Generating/optimising predicted models for MR

Adam Simpkin

Talk overview

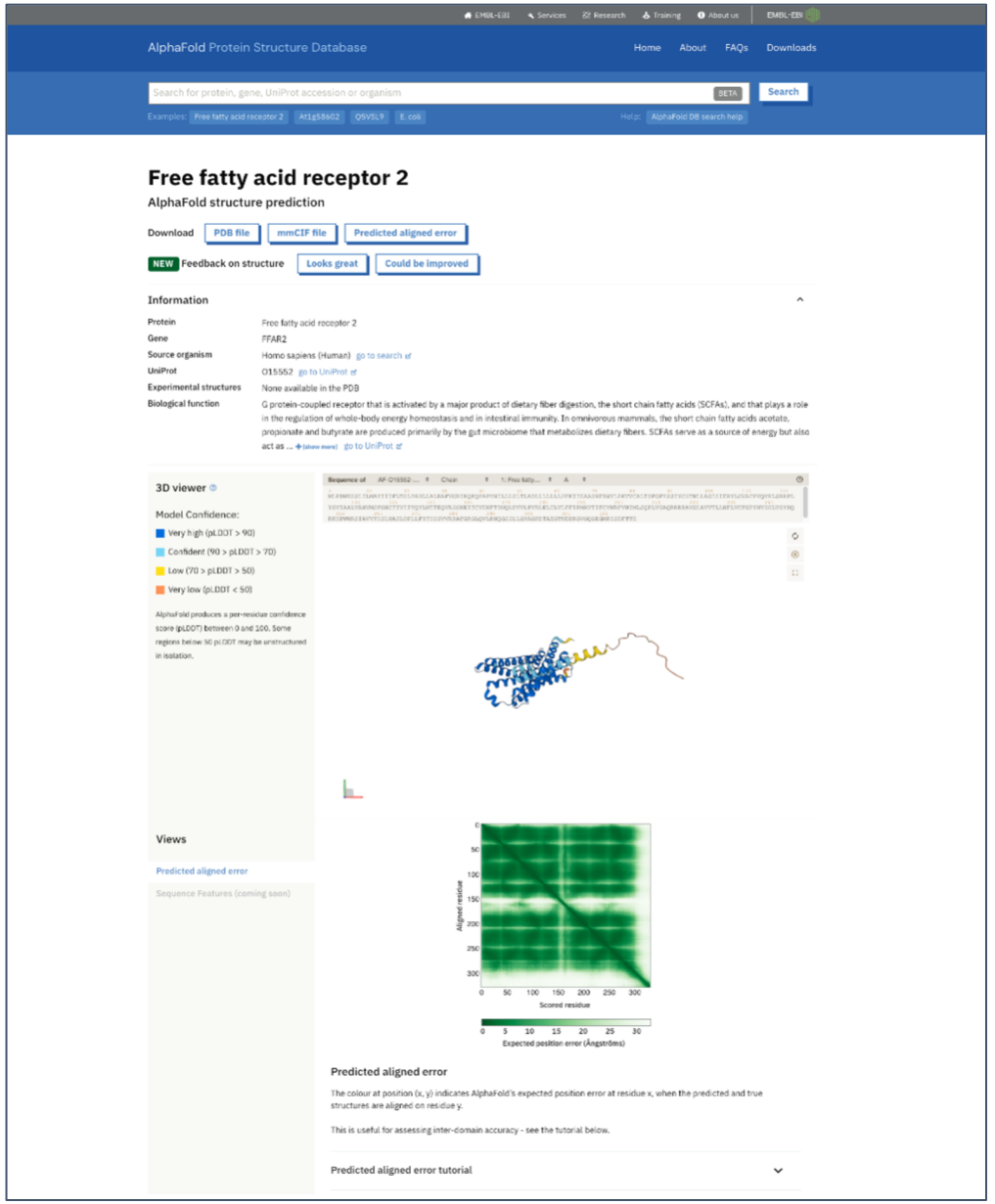
- The EBI AlphaFold database (AFDB)
  - How to query the AFDB for your target structure
  - How to assess the quality of hits in the AFDB
  - CCP4 tools that take advantage of the database
- Making your own models using Colab pages/online servers
  - ColabFold and AlphaFold's Colab page
  - Making complexes and other things to think about
  - Using RoseTTAFold's online server
  - Exploring conformational diversity with ColabFold
- ~~Making your own models locally~~
  - ~~What you need to run AlphaFold2 locally~~
- Optimising your models for MR
  - Problems with models
  - Slice'N'Dice

Obtaining models from AFDB

The EBI AlphaFold database

The AFDB should be your first port of call when looking for AlphaFold2 models for your target protein

- Initial release contained structure predictions for the human proteome and 20 other key organisms (~350,000 structures)
- December 2021, expanded to include manually curated UniProt (>400,000)
- Currently release contains >200 million models
  - Due to a compilation error, ~4% of these entries had low pLDDT entries
  - V4 of the database (releasing soon) will fix this

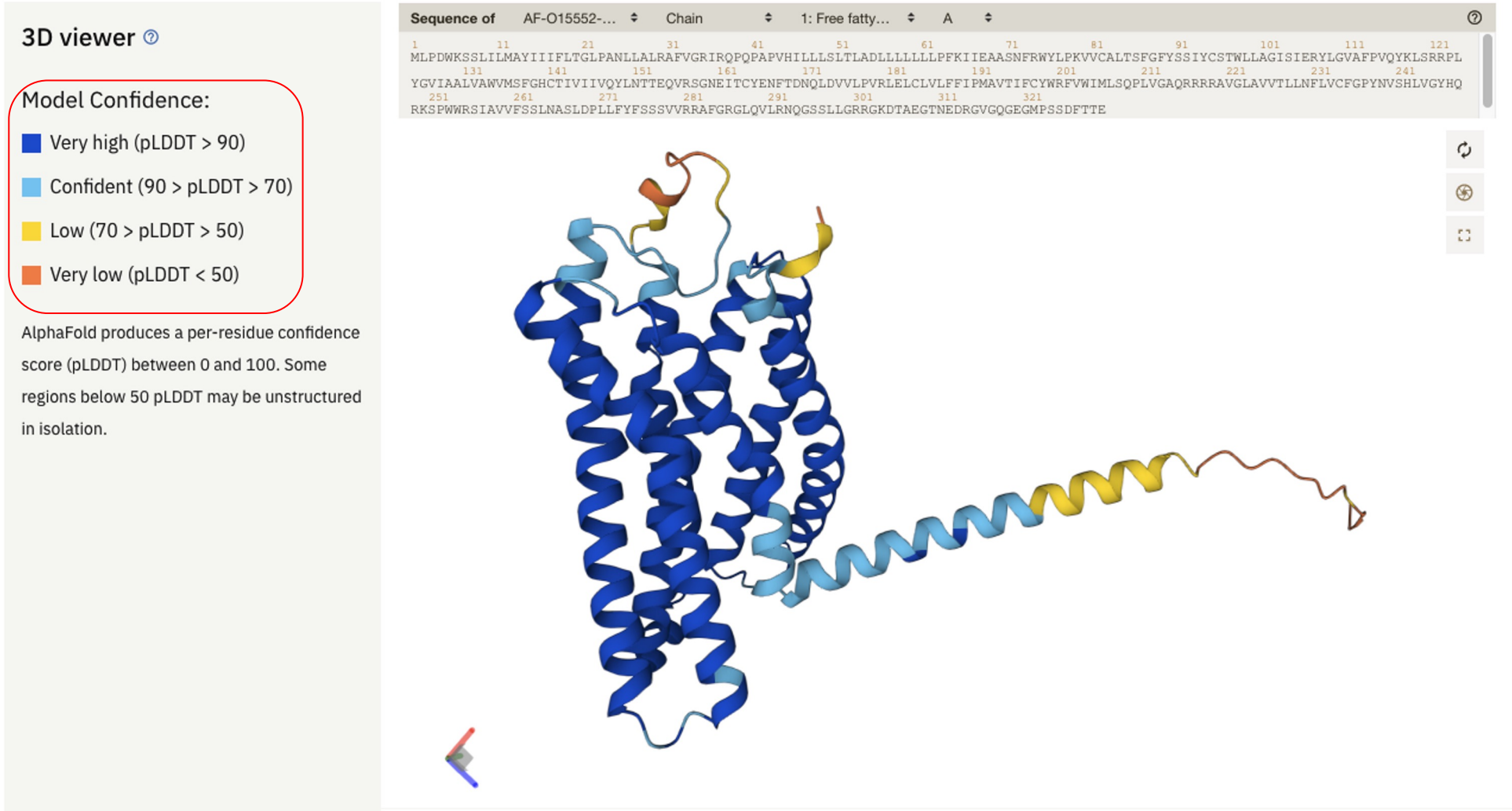


Varadi et al (2021) NAR, gkab1061

Searching the database

- Models can be obtained directly from https://alphafold.ebi.ac.uk/ using a UniProt ID or keywords relating to the protein
- For sequence search can use:
  - HMMER www.ebi.ac.uk/Tools/hmmer currently only searches older AFDB (2022-01-15)
  - MrParse in the CCP4 suite, searches UniprotKB and then finds which entries are in the AFDB (also does some model pre-processing to make them more effective in molecular replacement)
- For structural search can use:
  - DALI (http://ekhidna2.biocenter.helsinki.fi/dali/) is a well-established method, but only looks at one species at a time. Gives a Z-score. >2 non-random, ~>4 fold similarity
  - Foldseek (https://search.foldseek.com/search) uses Tmalign method to produce TM-score. Range is 0-1. Scans AFDB
  - RUPEE (https://ayoubresearch.com/) is a fast shape-based method. Maybe not the most sensitive. Scans all AFDB.

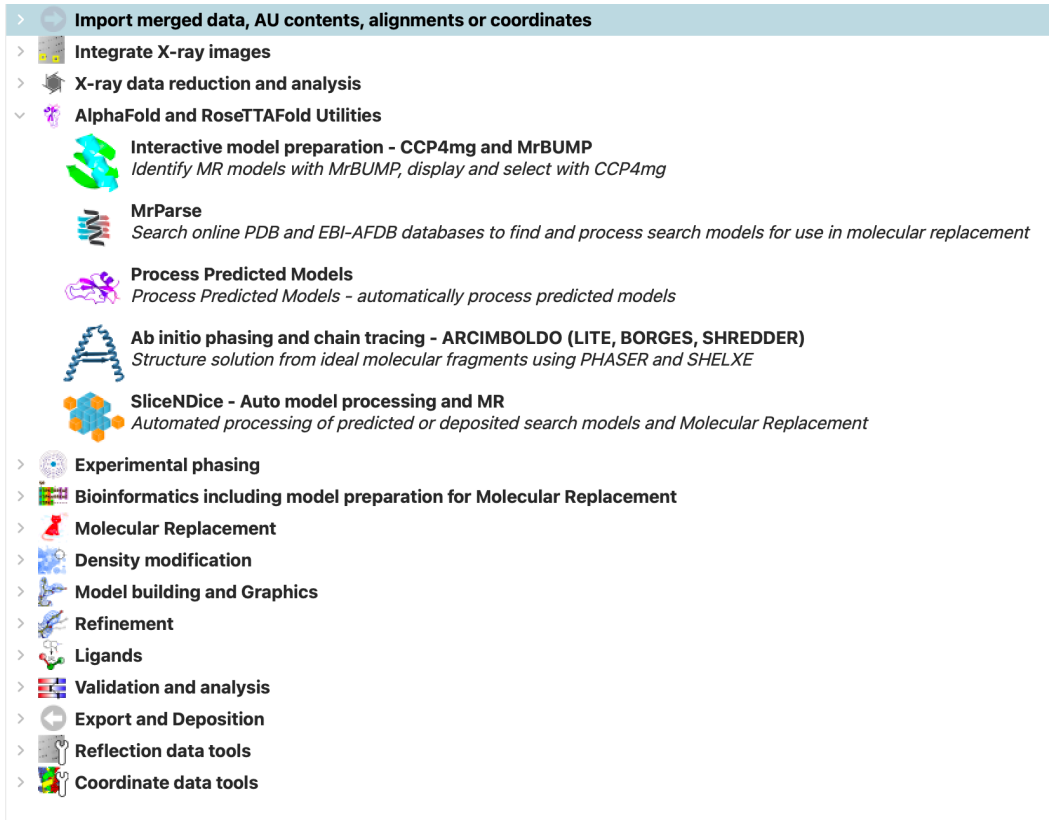
Assessing quality of models in the AFDB

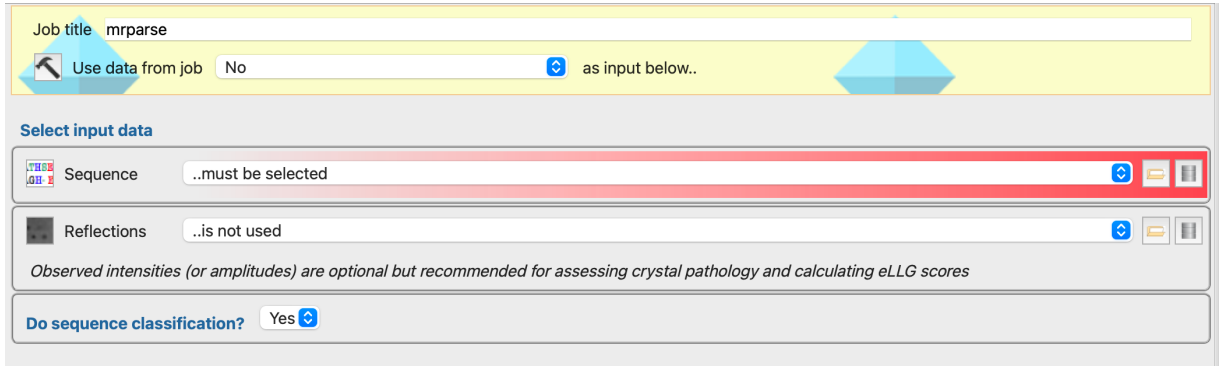


CCP4 programs that use the AFDB

MrParse

- Available through CCP4i2 and CCP4 Cloud
- Located under the AlphaFold and RoseTTAFold utilities tab
- Sequence must be provided, reflections are optional (used to calculate eLLG and provide a summary of the reflection data)





Avoiding the miscompiled models

MrParse

MrParse Analysis

Version: 0.3.5

MrParse: a program to find and analyse search models for crystallographic Molecular Replacement. The program is being developed by Dan Rigden's group 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 it's development, or ideas on how we could improve it, please get in touch.

Experimental structures from the PDB

Name	PDB	Resolution	Region	Range	Length	eLLG	Mol. Wt.	eRMSD	Seq. Ident.
4mo1_B	4mo1	2.10	1	2-144	141		15231		0.99

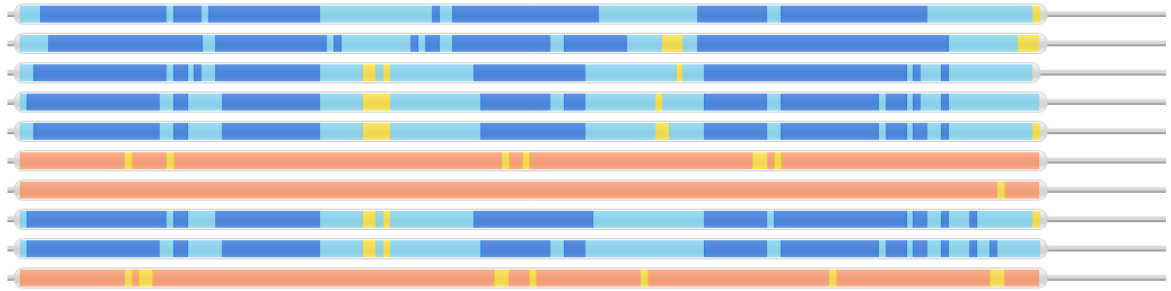
Visualisation of Regions

** Sequence Based Prediction step was skipped: append `--do_classify` argument to run **

Structure predictions from the EBI AlphaFold database

Name	model	Date Made	Region	Range	Length	Avg. pLDDT	H-score	Seq. Ident.
AF-A0A6B5D4Y7-F1	AF-A0A6B5D4Y7-F1	01-JUN-22	1	1-147	145	90.31	84	0.99
AF-A0A7R8VQK2-F1	AF-A0A7R8VQK2-F1	01-JUN-22	1	1-147	145	90.11	85	0.99
AF-A0A644RJ15-F1	AF-A0A644RJ15-F1	01-JUN-22	1	1-146	144	88.03	81	0.99
AF-A0A5U9QSB5-F1	AF-A0A5U9QSB5-F1	01-JUN-22	1	1-147	145	87.68	81	0.99
AF-A0A3U6ZW02-F1	AF-A0A3U6ZW02-F1	01-JUN-22	1	1-147	145	87.63	81	0.99
AF-A0A4Q8JJP8-F1	AF-A0A4Q8JJP8-F1	01-JUN-22	1	1-147	145	40.54	41	0.99
AF-A0A7U9FWD5-F1	AF-A0A7U9FWD5-F1	01-JUN-22	1	1-147	145	34.25	35	0.99
AF-A0A860YL56-F1	AF-A0A860YL56-F1	01-JUN-22	1	1-147	145	88.65	82	0.98
AF-A0A0K3YHM2-F1	AF-A0A0K3YHM2-F1	01-JUN-22	1	1-147	145	87.93	82	0.98
AF-A0A0A0FDZ1-F1	AF-A0A0A0FDZ1-F1	01-JUN-22	1	1-147	145	36.53	37	0.98

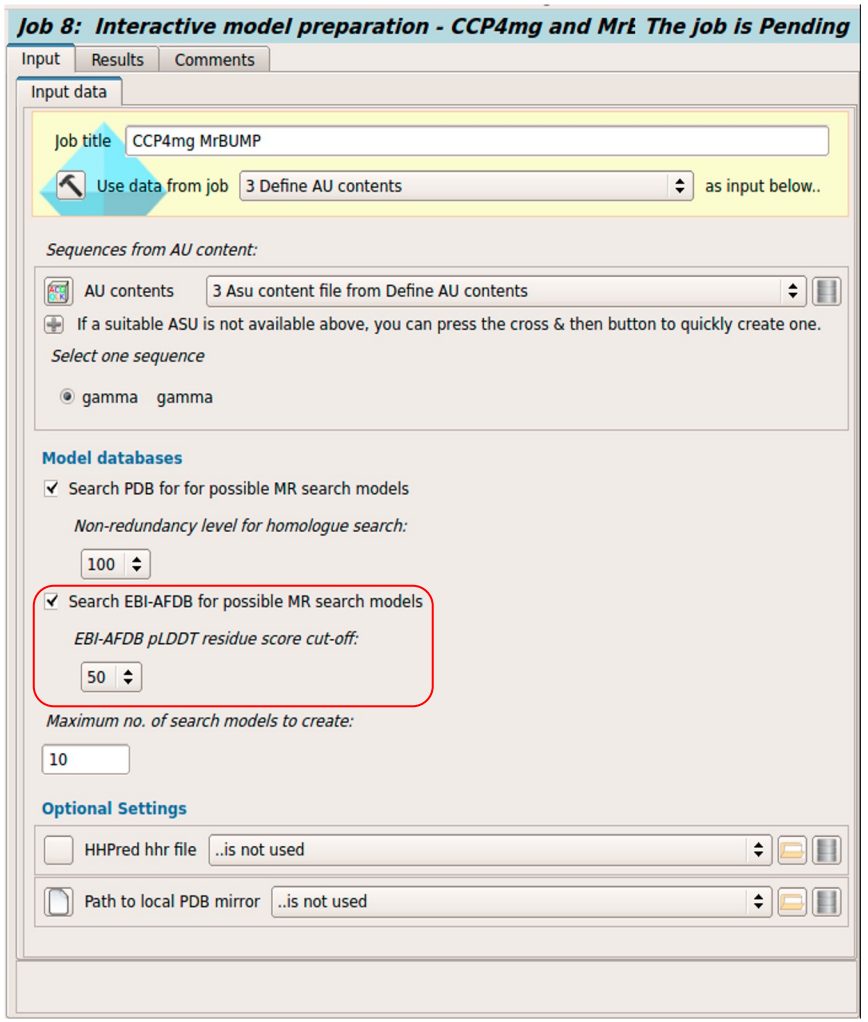
Visualisation of Regions



Problematic entries are very obvious

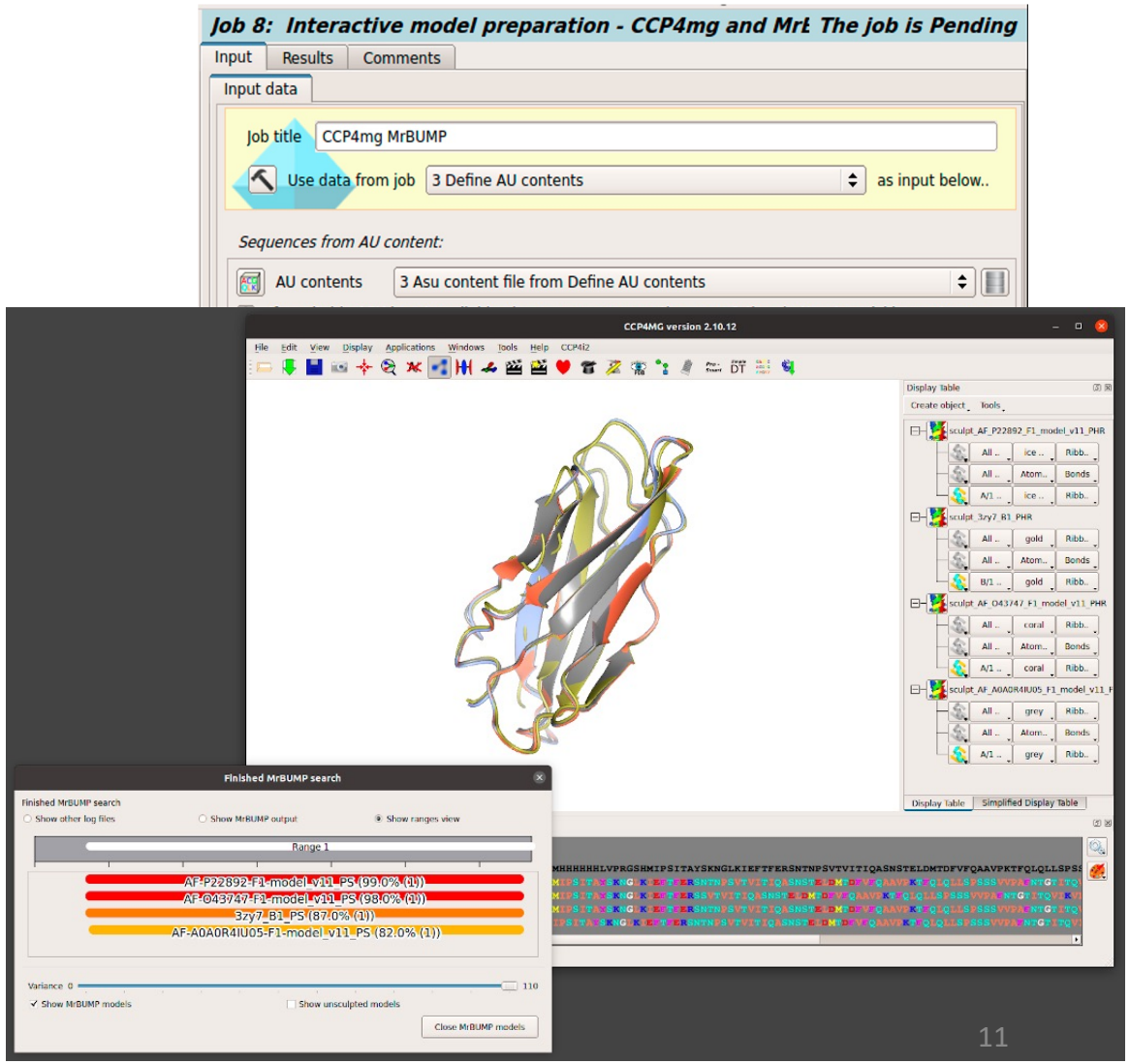
Using AlphaFold database models in MrBUMP

- AFDB models can now also be used in MrBUMP via CCP4mg
- Can select threshold score so that models are retrieved without low quality regions
- Ensembles can be made from PDB and AFDB entries combined
- Slider allows interactive truncation
- Full instructions available here: https://ccp4i2.gitlab.io/rstdocs/tutorials/alphafold/alphafold_tutorial.html



Using AlphaFold database models in MrBUMP

- AFDB models can now also be used in MrBUMP via CCP4mg
- Can select threshold score so that models are retrieved without low quality regions
- Ensembles can be made from PDB and AFDB entries combined
- Slider allows interactive truncation
- Full instructions available here: https://ccp4i2.gitlab.io/rstdocs/tutorials/alphafold/alphafold_tutorial.html



Pros and cons of AFDB

Pros	Cons
Immediately available	Not made with today's data
Likely to be a close match to your target	Only monomers
Large structures (up to ~2700 residues)	Only a single model, i.e. no conformational variation
Nice interactive output	Unable to vary AlphaFold2 settings to optimise modelling
PAE json file available	Temporarily, some models have errors

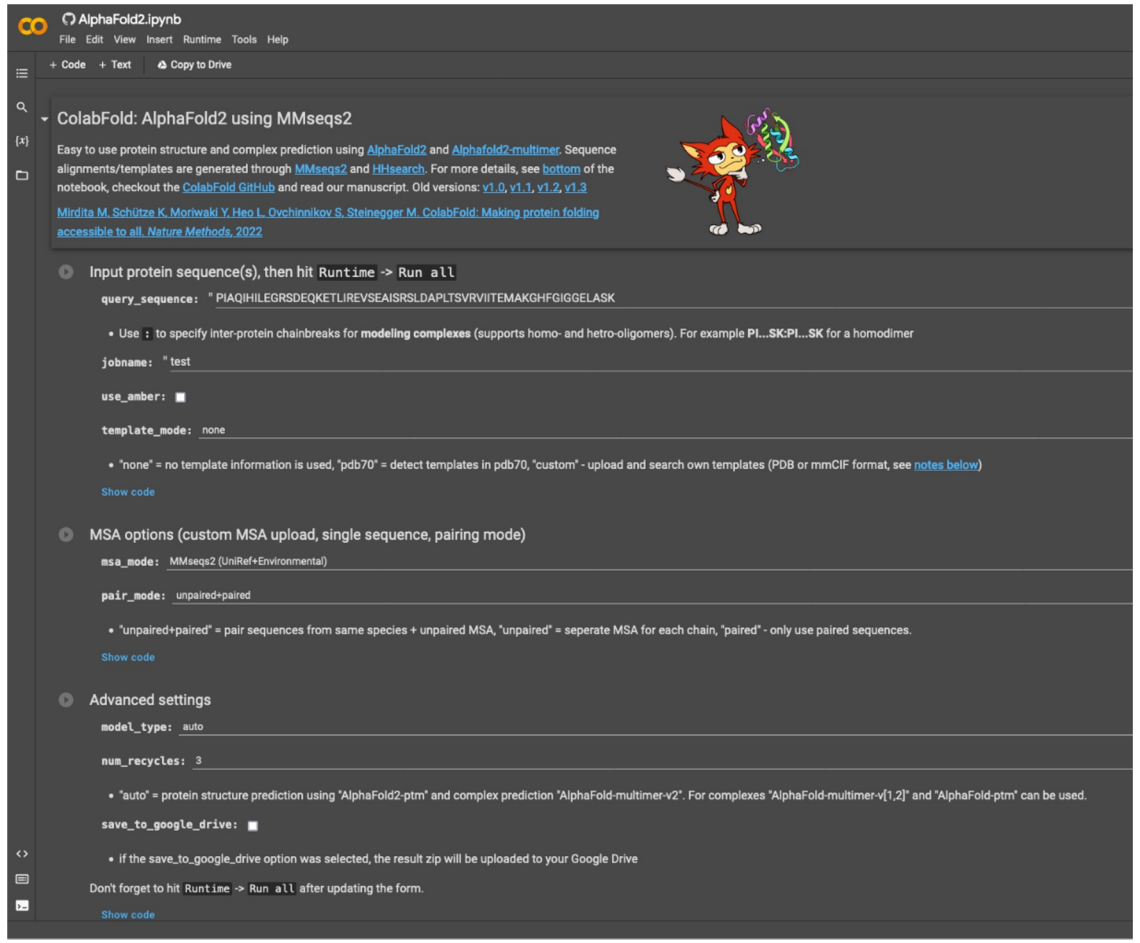
Making your own models using Colab pages

ColabFold and other Colab pages

A variety of Colab pages run AlphaFold2 on free Google resources

These differ in options for:

- MSA generation (MMSeqs2 vs JackHMMer)
- Multimer modelling
- Template usage

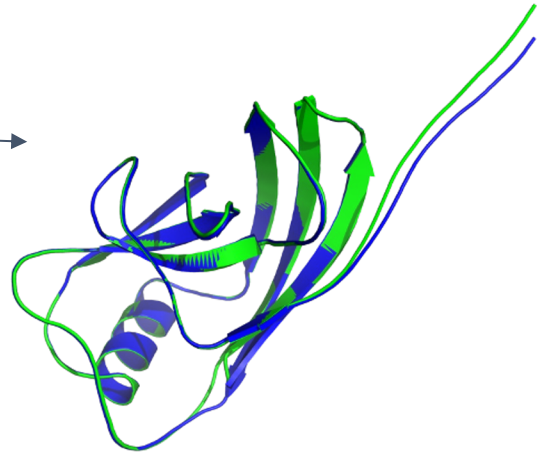


Mirdita, M., Schütze, K., Moriwaki, Y. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* (2022). https://doi.org/10.1038/s41592-022-01488-1

Colab comparison

Method	Monomers	Complexes	MMSeqs2	JackHMMer	Templates
AlphaFold colab	Yes	Yes	No	Yes	No
ColabFold	Yes	Yes	Yes	No	Yes
ColabFold Advanced	Yes	Yes	Yes	Yes	Yes

- MMSeqs2 runs significantly quicker than JackHMMer
  - Sequence coverage in the MSA is lower, however in the majority of cases this doesn't affect the quality of the output model
- ColabFold can accept template models uploaded by users
- Complexes are created differently:
  - AlphaFold colab and ColabFold use AlphaFold-multimer code to produce complexes.
  - ColabFold Advanced uses a residue-index mode (Similar to how RoseTTAFold makes complexes).
  - The AlphaFold-multimer code has been shown to produce the best complexes overall, however in certain instances the residue-index mode will create better complexes [1].



[1] Mirdita, M., Schütze, K., Moriwaki, Y. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* (2022). https://doi.org/10.1038/s41592-022-01488-1

Making models with RoseTTAFold online server

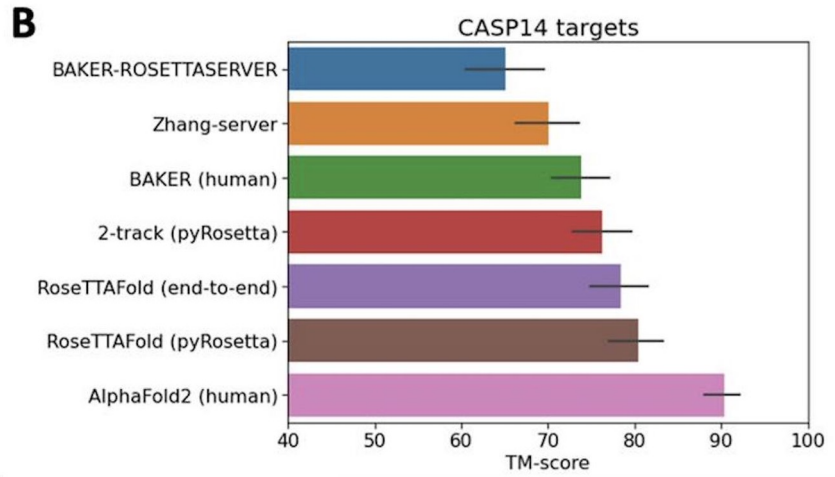
Shortly after CASP14, RoseTTAFold was released

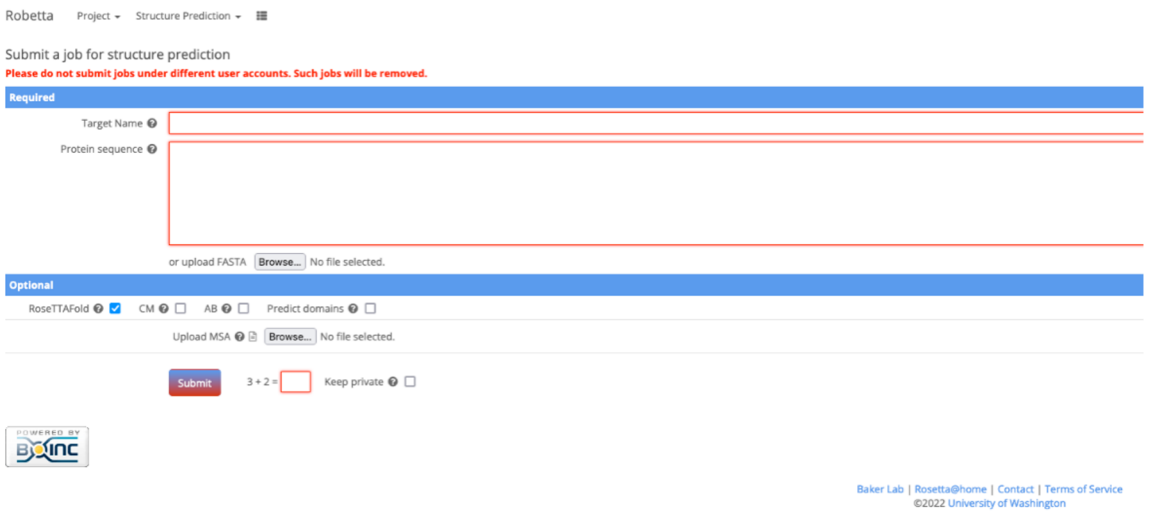
RoseTTAFold is also a deep learning based method that can produce highly accurate models

Takes template models

RoseTTAFold is available from https://robetta.bakerlab.org/

(nb registration required)



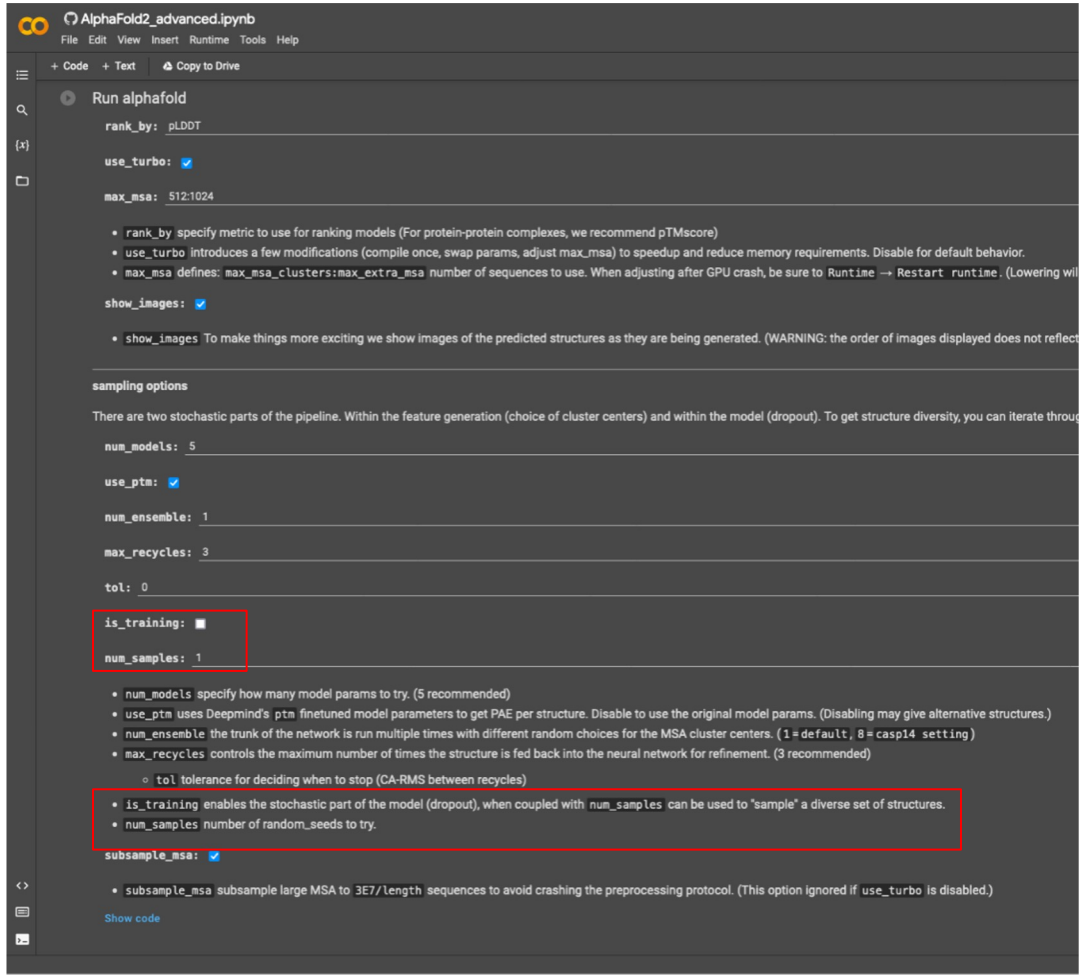


Exploring conformational diversity with ColabFold

The default pipeline often produces a set of similar models. If your protein has multiple conformations you may want to look for others

Ways to sample conformation more broadly

- Use the num_samples and/or is_training options on the advanced colab page
- Deliberately make the input MSA more shallow [1] (nb: this may reduce pLDDT scores)
- Feed AF2 templates in the 'right' conformation (and maybe ignore MSA features) [2]
- Edit the input MSA to mutate the Ala residue pairs that are driving the 'wrong' conformation [3] (have to provide own MSA)



[1] https://www.biorxiv.org/content/10.1101/2021.11.22.469536v1.full.pdf
[2] https://www.biorxiv.org/content/10.1101/2021.11.26.470086v1.full.pdf
[3] https://www.biorxiv.org/content/10.1101/2021.11.29.470469v1.full.pdf

Pros and cons of Colab pages

Pros	Cons
You can model your exact protein	Can only make models up to ~1400 residues (Depends on GPU assigned)
Can make complexes	
Can specify template proteins	
Made with latest databases	
Only requires an internet connection to run	
Lots of runtime options e.g. conformation diversity	
Nice looking output	

Advantages of making models locally

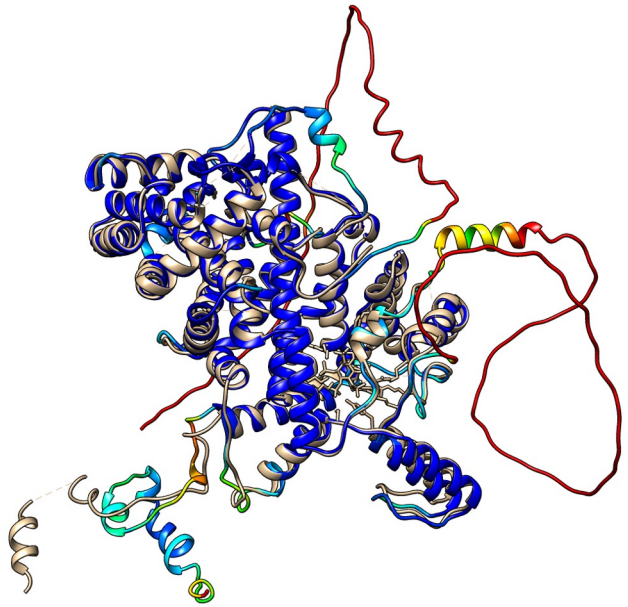
Pros and cons of local runs

Pros	Cons
You can model your exact protein	Less intuitive to run
Can typically make much bigger models	Requires decent hardware
Can make complexes	Less informative output?
Made with latest databases	

Optimising models for MR

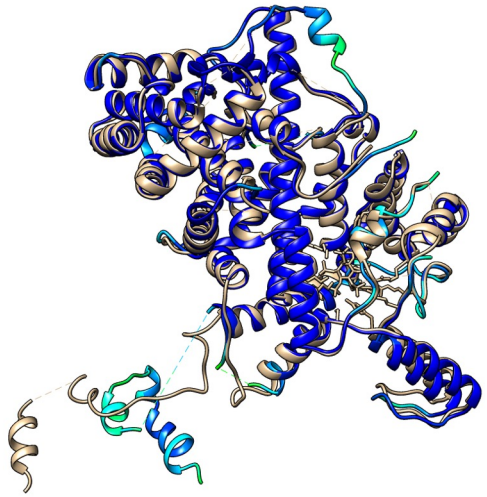
Problems with predicted models

- The quality of a predicted model is not uniform
- Some regions will be poorly predicted
  - These regions don't provide any useful information and can interfere with crystal packing
- Model quality assessment (MQA) scores such as pLDDT can be used as a guide to remove the poorly predicted regions



AFDB: Q5VSL9

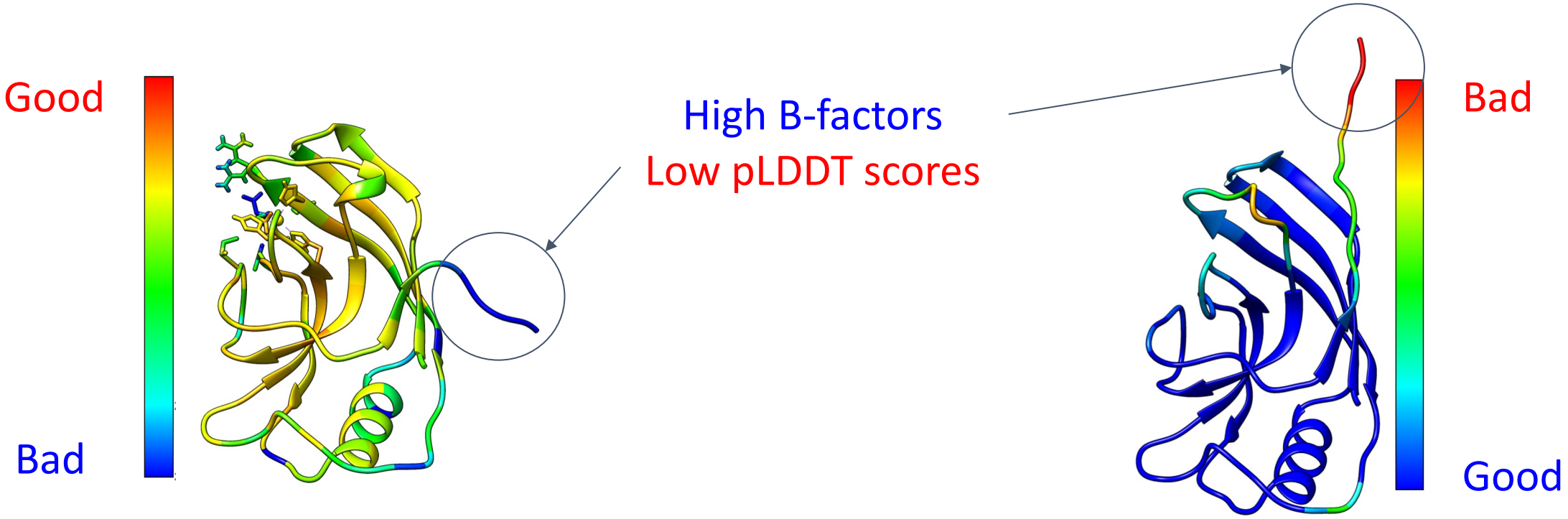
PDB: 7K36_I



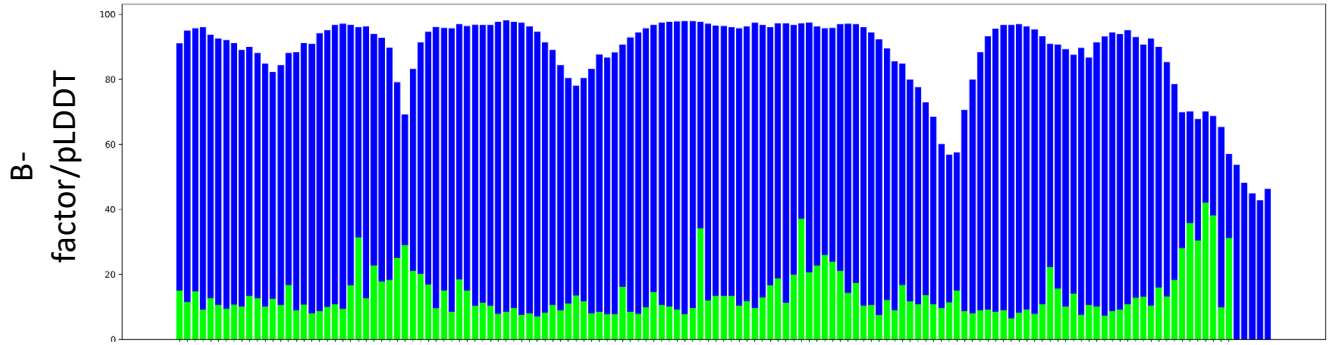
pLDDT threshold: 70

Problems with predicted models

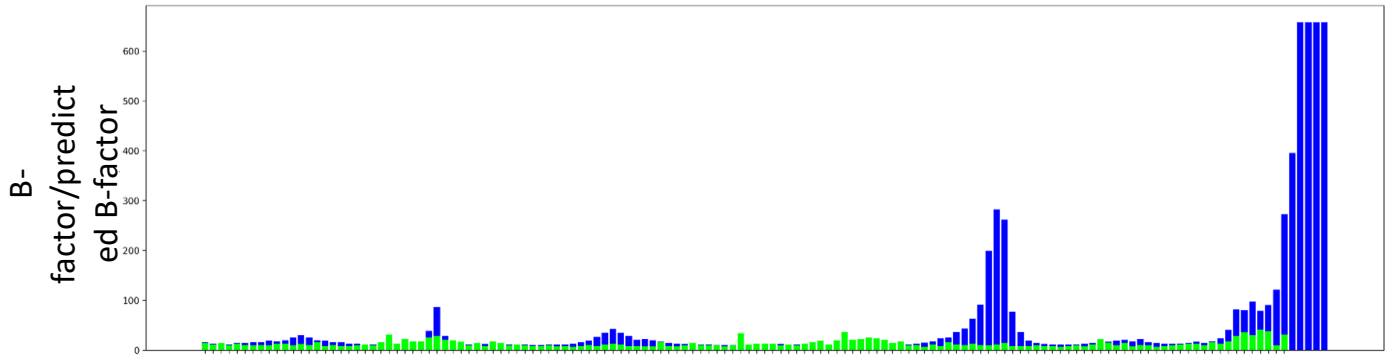
- MQA scores are often stored in the B-factor column of the PDB file
- B-factors are used weight residues in MR programs



Problems with predicted models



$$\frac{8 \times \pi^2}{3} \times \left(\frac{0.6}{pLDDT \times 0.01^3}\right)^2$$



Problems with predicted models

Many programs can now automatically remove low pLDDT regions and convert pLDDT scores to predicted B-factors, these include:

- CCP4 process predicted models
- Phenix.process_predicted_model
- MrParse (will correct for this when downloading models from the AFDB)
- MrBUMP (will correct for this when using models from the AFDB)
- Slice’N’Dice

Problems with predicted models

- Conformational differences between the predicted model and your target protein
- Individually, protein domains can be very accurately modelled, however if the orientation of these domains is incorrect relative to one another, the model may not work in molecular replacement

Alphafold2 model
Crystal structure


SLICENDICE

SLICENDICE

Composed of two elements

SLICE

Slices model up into distinct structural units using clustering methods.

Clustering input data:

- Predicted Aligned Error (PAE)
- C-alpha coordinates

Clustering methods fall into two categories:

- Manual clustering methods - number of clusters to produce must be specified
- Automatic clustering methods - number of clusters is automatically determined by the method

DICE

Performs automated Molecular replacement or Map fitting using the ‘sliced’ up model

Molecular Replacement uses:

- Phaser
- Molrep

Map Fitting uses:

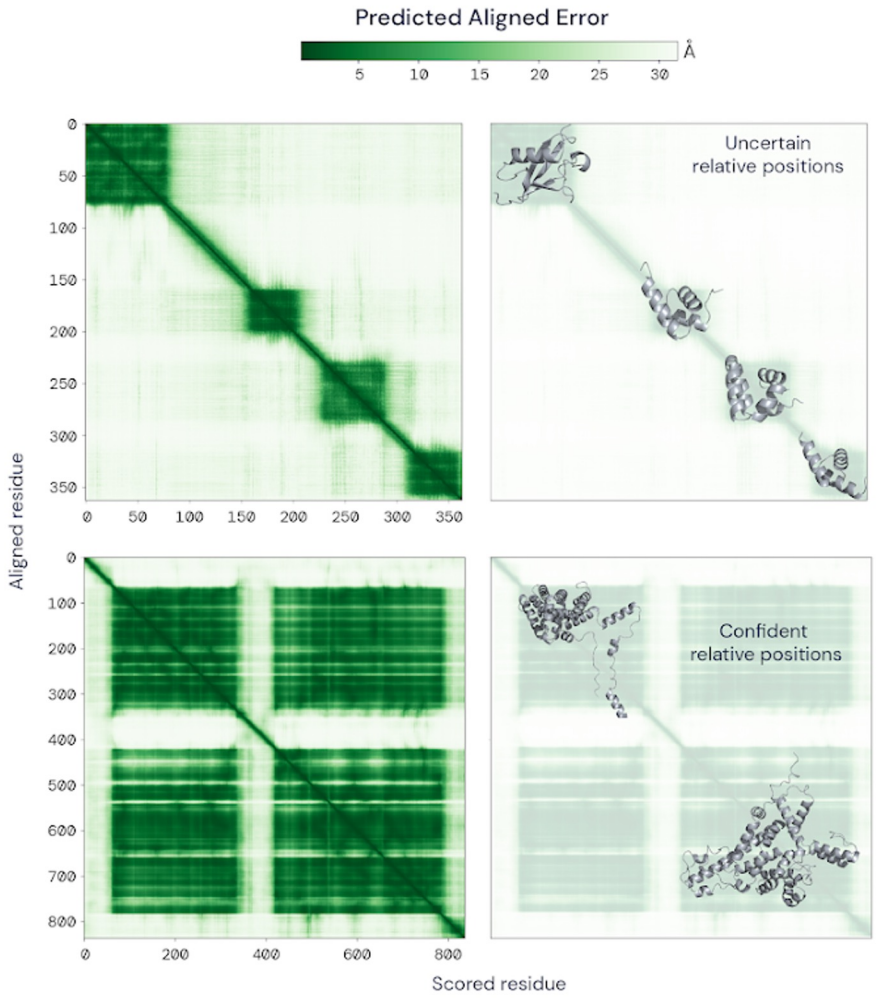
- Molrep
- Powerfit

Clustering Algorithms

Clustering on PAE

AlphaFold2 produces a metric called Predicted Aligned Error (PAE) that measure the confidence of residues relative to one another

Tristan Croll has created a tool in CCTBX that uses the information in the PAE file to identify individual structural units in an AlphaFold2 model

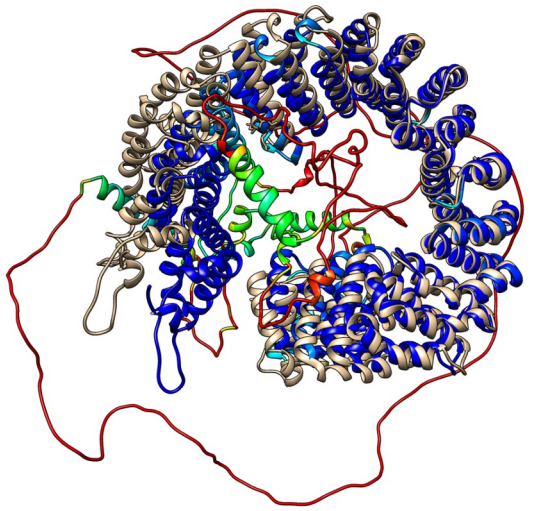


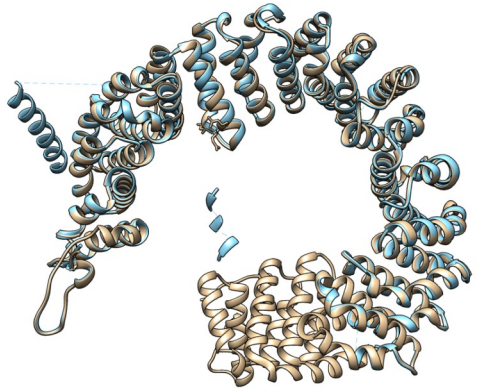
Clustering on C-alphas

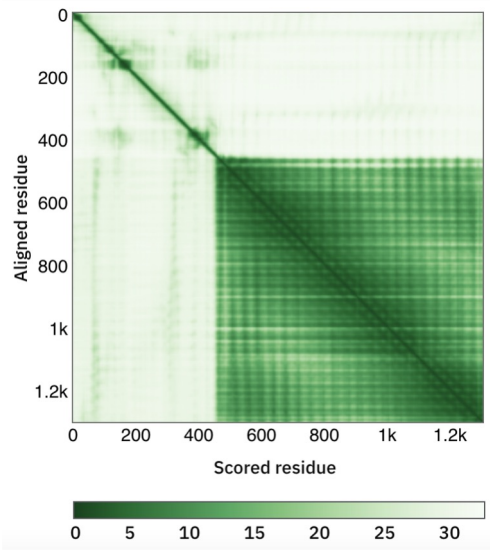
PAE is not always the best way to cluster models

SciKitLearn clustering algorithms that use the XYZ coordinates of C-alpha atoms can:

- Cluster predicted models from other sources (e.g. RoseTTAFold)
- Identify structural units when PAE is uninformative







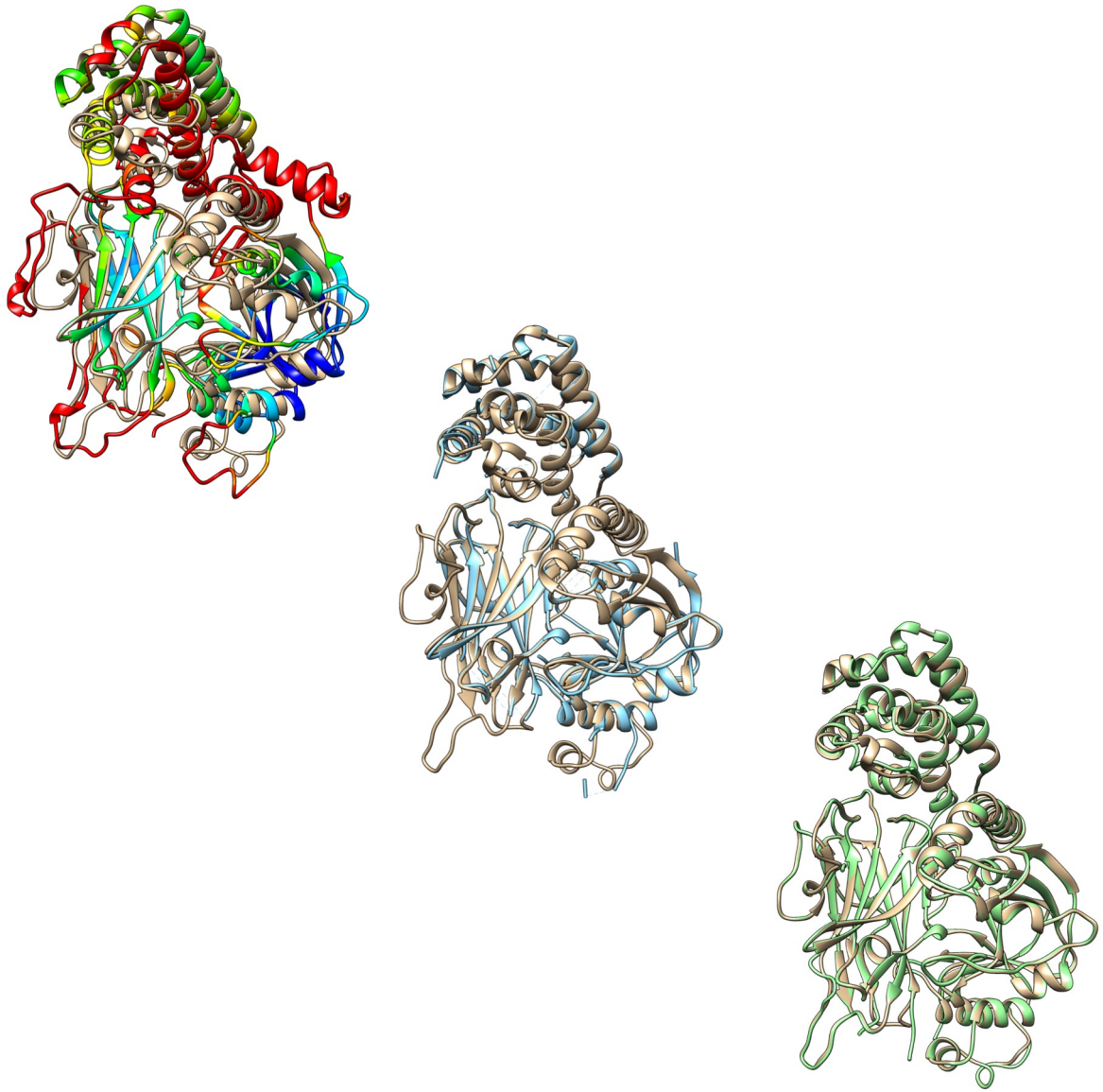
MX examples

Example 1 - 7OA7

- Crystal Structure solved by SAD
- Closest hit in the PDB, 12% seq ID
- AlphaFold2 model has avg pLDDT of 85.61
- Model was unable to solve the structure by MR
- Split into two using the Birch algorithm in Slice’N’Dice (C-alpha based)
- Phaser was able to solve the structure extremely easily (LLG: 1339, TFZ: 35.6)

Example 2 - 7RB4

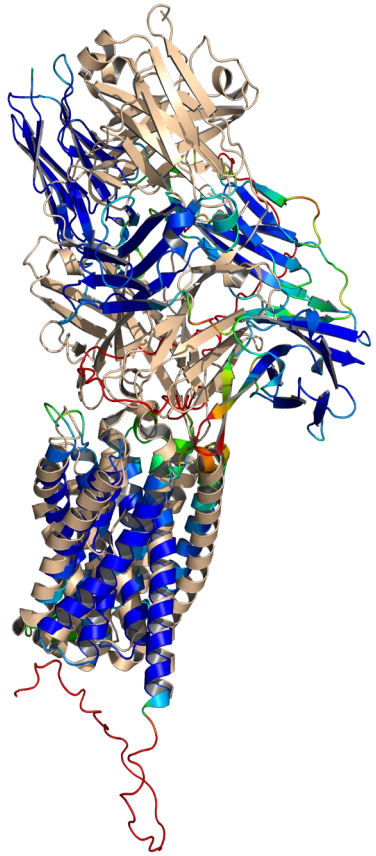
- Crystal Structure solved by SAD
- Closest hit in the PDB, 26% seq ID
- Only 26 sequences found in the MSA
- AlphaFold2 model has avg pLDDT of 61.02
- Model was unable to solve the structure by MR
- Split into three using the Birch algorithm in Slice’N’Dice (C-alpha based) with pLDDT threshold of 50
- Phaser was able to solve the structure (LLG: 114, TFZ: 13.2)
- Unable to solve with pLDDT threshold of 70

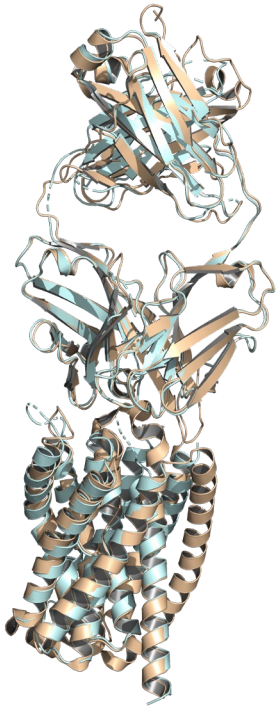


Cryo-EM example

Example 3 - 7FCI or EMD-31526

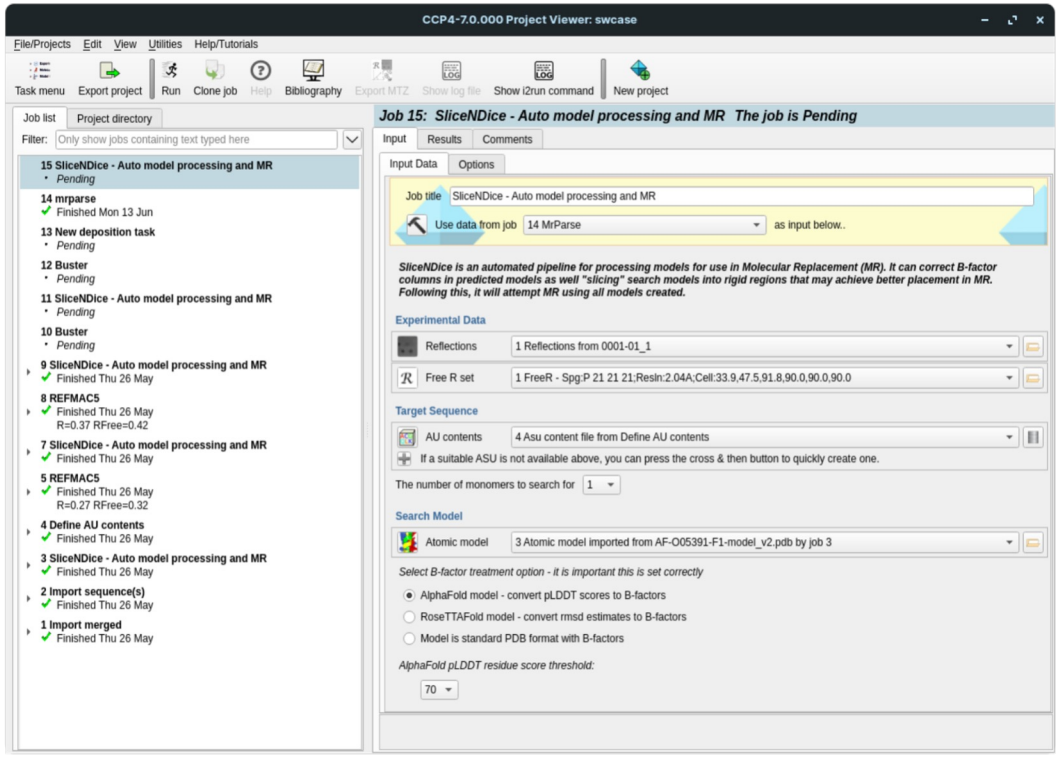
- NTCP in complex with YN69083 Fab
- Resolution: 3.30 Å map
- Initial AlphaFold2 model very different to PDB (RMSD: 5.1 Å)
- Split into four using the Birch algorithm in Slice’N’Dice (C-alpha based)
- Each slice placed using Molrep
- Placed structure was much more similar to PDB (RMSD: 0.6 Å)



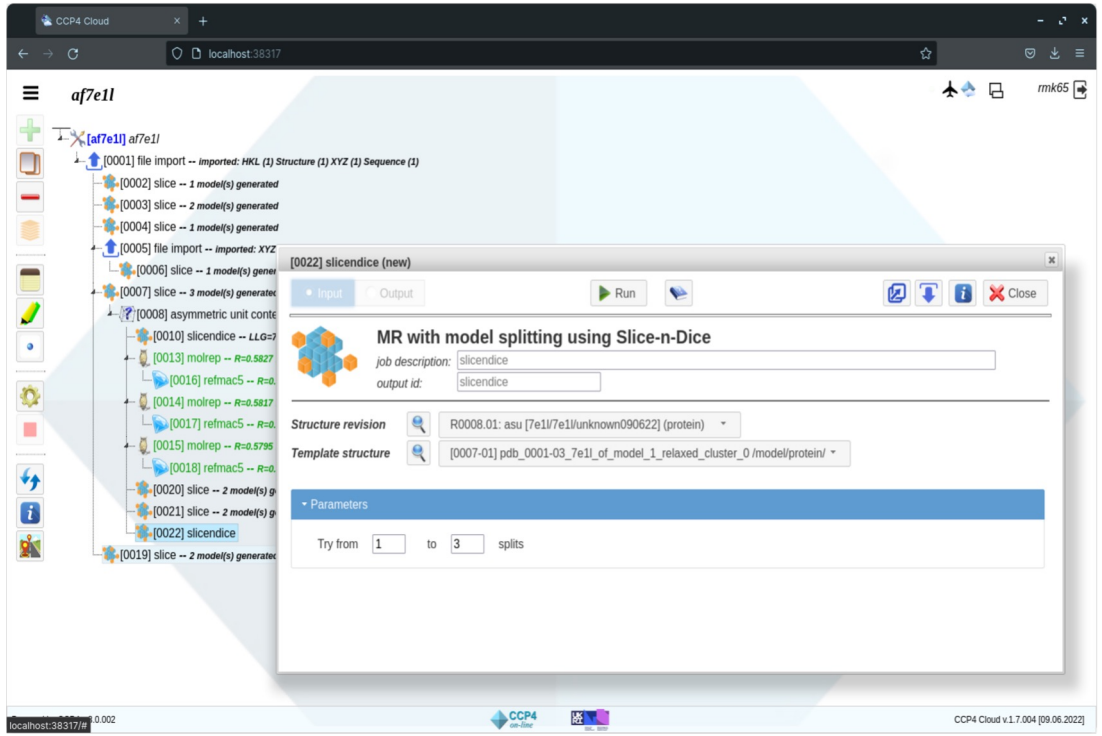


Availability

CCP4i2



CCP4cloud



Acknowledgements

Daniel Rigden

Ronan Keegan

Jens Thomas

Luc Elliott

Eugene Krissinel

Kyle Stevenson

Tom Burnley

hlasimpk@liverpool.ac.uk
Slack @Adam Simpkin

Slice'N'Dice: Maximising the value of predicted models for structural biologists
Adam J. Simpkin, Luc G. Elliott, Kyle Stevenson, Eugene Krissinel, Daniel J. Rigden, Ronan M. Keegan
bioRxiv 2022.06.30.497974; doi: https://doi.org/10.1101/2022.06.30.497974