

Molecular Replacement: *Assessing and improving the solution*

DLS CCP4 Workshop - 2022
Ronan Keegan CCP4 Group



UNIVERSITY OF
LIVERPOOL

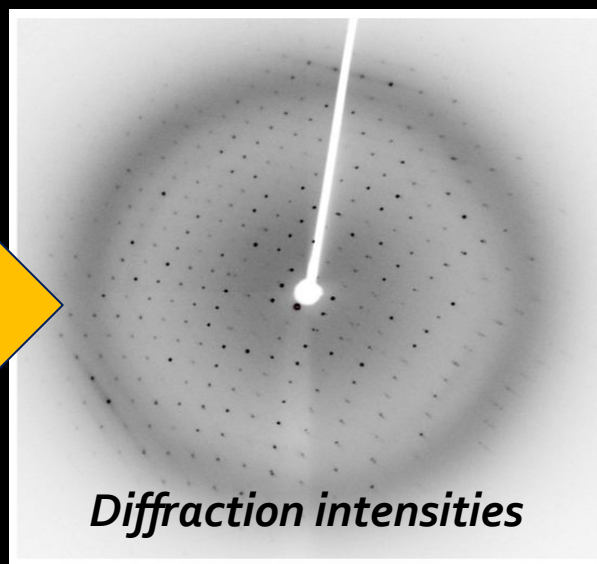
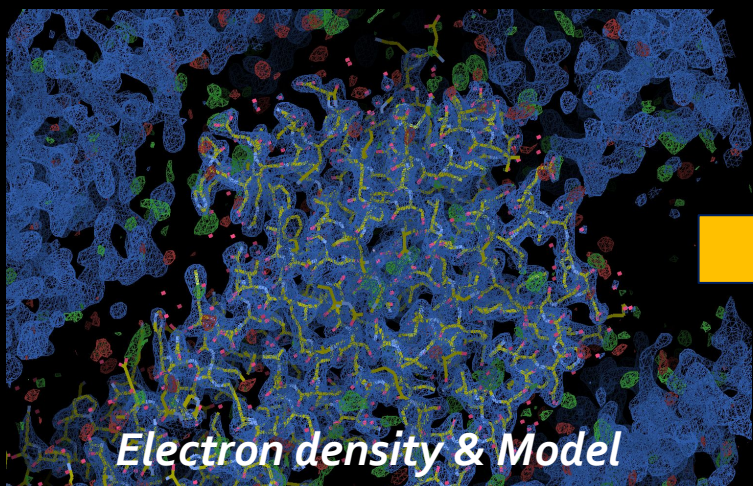
The Phase Problem

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

Electron density
at (x,y,z)

Amplitudes

Phases



?

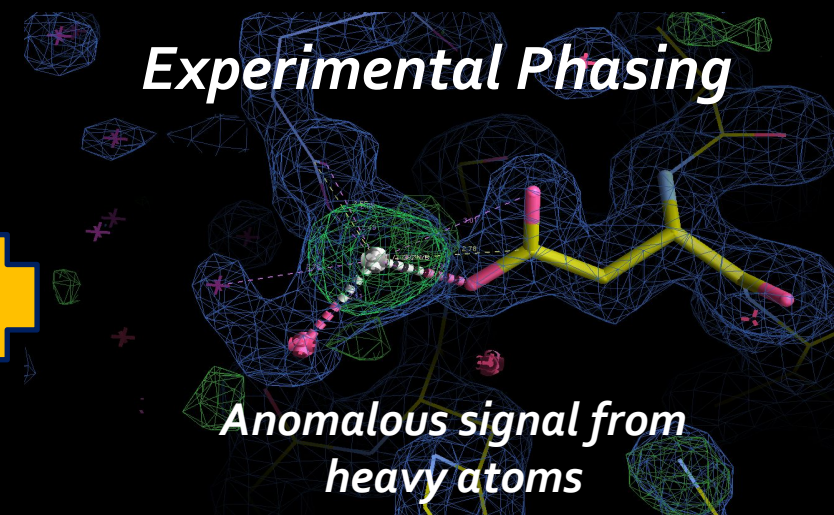
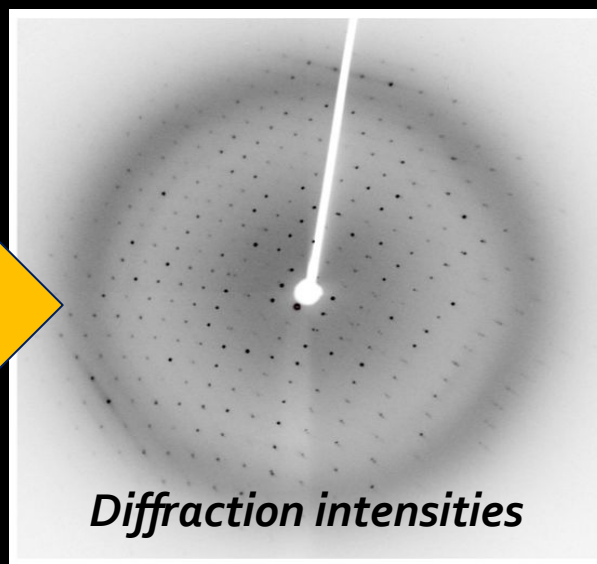
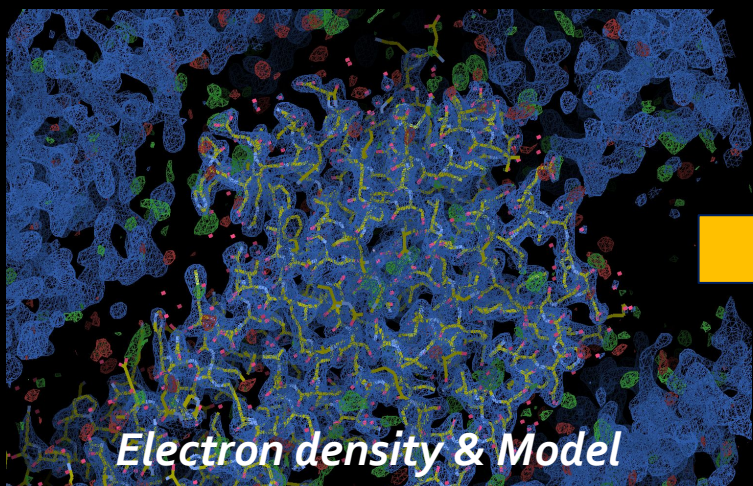
The Phase Problem: Experimental Phasing

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

Electron density
at (x,y,z)

Amplitudes

Phases



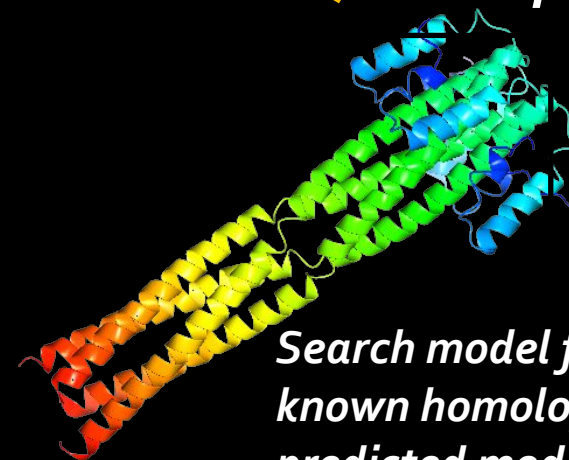
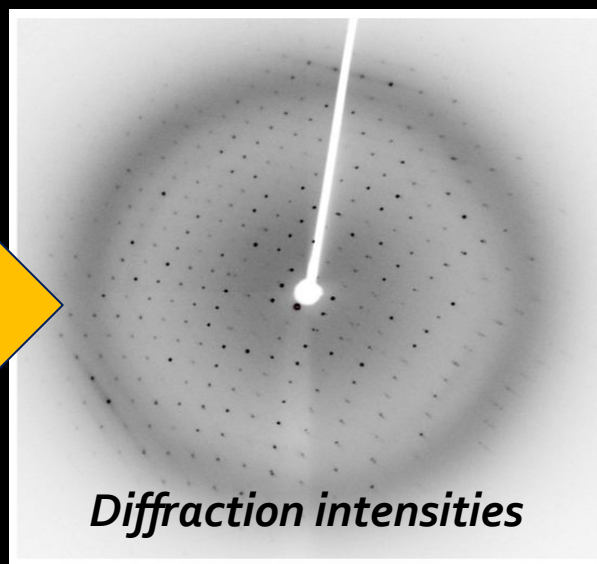
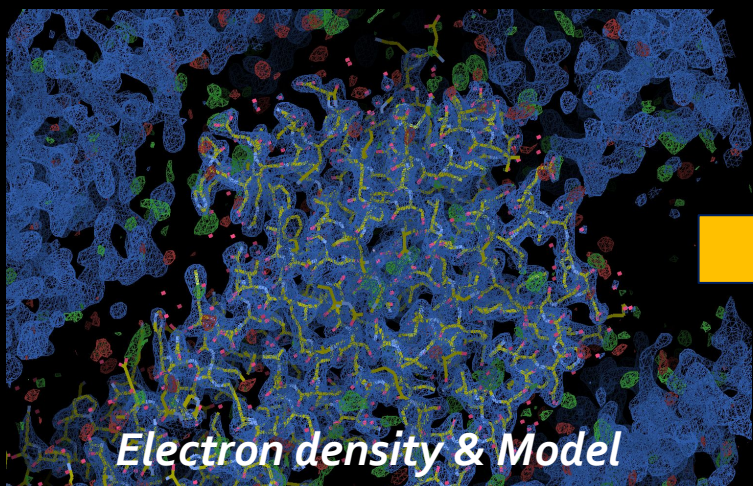
The Phase Problem: Molecular Replacement

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

Electron density
at (x,y,z)

Amplitudes

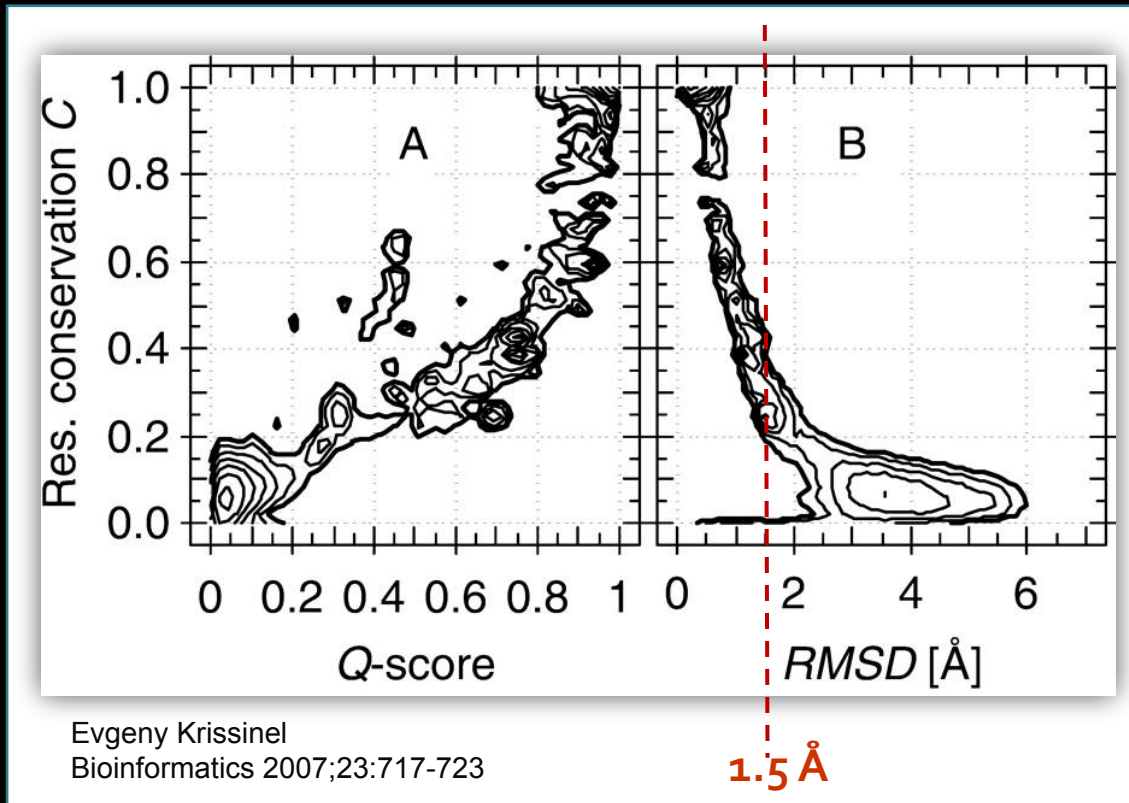
Molecular
Replacement



Search Models

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

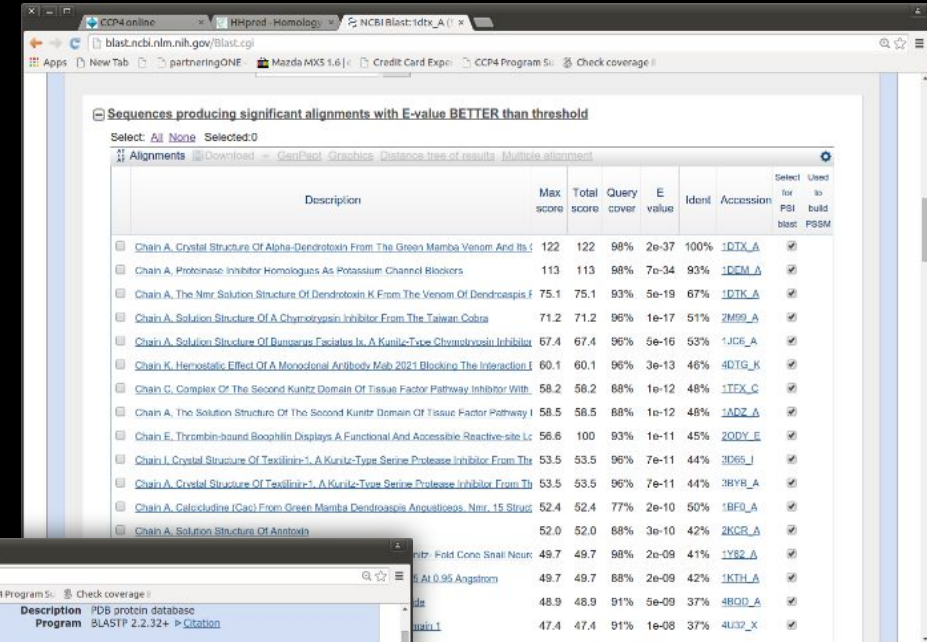
- How to find a search model?
 - Amino acid sequence similarity often correlates well with structural similarity



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

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%

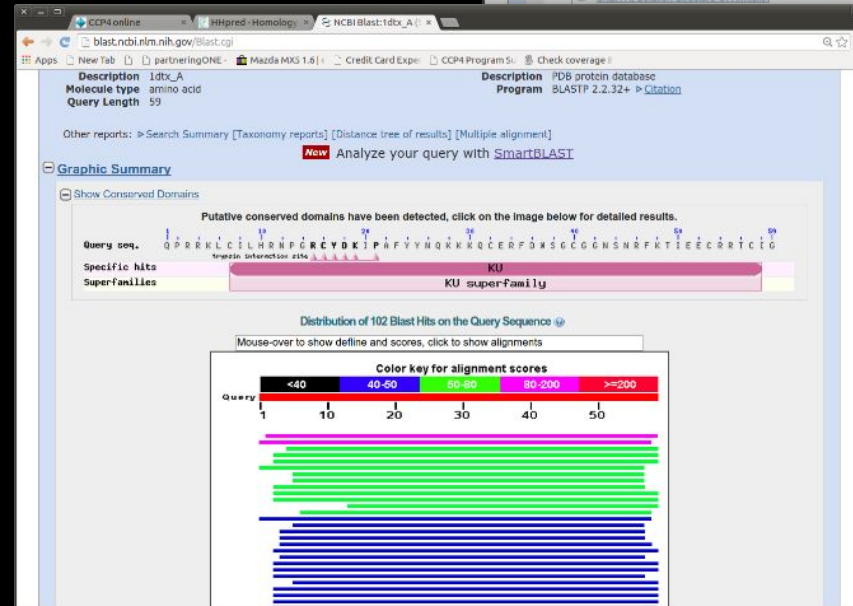


Sequences producing significant alignments with E-value BETTER than threshold

Select: All None Selected: 0

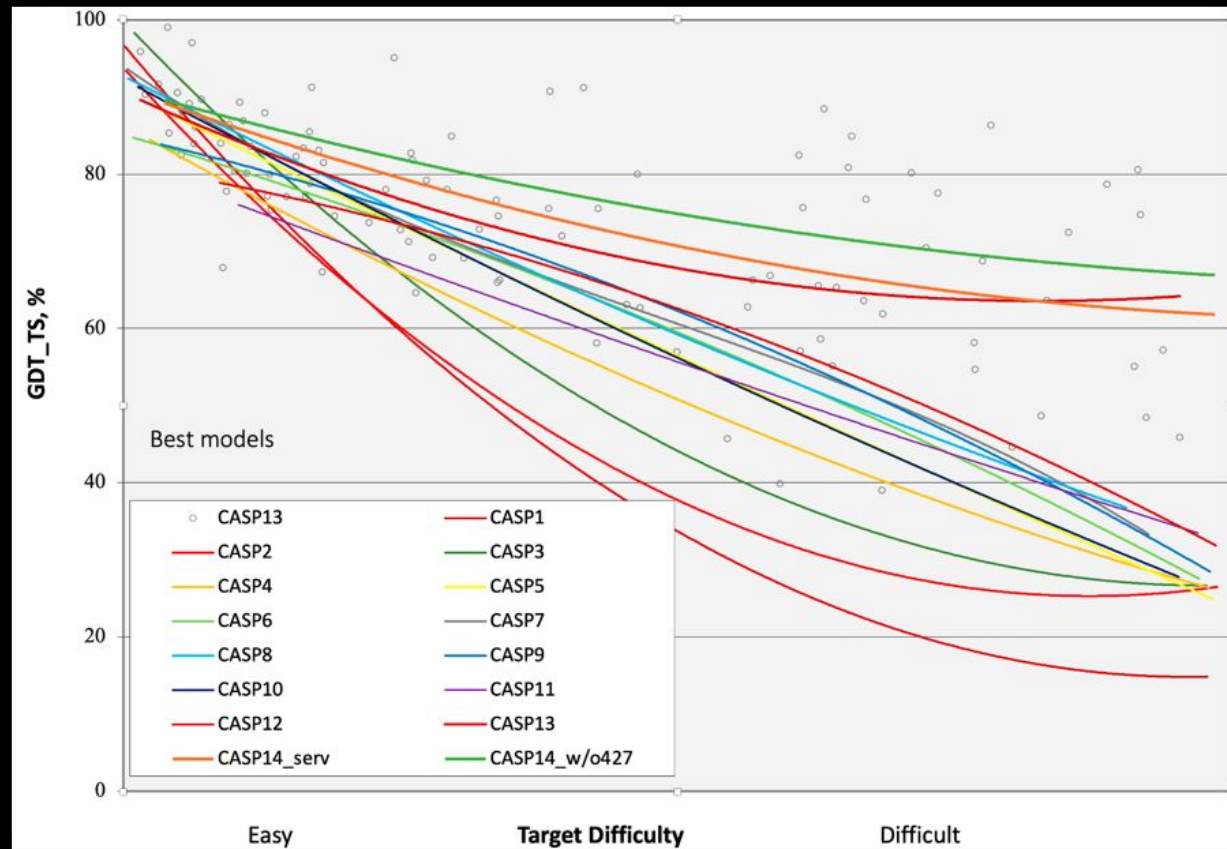
Alignments Download Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSM
Chain A, Crystal Structure Of Alpha-Dendrotoxin From The Green Mamba Venom And Its I	122	122	98%	2e-37	100%	1DTX_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Proteinase Inhibitor Homologues As Potassium Channel Blockers	113	113	98%	7e-34	93%	1DEM_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, The New Solution Structure Of Dendrotoxin K From The Venom Of Dendroaspis F	75.1	75.1	93%	5e-19	67%	1DTK_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of A Chymotrypsin Inhibitor From The Taiwan Cobra	71.2	71.2	96%	1e-17	51%	2M99_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of Bumpus Facialis Is A Kunitz-Type Chymotrypsin Inhibitor	67.4	67.4	96%	5e-16	53%	1JCE_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain K, Hemostatic Effect Of A Monoclonal Antibody Mab 2021 Blocking The Interaction I	60.1	60.1	96%	3e-13	46%	4DTG_K	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain C, Complex Of The Second Kunitz Domain Of Tissue Factor Pathway Inhibitor With	58.2	58.2	88%	1e-12	48%	1TFX_C	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, The Solution Structure Of The Second Kunitz Domain Of Tissue Factor Pathway I	58.5	58.5	88%	1e-12	48%	1ADZ_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain E, Thrombin-bound Bosphin Displays A Functional And Accessible Reactive-site Lc	56.6	100	93%	1e-11	45%	2ODY_E	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain J, Crystal Structure Of Tassinin-1, A Kunitz-Type Serine Protease Inhibitor From The	53.5	53.5	96%	7e-11	44%	3D65_J	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Crystal Structure Of Tassinin-1, A Kunitz-Type Serine Protease Inhibitor From The	53.5	53.5	96%	7e-11	44%	3HYH_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Calcicludine (Gsc) From Green Mamba Dendroaspis Angusticeps, Nerv. 15 Struc	52.4	52.4	77%	2e-10	50%	1BF0_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of Annotoxin	52.0	52.0	88%	3e-10	42%	2KCB_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of Annotoxin	49.7	49.7	98%	2e-09	41%	1Y82_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of Annotoxin	49.7	49.7	88%	2e-09	42%	1KTH_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of Annotoxin	48.9	48.9	91%	5e-09	37%	4BQD_A	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Chain A, Solution Structure Of Annotoxin	47.4	47.4	91%	1e-08	37%	4U37_X	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>



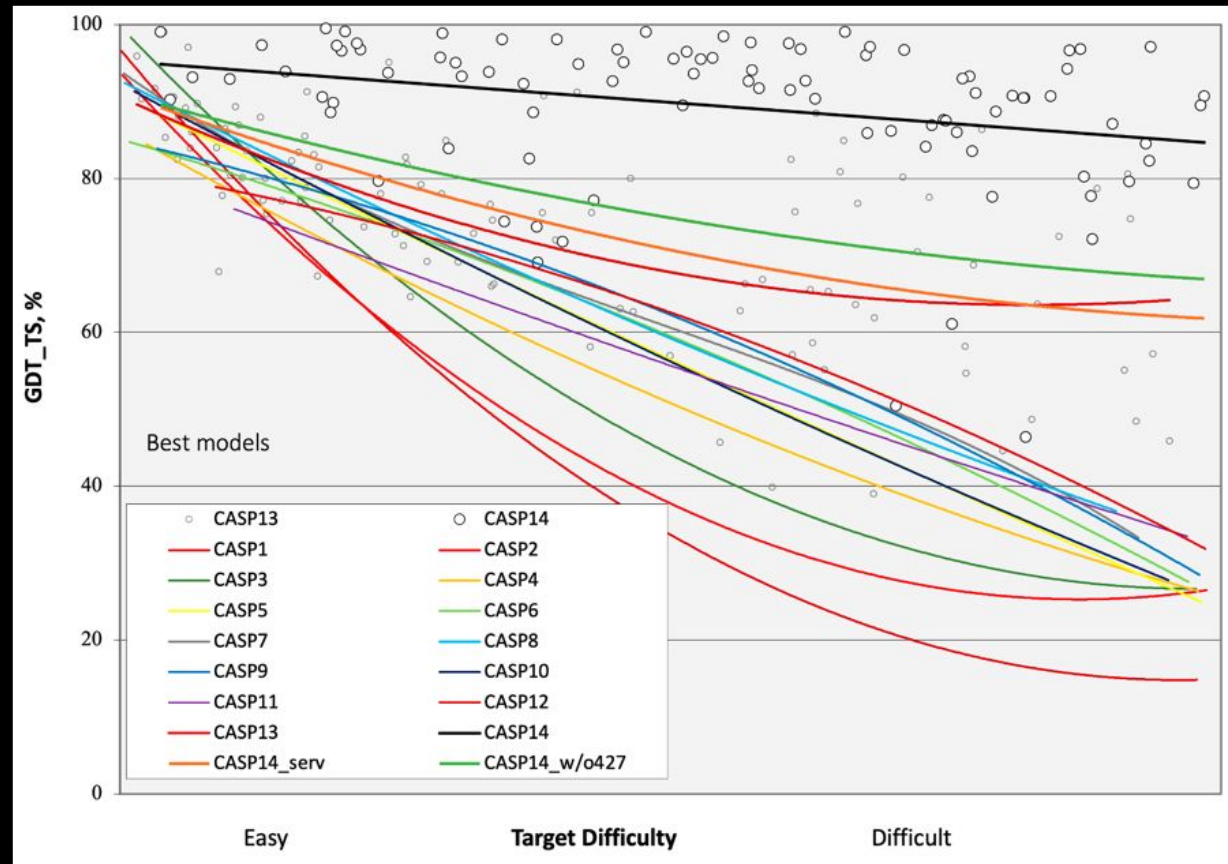
After CASP14 (2021): *Finding a search model*

CASP: Critical Assessment of Techniques for Protein Structure Prediction



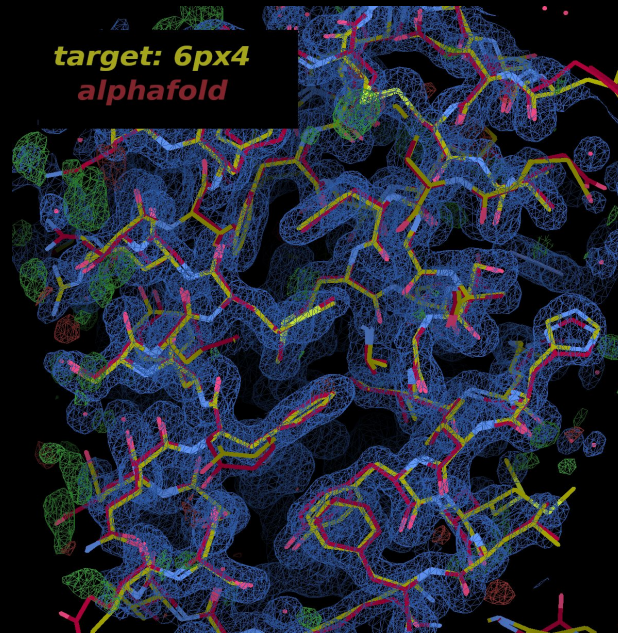
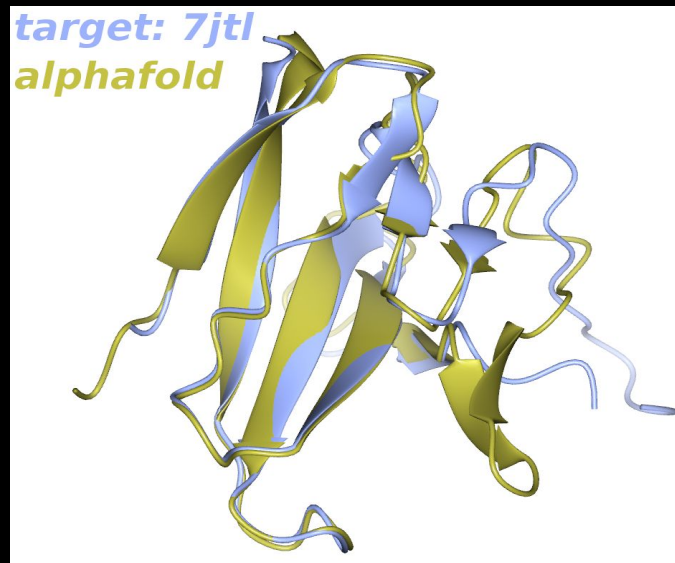
After CASP14 (2021): *Finding a search model*

CASP: Critical Assessment of Techniques for Protein Structure Prediction



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

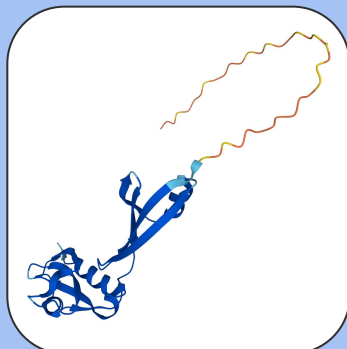
CASP: Critical Assessment of Techniques for Protein Structure Prediction



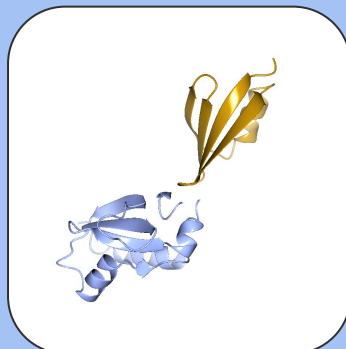
Molecular Replacement using predicted models

3 Step Process

Generate or
obtain
predicted
model

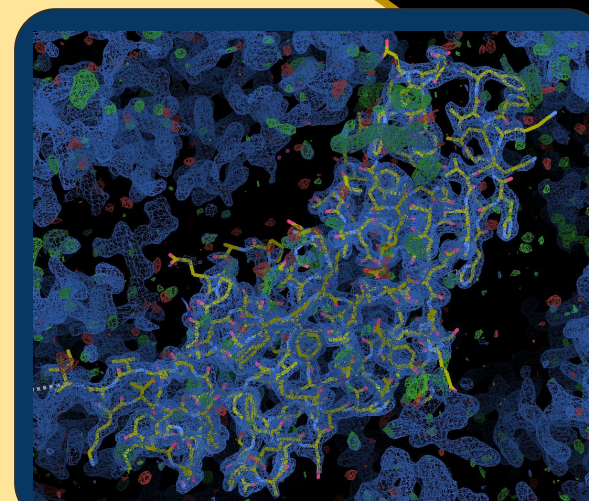


Process
predicted
model



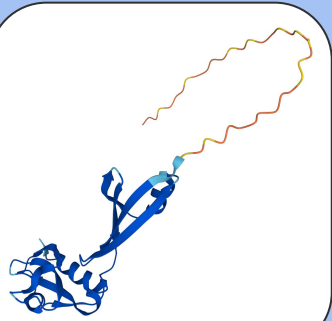
Manual or
automated
MR

```
[sw2022] zorn
  [0001] file import -- imported: HKL (1) XYZ (1) Sequence (1)
  [0002] asymmetric unit contents -- 1 molecule in ASU, Solv=35.8%
  [0003] structure prediction -- 1 structure predicted
  [0004] AlphaFold MR workflow -- received HKL (1), Sequences (1); workflow started
  [0005] AlphaFold MR workflow -- 1 structure predicted
  [0006] AlphaFold MR workflow -- prepare MR model(s) from xyz -- 1 model(s) generated (molrep protocol)
  [0007] AlphaFold MR workflow -- asymmetric unit contents -- 1 molecule in ASU, Solv=35.8%
  [0008] AlphaFold MR workflow -- phaser MR -- finished
  [0009] slice -- 2 model(s) generated
  [0010] molrep -- R=0.4996 Rfree=0.4787
  [0011] remlac5 -- R=0.4671 Rfree=0.4462
  [0012] molrep -- R=0.4827 Rfree=0.4599
  [0013] remlac5 -- R=0.2709 Rfree=0.3219
  [0014] sarservice -- LL=0.702, 0.712, 0.713 R=0.3415 Rfree=0.3367
  [0015] remlac5 -- R=0.2888 Rfree=0.3188
```



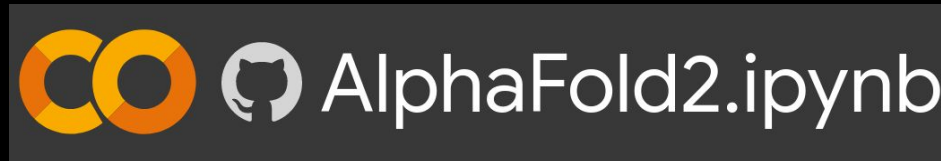
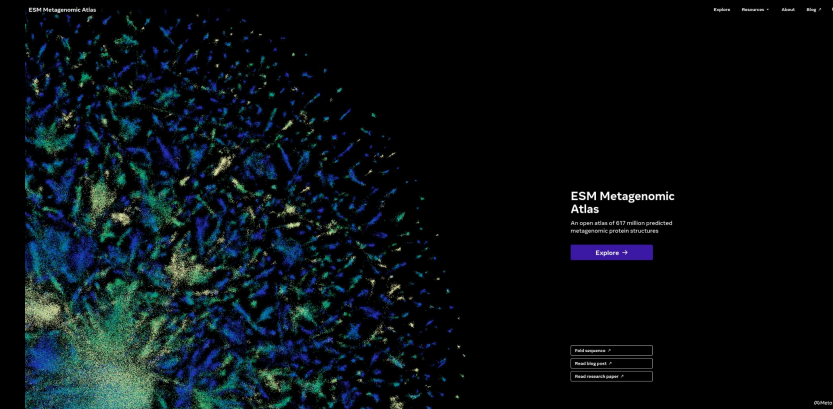
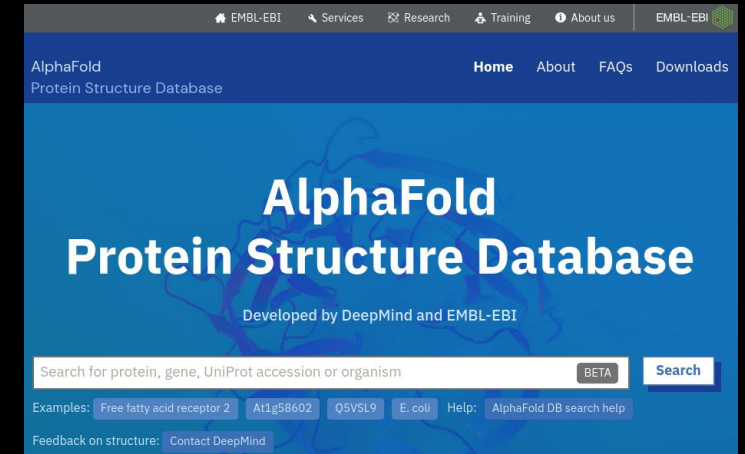
Generate or obtain predicted model

Generate or
obtain
predicted
model



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



Generate or obtain predicted model

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

The screenshot displays the CCP4 Cloud interface, which is used for protein structure analysis. The main window shows the MrParse analysis results for a protein structure. The analysis includes a summary of the input data, a table of experimental structures from the PDB, and a table of structure predictions from the EBI AlphaFold database. The 3D model of the protein is shown in the center, with various chains and regions highlighted in different colors. The interface also includes a sidebar with a list of tasks and a bottom panel showing the sequence viewer.

MrParse Analysis
Version: 0.2.5
MrParse: a program to find and analyse search models for crystallographic Molecular Replacement. The program is being developed by [Dan Suck](#) at the University of Liverpool. MrParse is currently under development and we are keen to make it as useful to the community as possible. If you have any suggestions for its development, or ideas on how we could improve it, please [get in touch](#).

HKL Info

Name	Resolution	Space Group	Has NCS?	Has Twinning?	Has Anisotropy?
reflections	1.60	R3	false	false	true

Experimental structures from the PDB

Name	PDB	Resolution	Region	Length	eLGL	Mol. Wt.	eRMSD	Seq. Ident.
4f3v_R_1	4f3v	2.00	1	31-274	242	284.1	26273	1.243

Structure predictions from the EBI AlphaFold database

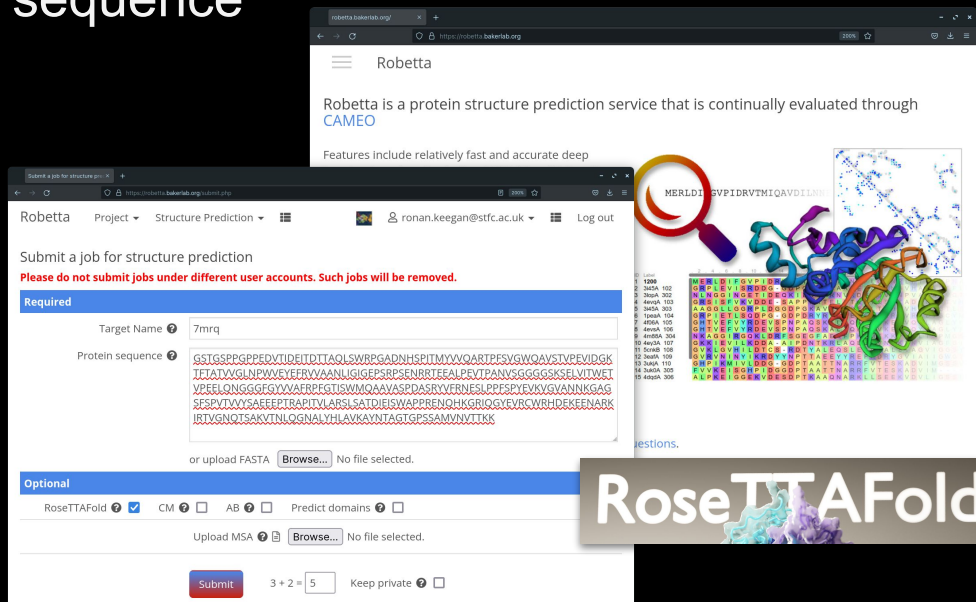
Name	model	Date Made	Region	Length	Avg. pLDDT	H-score	Seq. Ident.
PWWP3_1	PWWP3_1	01-JUL-21	2	2-278	275	95.28	90
PWWP1_1	PWWP1_1	01-JUL-21	1	10-280	269	94.32	90
PWWP4_1	PWWP4_1	01-JUL-21	1	32-275	242	94.89	89
O33329_1	O33329_1	01-JUL-21	2	5-235	229	92.44	88
PWWP7_1	PWWP7_1	01-JUL-21	1	26-278	251	95.91	92

Finished MBUMP search

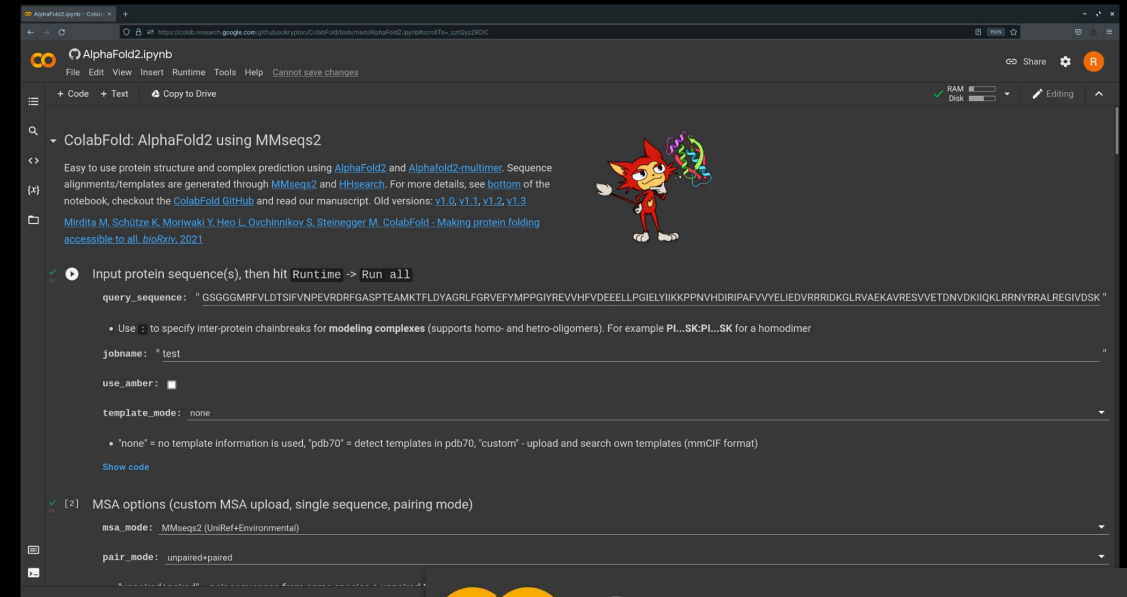
Model	Score	Seq. Ident.
AF-A0A0R4IMY7-F1-model_LV21-PC	91.15%	(1)
6d73_D1-PC	89.74%	(1)
6gn1_D1-PC	89.27%	(1)
6an1_D1-PC	89.24%	(1)
6puo_D1-PC	62.69%	(1)
AF-E9PTA2-F1-model_LV21-PC	60.10%	(1)
AF-Q91YD4-F1-model_LV21-PC	59.84%	(1)
AF-Q94759-F1-model_LV21-PC	58.70%	(1)
6puo_D2-PC	32.65%	(2)
AF-Q94759-F1-model_LV22-PC	29.55%	(2)

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



The screenshot shows the Robetta website interface. At the top, it says "Robetta" and "Robetta is a protein structure prediction service that is continually evaluated through CAMEO". Below this, it says "Features include relatively fast and accurate deep". The main section is "Submit a job for structure prediction" with a warning: "Please do not submit jobs under different user accounts. Such jobs will be removed." There are two sections: "Required" and "Optional". In the "Required" section, "Target Name" is set to "7mrq" and "Protein sequence" is a long amino acid sequence. In the "Optional" section, "RoseTTAFold" is checked, and "Predict domains" is also checked. There are buttons for "Submit" and "Keep private".

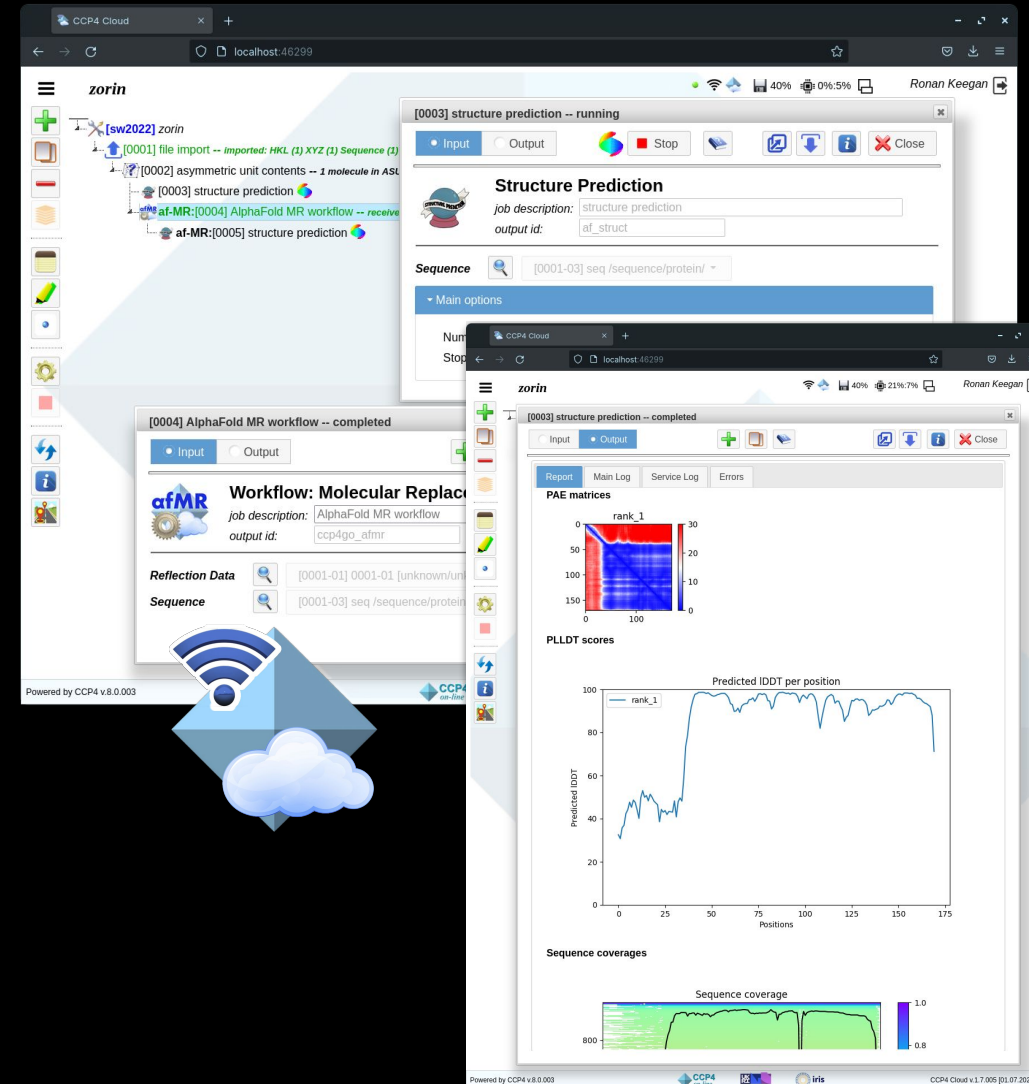


The screenshot shows the AlphaFold2.ipynb notebook in Google Colab. The notebook title is "AlphaFold2.ipynb". The first cell is a markdown cell with the title "ColabFold: AlphaFold2 using MMseqs2" and a description: "Easy to use protein structure and complex prediction using AlphaFold2 and AlphaFold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3". The second cell is a code cell with the title "Input protein sequence(s), then hit Runtime -> Run all". It contains a query sequence: "GSGGMRFLDTSIFVNPVDRFGASPTKMTFLDYAGRLFGVRFVYMPGGIYREVVFHVEDELLPGIELYIKKPPNVHDIRIPAFVYELIEDVRRRIDKGLRVAEKAVRESVETDNVDKIQKLRRNYRRALREGIVDSK". There are input fields for "jobname" (set to "test"), "use_amber" (set to "none"), and "template_mode" (set to "none"). The third cell is a code cell with the title "MSA options (custom MSA upload, single sequence, pairing mode)". It contains "msa_mode" (set to "MMseqs2 (UniRef+Environmental)") and "pair_mode" (set to "unpaired-paired").



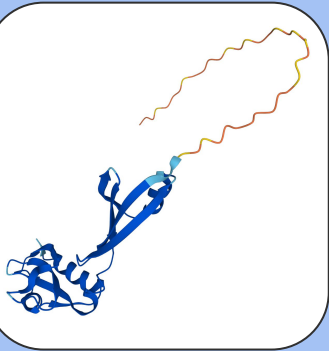
Generate or obtain predicted model

- 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

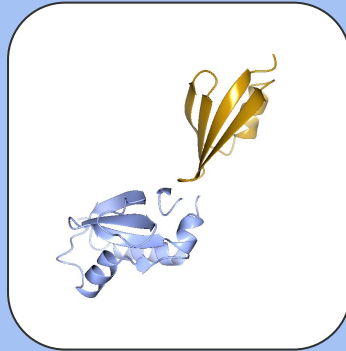


Process predicted model

Generate or
obtain
predicted
model



Process
predicted
model



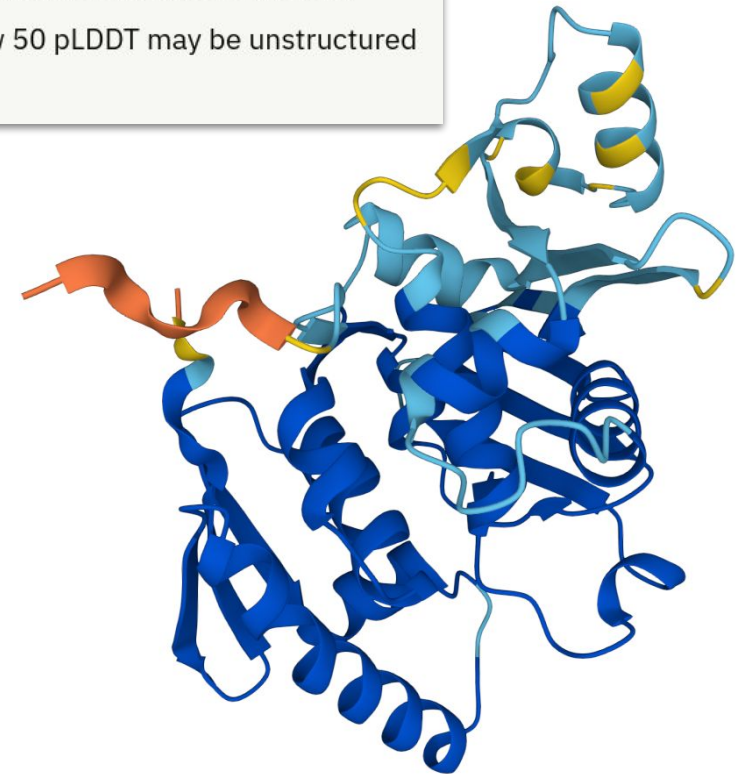
Process predicted model

- 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

Model Confidence:

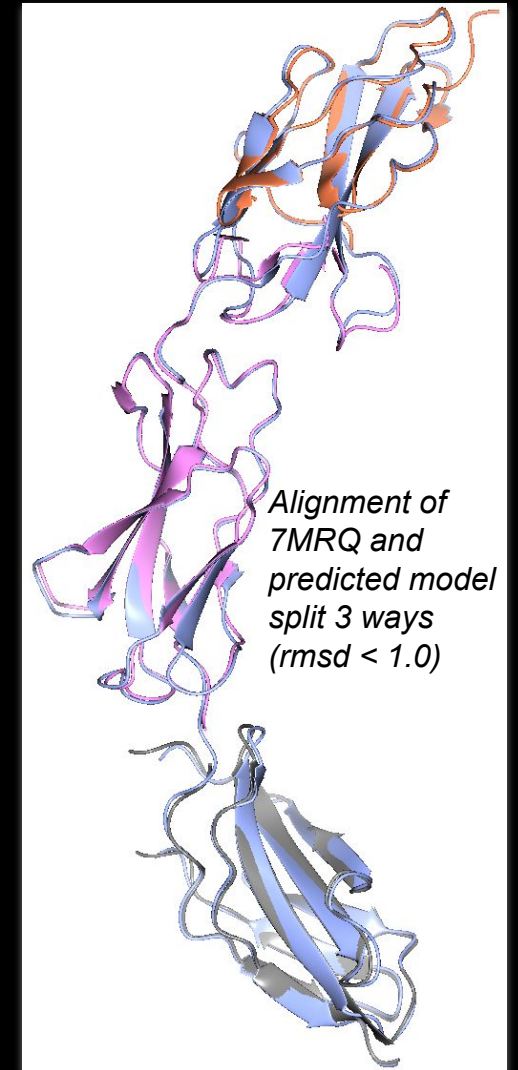
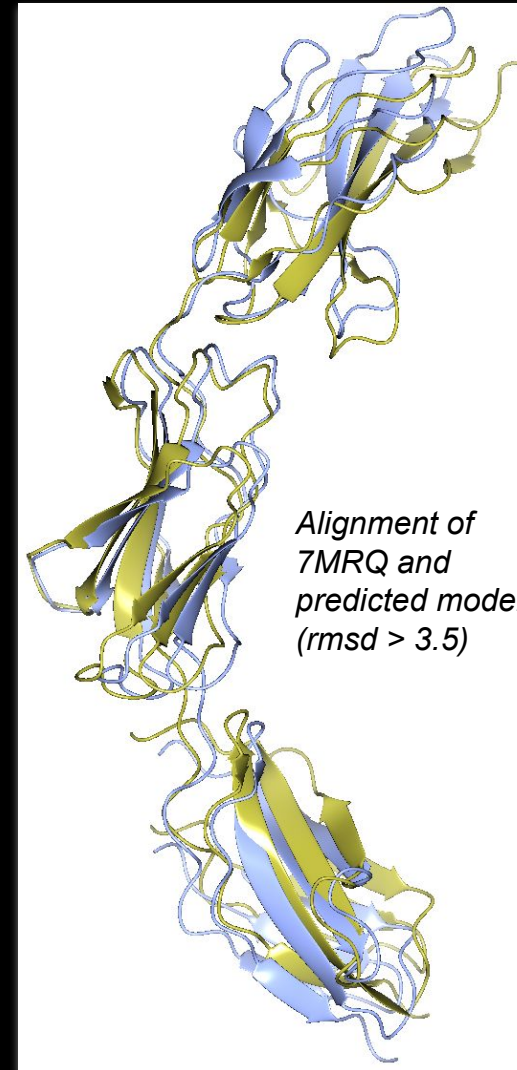
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

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



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



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

[0041] prepare MR model(s) from xyz (new)

Input Output Run

Prepare MR Model(s) from Coordinate data
job description: prepare MR model(s) from xyz

Sequence [0001-02] 7mrq_A /sequence/protein/

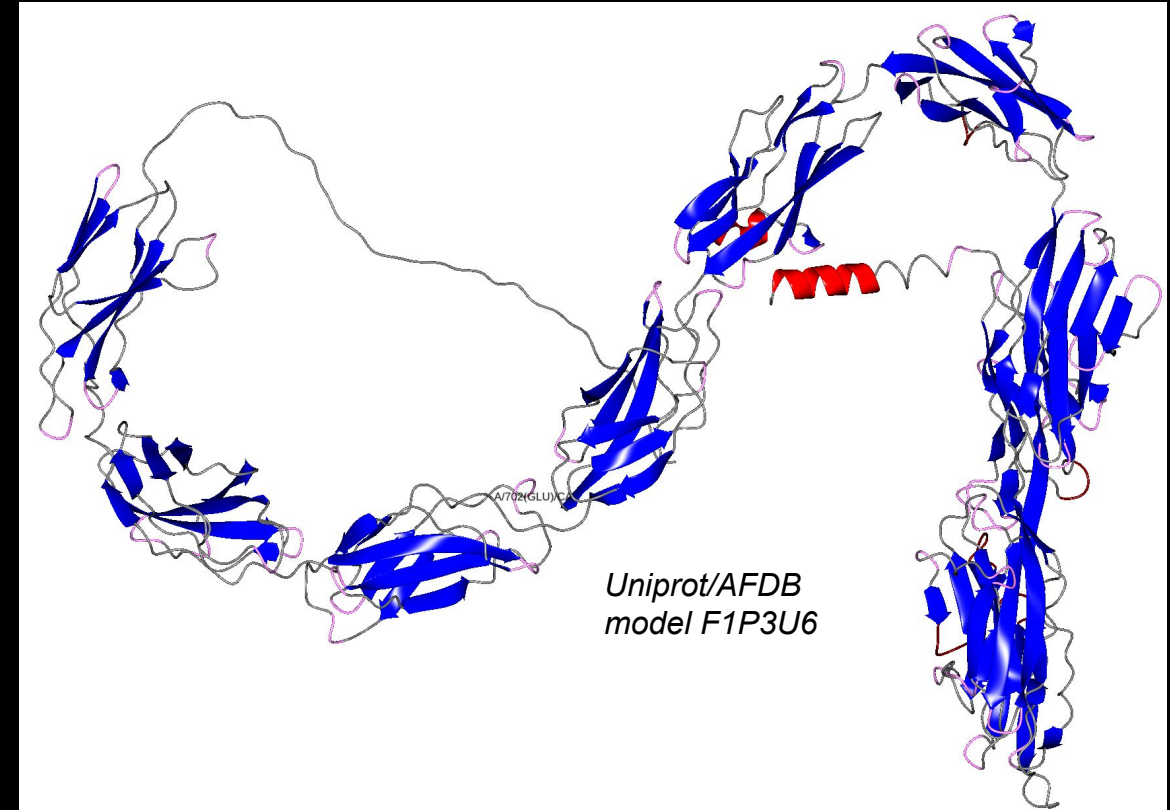
Coordinates (1) [0006-01] f1p3u6 /xyz/protein/

Select chain: A (protein)

Coordinates (2) [do not use]

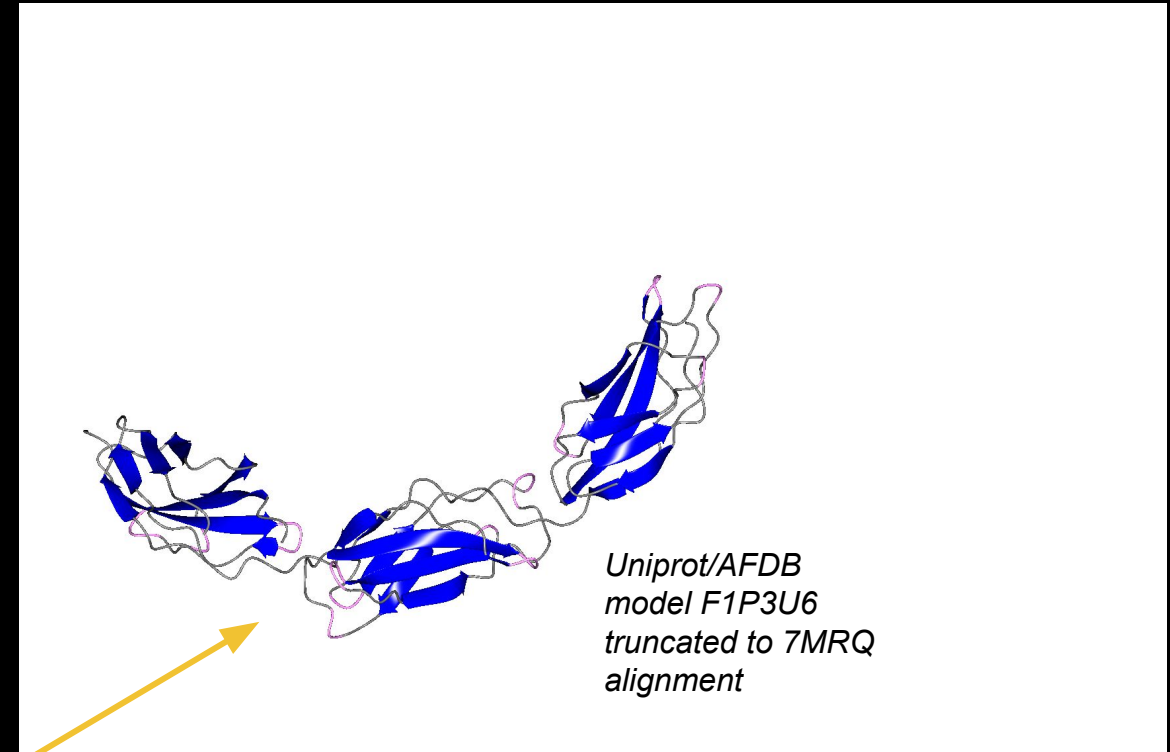
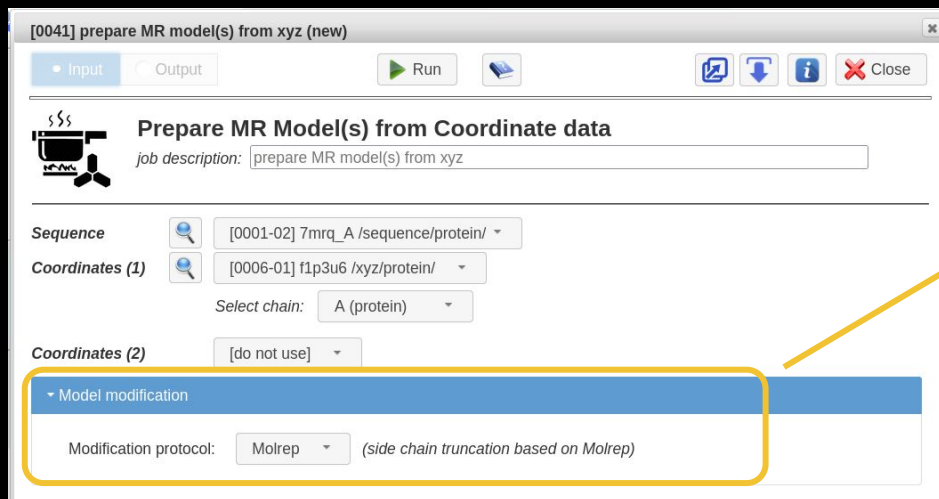
Model modification

Modification protocol: Molrep (side chain truncation based on Molrep)



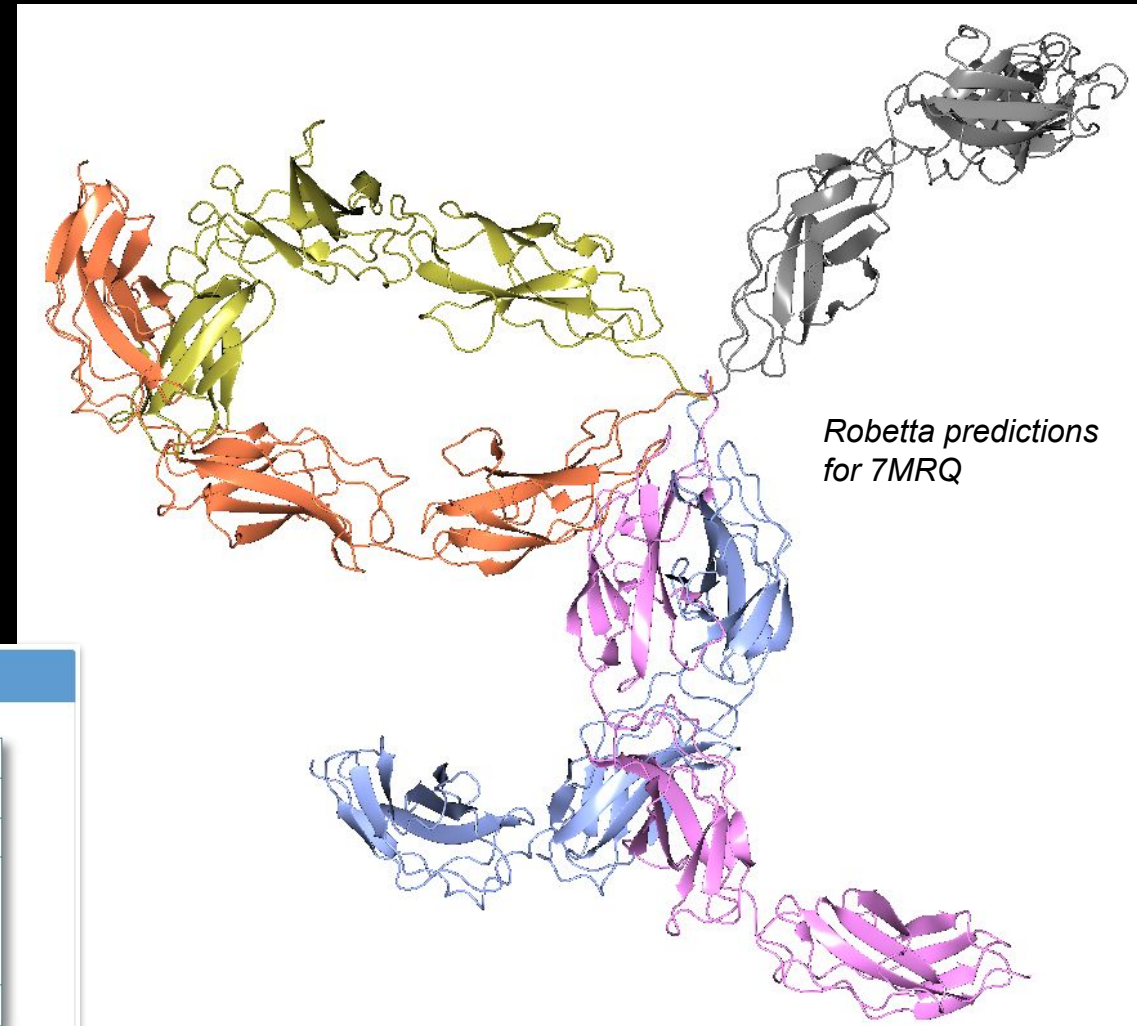
Process predicted model

- 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



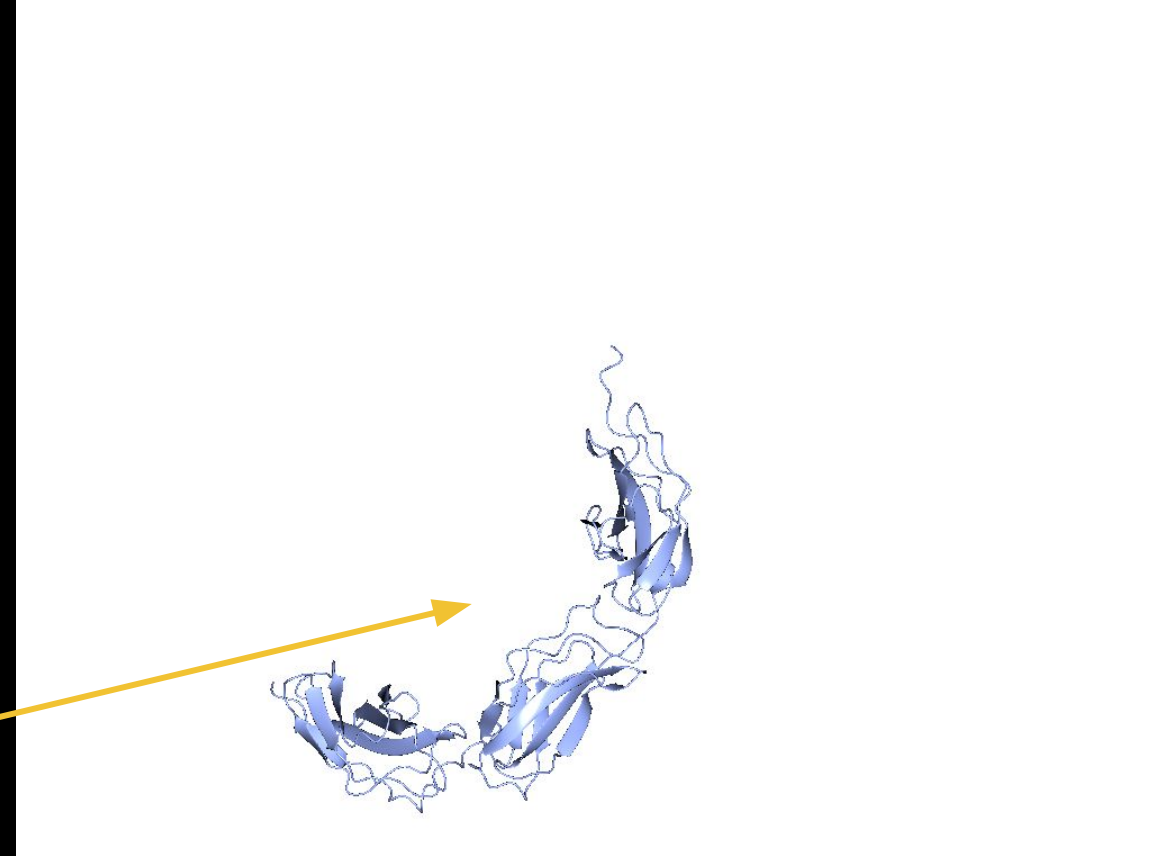
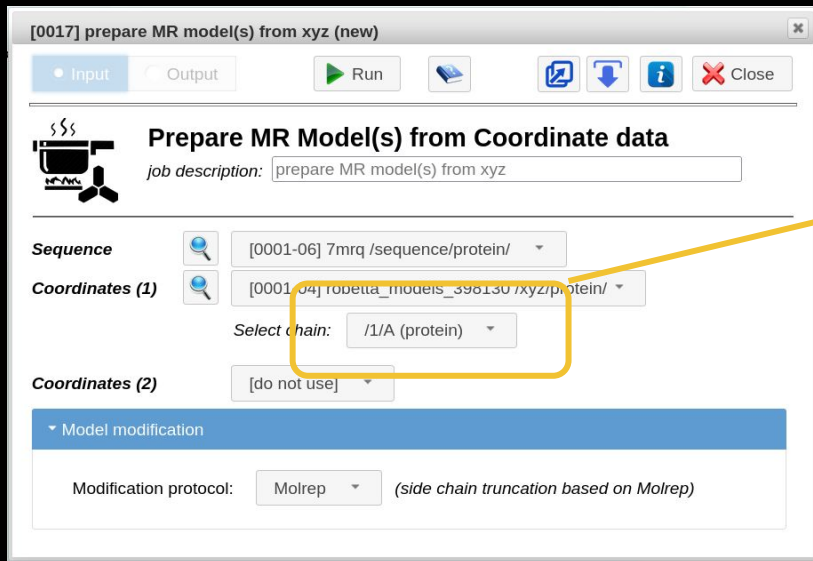
▼ Import robetta_models_398130.pdb

Assigned name	[0001-04] robetta_models_398130 /xyz/protein/
Space group	
Cell parameters	1.0 1.0 1.0 90.0 90.0 90.0
Contents	Model 1, chain A: 309 residues, type: Protein Model 2, chain A: 309 residues, type: Protein Model 3, chain A: 309 residues, type: Protein Model 4, chain A: 309 residues, type: Protein Model 5, chain A: 309 residues, type: Protein
B-factor correction	Assuming Rosetta model i File contents

▶ [0001-04] robetta_models_398130 /xyz/protein/ UglyMol ccp4mg Display

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

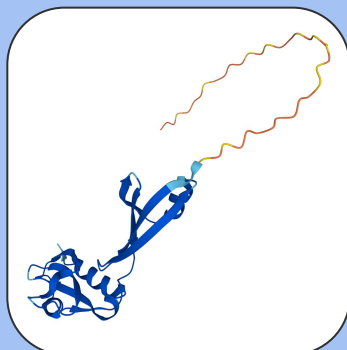
The screenshot displays the CCP4 Cloud web interface in a browser window. The main workspace shows a workflow for a project named 'zorin'. The workflow steps are:

- [0001] file import -- imported: HKL (1) XYZ (1) Sequence (1)
- [0002] asymmetric unit contents -- 1 molecule in ASU, Solv=35.8%
- [0003] structure prediction
- af-MR:[0004] AlphaFold MR workflow -- received HKL (1), Sequences (1); workflow started
- af-MR:[0005] structure prediction
- [0006] slice -- 2 model(s) generated

A modal window titled '[0006] slice -- completed' is open, showing the 'Split MR model with Slice-n-Dice' task. The window includes fields for 'job description' (slice) and 'output id' (slice). The 'Template structure' is set to '[0001-02] test_ec684_unrelaxed_rank_1_model_1/xyz/protein/'. The 'Parameters' section shows 'Number of splits' set to 2. The interface also features a sidebar with various tool icons, a top navigation bar with the user name 'Ronan Keegan', and a footer with logos for CCP4 on-line, iris, and the version 'CCP4 Cloud v.1.7.005 [01.07.2022]'.

Manual or automated MR

Generate or
obtain
predicted
model

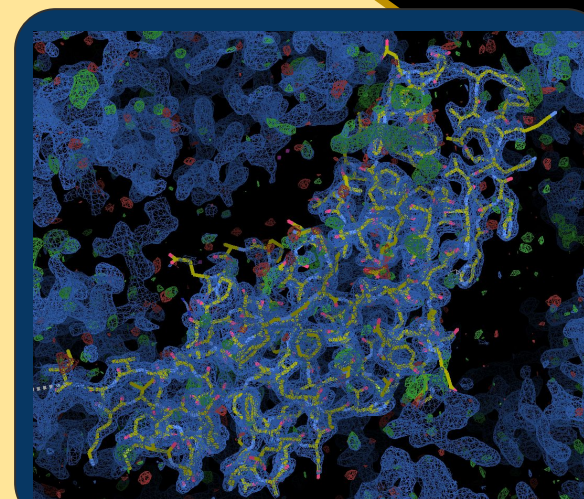


Process
predicted
model



Manual or
automated
MR

```
[sw2022] zorn
  [0001] file import -- imported: HKL (1) XYZ (1) Sequence (1)
  [0002] asymmetric unit contents -- 1 molecule in ASU, Solv=35.8%
  [0003] structure prediction -- 1 structure predicted
  [0004] AlphaFold MR workflow -- received HKL (1), Sequences (1); workflow started
  [0005] AlphaFold MR workflow -- 1 structure predicted
  [0006] AlphaFold MR workflow -- prepare MR model(s) from xyz -- 1 model(s) generated (molrep protocol)
  [0007] AlphaFold MR workflow -- asymmetric unit contents -- 1 molecule in ASU, Solv=35.8%
  [0008] AlphaFold MR workflow -- phaser MR -- finished
  [0009] slice -- 2 model(s) generated
  [0010] molrep -- R=0.4996 Rfree=0.4787
  [0011] reftmac5 -- R=0.4671 Rfree=0.4462
  [0012] molrep -- R=0.4027 Rfree=0.4099
  [0013] reftmac5 -- R=0.2709 Rfree=0.3209
  [0014] sarservice -- LL00702.8 PP2428.1 R=0.3415 Rfree=0.3367
  [0015] reftmac5 -- R=0.2868 Rfree=0.3188
```



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)



zorin

[0001] file import -- imported: HKL (1) XYZ (1) Sequence (1)

[0002] asymmetric unit contents -- 1 molecule in ASU, Solv=35.4

[0003] structure prediction -- 1 structure predicted

af-MR:[0004] AlphaFold MR workflow -- received HKL (1), 1

af-MR:[0005] structure prediction -- 1 structure predicted

af-MR:[0007] prepare MR model(s) from xyz

af-MR:[0008] asymmetric unit contents --

af-MR:[0009] phaser MR -- finished.

[0006] slice -- 2 model(s) generated

[0010] molrep -- R=0.4096 R_{free}=0.4787

[0011] reftmac5 -- R=0.4071 R_{free}=0.4462

[0013] molrep -- R=0.4027 R_{free}=0.4099

[0015] reftmac5 -- R=0.2709 R_{free}=0.3289

[0014] slicendice -- LLG=752.0 TFZ=25.1 R=0.3615 R_{free}=0

[0016] reftmac5 -- R=0.2663 R_{free}=0.3188

[0017] coot (model building) -- model saved from

[0004] AlphaFold MR workflow -- completed

Workflow: Molecular Replacement with AlphaFold Model

job description: AlphaFold MR workflow

output id: ccp4go_afmr

Reflection Data [0001-01] 0001-01 [unknown/unknown/unknown070122:12:18] /hkl/anom/

Sequence [0001-03] seq /sequence/protein/

CCP4 Cloud v.1.7.005 [01.07.2022]

mdm2

[0001] file import -- imported: HKL (1) Sequence (1)

[0002] asymmetric unit contents -- 1 molecule in ASU, Solv=35.4

[0003] ensemble preparation (ccp4mg)

[0004] phaser MR -- N_{ref}=1 LLG=328

[0005] mrparse -- MR model preparation from xyz

[0006] file import -- imported: XYZ (1)

[0007] prepare MR model(s) from xyz

[0008] slicendice -- failed.

[0009] slicendice -- failed.

[0010] slicendice -- LLG=3275.0 TFZ=25.1 R=0.3615 R_{free}=0

[0011] reftmac5 -- R=0.3036 R_{free}=0.3188

[0012] slicendice -- LLG=3275.0 TFZ=25.1 R=0.3615 R_{free}=0

[0013] slice -- 3 models generated

[0012] slicendice -- completed

MR with model splitting using Slice'N'Dice

job description: slicendice

output id: slicendice

Structure revision

Template structure

Parameters

Try from 1 to

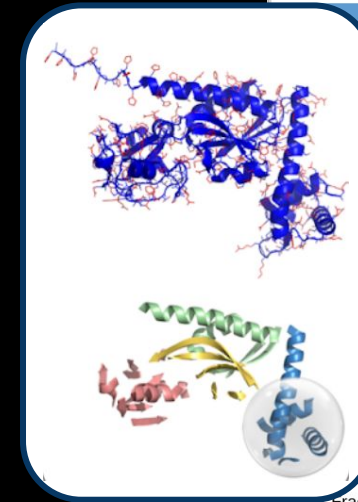
0012-02_slicendice Structure and electron density

This is not a cool, it shows help.

CCP4 Cloud v.1.7.006 [16.08.2022]

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





[0018] arcimboldo-shredder (new)

Input Output Run

Fragment Molecular Replacement with Arcimboldo-Shredder

job description: arcimboldo-shredder
output id: arcimboldo-shredder

Structure revision R0002.01: asu [unknown/unknown/unknown070122:12:12:18] (anom,protein) 

Homology model [0001-02] test_ec684_unrelaxed_rank_1_model_1 /xyz/protein/ 

Note: this task may take significant computational resources and put you outside your monthly quota.

Parameters

Fragment size 1.2 (Å) r.m.s.d. from target

Convert to polyaniline: Auto

Set all B-factors equal: Auto

Order mode: spherical

Chain coil in the model: Auto

Remove gyre refinement: Auto

Remove gimble refinement: Auto

Remove LLG-guided pruning: Auto

Define phases with alixe: Auto

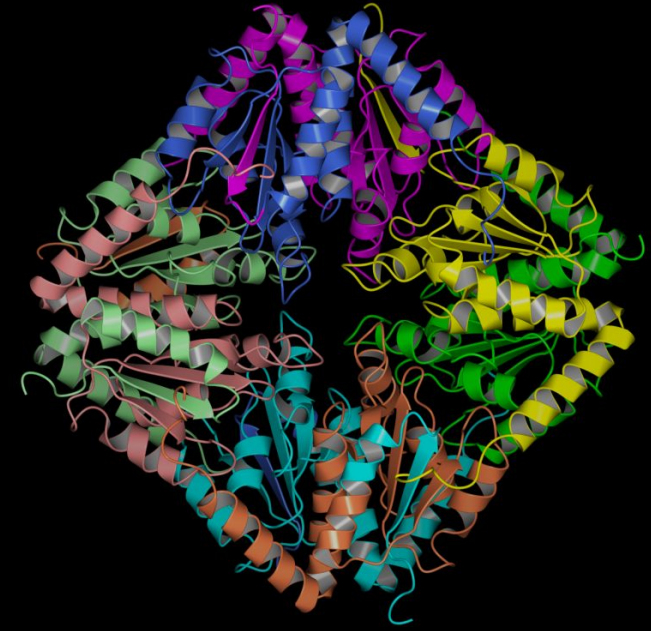
Fixed parameters

Fragment size



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



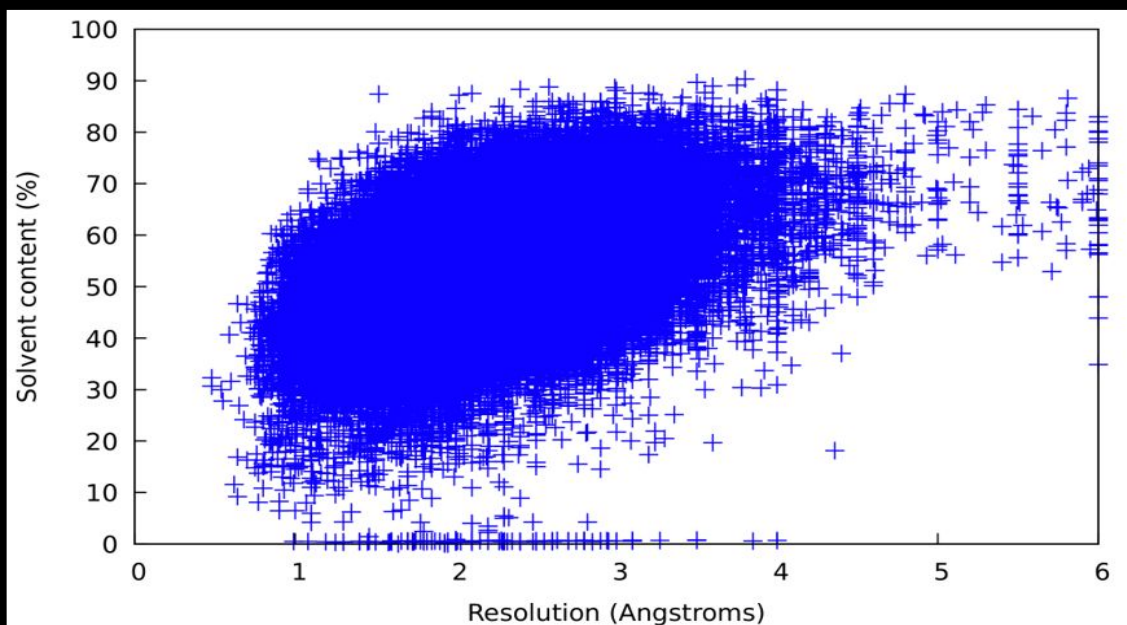
Experimental Data and Molecular Replacement

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

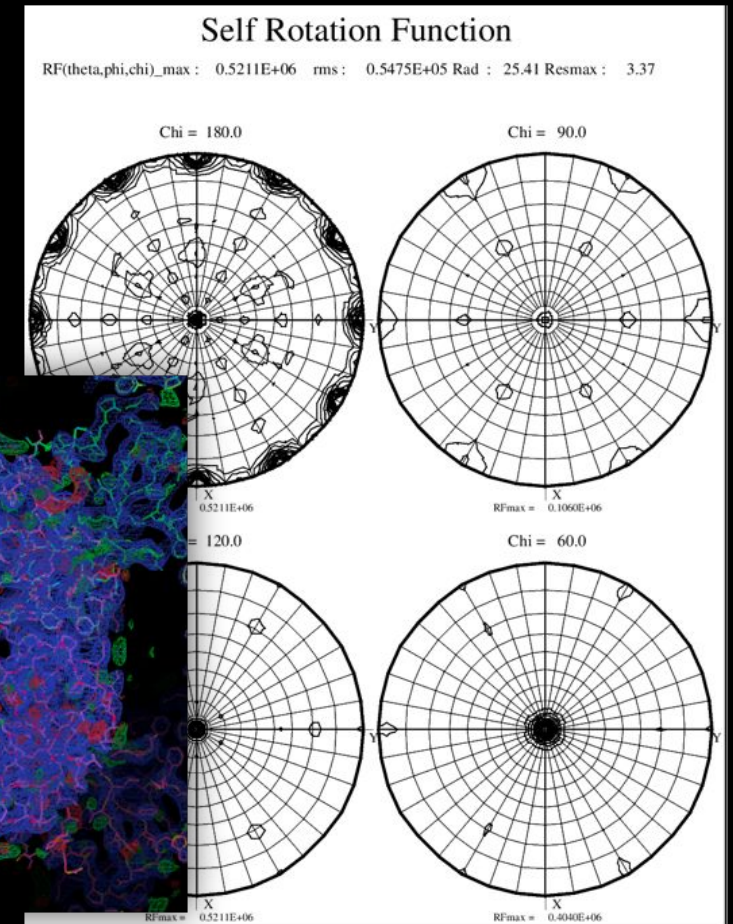


Distribution of solvent content across 156,000 X-ray datasets in the PDB

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

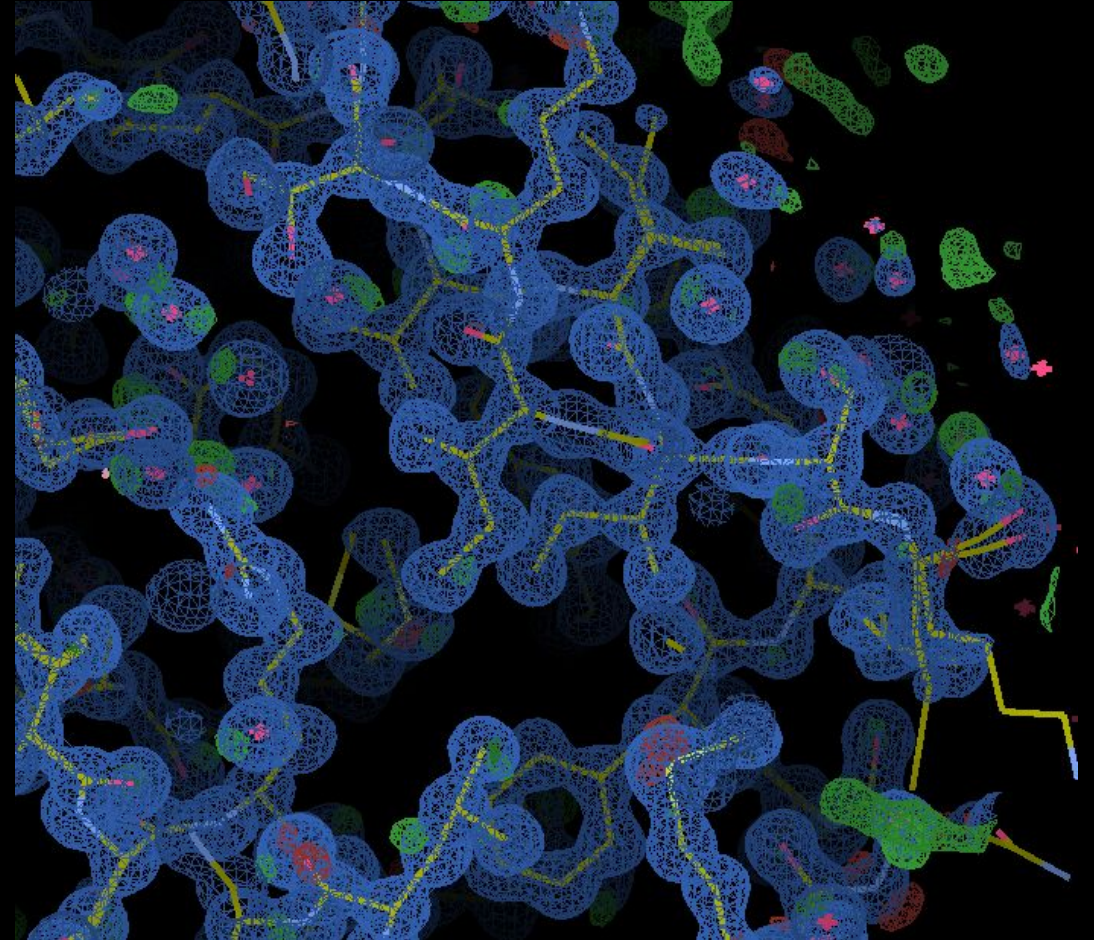


Experimental data

- 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?

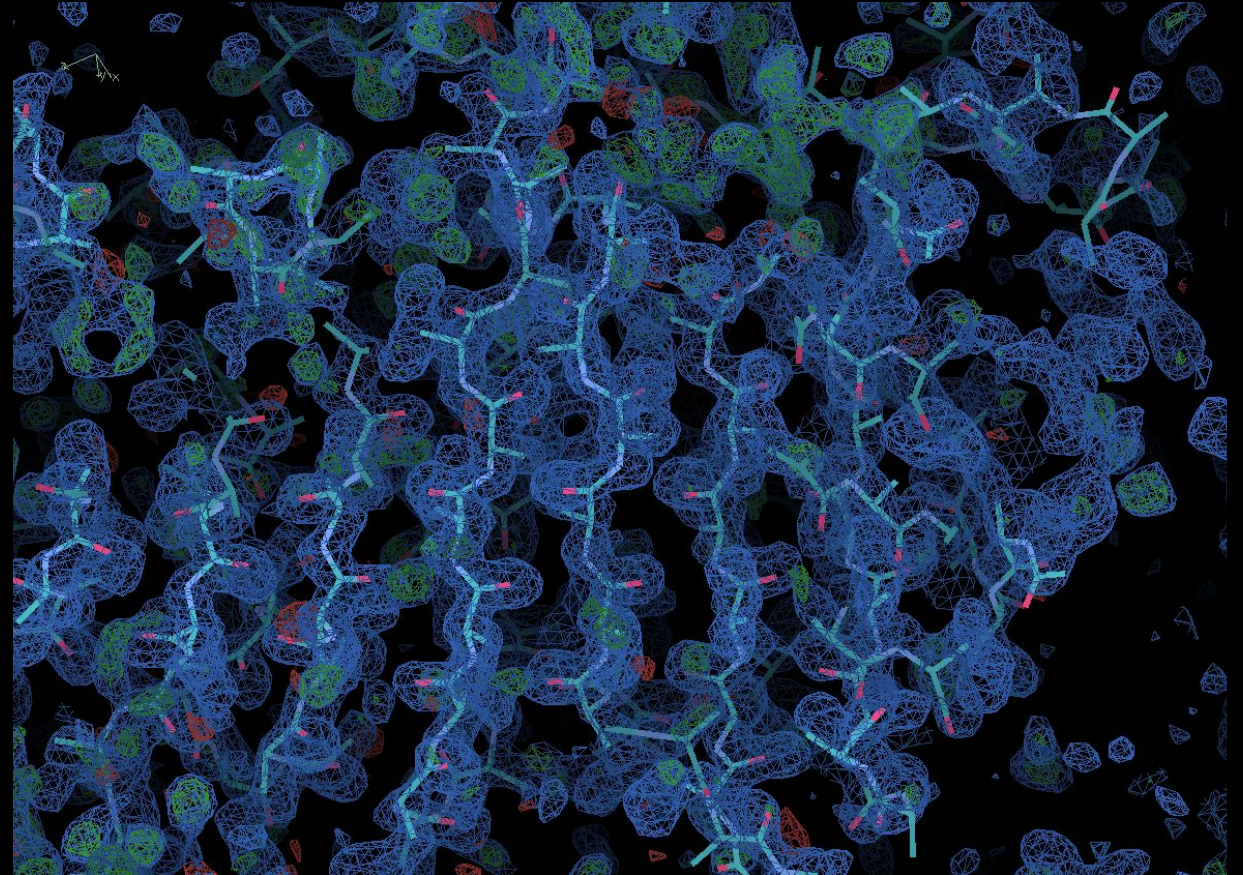
Experimental data: resolution and MR

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



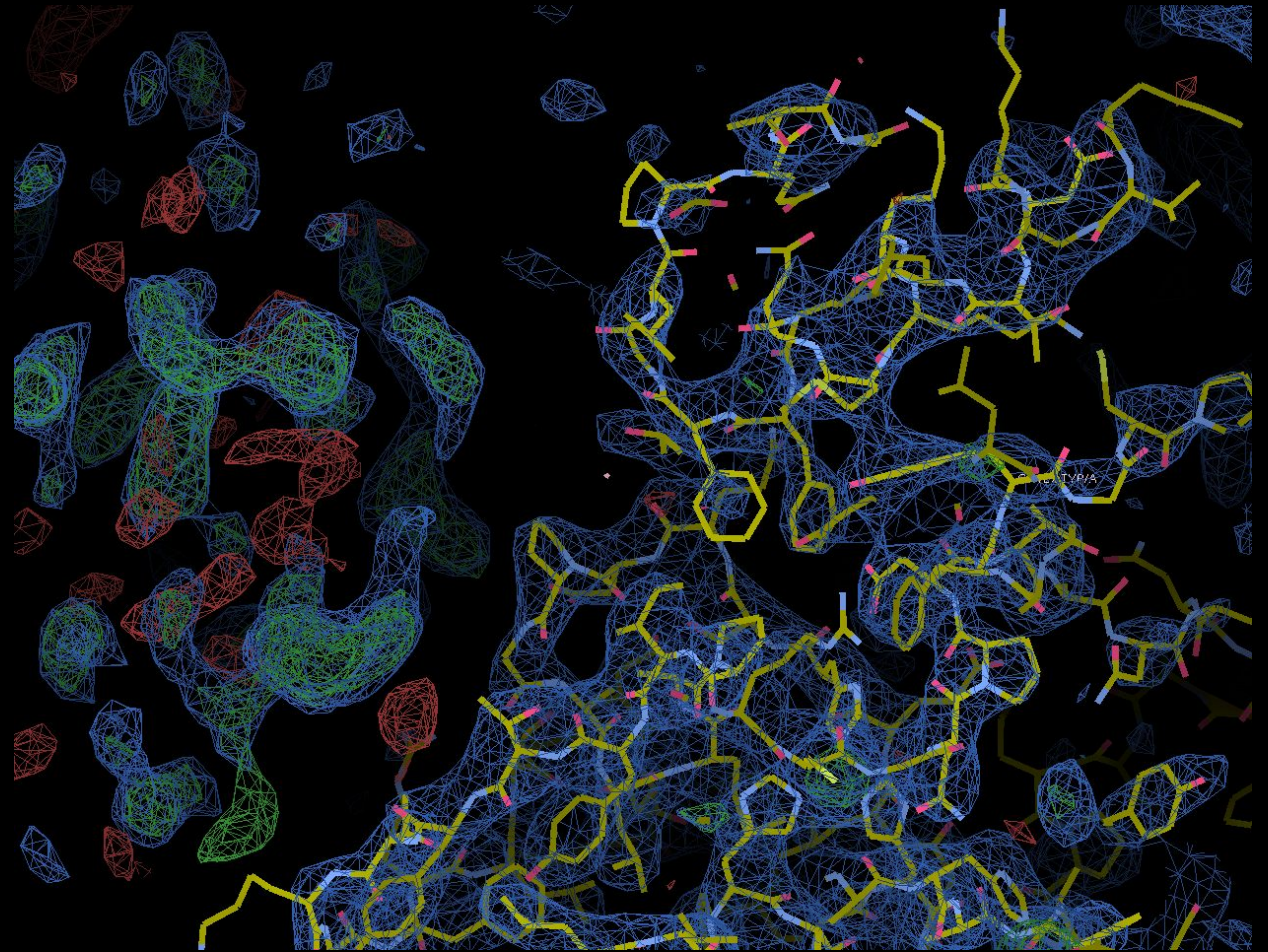
Experimental data: resolution and MR

- 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



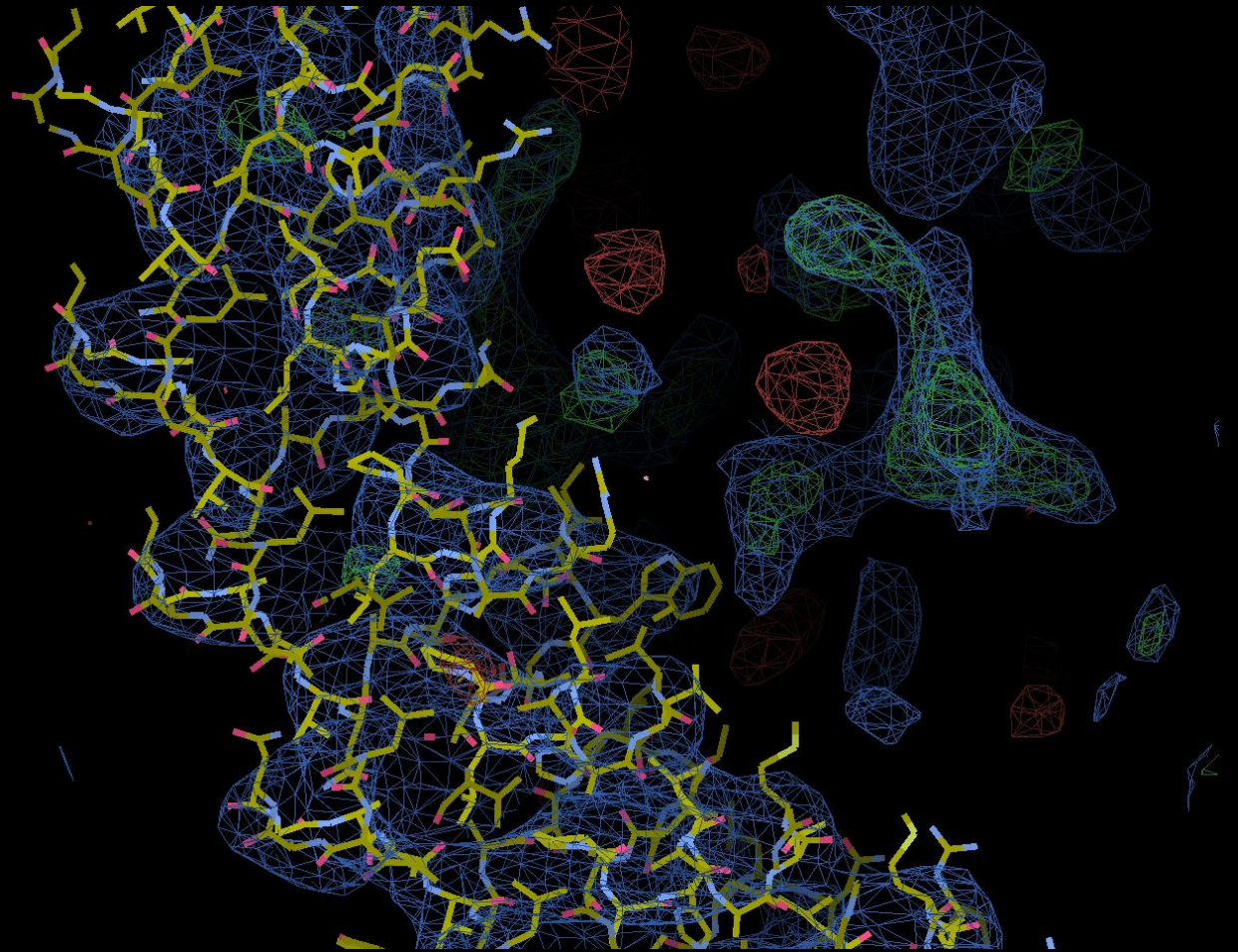
Experimental data: resolution and MR

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



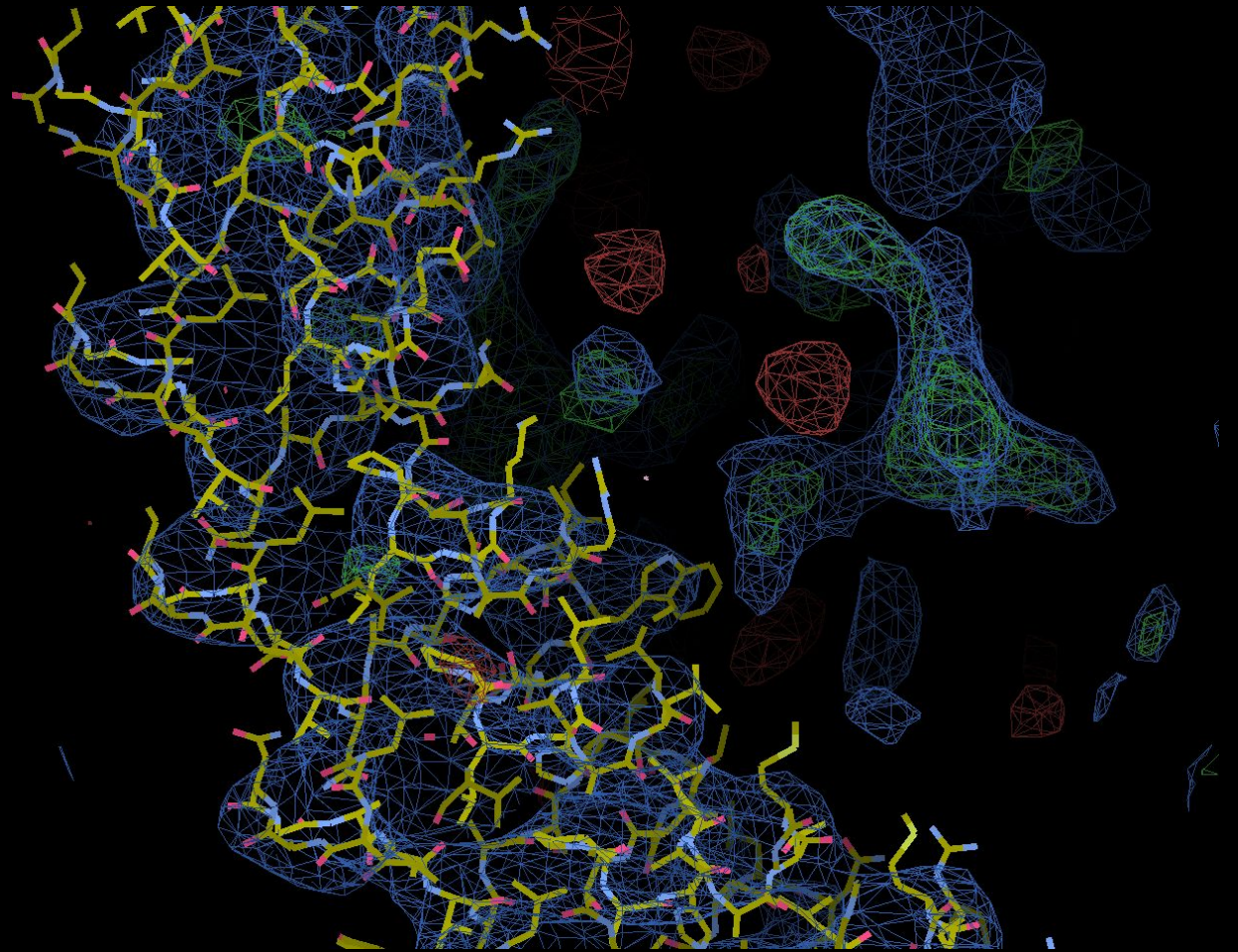
Experimental data: resolution and MR

- Below 4Å automated model building becomes difficult



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

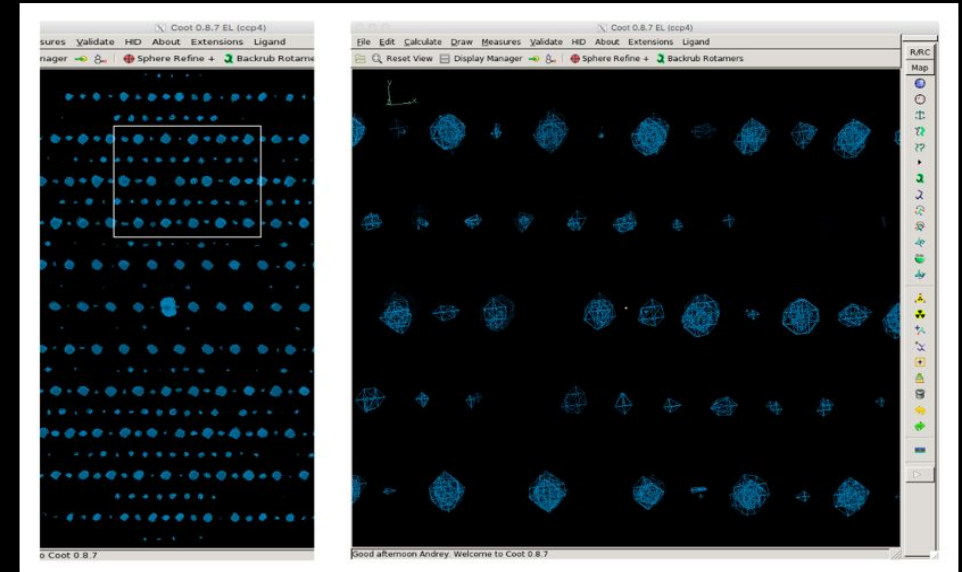


Experimental data

- 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:

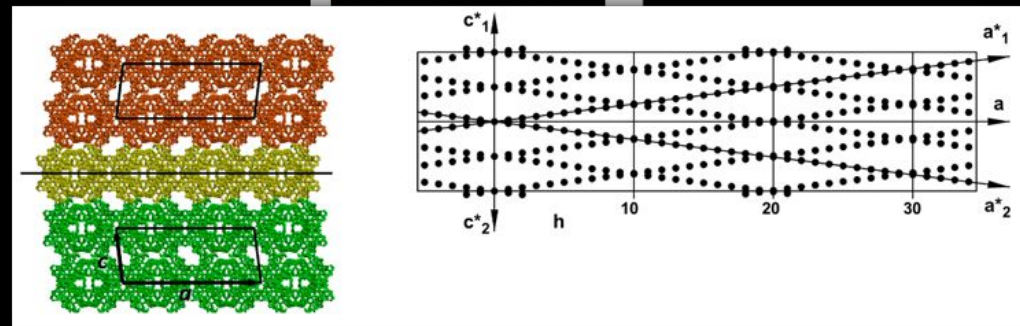
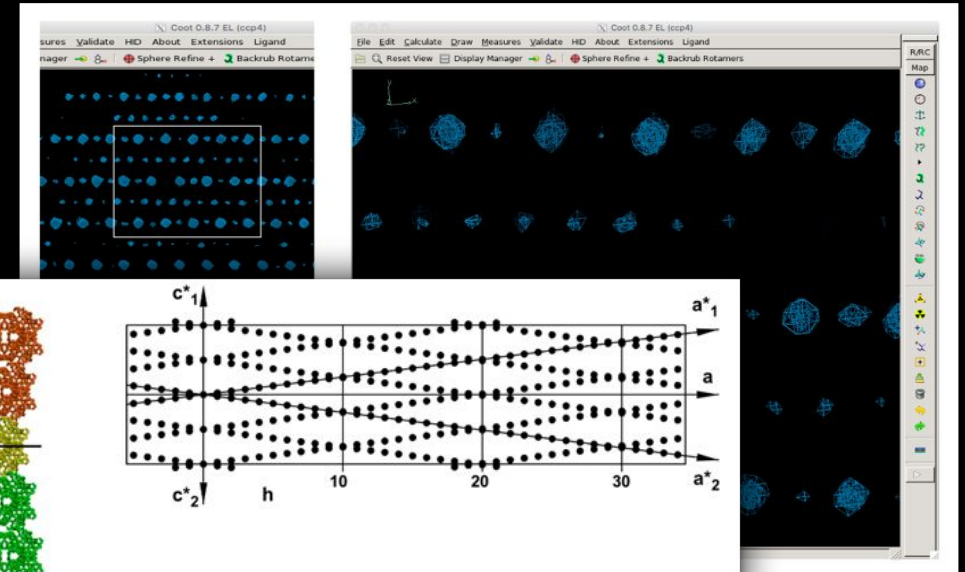
Experimental data

- 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



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



Molecular Replacement in CCP4

- **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 a Coot molecular model. The web interface, titled "SIMBAD Results", features the CCP4 on-line logo and navigation links for "Print" and "Refresh". It includes tabs for "Log file", "Summary", and "Lattice Parameter Search Results". The "Summary" tab is active, showing a "SIMBAD Summary" section with text about the best search model (1SMU) and its R/Rfree values. Below this is a "Best SIMBAD result Downloads" section with a table of files for download and export. A "Best SIMBAD result Log Files" section is also visible. Overlaid on the bottom right is a Coot 0.8.8 EL (ccp4) window showing a 3D molecular model with a blue mesh and yellow sticks. A "Refine/Regularize Control..." menu is open on the right side of the Coot window, listing various refinement and regularization options.

SIMBAD

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

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

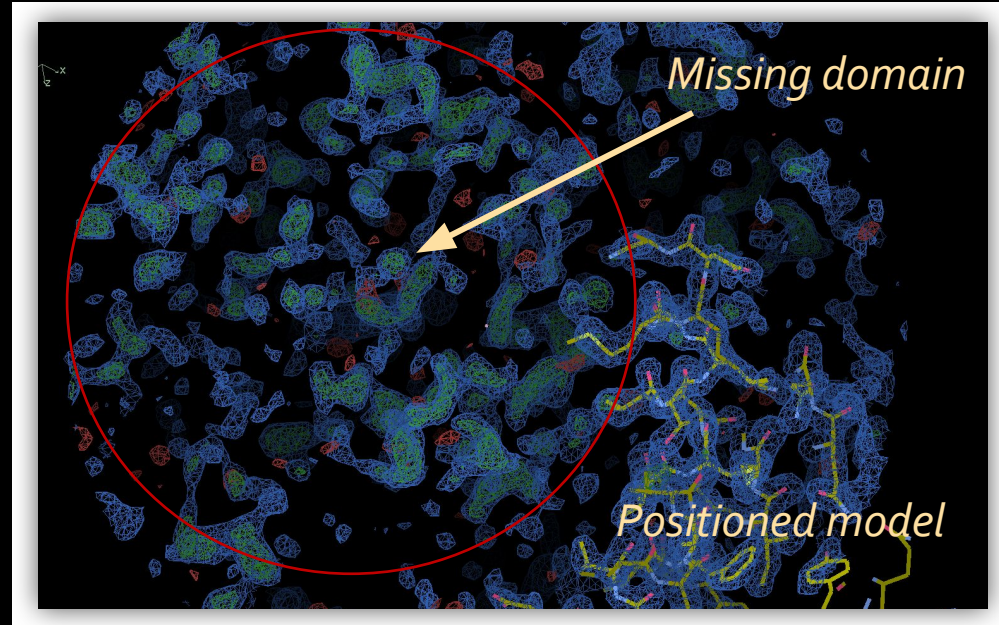
Molecular Replacement: Phaser

- 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

Phased Translation search



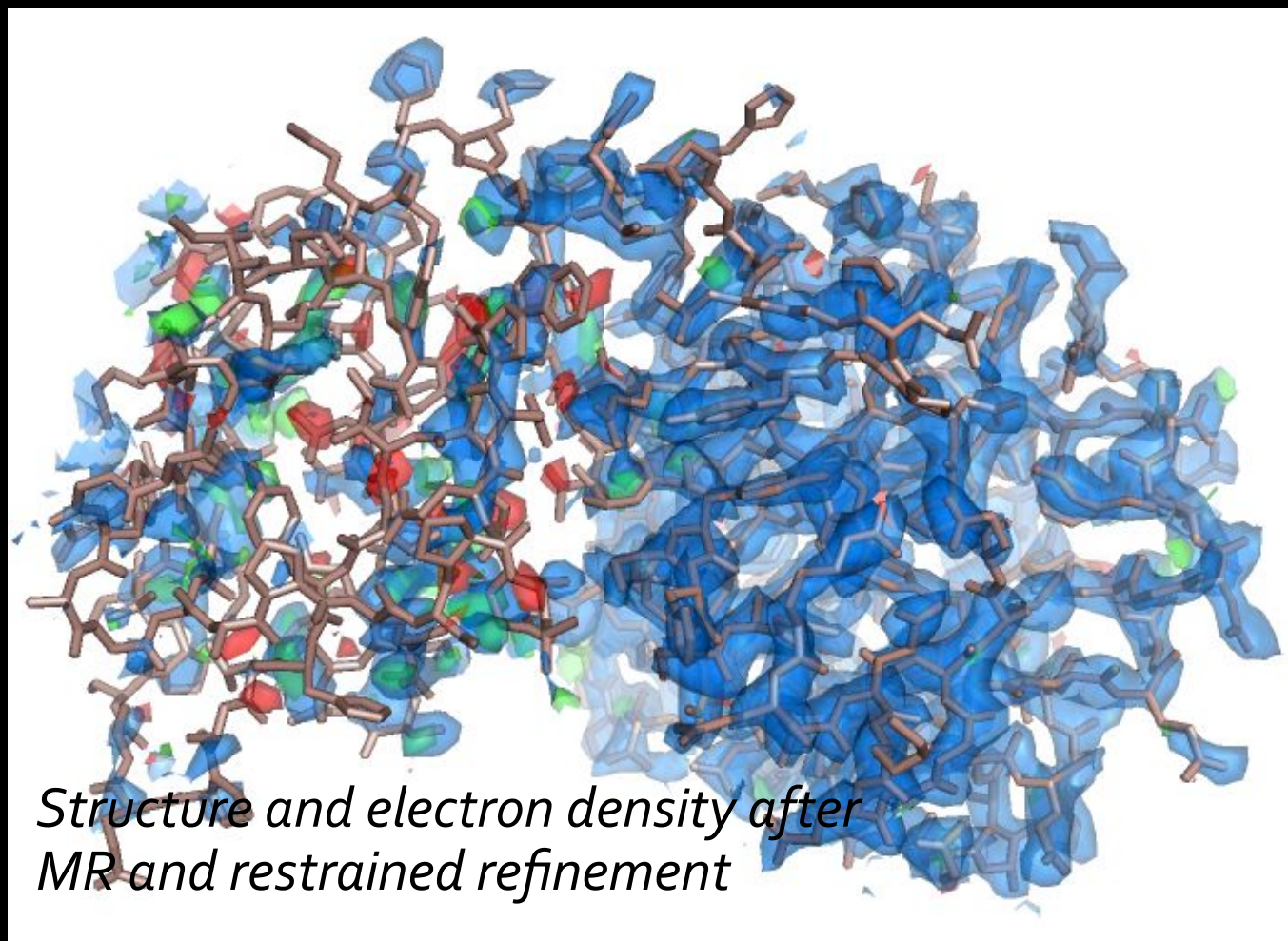
- Used when searching for several copies or dealing with a complex
- Available through MOLREP (3 protocols) and Phaser
- Often more successful than standard MR search approach particularly when looking for small domains

Phased Translation search and model Splitting

Phased translation search:

Example: 1tj3

Search model: 1s20, chain A



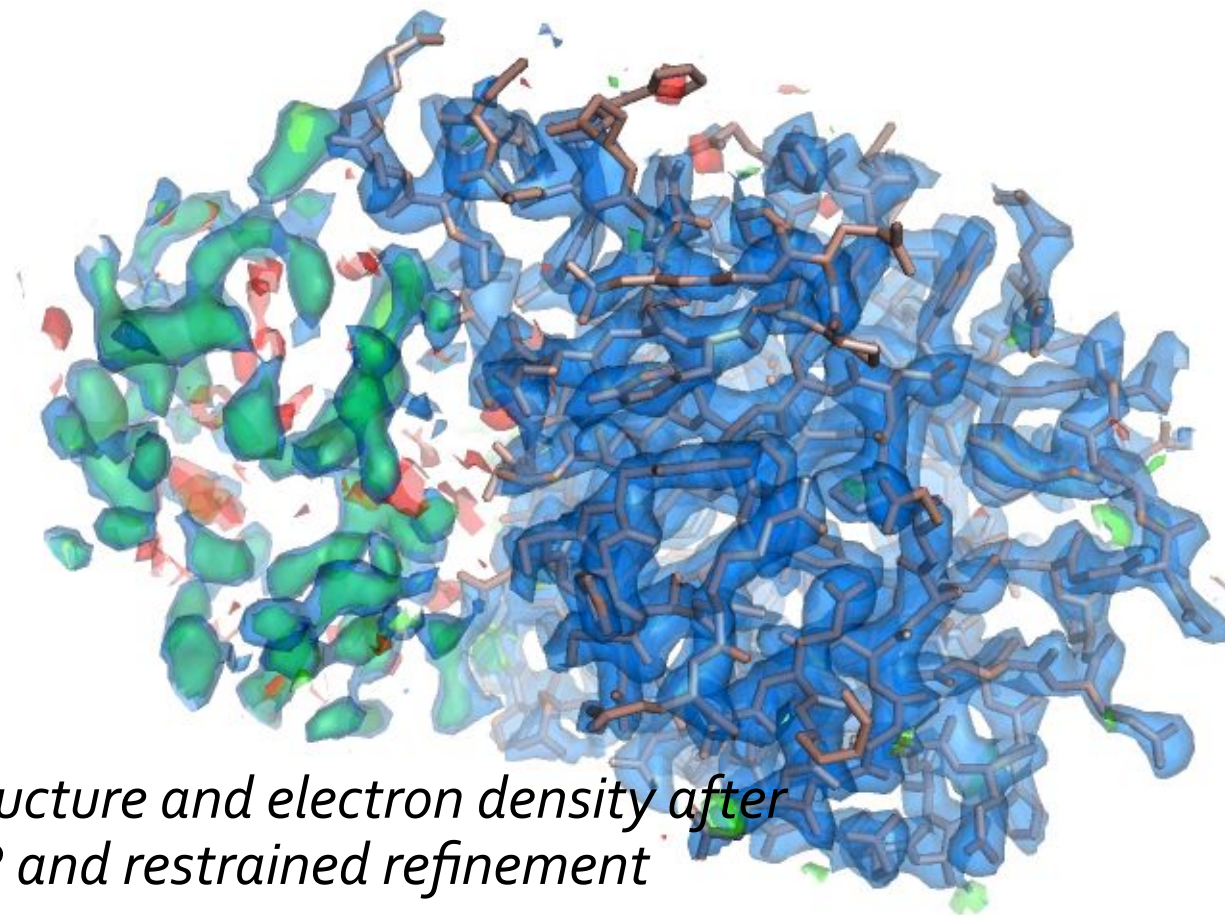
Phased Translation search and model Splitting

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

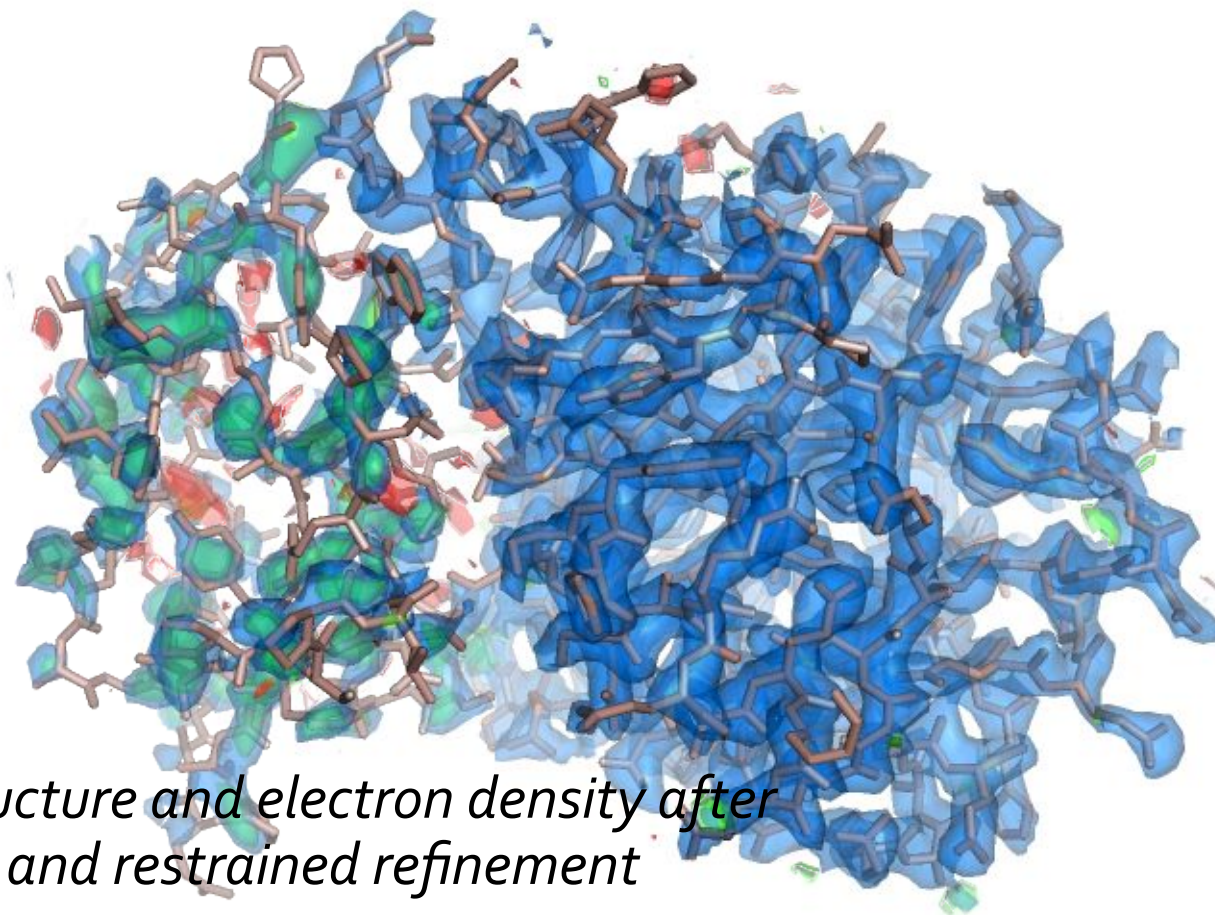
Search model: 1s20, chain A

Large Domain fixed

Small Domain search

Fit into density

(Molrep & Phaser)



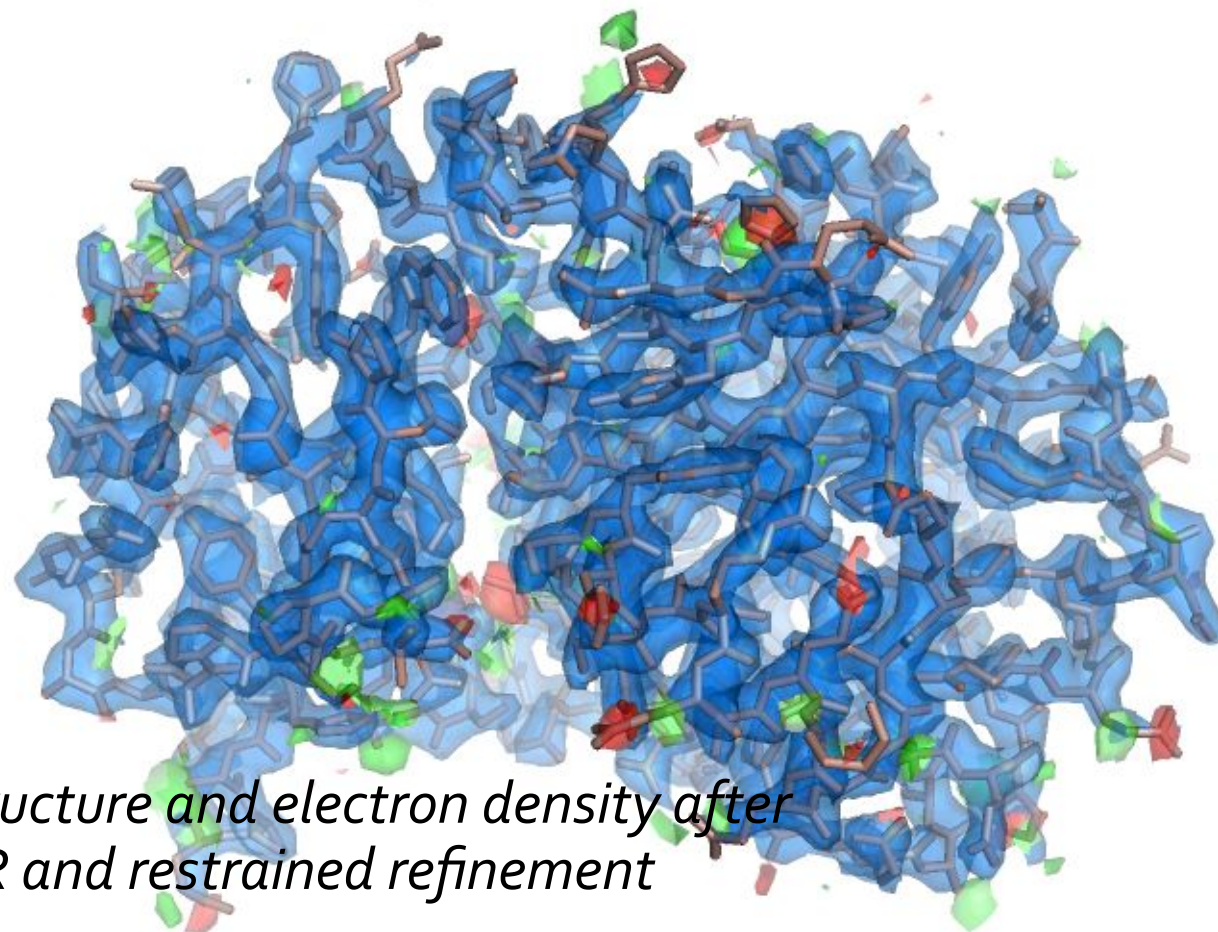
*Structure and electron density after
MR and restrained refinement*

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*

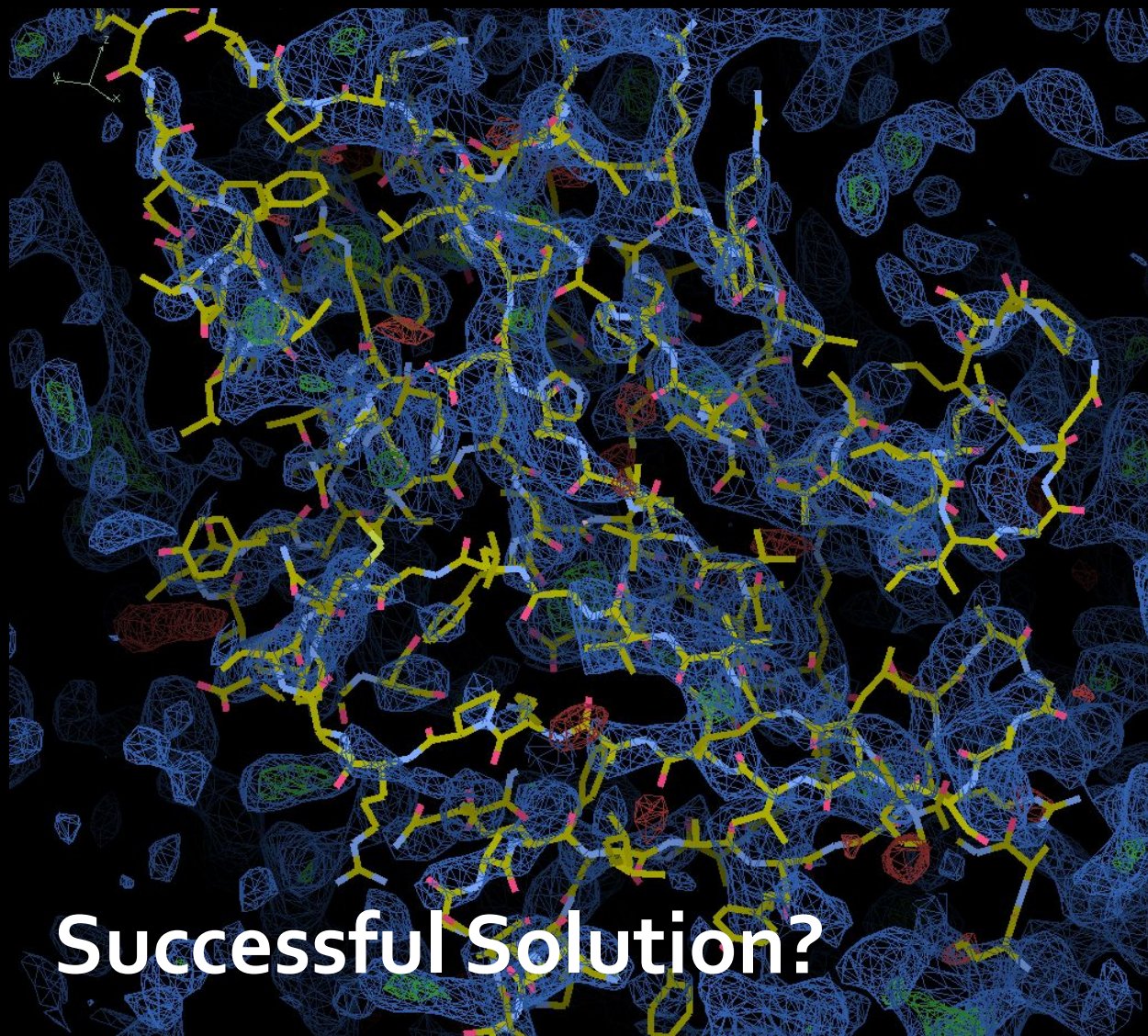
How do I know my MR solution is successful?

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



With 50 cycles of jelly body refinement

Successful Solution? Yes

Assessing the solution: Molecular Replacement Scoring

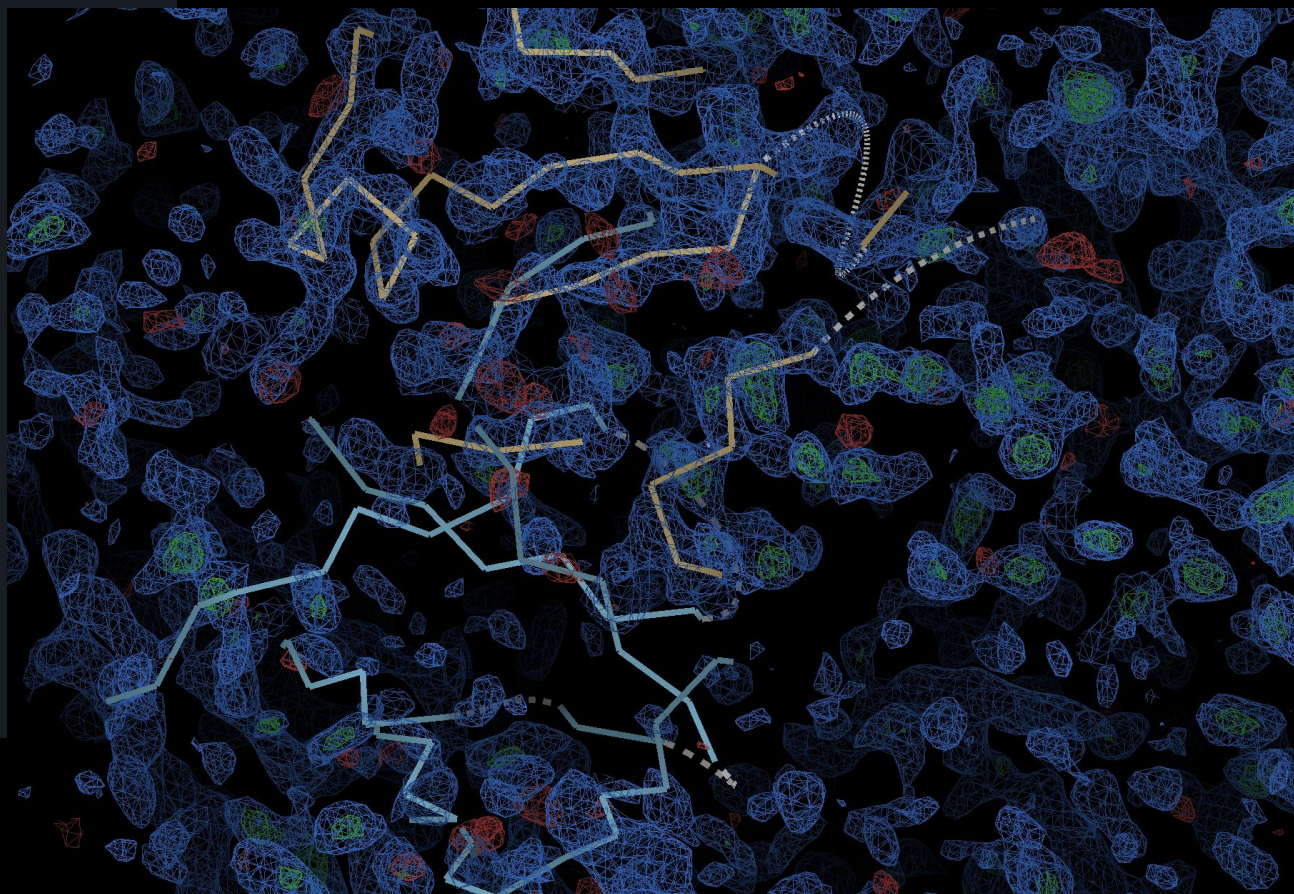
- Rough guide to MR program scoring

- 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
```



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?

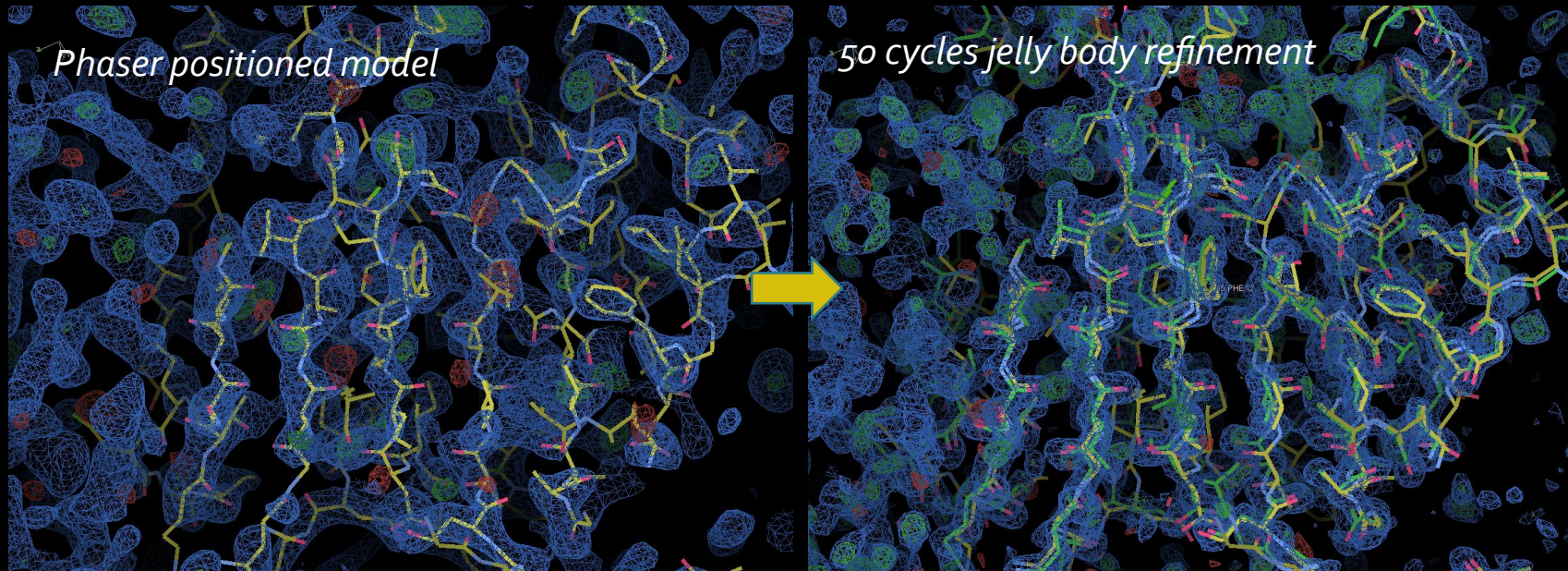
- Use 50 cycles of jelly body refinement option in Refmac post MR

- Refinement

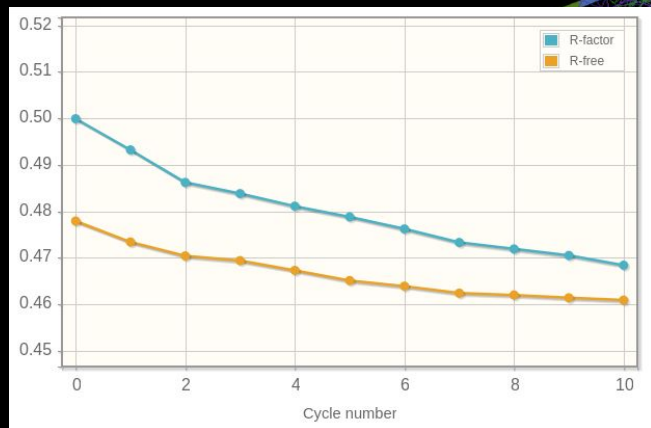
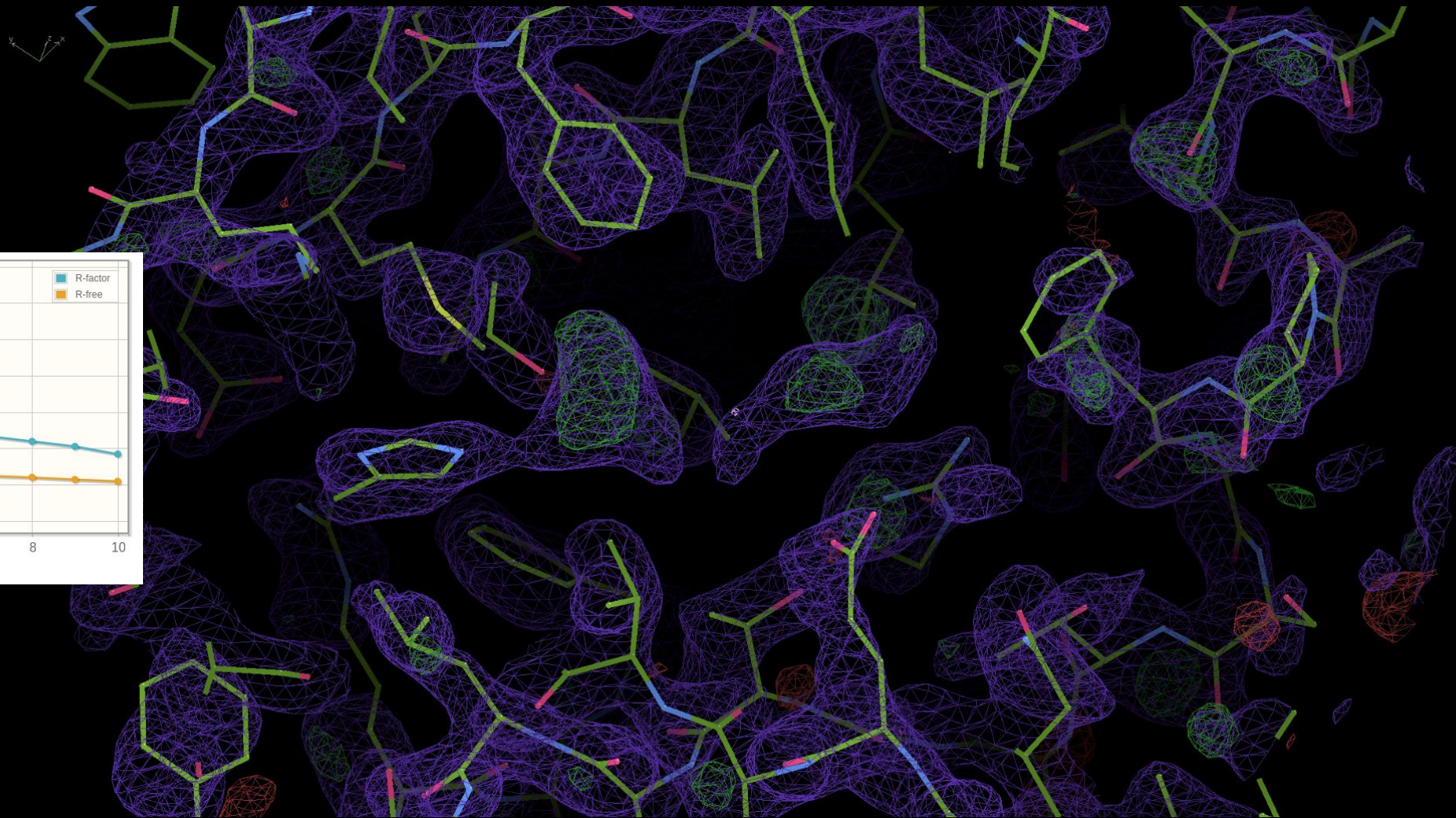
- Look at Rfactor/Rfree

- are they falling? Is Rfree below 0.5?

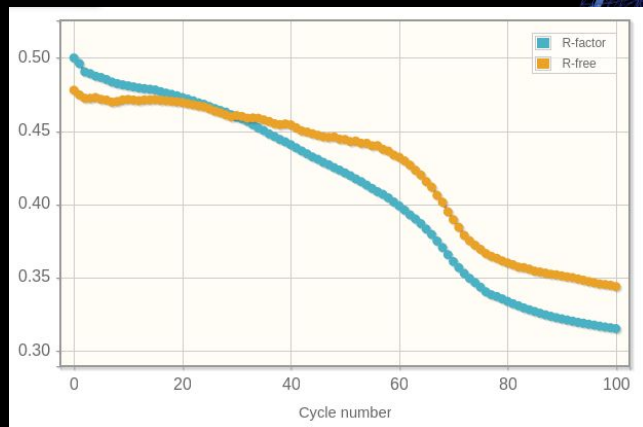
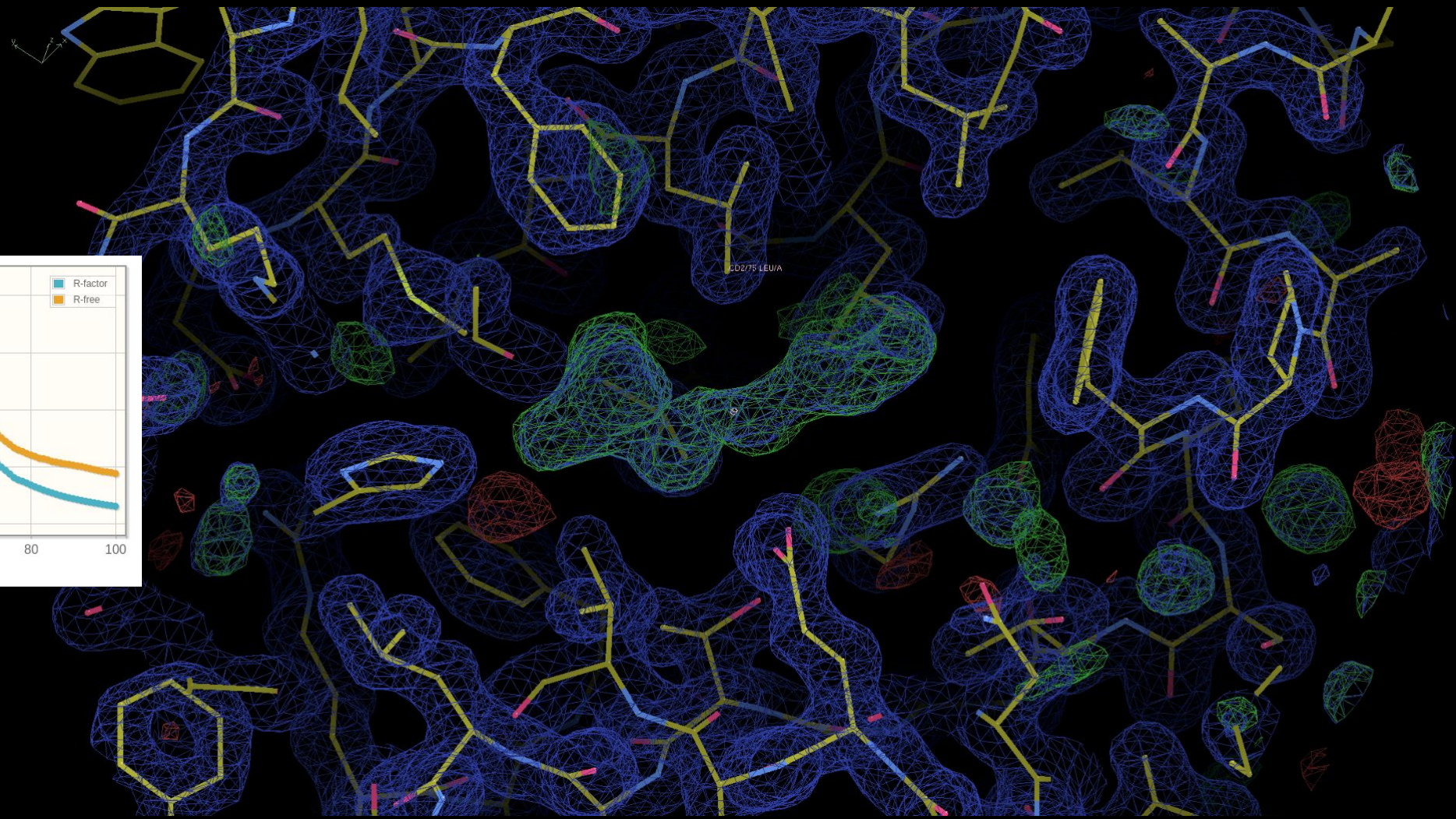
- Use 50 cycles of jelly body refinement option in Refmac post MR



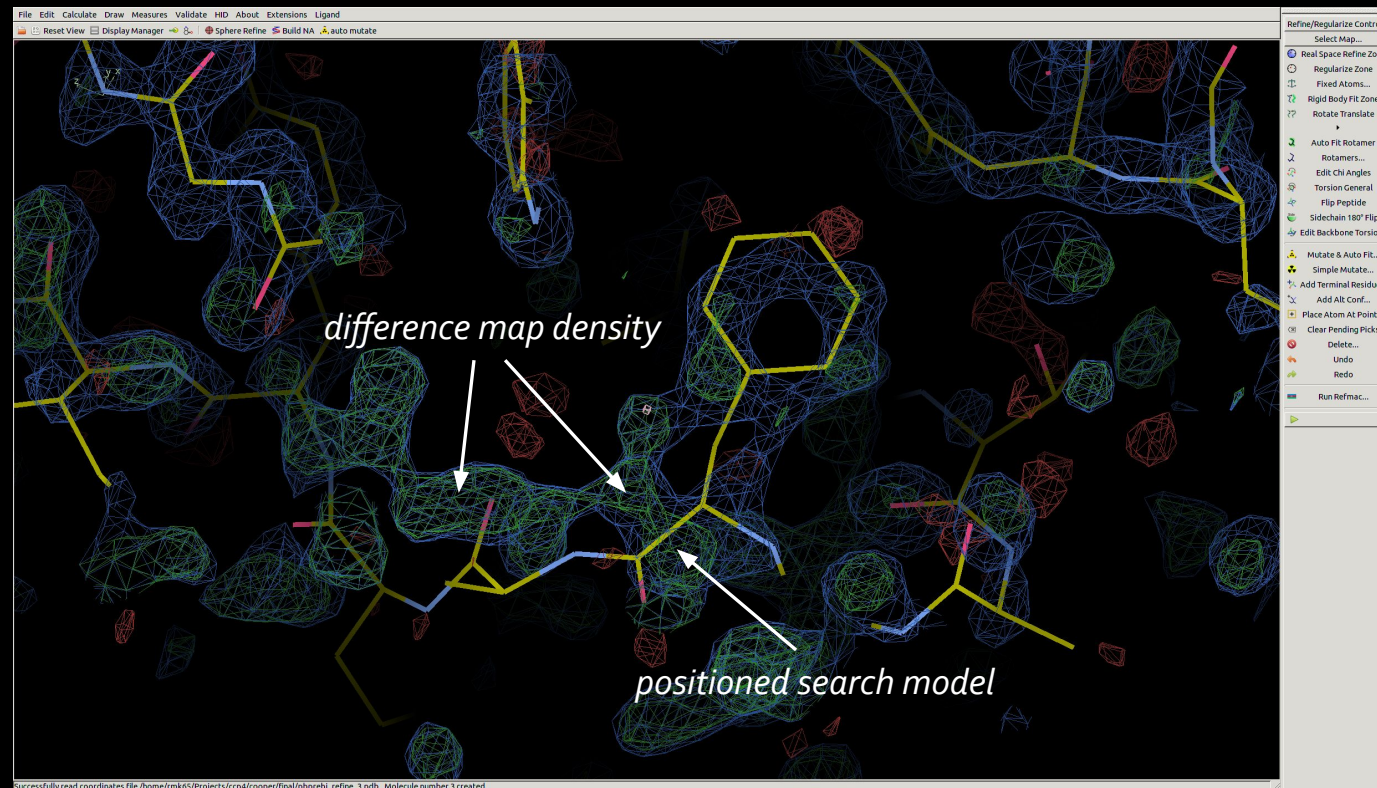
Refining predicted models



Refining predicted models

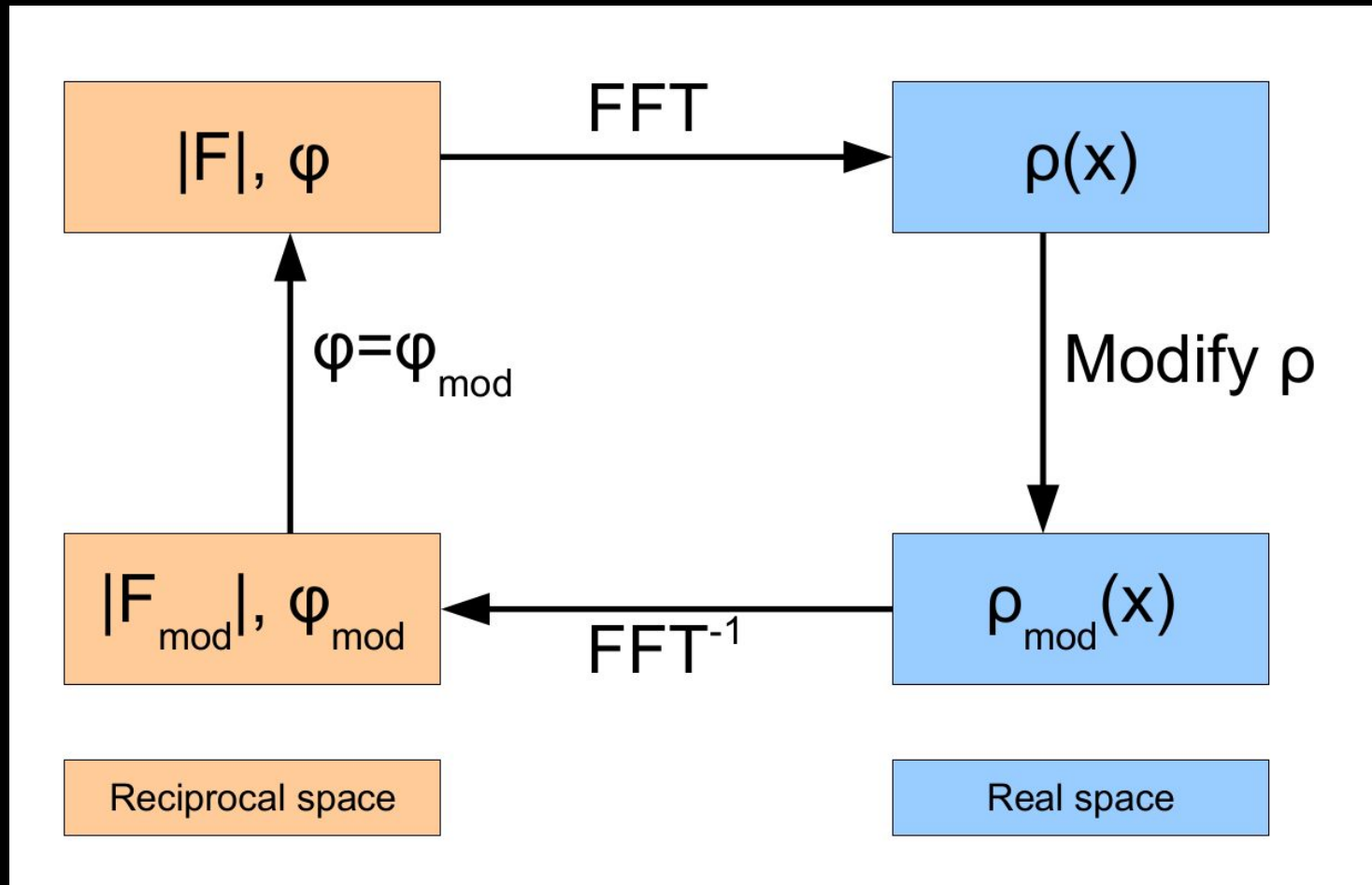


- Examine solution by eye
 - Use Coot to examine positioned models & maps



Assessing and improving the solution: Density Modification

Density modification

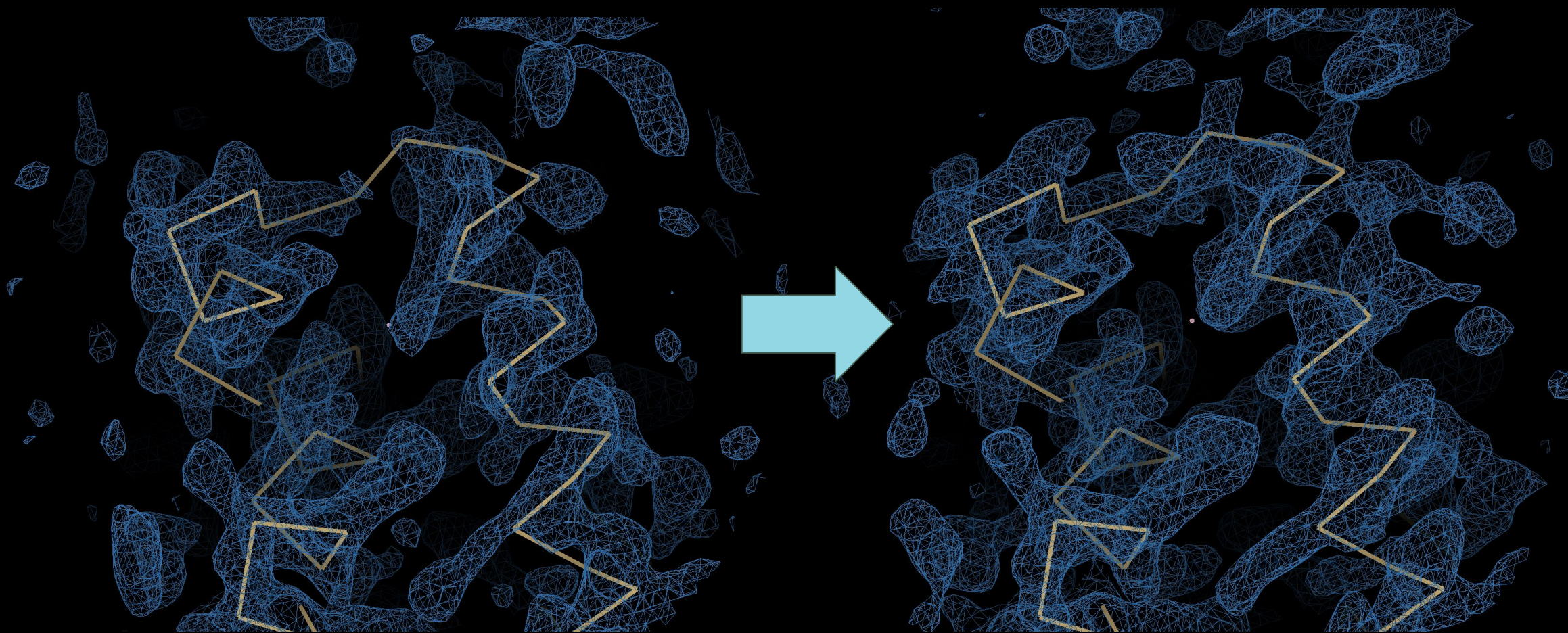


Main techniques:

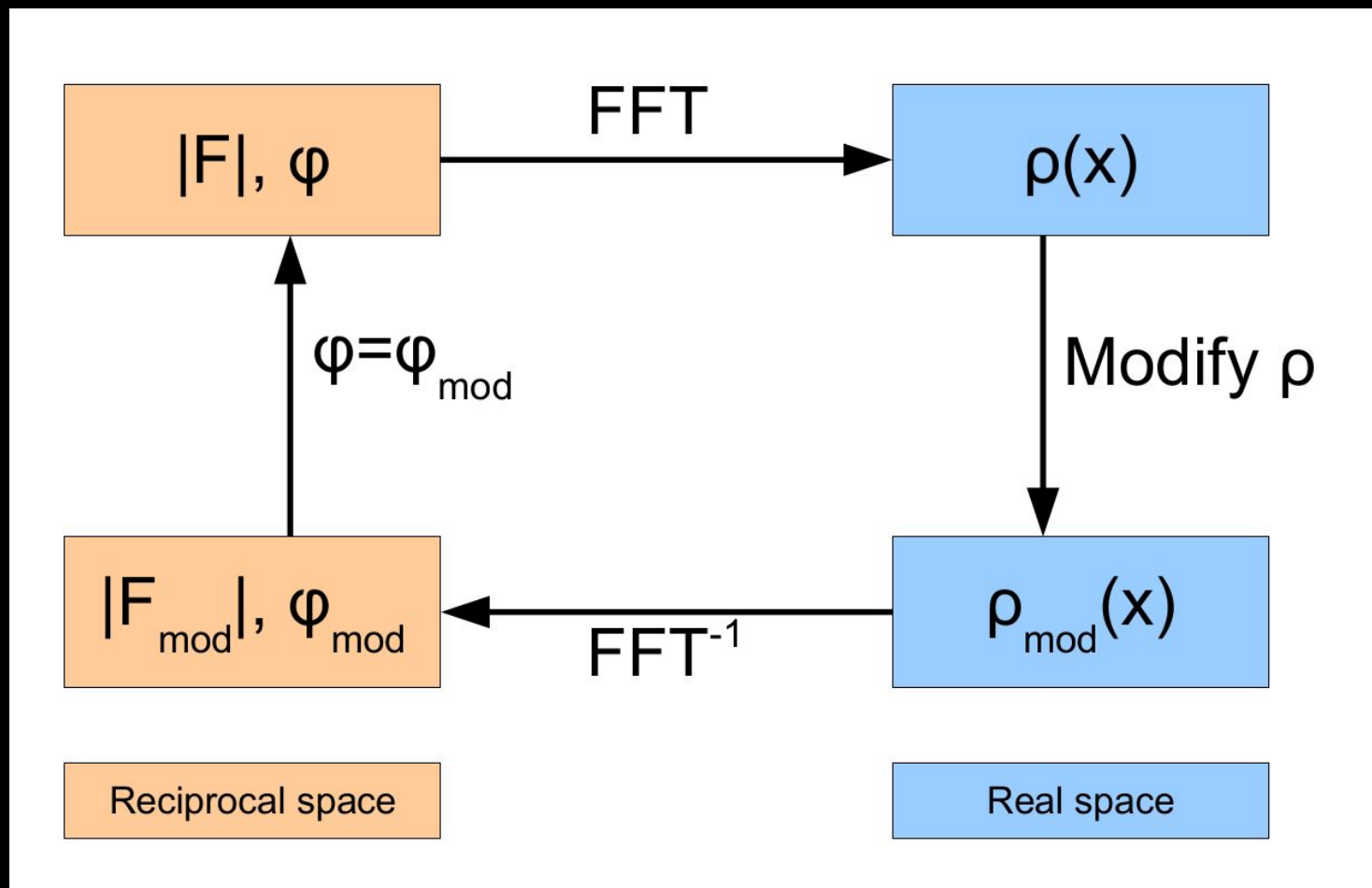
1. Solvent flattening
2. Histogram matching
3. NCS averaging
4. C-alpha tracing

(slide from Kevin Cowtan)

Density modification



Density modification



CCP4 Applications:

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

(slide from Kevin Cowtan)

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

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