# Complex, multiple copy molecular replacement example

Crystal structure of the RhsP2 C-terminal toxin domain in complex with its immunity protein, RhsI2

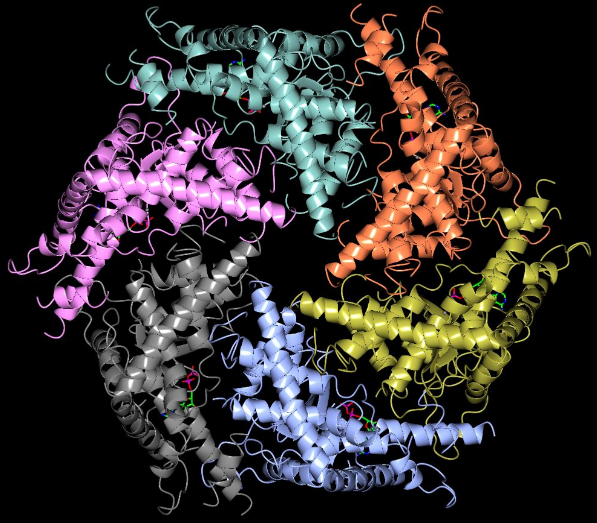
*Using complex/multimeric predictions in structure determination*

Ronan Keegan, CCP4

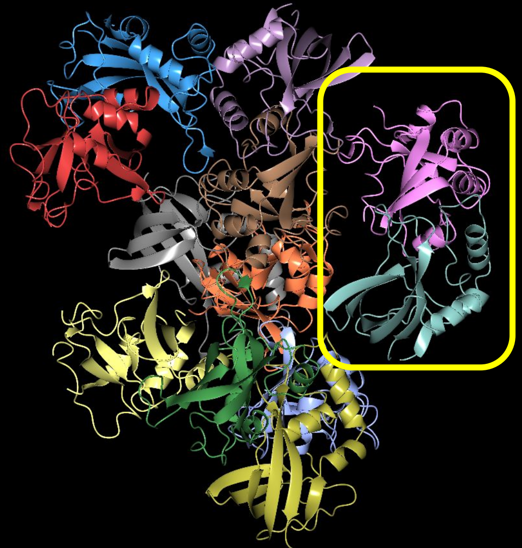
# Multimers and Complexes

# Multimers and Complexes

- In some cases, the Matthews Coefficient calculation can indicate the presence of many copies of the target structure in the asymmetric unit
- MR can be difficult in such cases
- A single monomer search model may not suffice despite being accurate due to it being a small fraction of the scattering content of the crystal. It can have too weak a signal in the MR search for the correct placement to be identified against the noise inherent in any diffraction dataset
- In these cases, a good strategy is to create a multimeric or complex model, increasing the signal of the search model and aiding its correct placement in MR



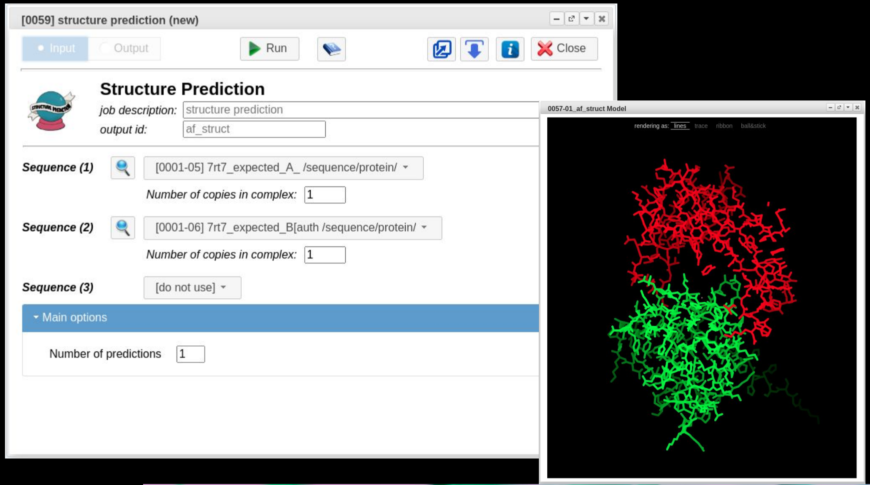
***7zbh - ATP-dependent zinc metalloprotease** – 6 fold symmetry (hexamer)*



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

# Multimers and Complexes

- Generating predicted multimeric and complex models:
  a. *CCP4Cloud* Structure Prediction task
  b. *AlphaFold3*/*Boltz-1*/*Chai-1* servers
  c. *Uni-fold* Colab Notebook


[0059] structure prediction (new)
Input
Output
Run
Close
0057-01_af_struct Model
rendering as: lines trace ribbon ball&stick

**Structure Prediction**

*job description:* structure prediction

*output id:* af_struct

***Sequence (1)*** [0001-05] 7rt7_expected_A_ /sequence/protein/

*Number of copies in complex:* 1

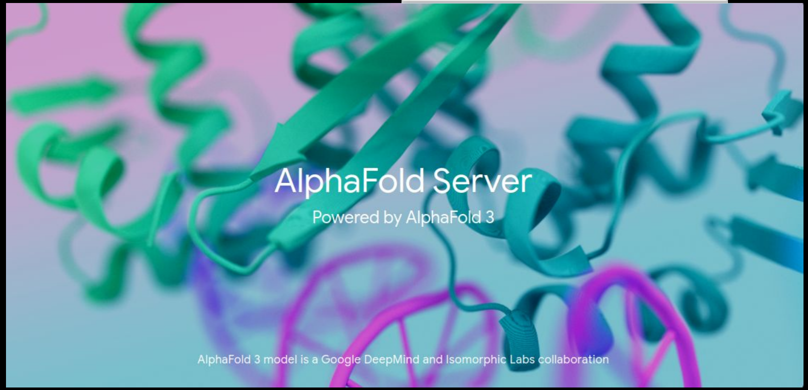
***Sequence (2)*** [0001-06] 7rt7_expected_B[auth /sequence/protein/

*Number of copies in complex:* 1

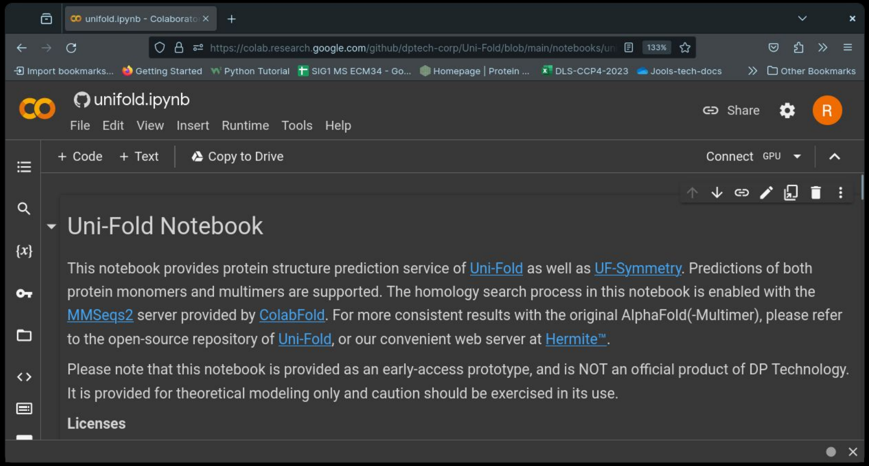
***Sequence (3)*** [do not use]

Main options

Number of predictions 1



AlphaFold 3 model is a Google DeepMind and Isomorphic Labs collaboration

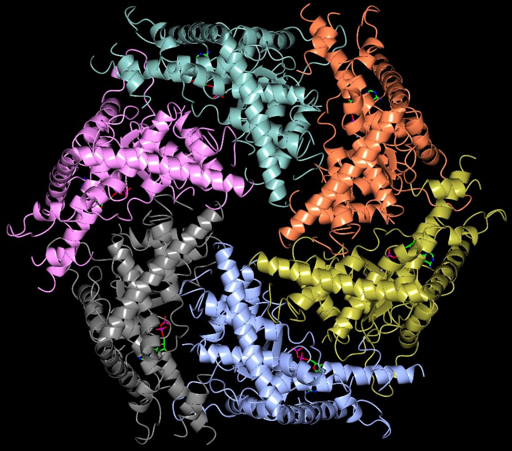


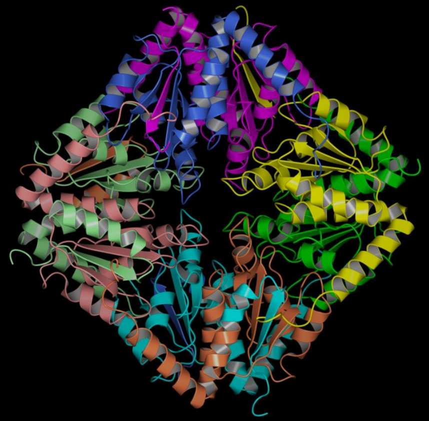
## Uni-Fold Notebook

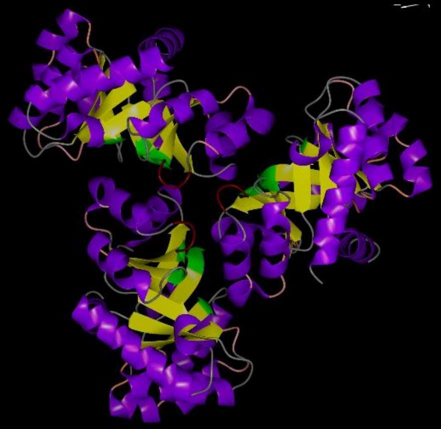
This notebook provides protein structure prediction service of Uni-Fold as well as UF-Symmetry. Predictions of both protein monomers and multimers are supported. The homology search process in this notebook is enabled with the MMSeqs2 server provided by ColabFold. For more consistent results with the original AlphaFold(-Multimer), please refer to the open-source repository of Uni-Fold, or our convenient web server at Hermite™.

Please note that this notebook is provided as an early-access prototype, and is NOT an official product of DP Technology. It is provided for theoretical modeling only and caution should be exercised in its use.

**Licenses**



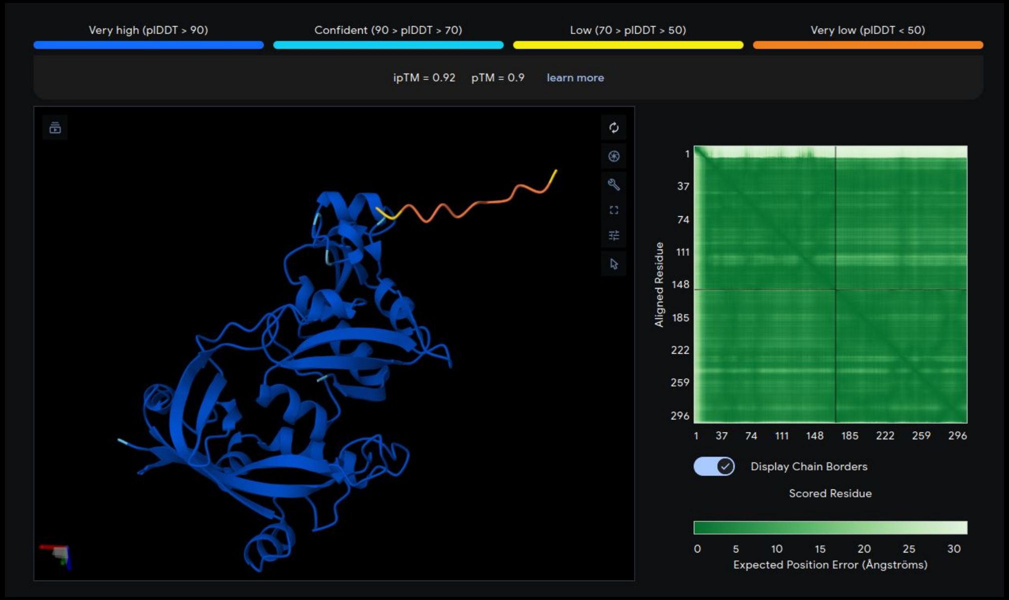




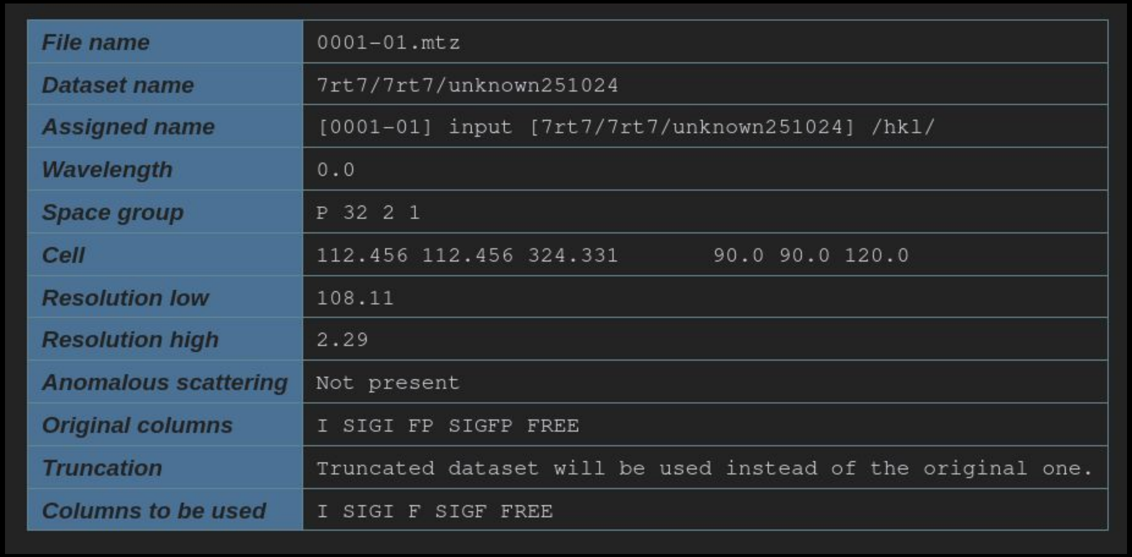
# Tutorial

# Data

- Start from merged reflection data (mtz file) and sequence information for protein
- Sequence contains two chains with multiple copies in the asymmetric unit
- Structure can be determined using molecular replacement with a predicted model for the protein complex



*AlphaFold3 server report page*



| | |
|---|---|
| *File name* | 0001-01.mtz |
| *Dataset name* | 7rt7/7rt7/unknown251024 |
| *Assigned name* | [0001-01] input [7rt7/7rt7/unknown251024] /hkl/ |
| *Wavelength* | 0.0 |
| *Space group* | P 32 2 1 |
| *Cell* | 112.456 112.456 324.331 90.0 90.0 120.0 |
| *Resolution low* | 108.11 |
| *Resolution high* | 2.29 |
| *Anomalous scattering* | Not present |
| *Original columns* | I SIGI FP SIGFP FREE |
| *Truncation* | Truncated dataset will be used instead of the original one. |
| *Columns to be used* | I SIGI F SIGF FREE |

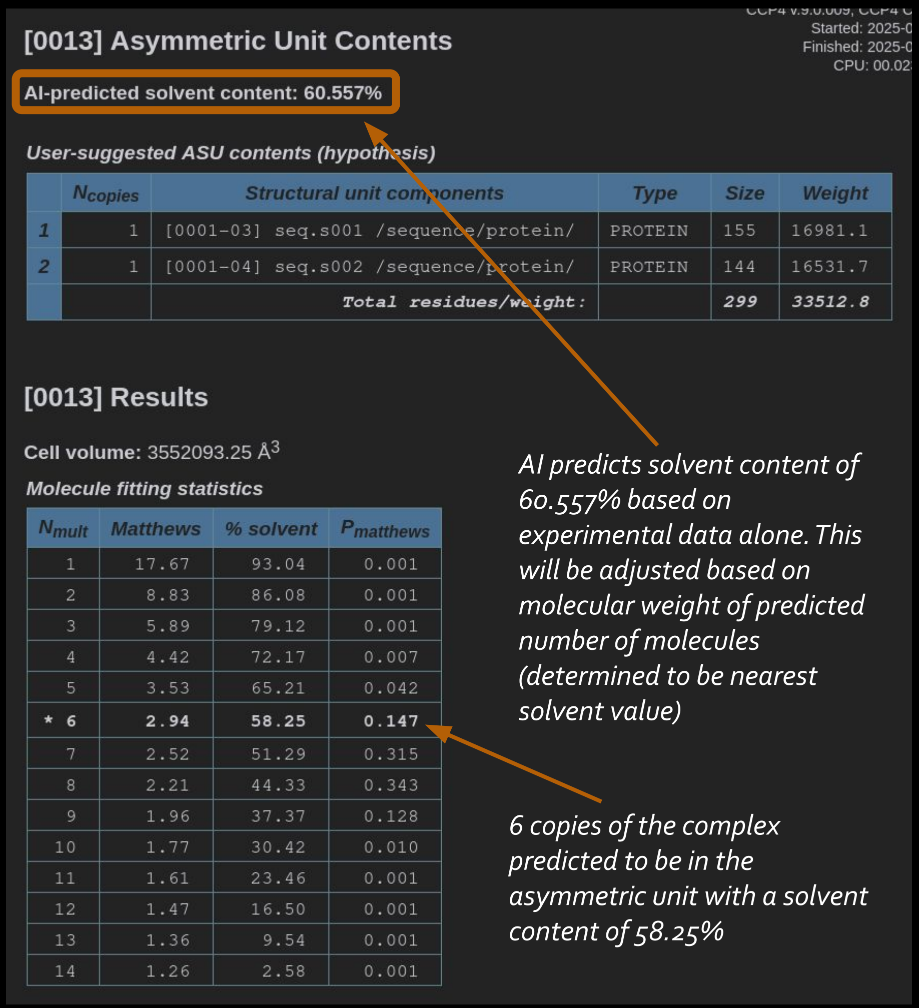
*Summary of the reflection data set*

- The data also includes a prediction of the complex from the AlphaFold3 server that we can use in molecular replacement
- AlphaFold3 is better in some instances than AlphaFold2 for predicting complexes and larger proteins

# Project setup and cell contents

- Create a **new project "MR-example-3"** and import the files from the Cloud storage "Tutorials" -> "Data" -> "2_phasing", "mr-3-complexes" folder. The files include a merged reflection mtz, fasta sequence file as well as a predicted complex model from the AlphaFold3 server. Note that the sequence file contains two chains, A and B, which are separated by CCP4Cloud.

- After the import, add an **asymmetric unit contents** task to determine the number of molecules to search for. Using the **AI-predicted** target solvent content will predict 6 copies of the complex in the ASU, whereas using the **hypothesized** (Matthews Coefficient) target solvent content will predict 7. The true number is actually 6 copies of the complex. It is common for the Matthews Coefficient to predict the contents inaccurately in cases where there are several copies of the molecule in the ASU, whereas the new AI-based method tends to be more accurate. In both instances, the number of copies is an estimate and the possibility that it is incorrect should be kept in mind throughout the structure solution process.



CCP4 v.9.0.009, CCP4 C
Started: 2025-0
Finished: 2025-0
CPU: 00.02

## [0013] Asymmetric Unit Contents

**AI-predicted solvent content: 60.557%**

*User-suggested ASU contents (hypothesis)*

| | $N_{copies}$ | Structural unit components | Type | Size | Weight |
|---|---|---|---|---|---|
| 1 | 1 | [0001-03] seq.s001 /sequence/protein/ | PROTEIN | 155 | 16981.1 |
| 2 | 1 | [0001-04] seq.s002 /sequence/protein/ | PROTEIN | 144 | 16531.7 |
| | | *Total residues/weight:* | | *299* | *33512.8* |

## [0013] Results

**Cell volume:** 3552093.25 Å$^3$

*Molecule fitting statistics*

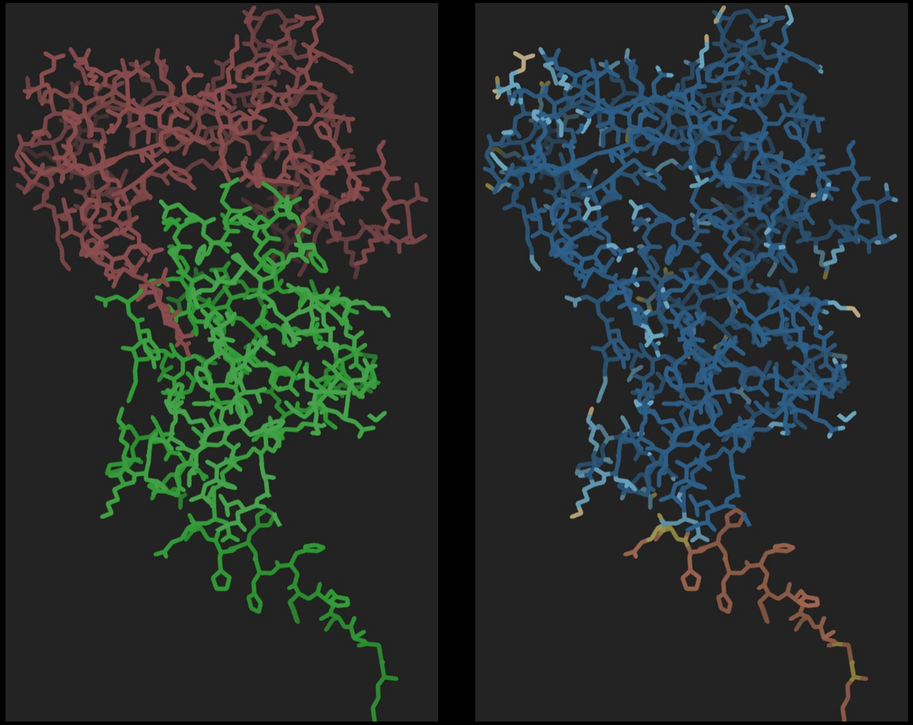
| $N_{mult}$ | Matthews | % solvent | $P_{matthews}$ |
|---|---|---|---|
| 1 | 17.67 | 93.04 | 0.001 |
| 2 | 8.83 | 86.08 | 0.001 |
| 3 | 5.89 | 79.12 | 0.001 |
| 4 | 4.42 | 72.17 | 0.007 |
| 5 | 3.53 | 65.21 | 0.042 |
| * 6 | **2.94** | **58.25** | **0.147** |
| 7 | 2.52 | 51.29 | 0.315 |
| 8 | 2.21 | 44.33 | 0.343 |
| 9 | 1.96 | 37.37 | 0.128 |
| 10 | 1.77 | 30.42 | 0.010 |
| 11 | 1.61 | 23.46 | 0.001 |
| 12 | 1.47 | 16.50 | 0.001 |
| 13 | 1.36 | 9.54 | 0.001 |
| 14 | 1.26 | 2.58 | 0.001 |

*AI predicts solvent content of 60.557% based on experimental data alone. This will be adjusted based on molecular weight of predicted number of molecules (determined to be nearest solvent value)*

*6 copies of the complex predicted to be in the asymmetric unit with a solvent content of 58.25%*

*Asymmetric Unit Contents CCP4Cloud report*

# Preparing search model for MR



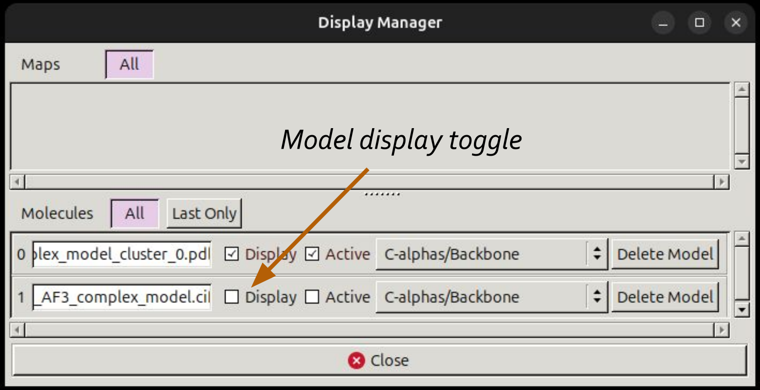
*The AlphaFold3 complex prediction as shown in Uglymol (coloured by chain (left) and pLDDT (right))*

- **Examine the AlphaFold3** predicted model in Uglymol
  - Colour it by chain ("C" key) to see the two chains present in the prediction
  - Multimeric or complex predictions do not always get the relative orientations of the two molecules correct but in many instances it can be accurate and it is always worth testing in cases like this
  - Colour the model by pLDDT and note that there are some residues that are low confidence (orange). These will need to be removed by the "Slice" task before we do MR
- Next add **a "Slice" task to generate a search model** for molecular replacement
  - Provide the AlphaFold3 prediction and ensure the number of splits is set to 1, we want to make the entire complex a single search model
  - Set the option "Correct B-factors" to "assuming Alphafold model" to convert pLDDT values to B-factor estimates (required by Phaser)
  - Compare the input model with the generated search model in Coot. To do this add an "Edit coordinates with Coot" from the "Coot" menu. Select both the input and output models from the Slice task. Set the display of both models to be "C-alpha/backbone" in "Display Settings". Toggle the output model from Slice on and off. The low confidence residues (pLDDT < 70) should have been removed

# Comparing models in Coot

Coot can be a valuable tool at many stages in the structure solution process. Here we can use it to compare the original AF3 model with the prepared search model generated from it using the Slice task

- From the previous "Slice" task add an "Edit coordinates in Coot" task from the "Coot" menu. Select both the input and output models from the Slice task (Structure to edit (1) and (2)). Press "Run"
- The Coot application will launch in a separate window. On this occasion press "yes" or "ok" in any dialogue boxes that appear. These will be important in finalising the structure before deposition but here we are just using Coot to view the models.
- Press the "Display Manager" button to control the model view. You can switch the models view to "C-alpha/backbone" to view them more easily. Use the "toggle" box to hide and show the output model from Slice. You should notice that the low confidence (pLDDT<70) terminal region in one of the chains has been removed by Slice.



Display Manager

Maps All

*Model display toggle*

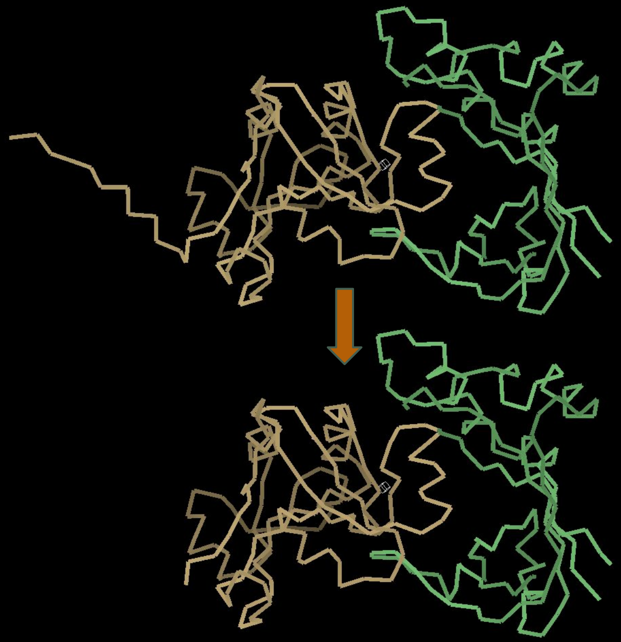
Molecules All Last Only

0 plex_model_cluster_0.pdl ☑ Display ☑ Active C-alphas/Backbone Delete Model

1 _AF3_complex_model.cif □ Display □ Active C-alphas/Backbone Delete Model

Close

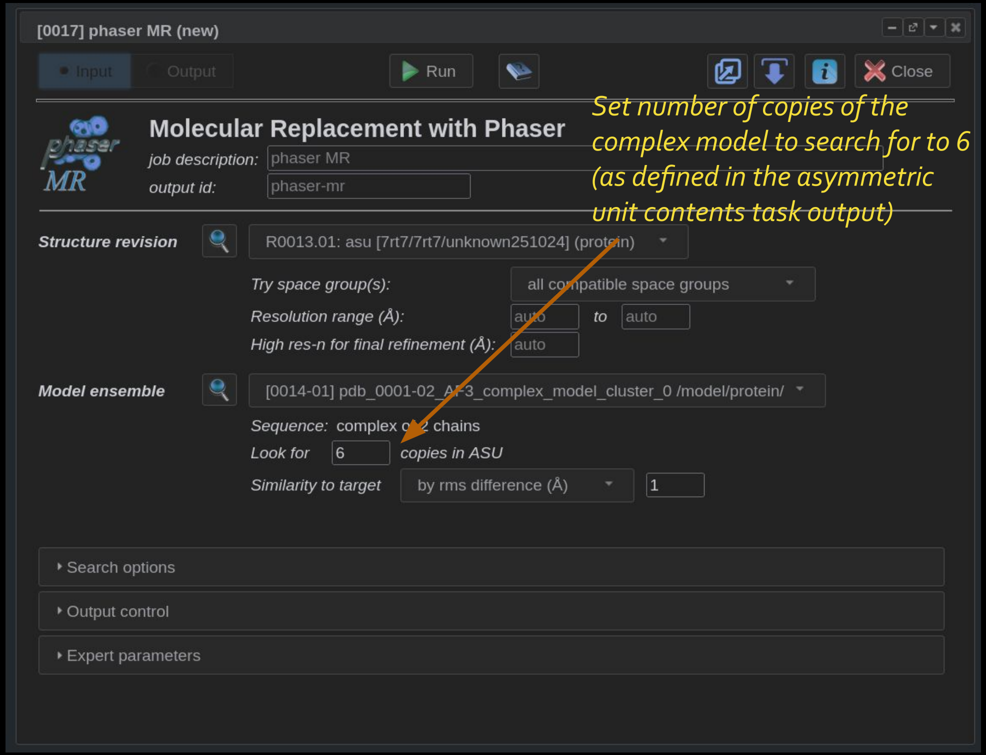
*The Coot Display Manager*



*Alphafold3 Complex as displayed in Coot before (top) and after preparation using Slice (bottom)*

# Molecular Replacement

- Follow on with a **Phaser** task using the search model output by the **Split** task
  - Search for 6 copies of the search model as predicted by the Matthews Coefficient using the AI solvent prediction.
  - Leave everything else as default. Note that Phaser will search all possible spacegroups. In this case the possibilities are P3221 and its enantiomorph P3121, as well as P321
  - Phaser has three main steps when searching with the model - the rotation search, translation search and a rigid body refinement step to optimize the placement following the translation step. The LLG values are updated after each step as components are placed.



[0017] phaser MR (new)

**Molecular Replacement with Phaser**

*job description:* phaser MR

*output id:* phaser-mr

***Structure revision*** R0013.01: asu [7rt7/7rt7/unknown251024] (protein)

*Try space group(s):* all compatible space groups

*Resolution range (Å):* auto *to* auto

*High res-n for final refinement (Å):* auto

***Model ensemble*** [0014-01] pdb_0001-02_AF3_complex_model_cluster_0 /model/protein/

*Sequence:* complex of 2 chains

*Look for* 6 *copies in ASU*

*Similarity to target* by rms difference (Å) 1

‣ Search options

‣ Output control

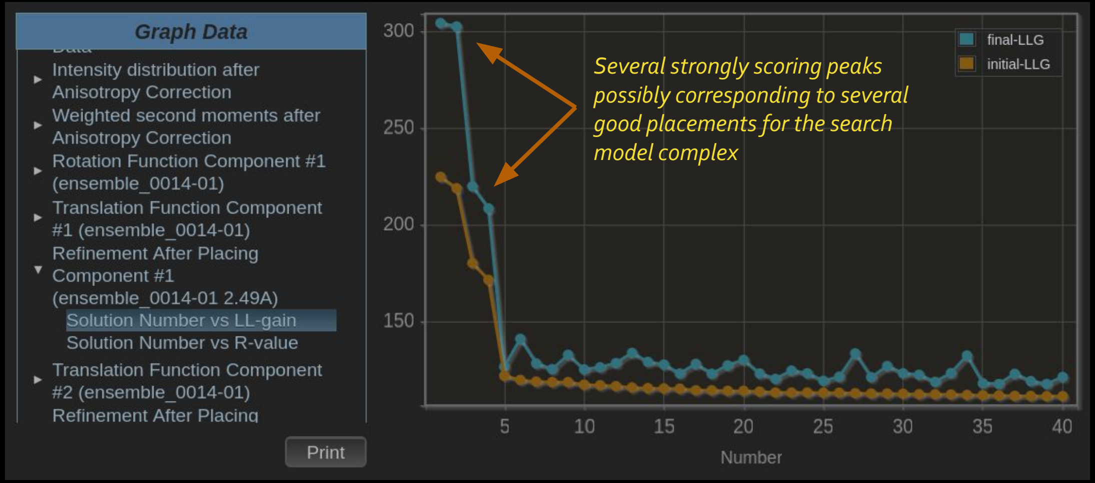
‣ Expert parameters

*Set number of copies of the complex model to search for to 6 (as defined in the asymmetric unit contents task output)*

*Phaser task interface*

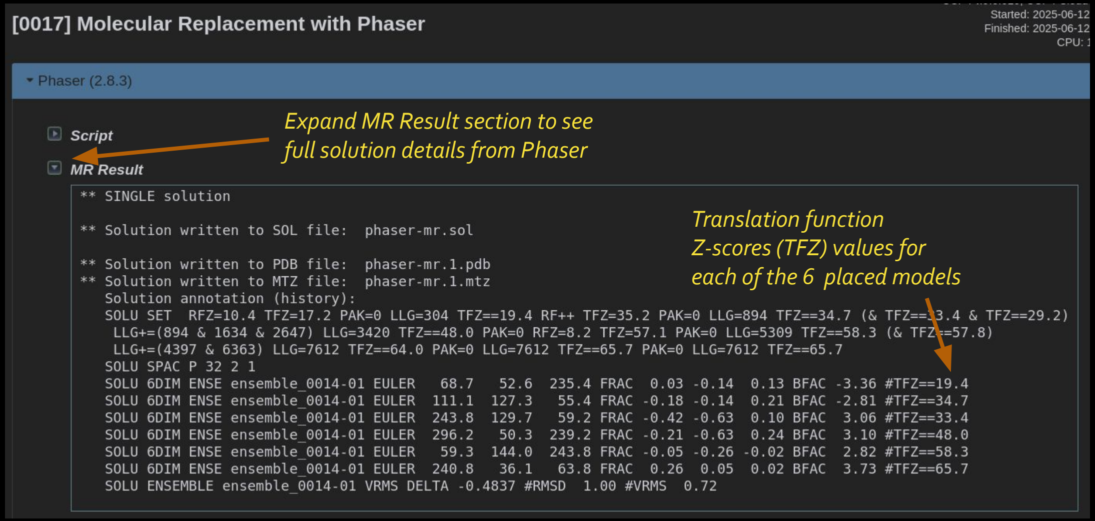
# Molecular Replacement

- Watch the refinement plots under “Graph data” as each component is searched for. For correct placement we expect the LLG to rise by at least 60 for each placed component.
- Note also that the number of peaks in each plot reduces as the components are placed. This indicates that the correct solution is more and more apparent as the job progresses. Phaser should find all 6 copies and give a final LLG of several thousand.
- It’s possible to see the full solution details by expanding the “MR Result” section in the Phaser output report (see below). Note that each model placed also has a TFZ score. We expect values above 8 for each of these. If some of the placed models have lower than this value, it may indicate that they have been incorrectly placed.
- Note too that there is a “SINGLE solution”. This is another indication of correct placement for the search models. Many possible solutions can mean that Phaser was unable to distinguish the correct placement from incorrect placement of the models.


Graph Data
Intensity distribution after Anisotropy Correction
Weighted second moments after Anisotropy Correction
Rotation Function Component #1 (ensemble_0014-01)
Translation Function Component #1 (ensemble_0014-01)
Refinement After Placing Component #1 (ensemble_0014-01 2.49A)
Solution Number vs LL-gain
Solution Number vs R-value
Translation Function Component #2 (ensemble_0014-01)
Refinement After Placing
Print
300
250
200
150
5
10
15
20
25
30
35
40
Number
final-LLG
initial-LLG

*Several strongly scoring peaks possibly corresponding to several good placements for the search model complex*

*Phaser report plot showing LLG scores after the refinement of the first placed component*



*Expand MR Result section to see full solution details from Phaser*

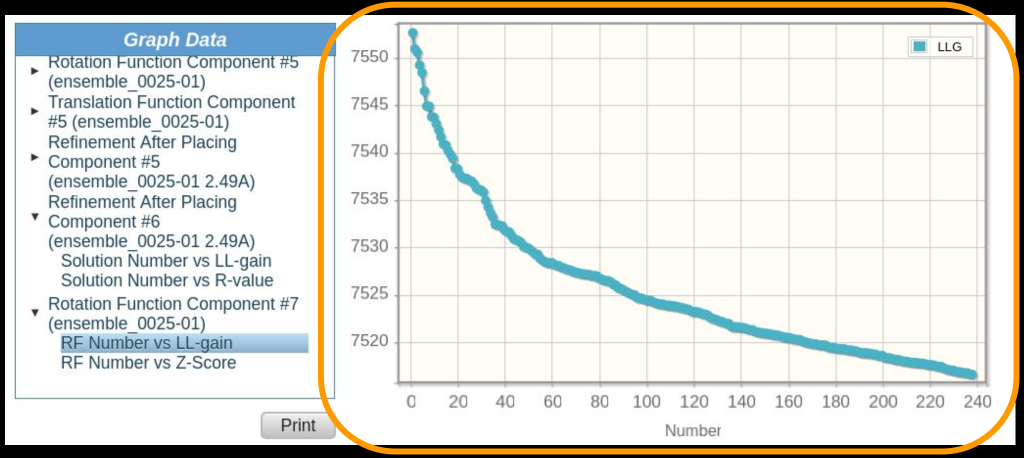
*Translation function Z-scores (TFZ) values for each of the 6 placed models*

```
** SINGLE solution

** Solution written to SOL file:  phaser-mr.sol

** Solution written to PDB file:  phaser-mr.1.pdb
** Solution written to MTZ file:  phaser-mr.1.mtz
   Solution annotation (history):
   SOLU SET  RFZ=10.4 TFZ=17.2 PAK=0 LLG=304 TFZ==19.4 RF++ TFZ=35.2 PAK=0 LLG=894 TFZ==34.7 (& TFZ==33.4 & TFZ==29.2)
    LLG+=(894 & 1634 & 2647) LLG=3420 TFZ==48.0 PAK=0 RFZ=8.2 TFZ=57.1 PAK=0 LLG=5309 TFZ==58.3 (& TFZ==57.8)
    LLG+=(4397 & 6363) LLG=7612 TFZ==64.0 PAK=0 LLG=7612 TFZ==65.7 PAK=0 LLG=7612 TFZ==65.7
   SOLU SPAC P 32 2 1
   SOLU 6DIM ENSE ensemble_0014-01 EULER   68.7   52.6  235.4 FRAC  0.03 -0.14  0.13 BFAC -3.36 #TFZ==19.4
   SOLU 6DIM ENSE ensemble_0014-01 EULER  111.1  127.3   55.4 FRAC -0.18 -0.14  0.21 BFAC -2.81 #TFZ==34.7
   SOLU 6DIM ENSE ensemble_0014-01 EULER  243.8  129.7   59.2 FRAC -0.42 -0.63  0.10 BFAC  3.06 #TFZ==33.4
   SOLU 6DIM ENSE ensemble_0014-01 EULER  296.2   50.3  239.2 FRAC -0.21 -0.63  0.24 BFAC  3.10 #TFZ==48.0
   SOLU 6DIM ENSE ensemble_0014-01 EULER   59.3  144.0  243.8 FRAC -0.05 -0.26 -0.02 BFAC  2.82 #TFZ==58.3
   SOLU 6DIM ENSE ensemble_0014-01 EULER  240.8   36.1   63.8 FRAC  0.26  0.05  0.02 BFAC  3.73 #TFZ==65.7
   SOLU ENSEMBLE ensemble_0014-01 VRMS DELTA -0.4837 #RMSD  1.00 #VRMS  0.72
```

*Final report summary from Phaser after placing 6 copies of the search model complex*

# Molecular Replacement

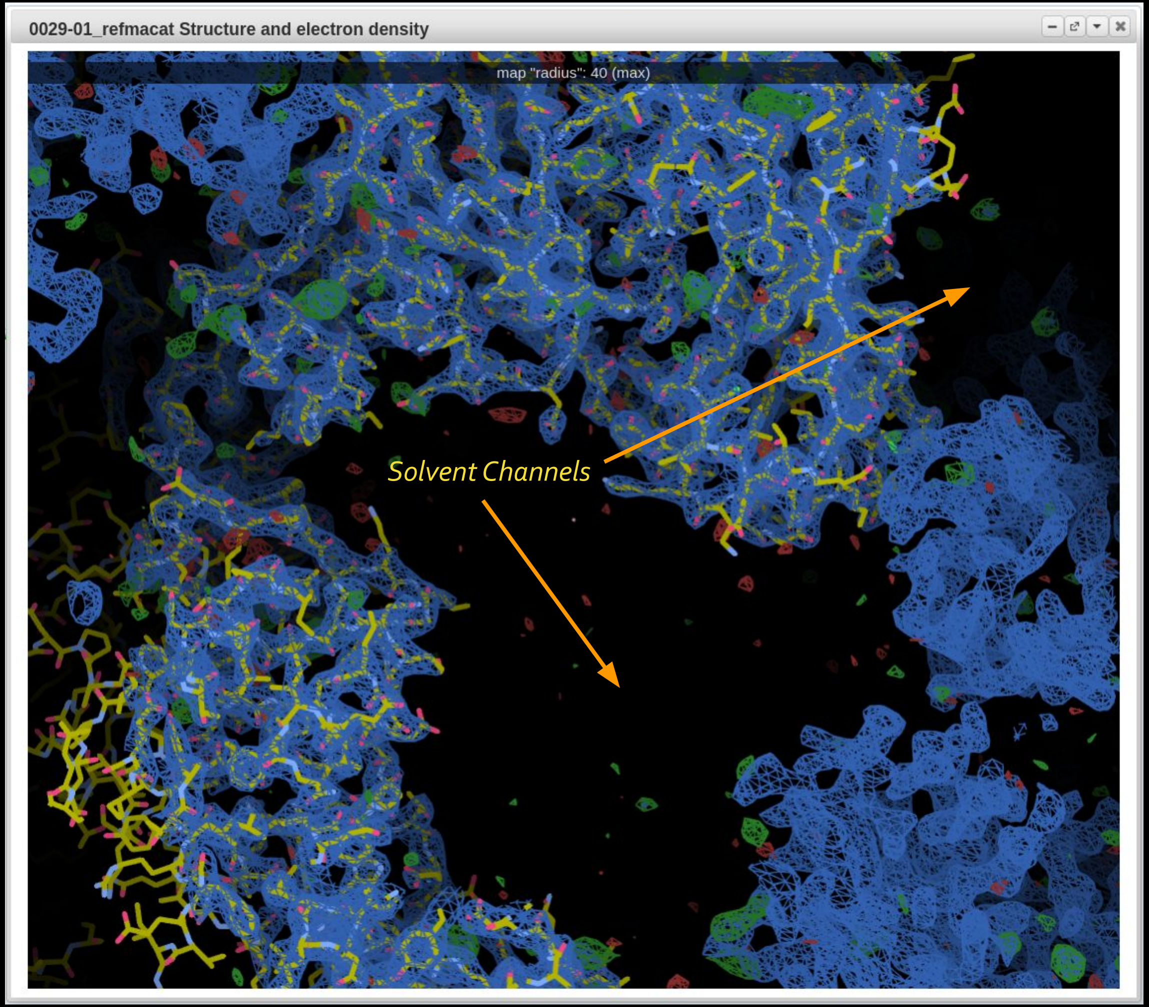
- To see what effect predicting the asymmetric unit contents incorrectly has, try running Phaser again but set the number of copies to search for to 7 (as predicted by the non-AI cell contents task). In a novel case we would have to assume the prediction is correct.
- Leave everything else as default. Note that as before, Phaser will search all possible spacegroups
- Phaser should find 6 copies relatively quickly but struggle to find a 7th. This will be obvious from the Rotation function plot for the 7th copy. It will have many similar scoring points indicating no clear rotation scores well (see figure)
- Eventually Phaser will give up trying to place a 7th copy and give the (correct) 6 copy solution


Graph Data
Rotation Function Component #5 (ensemble_0025-01)
Translation Function Component #5 (ensemble_0025-01)
Refinement After Placing Component #5 (ensemble_0025-01 2.49A)
Refinement After Placing Component #6 (ensemble_0025-01 2.49A)
Solution Number vs LL-gain
Solution Number vs R-value
Rotation Function Component #7 (ensemble_0025-01)
RF Number vs LL-gain
RF Number vs Z-Score
Print
LLG
7550
7545
7540
7535
7530
7525
7520
0
20
40
60
80
100
120
140
160
180
200
220
240
Number

*The rotation function LLG for the 7th copy search in Phaser. Many points with similar scores here indicate that a rotation solution for this copy cannot be found*

# Initial refinement

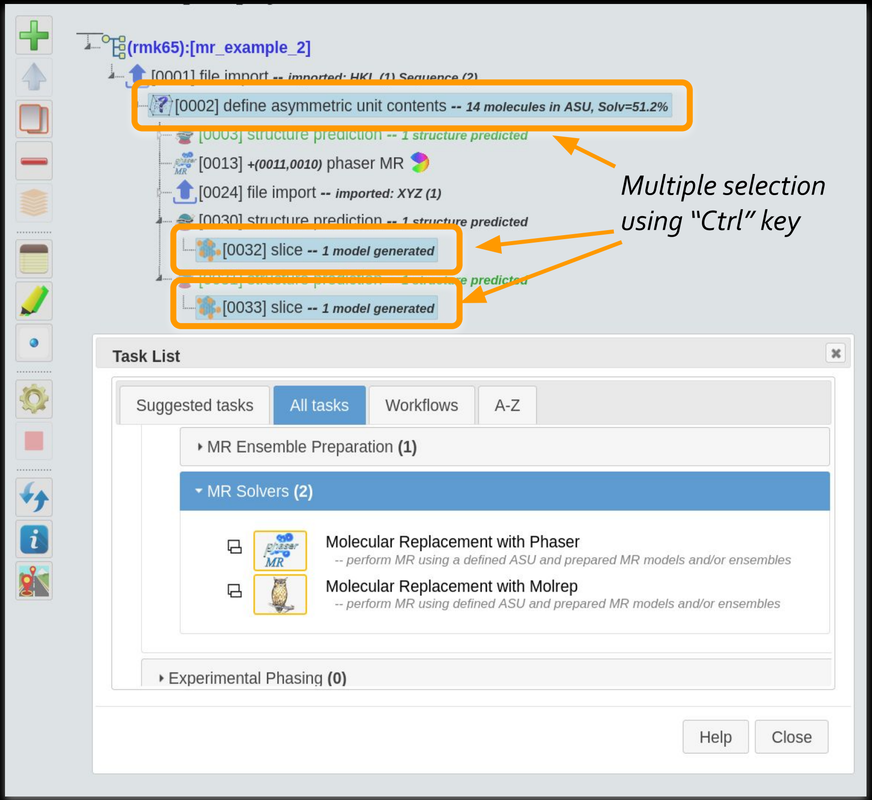
- Follow on from **Phaser** with a 100 cycles of jelly-body refinement in **Refmacat**
  - Note that the refinement scores drop sharply until they plateau at about and R/Rfree of 0.26/0.31
  - View the resulting map and model in **Uglymol**. Use the "[" and "]" keys to adjust the map radius. Note that there are large areas with no density. These are the solvent channels between the molecules. Seeing clear solvent channels in a solution is another strong indicator of successful MR. In this case the crystal is more than 58% solvent
  - At this point the remaining model building should be done manually in **Coot** or **Moorhen**



*The refined model and the resulting electron density map shown in Uglymol*

# Molecular replacement with single chain models

- To appreciate the advantage of the complex search model, run the "Structure prediction" task to generate predictions for each of the two sequences separately (Chain A and chain B). We will use these as seperate input search models for Phaser
- After each prediction, run "Slice" to remove the low confidence residues
- To give both output search models from "Slice" to Phaser we need to create a new branch in the CCP4Cloud project starting from the "Asymmetric unit contents" task.
  - Using the "Ctrl" key, select the two slice tasks and the asymmetric unit task (see figure)
  - With all 3 selected add a new **Phaser** job to the asymmetric unit task
  - Provide both input search models (chain A and B) and click run
  - The job should find most of the molecules but is likely to take several hours



(rmk65):[mr_example_2]

[0001] file import -- *imported: HKL (1) Sequence (2)*

[0002] define asymmetric unit contents -- ***14 molecules in ASU, Solv=51.2%***

[0003] structure prediction -- *1 structure predicted*

[0013] *+(0011,0010)* phaser MR

[0024] file import -- *imported: XYZ (1)*

[0030] structure prediction -- *1 structure predicted*

[0032] slice -- ***1 model generated***

[0031] structure prediction -- *1 structure predicted*

[0033] slice -- ***1 model generated***

*Multiple selection using "Ctrl" key*

**Task List**

Suggested tasks | All tasks | Workflows | A-Z

MR Ensemble Preparation **(1)**

MR Solvers **(2)**

Molecular Replacement with Phaser
-- *perform MR using a defined ASU and prepared MR models and/or ensembles*

Molecular Replacement with Molrep
-- *perform MR using defined ASU and prepared MR models and/or ensembles*

Experimental Phasing **(0)**