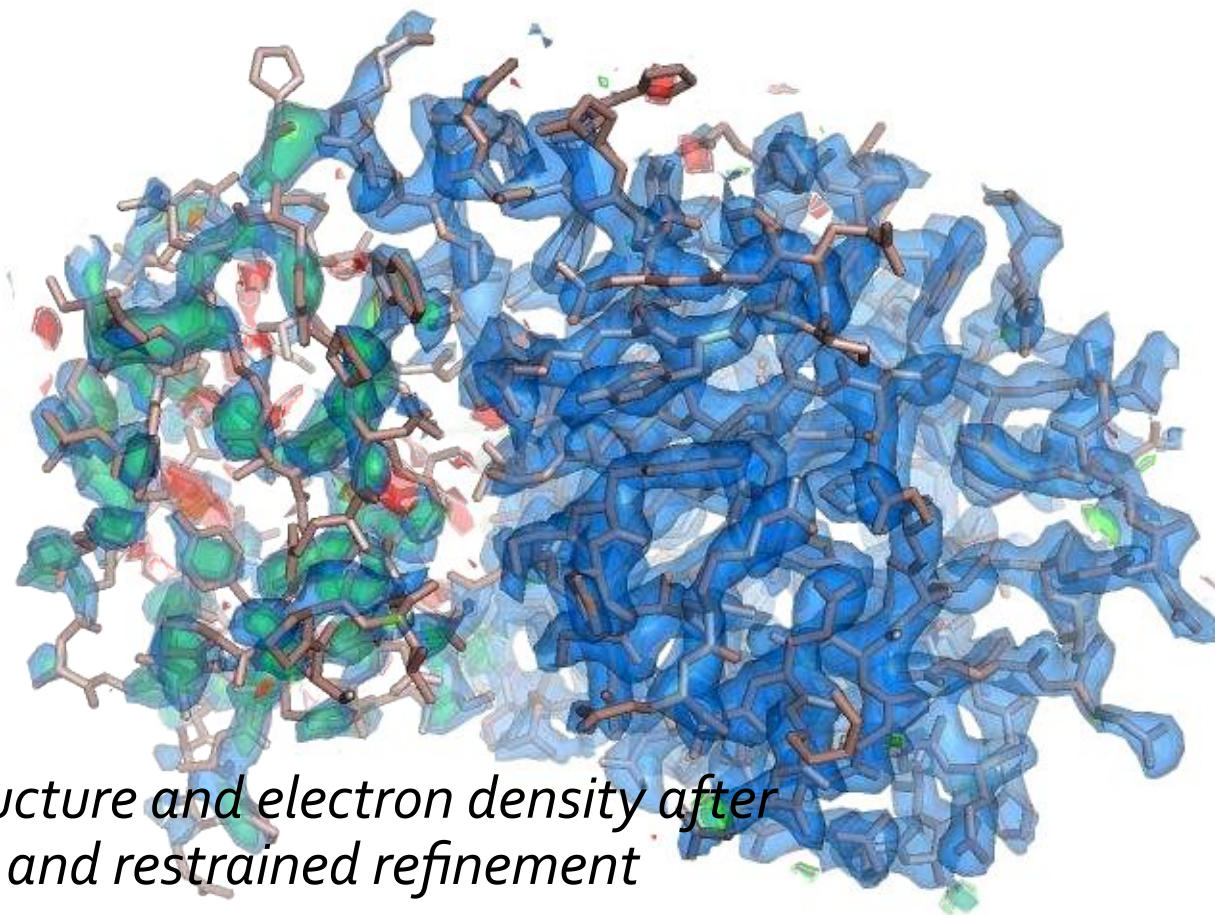
Search model: 1s20, chain A

Large Domain fixed

Small Domain search

Fit into density

(Molrep & Phaser)



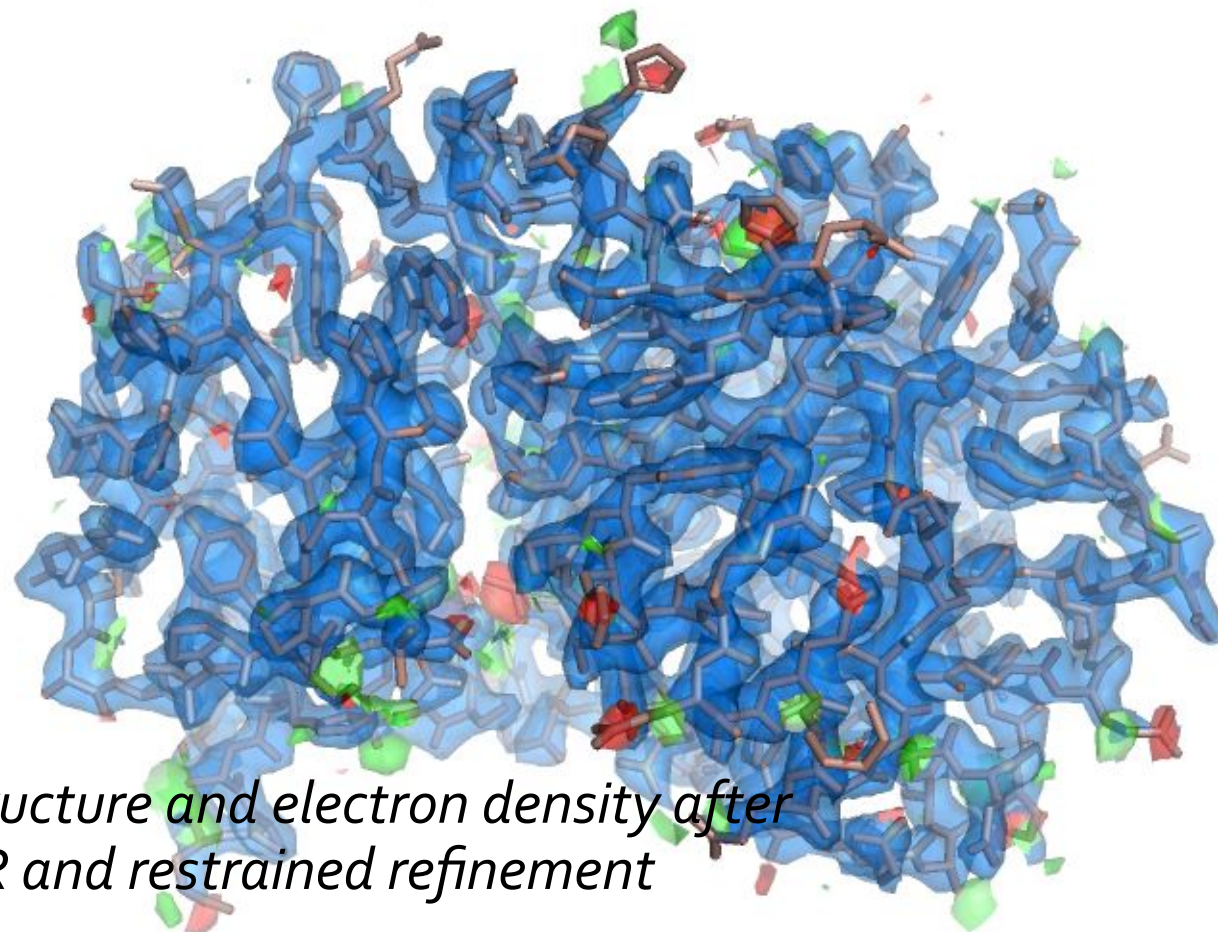
*Structure and electron density after
MR and restrained refinement*

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*

**Low confidence or inaccurate
predicted models**

Predicted model scoring:

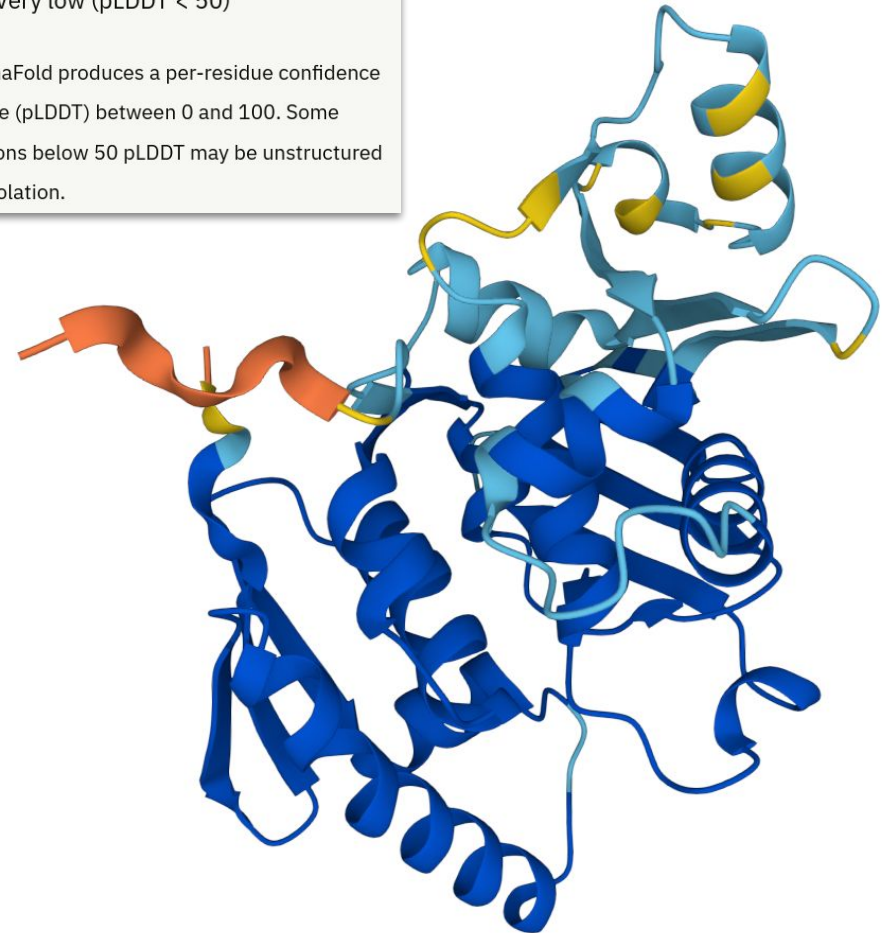
- Alphafold2 convention is a per-residue confidence score (pLDDT) between 0 and 100
- Newer convention is per-atom pLDDT
- Other prediction applications use same convention, although you may see fractional values (0 to 1.0) or r.m.s.d estimates

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

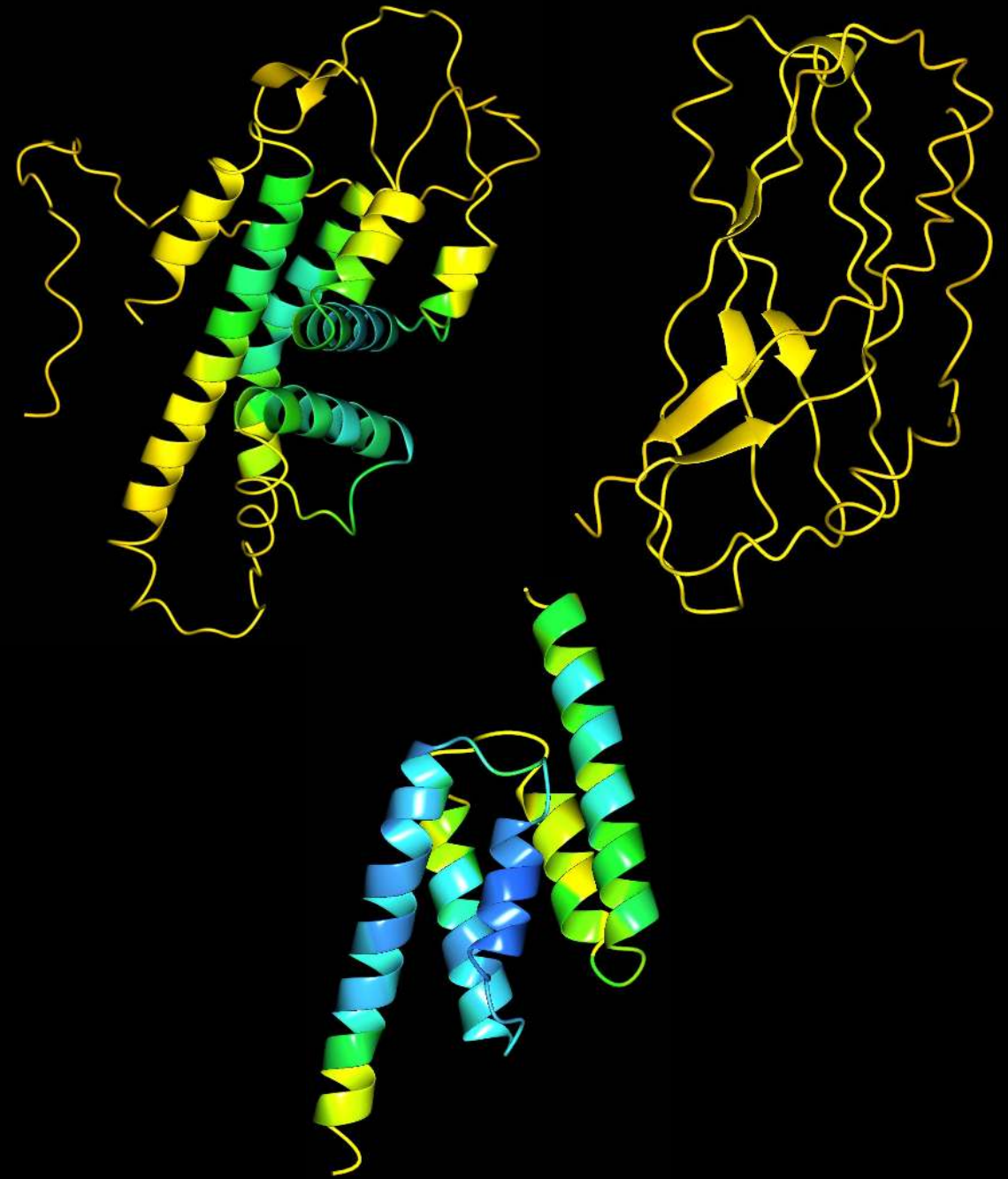
Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

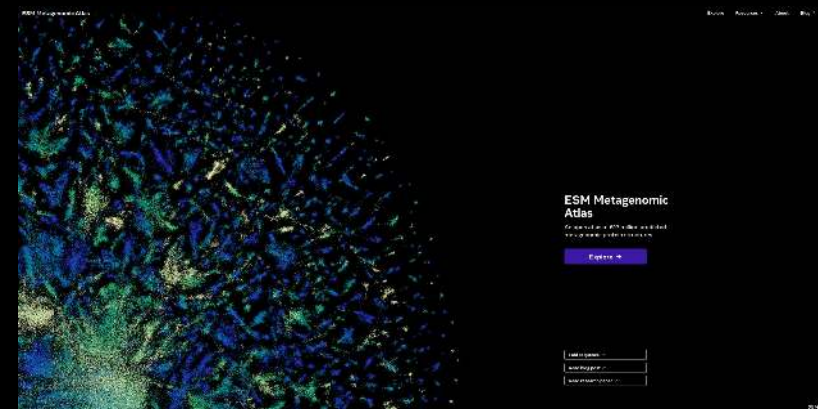
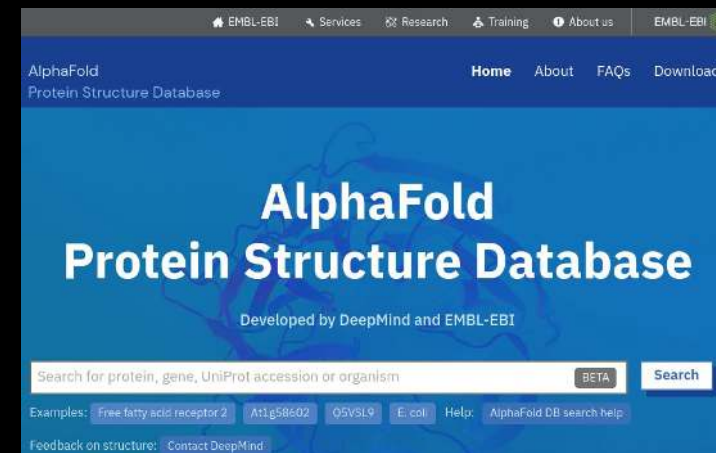


- Poor predictions will have low confidence scores (pLDDT<50) across much of the structure
- Low confidence predictions are nearly always poor MR search models
- High confidence predictions can sometimes differ significantly from the crystallised structure



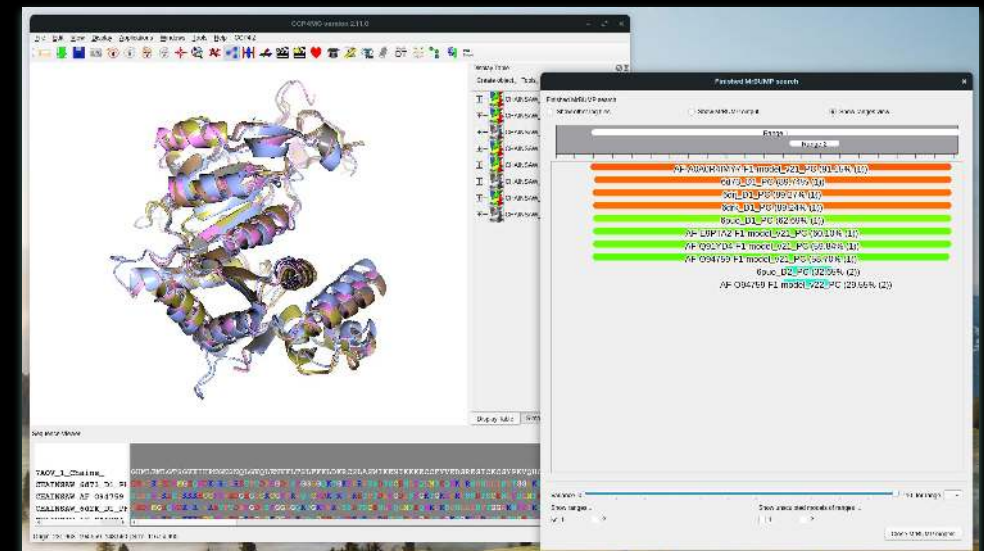
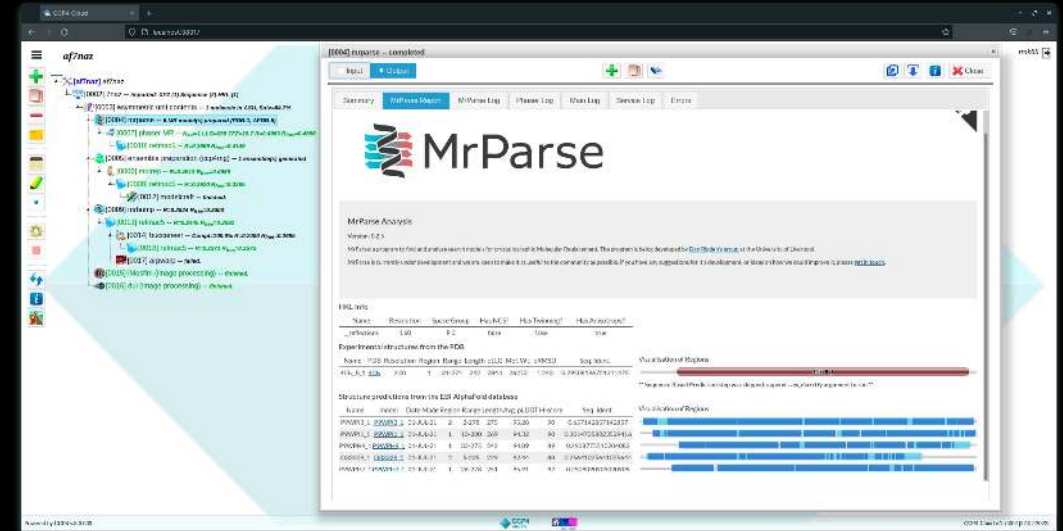
What to do in these cases?

- 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



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/>



Fragment Molecular Replacement with Arcimboldo-Lite

-- performs *ab initio* phasing using polyalanine helices or other single search fragments for MR



Fragment Molecular Replacement with Arcimboldo-Borges

-- performs *ab initio* phasing with nonspecific libraries of small folds



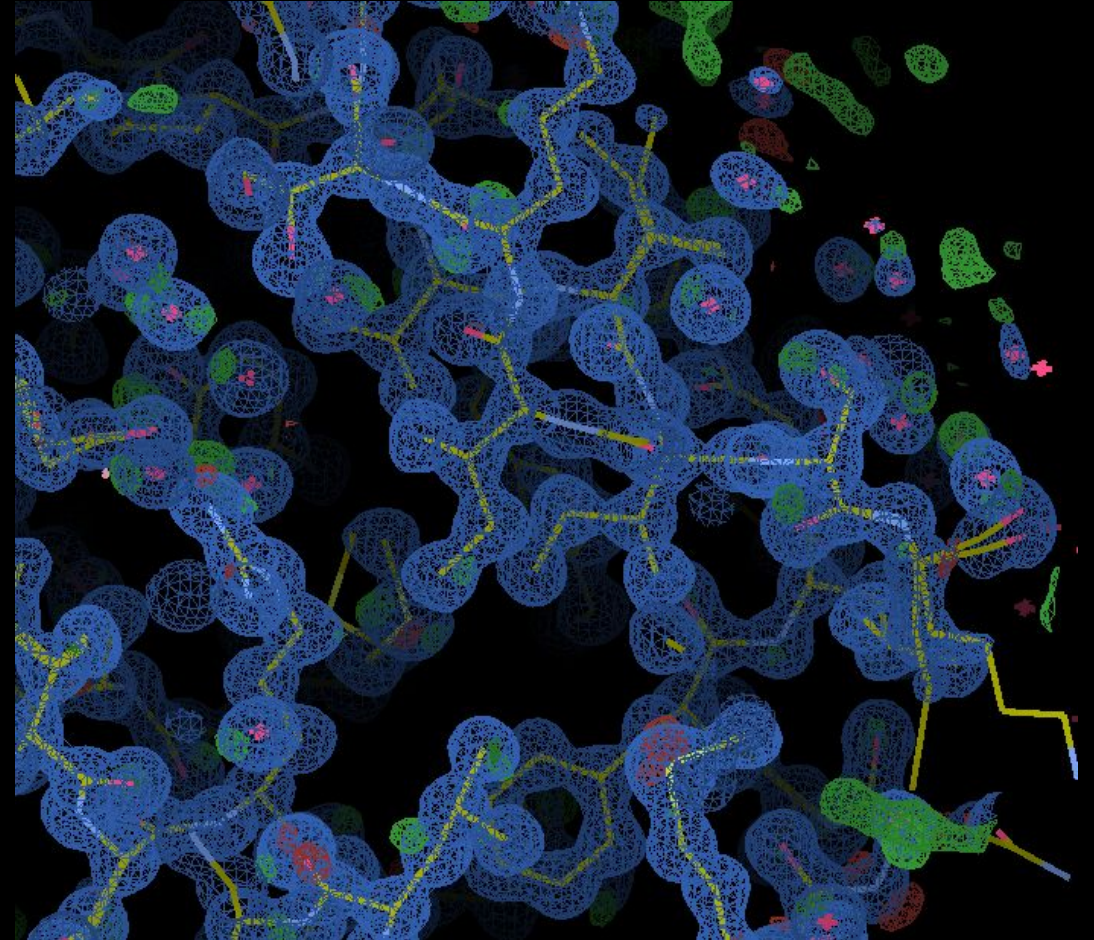
Fragment Molecular Replacement with Arcimboldo-Shredder

-- uses fragments derived from a distant homologue template for MR

Data resolution

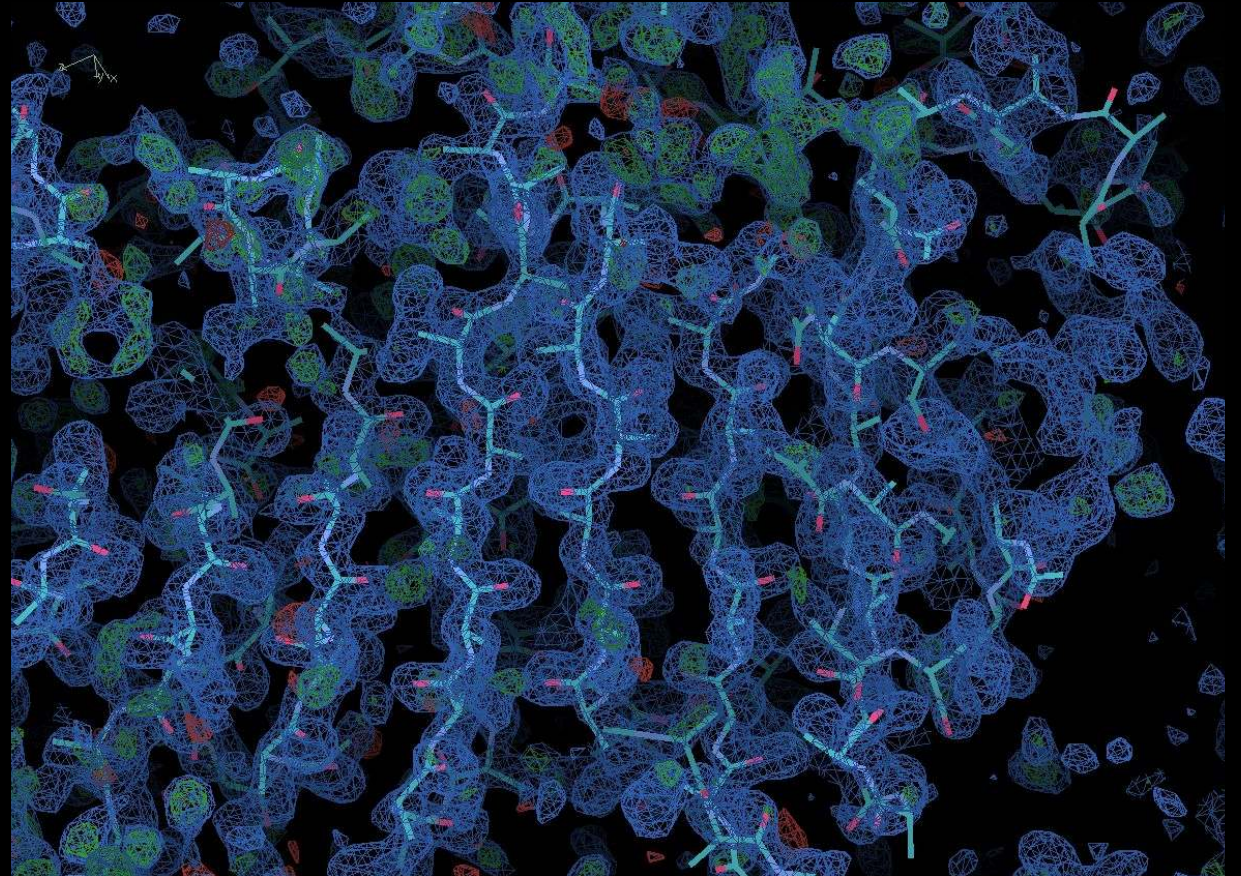
Resolution and Molecular Replacement

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



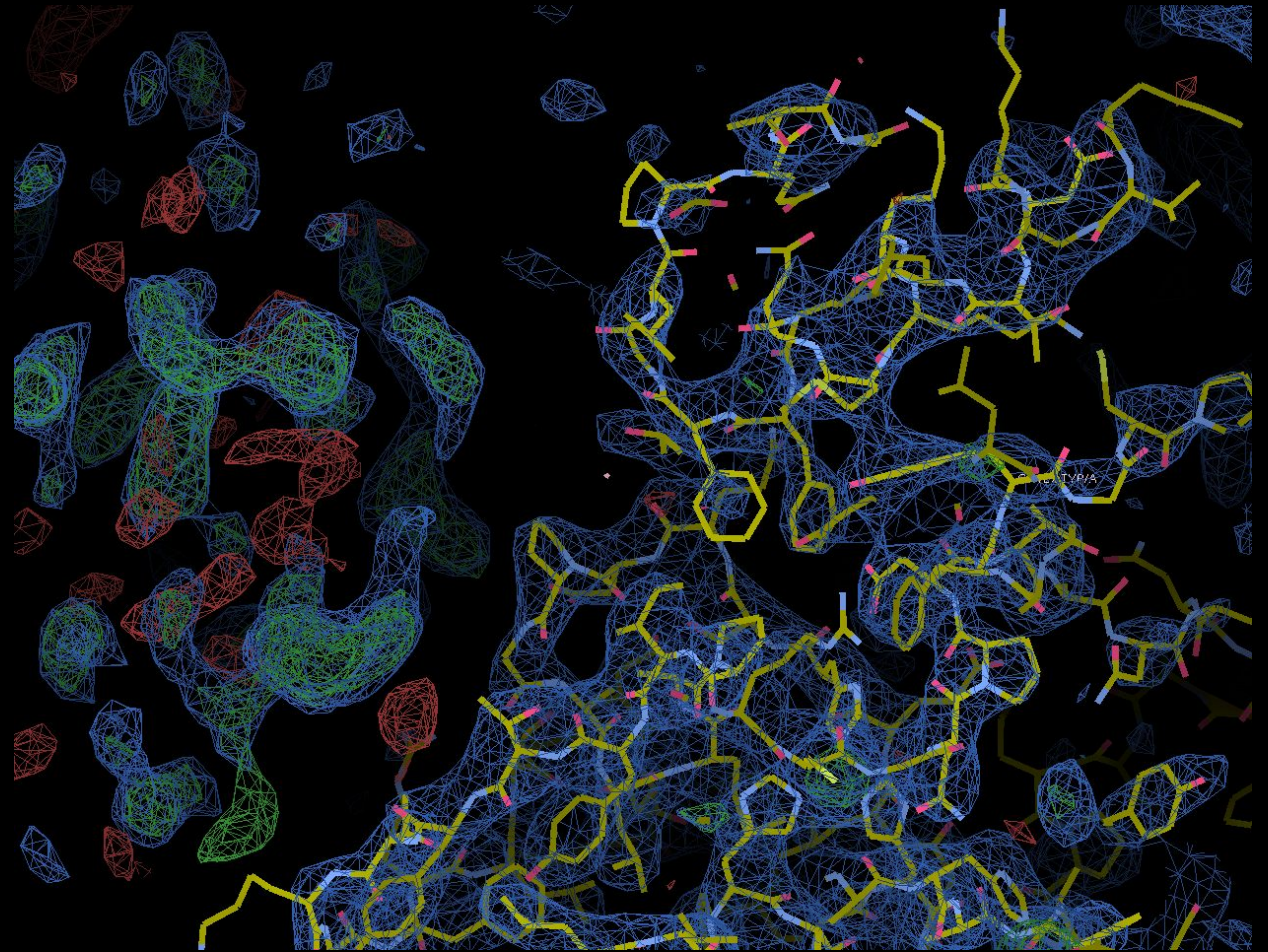
Resolution and Molecular Replacement

- 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



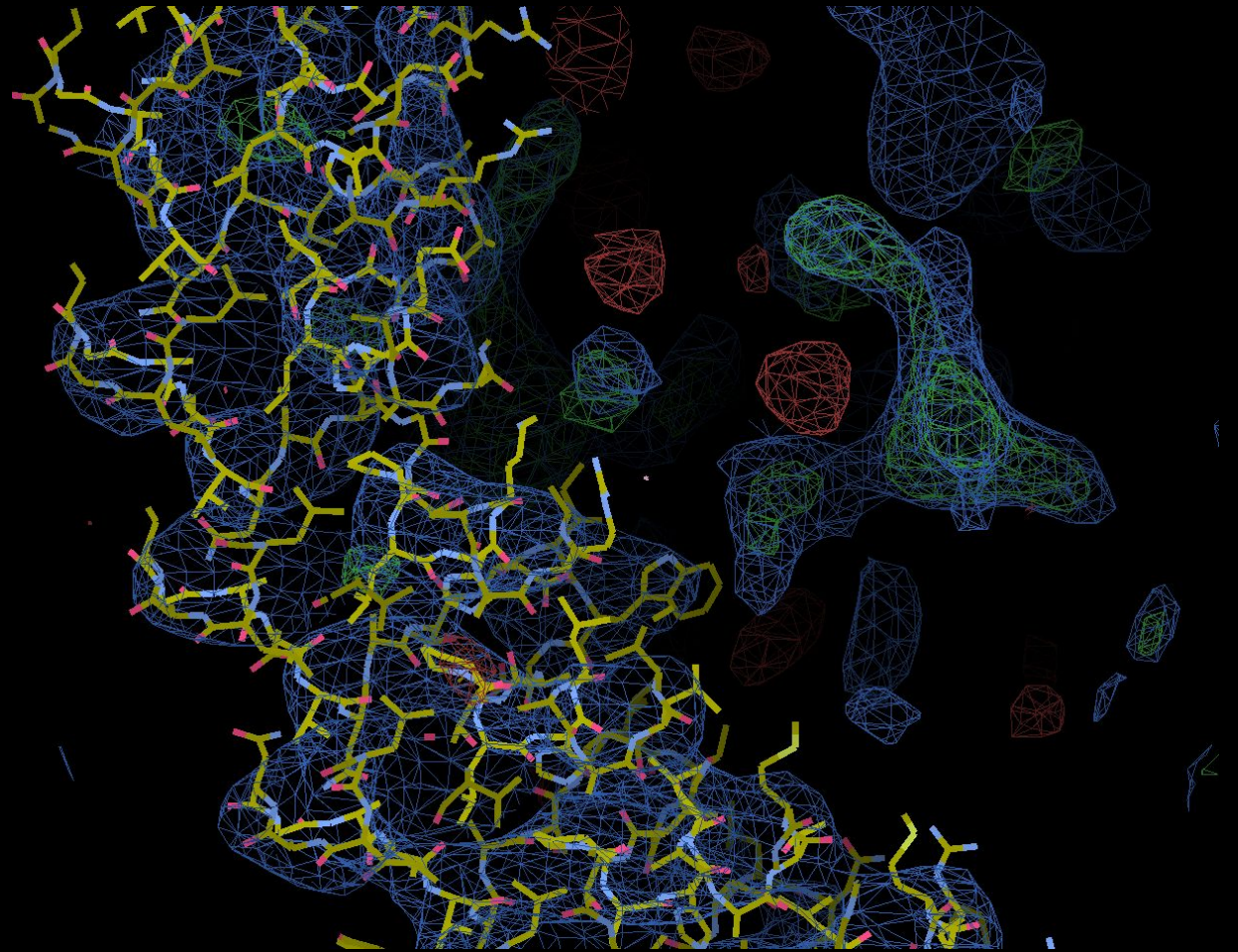
Resolution and Molecular Replacement

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



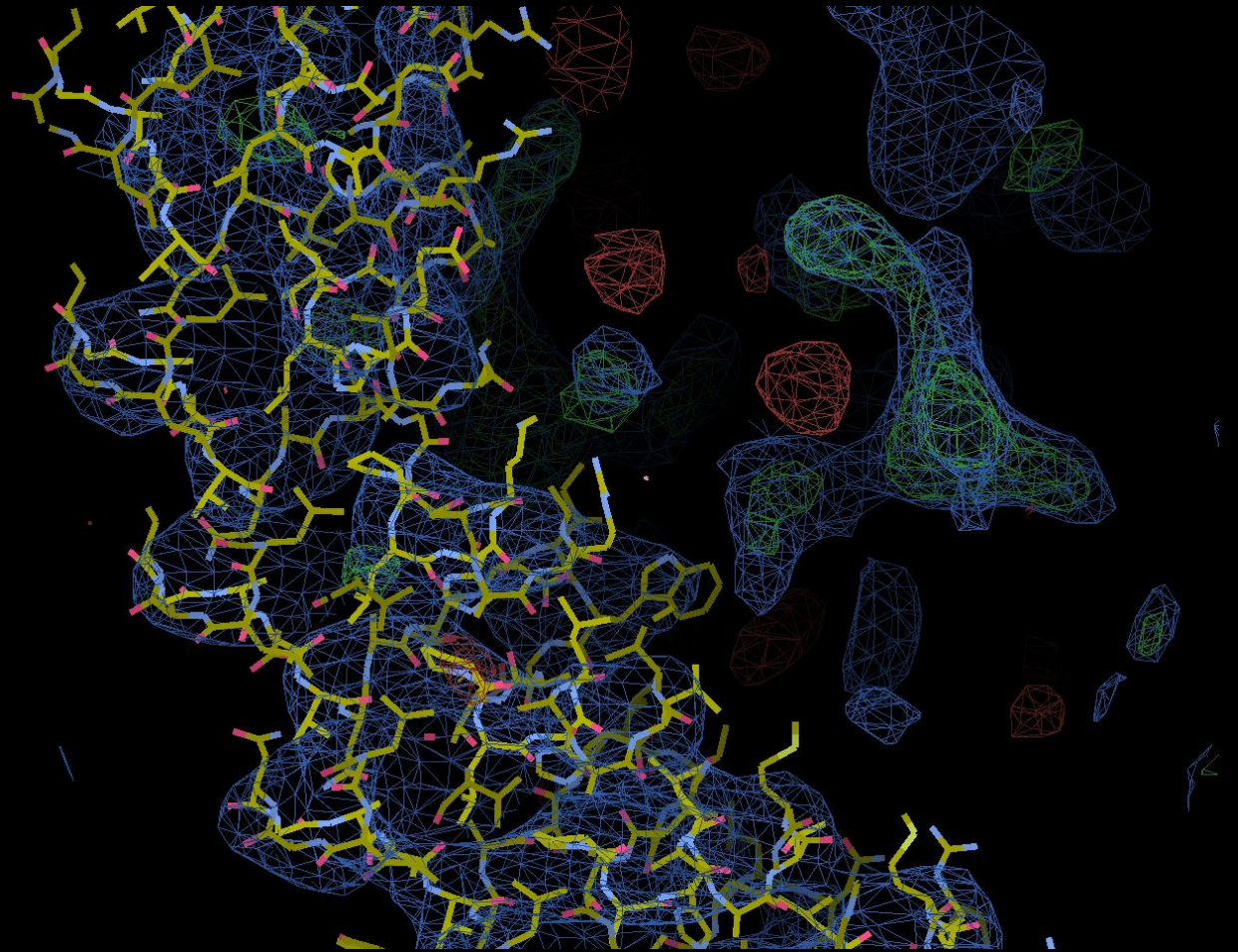
Resolution and Molecular Replacement

- Below 4Å automated model building becomes difficult



Resolution and Molecular Replacement

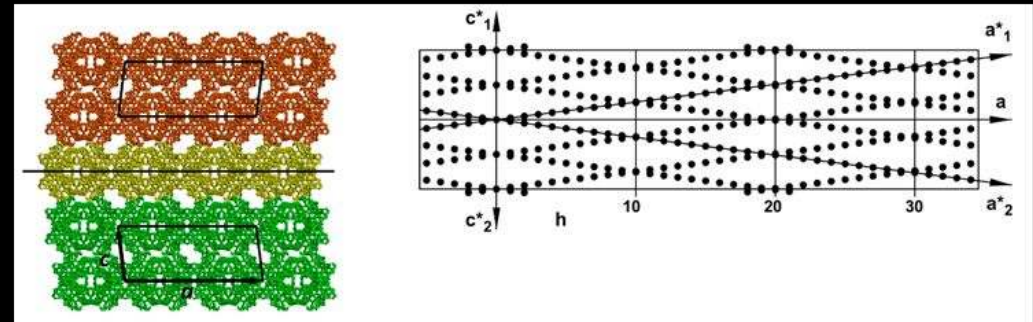
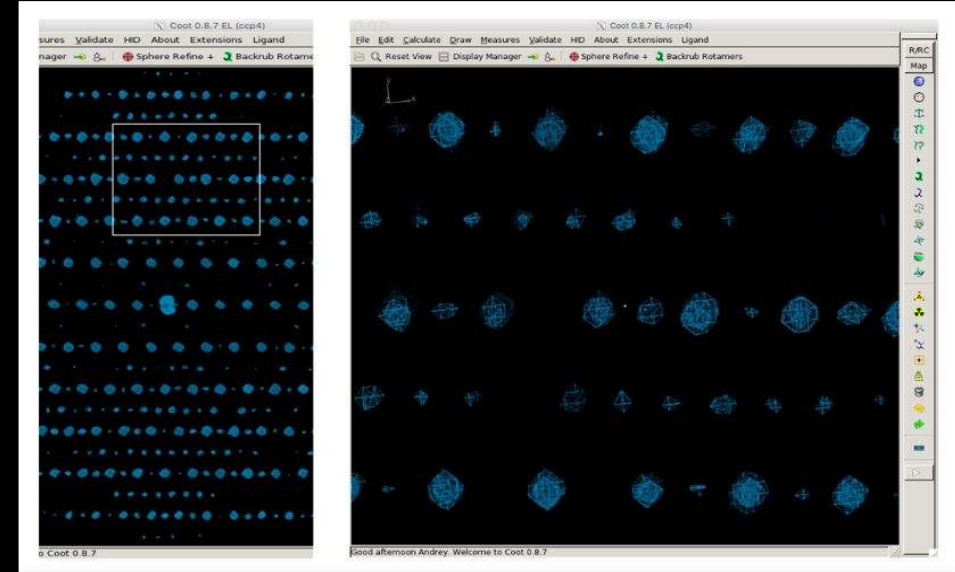
- Below 4Å automated model building becomes difficult
- However, given suitable search models, MR can be used to build up a model structure



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



Crystal contamination

- **SIMBAD: *Sequence-less MR***

1. nearest cell (lattice)
2. contaminants
3. entire non-redundant domain database (90000 models)

The image displays the SIMBAD Results web interface and the Coot graphical map viewer. The SIMBAD Results window shows the CCP4 on-line logo, SIMBAD Results title, and navigation tabs for Log file, Summary, and Lattice Parameter Search Results. The Summary tab is active, displaying the SIMBAD Summary and Best SIMBAD result Downloads. The summary text states: "The best search model found by SIMBAD was 1SMU. This gave an R/Rfact of 0.267 and an R/Rfree of 0.258. An R/Rfree lower than 0.450 is indicative of a solution. Values above this may also be indicative of a correct solution but you should examine the maps through the graphical map viewer for verification". The Downloads section lists four files for Electron density for 1SMU: 1SMU_refinement_output.pdb, 1SMU_refinement_output.mtz, 1SMU_refmac_2fofcwt.map, and 1SMU_refmac_fofcwt.map, each with an Export button. The Best SIMBAD result Log Files section is also visible. The Coot 0.8.8 EL (ccp4) window is overlaid on the bottom right, showing a 3D map of the protein structure with a blue mesh and yellow sticks. The Coot menu bar includes File, Edit, Calculate, Draw, Measures, Validate, HID, About, Extensions, and Ligand. The Refine/Regularize Control... menu is open, showing options like Real Space Refine Zone, Regularize Zone, Fixed Atoms..., Rigid Body Fit Zone, Rotate Translate, Auto Fit Rotamer, Rotamers..., Edit Chi Angles, Torsion General, Flip Peptide, Sidechain 180° Flip, Edit Backbone Torsions, Mutate & Auto Fit..., and Simple Mutate... The status bar at the bottom of Coot shows the file path: ...latt/mr_search/1SMU/mr/molrep/refine/1SMU_refinement_output.pdb. Molecule nu...

