CCP4 DLS Workshop 2024

VERIFICATION:
can you prove yourself wrong?

Iracema Caballero

icmcri@ibmb.csic.es

Arcimboldo team

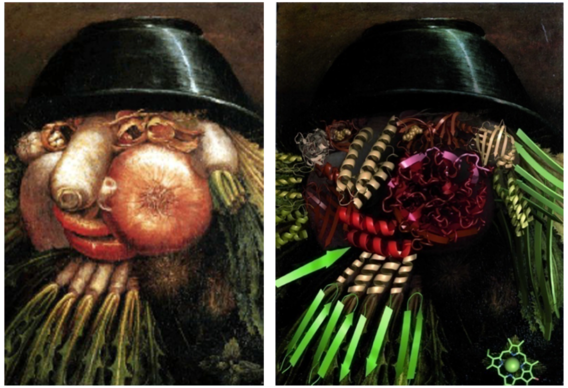
Prof. Dr. Isabel Usón

chango.ibmb.csic.es

ibmb
Institut de Biologia Molecular de Barcelona
Molecular Biology Institute of Barcelona CSIC

CCP4

ARCIMBOLDO: *ab initio* phasing approach



ARCIMBOLDO combines:

- Fragment location with *Phaser* [1]
- Density modification and autotracing with *SHELXE* [2]

Main idea:

- Compose minimal hypotheses from fragments.
- Extend partial structures into full structures with density modification and model building.
- Correct solutions are identified when the minimal starting hypothesis reveals previously unknown structural features, while incorrect ones do not progress into meaningful solutions.

1. A.J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, *et al.* “Phaser crystallographic software” (2007) *J.Appl.Cryst, 40, 658–674*
2. I. Usón and G.M. Sheldrick. “Modes and model building in SHELXE”. (2024) *Acta Cryst. D80*, 4–15

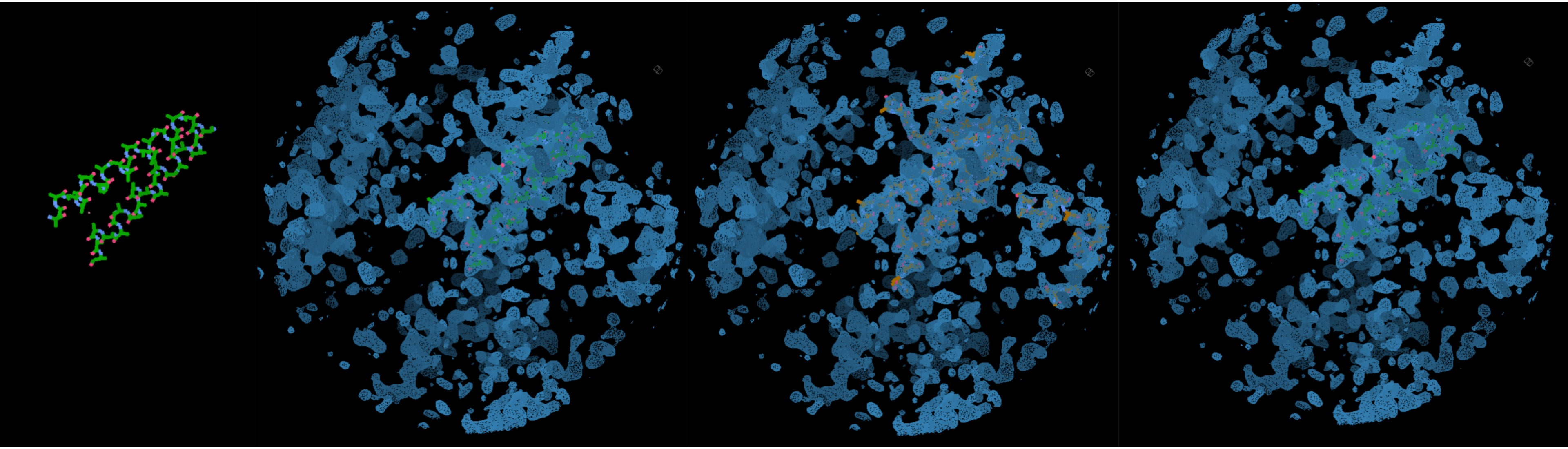
Effect of density modification from a small, accurate fraction of ASU

Phases calculated from a partial solution

Density modification

Autotracing of new polyalanine residues

Iterate: perform several cycles



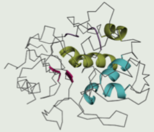
ARCIMBOLDO programs & modes

LITE

Single models such as ideal polyalanine **α**-helices

- `coiled_coil mode`

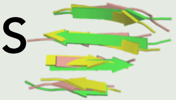
SHREDDER



Fragment library from homologs or predicted models

- `coiled_coil mode`
- `predicted_model mode`

BORGES



Fragment library representing a fold

LITE

Sequential search for single models

Signal to noise will be better upon each placement

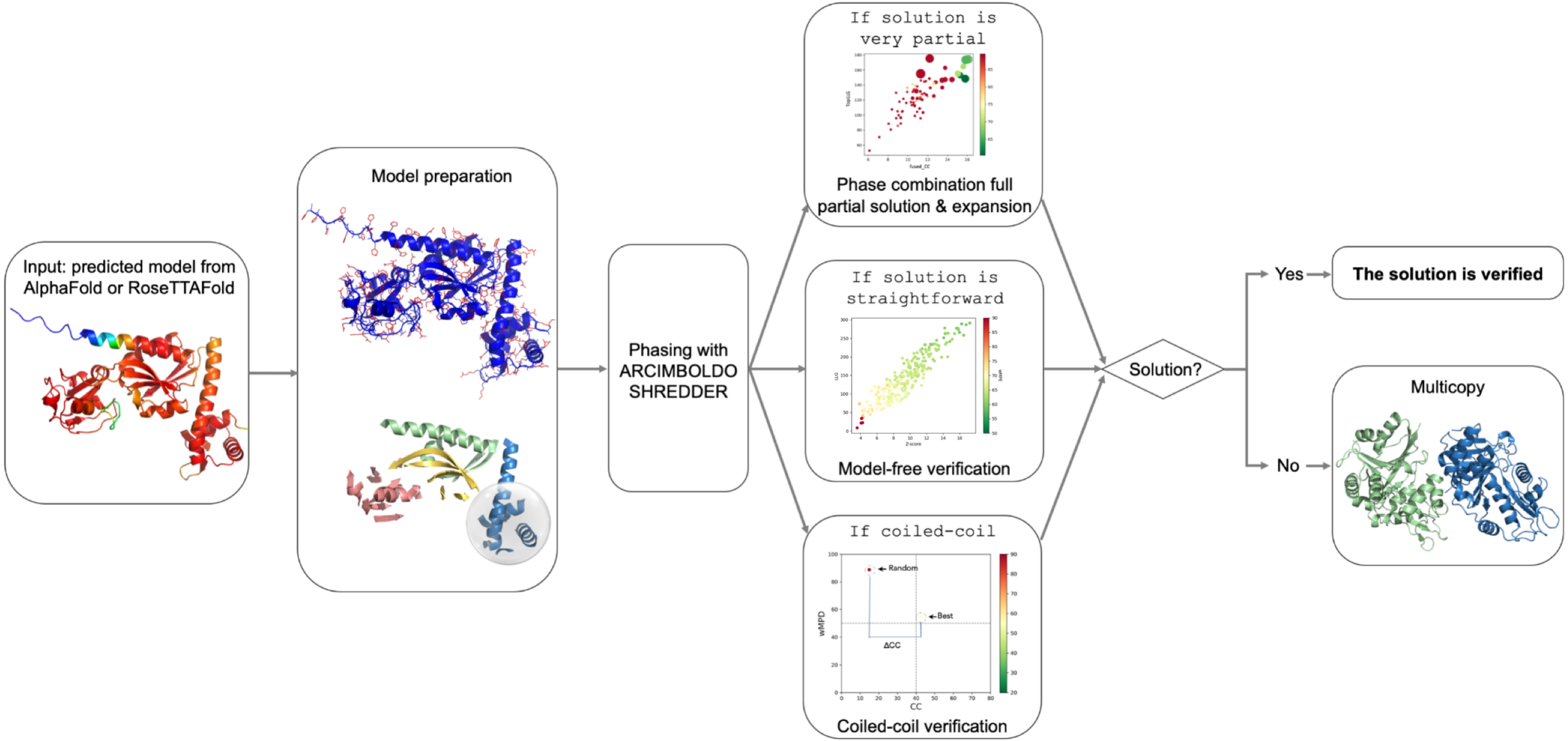
SHREDDER & BORGES

Simultaneous search for multiple models of same size

Results are considered jointly

predicted_model mode in
ARCIMBOLDO_SREDDER

predicted_model mode in ARCIMBOLDO_SHREDDER



Input: predicted model from AlphaFold or RoseTTAFold

Model preparation

Phasing with ARCIMBOLDO SHREDDER

`If solution is very partial`

Phase combination full partial solution & expansion

`If solution is straightforward`

Model-free verification

`If coiled-coil`

Coiled-coil verification

Solution?

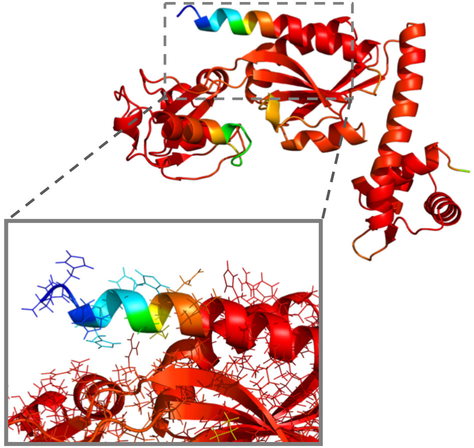
Yes → **The solution is verified**

No → Multicopy

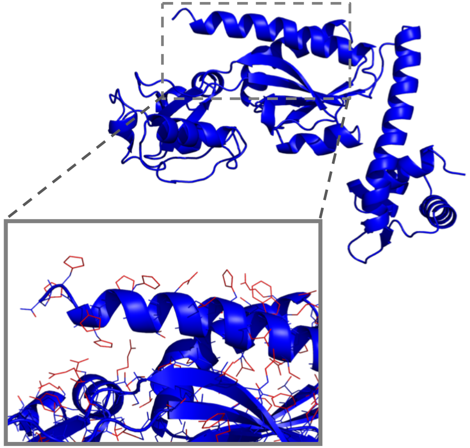
Medina, A., Jiménez, E., Caballero, I., *et al.* “Verification: model-free phasing with enhanced predicted models in ARCIMBOLDO_SHREDDER” (2022). Acta Cryst. D 78(11), 1283–1293

Model preparation of predicted models

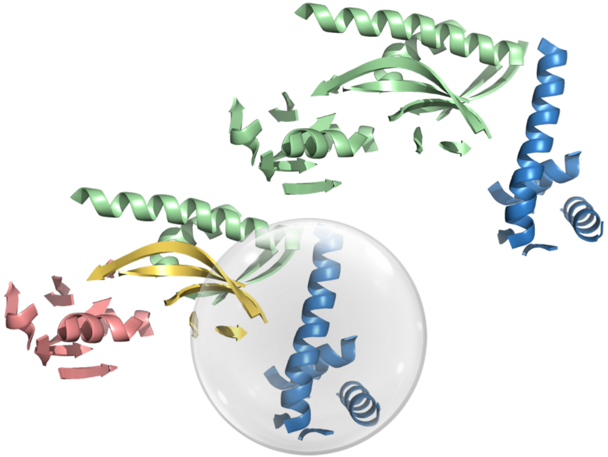
Predicted model → Pre-processing → Annotation & Decomposition



ARCIMBOLDO_SHREDDER performs an **automatic** preparation of the input template



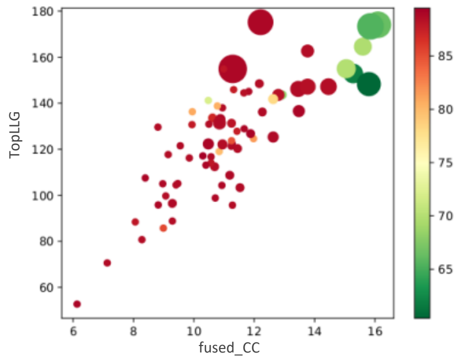
- B-factors are set to a common value of 25Å for the main chain and 50Å for the side chains.
- Side chains from predicted models are preserved.
- Hydrogen atoms are removed.



- Unstructured areas predicted with low confidence are removed.
- Domains are differentiated through hierarchical decomposition.
- The model is shred into equal-sized spheres based on eLLG.

Verification

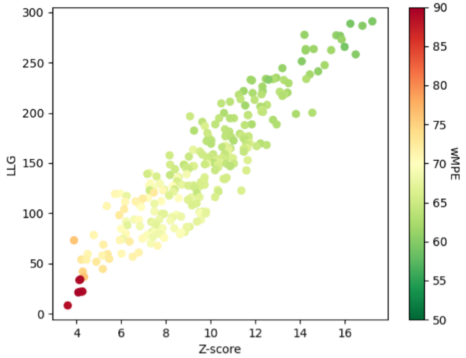
If solution is very partial



Difficult phasing is self-validating as wrong solutions impede expansion

Phase combination of partial solutions & expansion

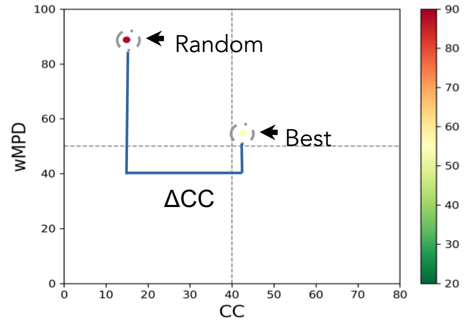
If solution is straightforward



Straightforward returning not much beyond the original model

Model-free verification

If coiled coil

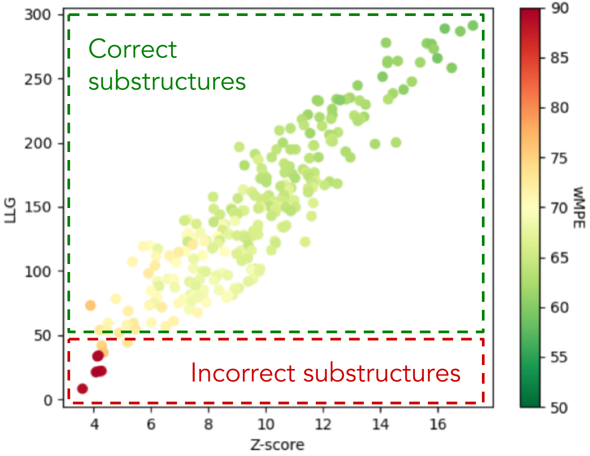


Deceptive phasing as in coiled-coil structures

Coiled-coil verification

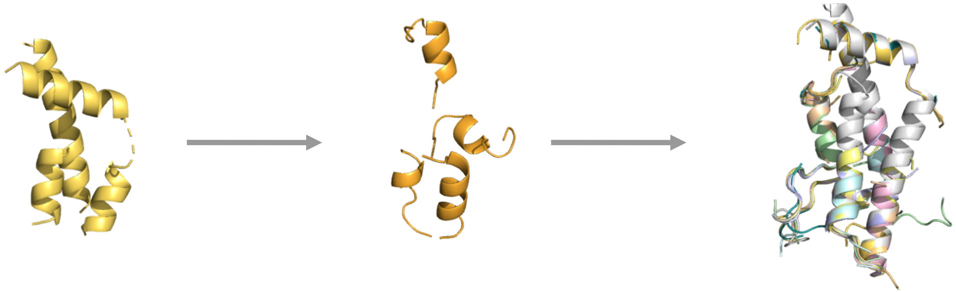
Model-free verification: eliminating the model bias

If solution is straightforward



In predicted model fragments are now more accurate, but this also introduces model bias

Model-free verification



Partial solutions are expanded by density modification with SHELXE

One cycle of auto-tracing with SHELXE omits the starting model

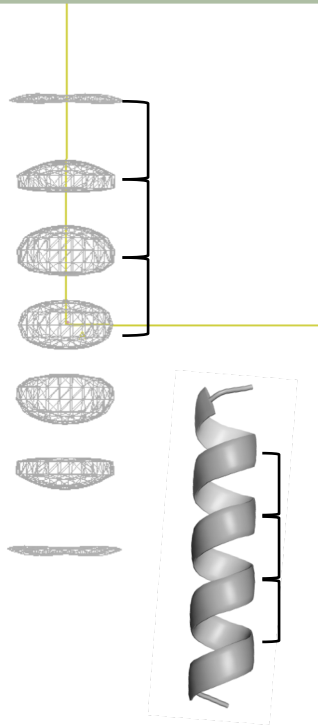
All consistent traces are combined in reciprocal space with ALIXE, this new map will be iteratively modified and traced

Also used to verify external MR solution

coiled_coil mode in
ARCIMBOLDO_LITE &
ARCIMBOLDO_SREDDER

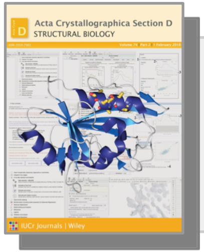
Coiled coils difficulties for phasing

- Coiled coil structures typically present intrinsic difficulties for phasing:
    - Tendency to aggregate into fibers complicating the formation of well diffracting crystals
    - Large dimensions of their asymmetric units
    - Anisotropic diffraction
    - Highly modulated diffraction patterns
- Phasing of coiled-coil with fragments has been implemented in the AMPLE[1] and CCsolve[2] pipelines.



1. Thomas, J. M. H., Keegan, R. M., Bibby, J., Winn, M. D., Mayans, O. & Rigden, D. J. (2015). IUCrJ, 2, 198-206.
2. Rämisch, S., Lizatović, R. & André, I. (2015). Acta Cryst. D71, 606-614.

coiled_coil mode



ARCIMBOLDO on coiled coils
Caballero, I., Sammito, M., Millán, C., *et al.*
Acta Cryst. **2018**; D74, 194–204

Up to 3Å

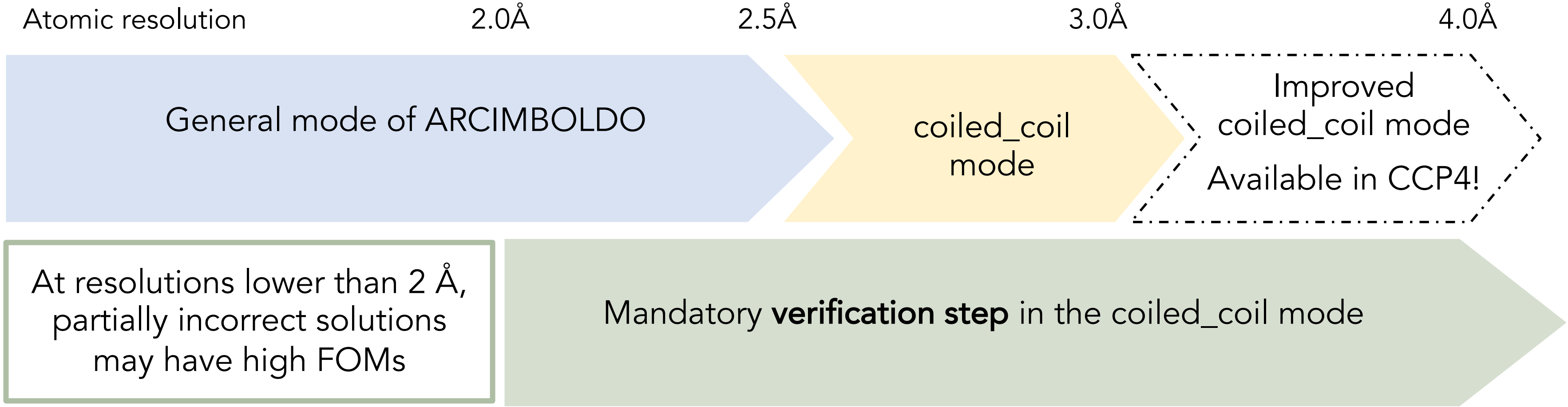
ARCIMBOLDO at low resolution: Verification for coiled coils and globular proteins
Caballero I, Castellví A, Triviño J, *et al.*
Protein Sci **2024**; 33(9):e5136.

Up to 4Å

PROBLEMS	SOLUTIONS
Difficulties to differentiate genuine intermolecular tNCS from Patterson artefacts.	Disable the placement of pairs of tNCS related helices.
Overlapping solutions with high FOMs.	Activate the *Phaser* packing filter during the translation search.
The geometry of helical trace can degrade.	SHELXE with helical sliding that anchors ahead of the trace, improving the phases.
Placement of helices in the correct position but in reversed direction.	Generate and test reversed helices in the same positions (only in LITE).
Wrong solutions can have high figures of merit.	Final verification step.

Resolution limits

Resolution in ARCIMBOLDO is **auto-limited** in order to prevent false positives



Atomic resolution 2.0Å 2.5Å 3.0Å 4.0Å

General mode of ARCIMBOLDO

coiled_coil mode

Improved coiled_coil mode

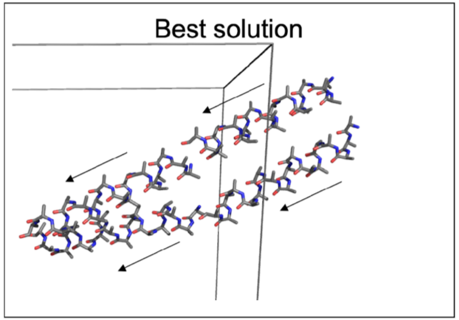
Available in CCP4!

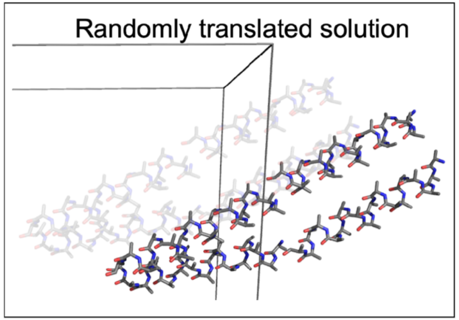
At resolutions lower than 2 Å, partially incorrect solutions may have high FOMs

Mandatory **verification step** in the coiled_coil mode

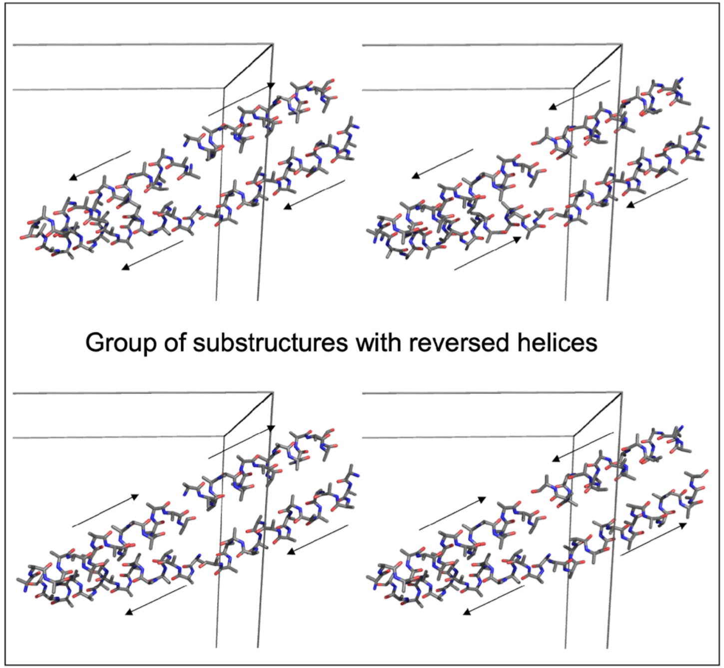
Verification step - prove ourselves wrong

To verify the most promising solution, its original substructure will be perturbed by a random translation and by reversing the direction of the helices


Best solution


Randomly translated solution

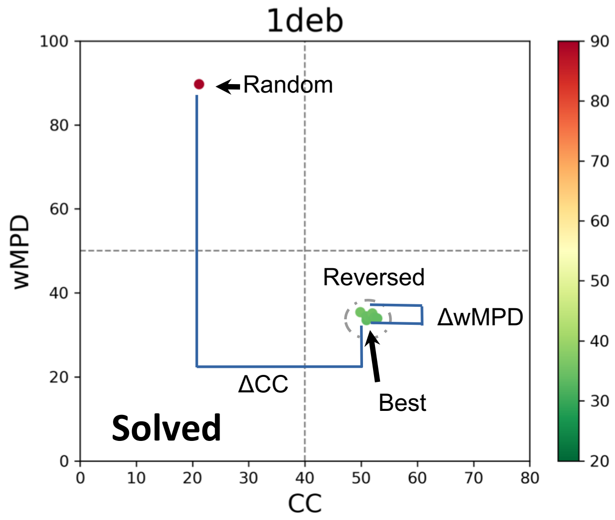
In LITE & SHREDDER →

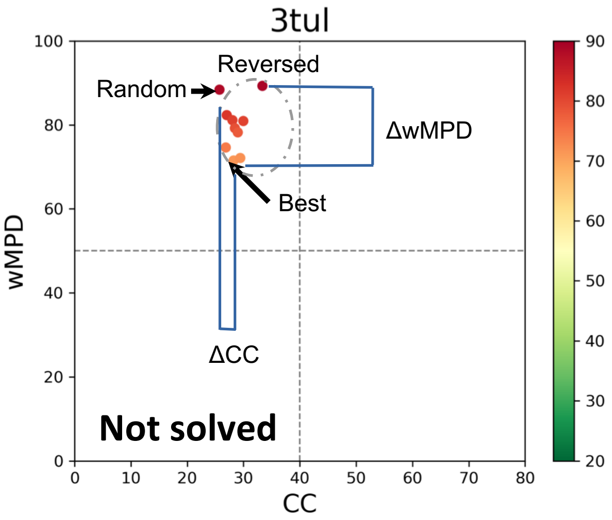

Group of substructures with reversed helices

← Only in LITE

Verification in ARCIMBOLDO_LITE

The extensions are compared for evidence of discrimination between the randomly translated solution or groups of consistent solutions



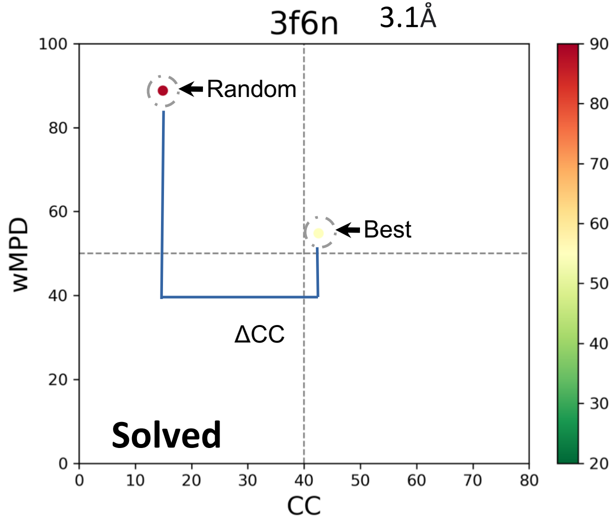


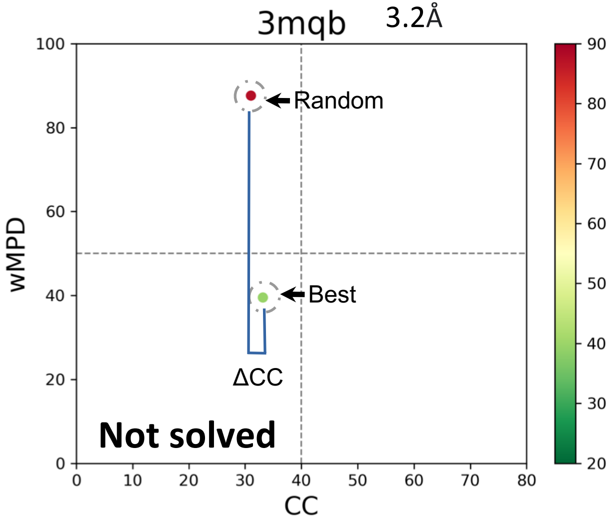
The best solution is **discriminated** from the random solution or the perturbations develop into a group of structurally **equivalent solutions**

The best solution is **not discriminated** from the random solution or the perturbations develops into structurally **different solutions**

Verification in ARCIMBOLDO_SHREDDER

The extensions are compared for evidence of discrimination between the best solution and the randomly translated solution



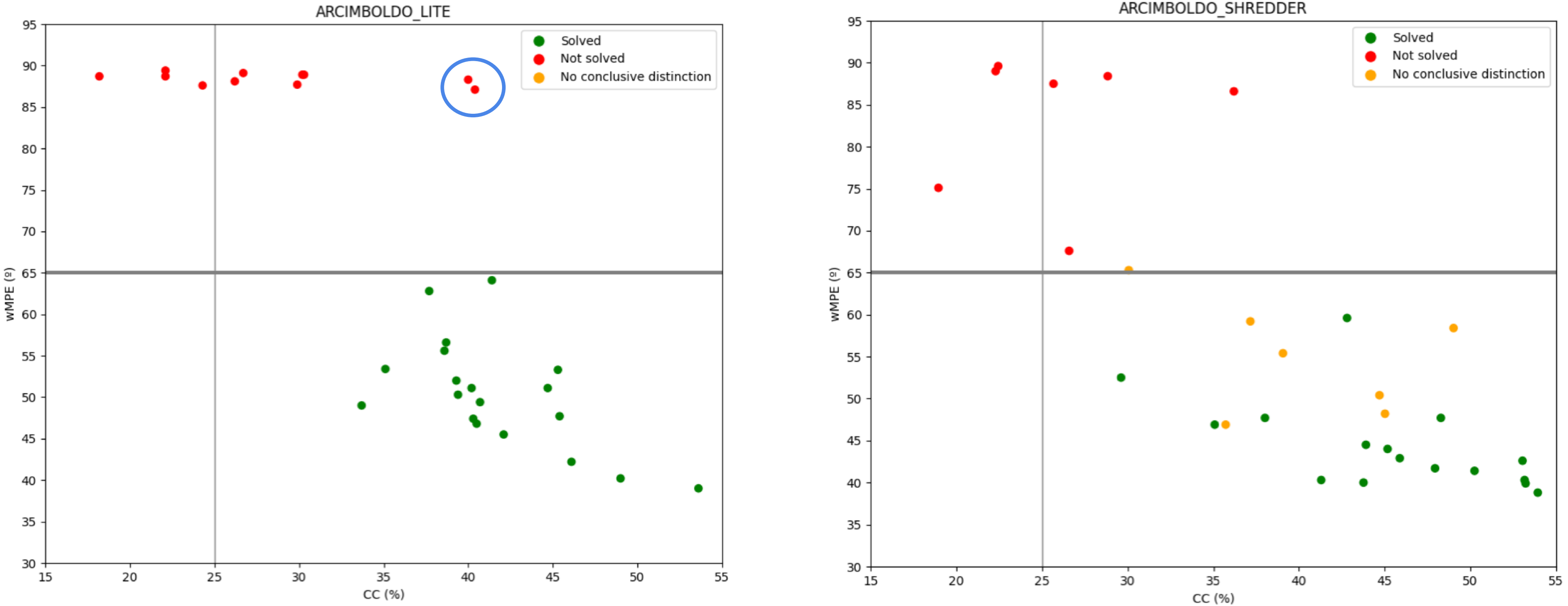


The best solution is discriminated from the random solution

The best solution is not discriminated from the random solution

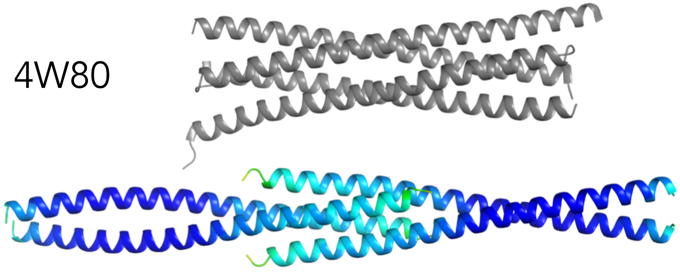
Verification as a key tool to eliminate false positives

Verification ruled out 7 false positives in the case of ARCIMBOLDO_LITE and 4 false positives in the case of ARCIMBOLDO_SHREDDER

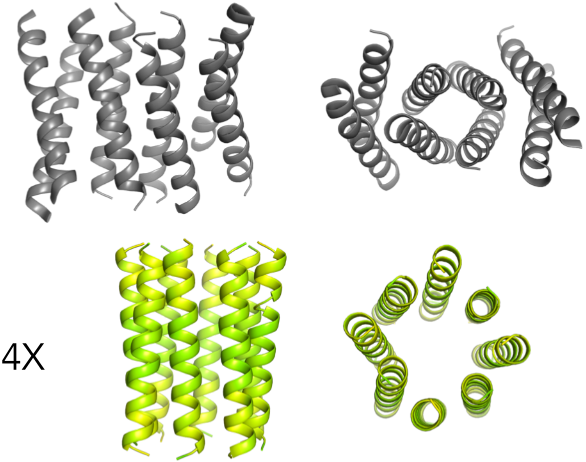


Choices for phasing coiled coils

4W80



6C4X



ARCIMBOLDO_LITE in coiled_coil mode

Choice for searching single helices

It performs a sequential search for helices of a given length with specific features to avoid coiled-coil pitfalls.

ARCIMBOLDO_SHREDDER
in coiled_coil mode & predicted_model mode

Choice for searching multimeric predictions

It extracts compact fragments from the starting model and gives them degrees of freedom to improve the models.

Low resolution verification for globular proteins

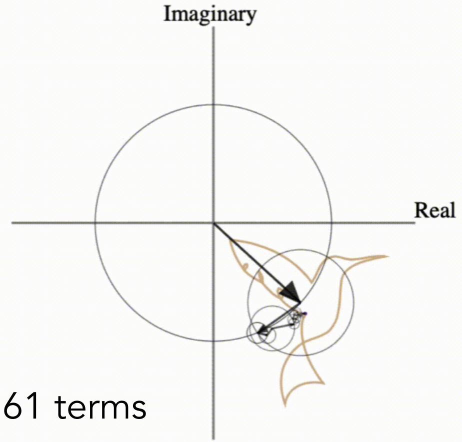
Structure approximation by a Fourier series

- Structure determination by crystallographic diffraction can be expressed through the Fourier series.
- This animation uses a line drawing of a hummingbird to define a periodic path, where the coordinates in the complex plane are a function of time. The path can be approximated by a Fourier series equation:

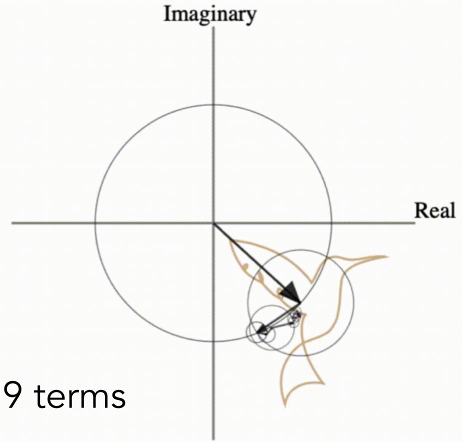
$$F(t) = \sum coef_n \cdot \exp\left(i \cdot 2\pi \cdot n \cdot t\right)$$

- The Fourier series approximation is only exact with infinite terms.
- With fewer terms (low resolution), we not only lose detail but may also introduce artifacts that do not exist in the original structure.

http://chango.ibmb.csic.es/colibri



61 terms



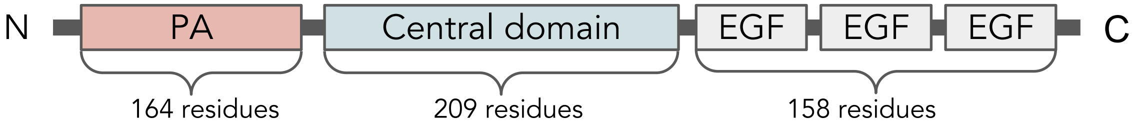
9 terms

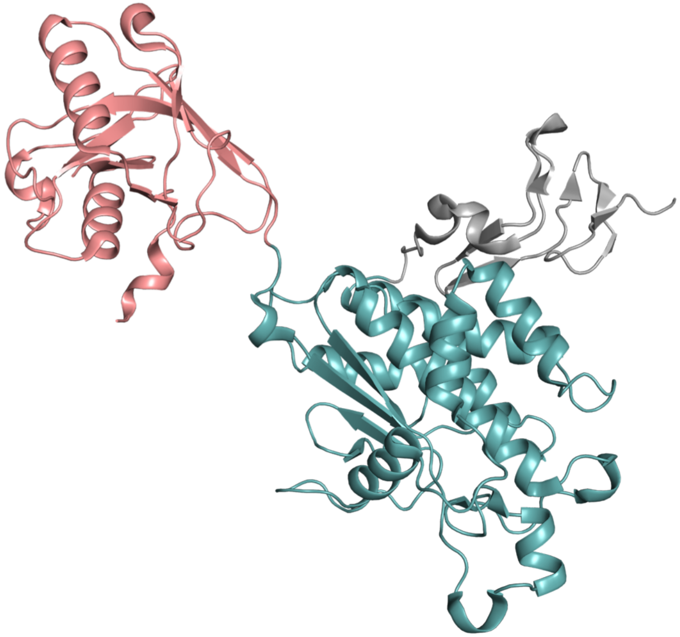
LLG-based fragment verification from density modified maps

1. <u>Constrained map interpretation</u>: Use SHELXE density-modified maps from partial, reliable solutions to build hypotheses consistent with prior knowledge.

1. <u>Comparative scoring of alternative hypotheses</u>: Evaluate fragments from remote homologs or secondary structure predictions:
   - Fragments were rigid-body refined, allowing internal degrees of freedom using *Gimble* in *Phaser*.
   - Fragments were scored by their $\mathrm{LLG}_{\mathrm{contribution}}$ to the overall value, estimated from LLG change upon fragment omission.
   - Relative differences in the $\mathrm{LLG}_{\mathrm{contribution}}$ of each fragment will guide the choice between different possibilities: size, orientation, secondary structure or side-chain inclusion.

Multidomain protein VSR1 at 3.5Å

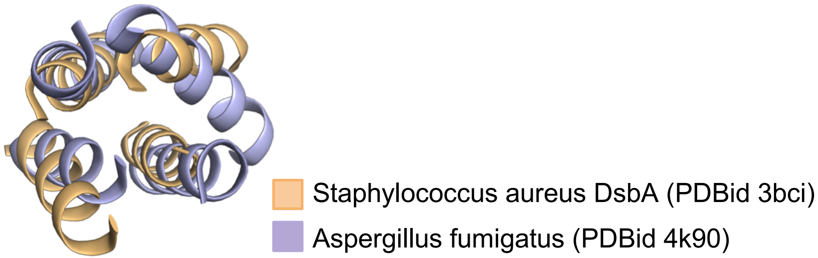




The complex state of the PA domain 4txj rendered a partial solution of VSR1 by MR with *Phaser*.

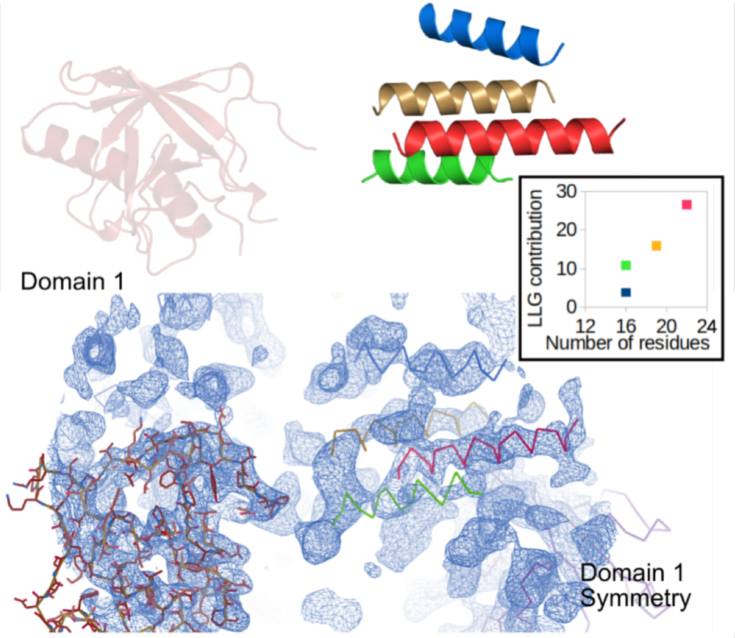
No models were available for the Central domain, but has two remote homologs with 15% identity.

Both structures share a roughly parallel four-helix bundle orientation.

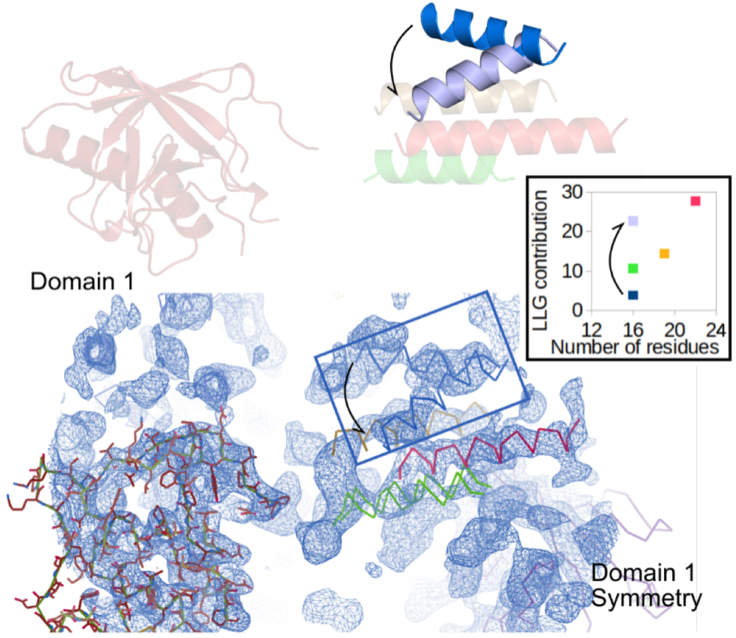


Multidomain protein VSR1 at 3.5Å

Density-modified map calculated using the coordinates of the PA domain

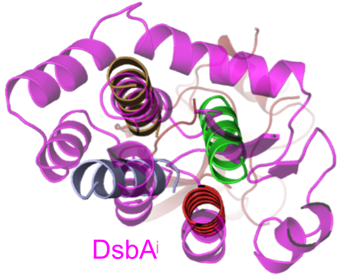


The four-helix bundle was placed in the Central domain region and scored on $\text{LLG}_{\text{contribution}}$

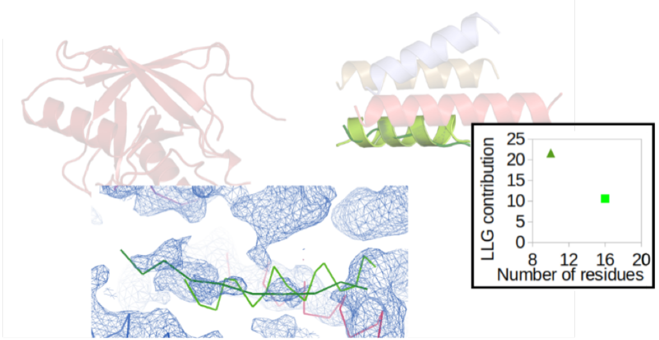


Optimization of the blue helix orientation increases its $\text{LLG}_{\text{contribution}}$

Multidomain protein VSR1 at 3.5Å

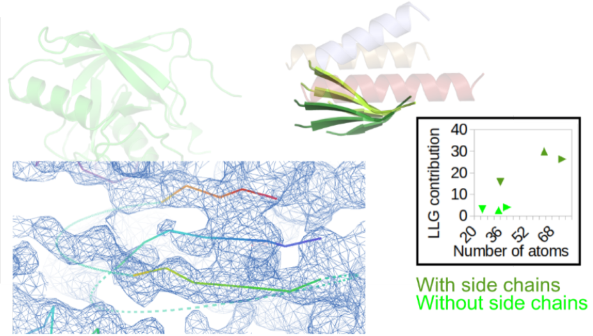


Hypothesis of four helices of the VSR1 superimposed to DsbA (15% identity), the green helix in the central domain corresponds to a strand of DsbA.

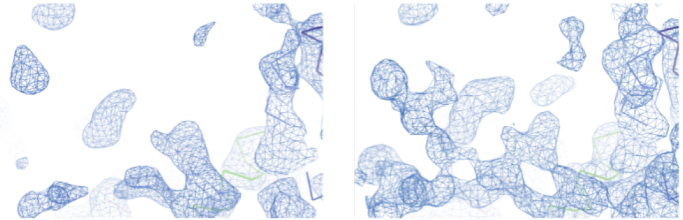


The green helix was replaced by a strand and its LLG$_{contribution}$ increases.

We included the central domain fragments in the map calculation



The green strand was expanded to a sheet of 3 strands, and adding side chains improved their LLG$_{contribution}$.

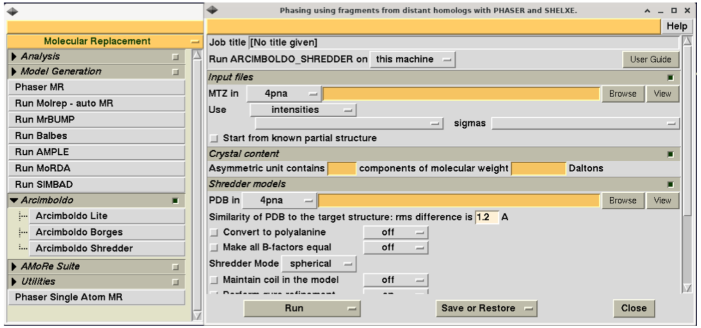


Density-modified map at the EGF region generated solely with the PA domain vs. using also fragments from the Central domain. **Emergence of new features in unmodeled map regions confirmed partial model correctness.**

Tutorial

ARCIMBOLDO Interfaces

CCP4i



Job title [No title given]

Run ARCIMBOLDO_SHREDDER on this machine

Input files

MTZ in 4pna

Use intensities

sigmas

☐ Start from known partial structure

Crystal content

Asymmetric unit contains components of molecular weight Daltons

Shredder models

PDB in 4pna

Similarity of PDB to the target structure: rms difference is 1.2 A

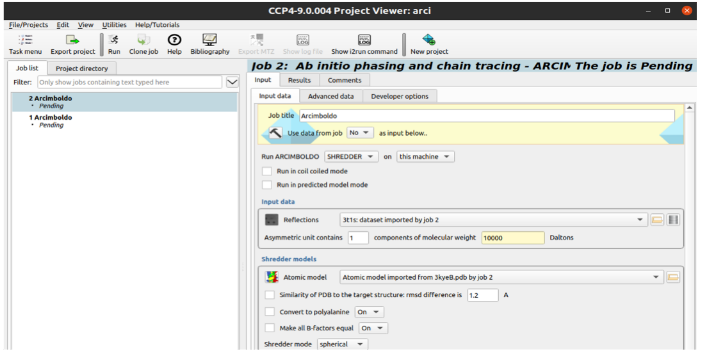
☐ Convert to polyalanine off

☐ Make all B-factors equal off

Shredder Mode spherical

☐ Maintain coil in the model off

CCP4i2



Filter: Only show jobs containing text typed here

2 Arcimboldo
- *Pending*

1 Arcimboldo
- *Pending*

Job 2: Ab initio phasing and chain tracing - ARCIM The job is Pending

Job title Arcimboldo

Use data from job No as input below..

Run ARCIMBOLDO SHREDDER on this machine

☐ Run in coil coiled mode

☐ Run in predicted model mode

Input data

Reflections 3t1s: dataset imported by job 2

Asymmetric unit contains 1 components of molecular weight 10000 Daltons

Shredder models

Atomic model Atomic model imported from 3kyeB.pdb by job 2

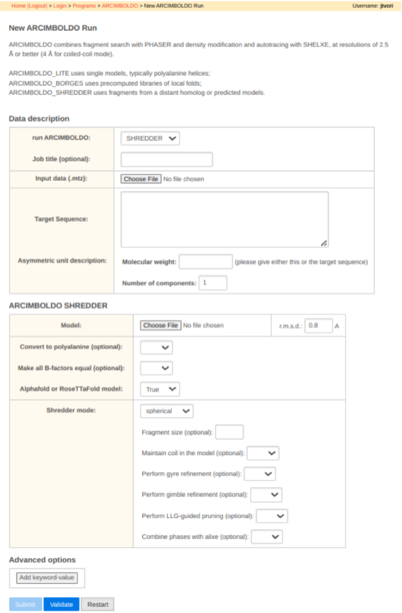
☐ Similarity of PDB to the target structure: rmsd difference is 1.2 A

☐ Convert to polyalanine On

☐ Make all B-factors equal On

Shredder mode spherical

CCP4 online (Ville Uski)



New ARCIMBOLDO Run

ARCIMBOLDO combines fragment search with PHASER and density modification and autotracing with SHELXE, at resolutions of 2.5 Å or better (4 Å for coiled-coil mode).

ARCIMBOLDO_LITE uses single models, typically polyalanine helices;
ARCIMBOLDO_BORGES uses precomputed libraries of local folds;
ARCIMBOLDO_SHREDDER uses fragments from a distant homolog or predicted models.

Data description

run ARCIMBOLDO: SHREDDER

Job title (optional):

Input data (.mtz): No file chosen

Target Sequence:

Asymmetric unit description: Molecular weight: (please give either this or the target sequence)

Number of components: 1

ARCIMBOLDO SHREDDER

Model: No file chosen r.m.s.d.: 0.8 A

Convert to polyalanine (optional):

Make all B-factors equal (optional):

Alphafold or RoseTTaFold model: True

Shredder mode: spherical

Fragment size (optional):

Maintain coil in the model (optional):

Perform gyre refinement (optional):

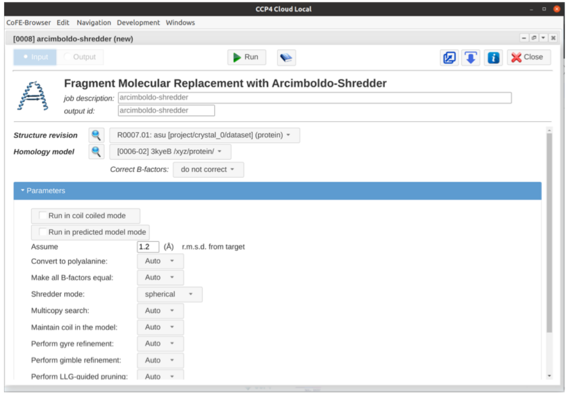
Perform gimble refinement (optional):

Perform LLG-guided pruning (optional):

Combine phases with alixe (optional):

Advanced options

CCP4 Cloud



[0008] arcimboldo-shredder (new)

Fragment Molecular Replacement with Arcimboldo-Shredder

job description: arcimboldo-shredder

output id: arcimboldo-shredder

Structure revision R0007.01: asu [project/crystal_0/dataset] (protein)

Homology model [0006-02] 3kyeB /xyz/protein/

Correct B-factors: do not correct

☐ Run in coil coiled mode

☐ Run in predicted model mode

Assume 1.2 (Å) r.m.s.d. from target

Convert to polyalanine: Auto

Make all B-factors equal: Auto

Shredder mode: spherical

Multicopy search: Auto

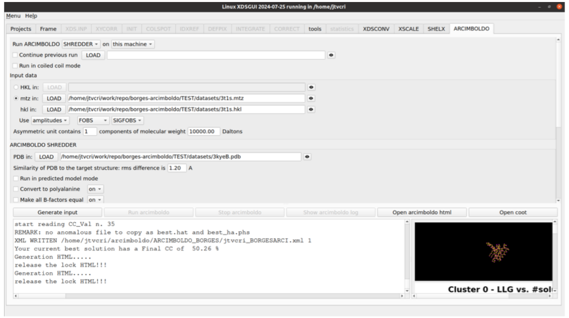
Maintain coil in the model: Auto

Perform gyre refinement: Auto

Perform gimble refinement: Auto

Perform LLG-guided pruning: Auto

XDSGUI (Kay Diederichs)



Run ARCIMBOLDO SHREDDER on this machine

☐ Continue previous run

☐ Run in coiled coil mode

Input data

○ HKL in:

● mtz in: /home/jtvcri/work/repo/borges-arcimboldo/TEST/datasets/3t1s.mtz

hkl in: /home/jtvcri/work/repo/borges-arcimboldo/TEST/datasets/3t1s.hkl

Use amplitudes FOBS SIGFOBS

Asymmetric unit contains 1 components of molecular weight 10000.00 Daltons

ARCIMBOLDO SHREDDER

PDB in: /home/jtvcri/work/repo/borges-arcimboldo/TEST/datasets/3kyeB.pdb

Similarity of PDB to the target structure: rms difference is 1.20 A

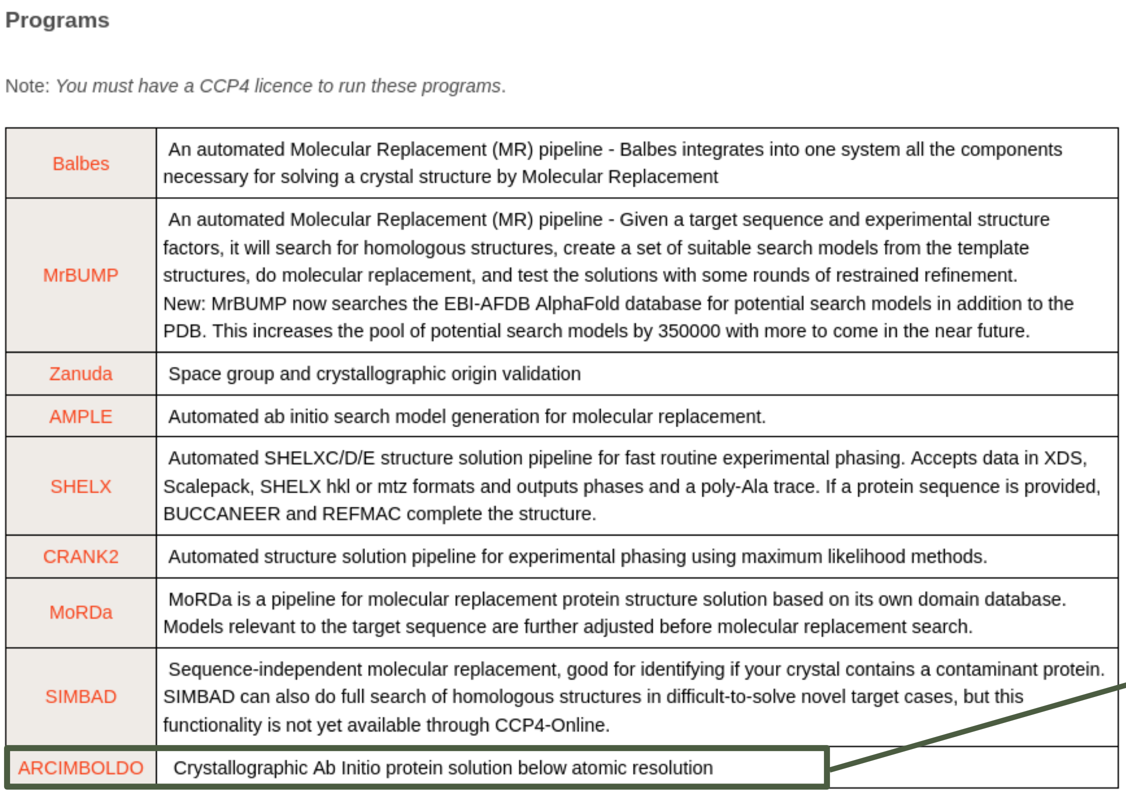
☐ Run in predicted model mode

☐ Convert to polyalanine on

☐ Make all B-factors equal on

```
start reading CC_Val n. 35
REMARK: no anomalous file to copy as best.hat and best_ha.phs
XML WRITTEN /home/jtvcri/arcimboldo/ARCIMBOLDO_BORGES/jtvcri_BORGESARCI.xml 1
Your current best solution has a Final CC of  50.26 %
Generation HTML.....
release the lock HTML!!!
Generation HTML.....
release the lock HTML!!!
```

Setting the job in CCP4 online



Programs

Note: *You must have a CCP4 licence to run these programs.*

Balbes	An automated Molecular Replacement (MR) pipeline - Balbes integrates into one system all the components necessary for solving a crystal structure by Molecular Replacement
MrBUMP	An automated Molecular Replacement (MR) pipeline - Given a target sequence and experimental structure factors, it will search for homologous structures, create a set of suitable search models from the template structures, do molecular replacement, and test the solutions with some rounds of restrained refinement. New: MrBUMP now searches the EBI-AFDB AlphaFold database for potential search models in addition to the PDB. This increases the pool of potential search models by 350000 with more to come in the near future.
Zanuda	Space group and crystallographic origin validation
AMPLE	Automated ab initio search model generation for molecular replacement.
SHELX	Automated SHELXC/D/E structure solution pipeline for fast routine experimental phasing. Accepts data in XDS, Scalepack, SHELX hkl or mtz formats and outputs phases and a poly-Ala trace. If a protein sequence is provided, BUCCANEER and REFMAC complete the structure.
CRANK2	Automated structure solution pipeline for experimental phasing using maximum likelihood methods.
MoRDa	MoRDa is a pipeline for molecular replacement protein structure solution based on its own domain database. Models relevant to the target sequence are further adjusted before molecular replacement search.
SIMBAD	Sequence-independent molecular replacement, good for identifying if your crystal contains a contaminant protein. SIMBAD can also do full search of homologous structures in difficult-to-solve novel target cases, but this functionality is not yet available through CCP4-Online.
ARCIMBOLDO	Crystallographic Ab Initio protein solution below atomic resolution

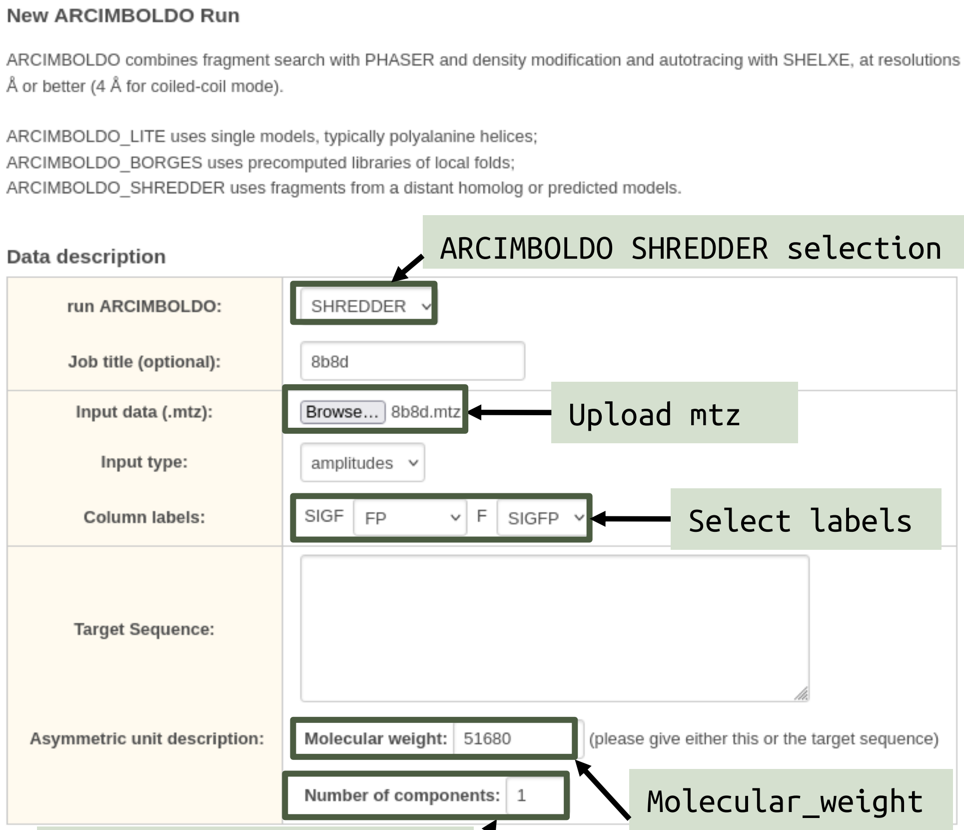
ARCIMBOLDO

CLICK HERE TO START A NEW ARCIMBOLDO RUN

The table below shows your current ARCIMBOLDO runs - the results will be stored on the server for at least 30 days

id	Job Title	Delete/Stop	Date	Status	View Results

Setting the job in CCP4 online



New ARCIMBOLDO Run

ARCIMBOLDO combines fragment search with PHASER and density modification and autotracing with SHELXE, at resolutions of 2.5 Å or better (4 Å for coiled-coil mode).

ARCIMBOLDO_LITE uses single models, typically polyalanine helices;
ARCIMBOLDO_BORGES uses precomputed libraries of local folds;
ARCIMBOLDO_SHREDDER uses fragments from a distant homolog or predicted models.

Data description

run ARCIMBOLDO: SHREDDER
Job title (optional): 8b8d
Input data (.mtz): Browse... 8b8d.mtz
Input type: amplitudes
Column labels: SIGF FP F SIGFP
Target Sequence:
Asymmetric unit description: Molecular weight: 51680 (please give either this or the target sequence)
Number of components: 1

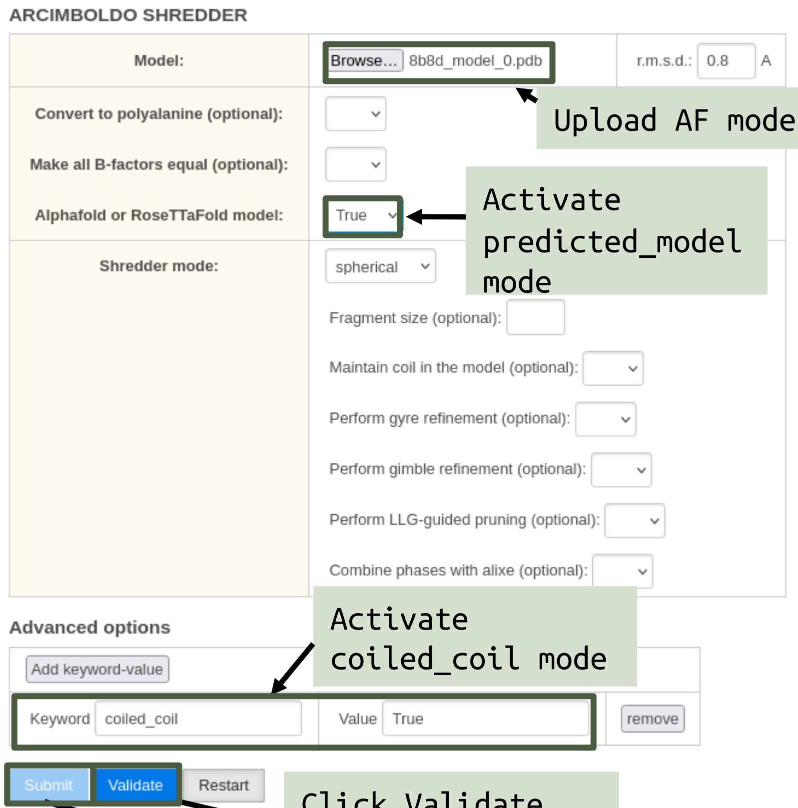
ARCIMBOLDO SHREDDER selection

Upload mtz

Select labels

Molecular_weight = 51680 Da

number_of_component = 1



ARCIMBOLDO SHREDDER

Model: Browse... 8b8d_model_0.pdb r.m.s.d.: 0.8 A
Convert to polyalanine (optional):
Make all B-factors equal (optional):
Alphafold or RoseTTaFold model: True
Shredder mode: spherical
Fragment size (optional):
Maintain coil in the model (optional):
Perform gyre refinement (optional):
Perform gimble refinement (optional):
Perform LLG-guided pruning (optional):
Combine phases with alixe (optional):

Advanced options

Keyword coiled_coil Value True

Upload AF model

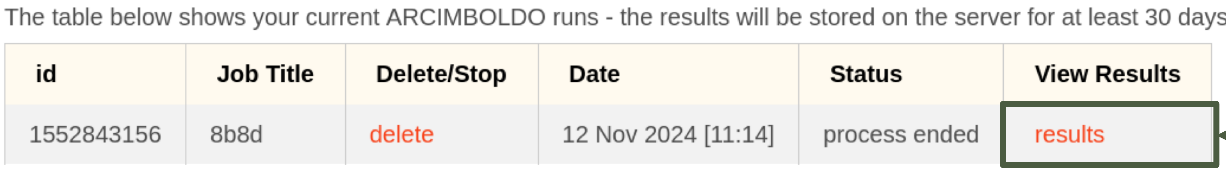
Activate predicted_model mode

Activate coiled_coil mode

Click Validate

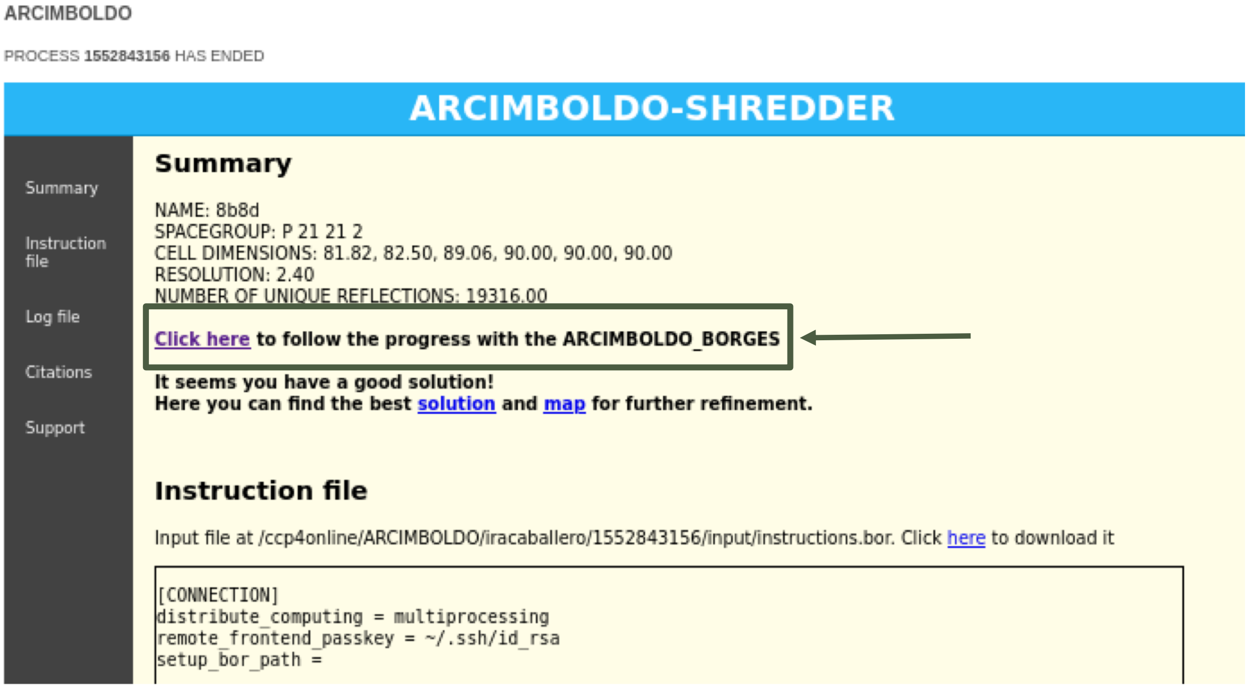
Then click submit

Running the job: HTML results



The table below shows your current ARCIMBOLDO runs - the results will be stored on the server for at least 30 days

id	Job Title	Delete/Stop	Date	Status	View Results
1552843156	8b8d	delete	12 Nov 2024 [11:14]	process ended	results



ARCIMBOLDO

PROCESS **1552843156** HAS ENDED

ARCIMBOLDO-SHREDDER

Summary

NAME: 8b8d
SPACEGROUP: P 21 21 2
CELL DIMENSIONS: 81.82, 82.50, 89.06, 90.00, 90.00, 90.00
RESOLUTION: 2.40
NUMBER OF UNIQUE REFLECTIONS: 19316.00

Click here to follow the progress with the ARCIMBOLDO_BORGES

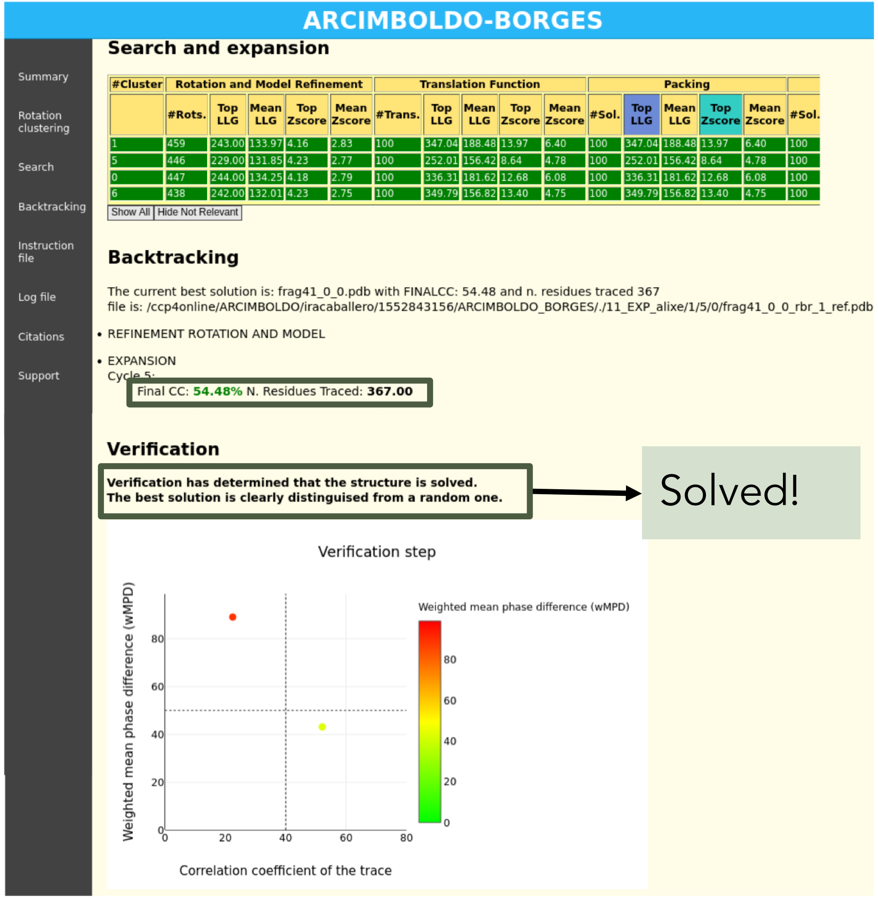
It seems you have a good solution!
Here you can find the best solution and map for further refinement.

Instruction file

Input file at /ccp4online/ARCIMBOLDO/iracaballero/1552843156/input/instructions.bor. Click here to download it

```
[CONNECTION]
distribute_computing = multiprocessing
remote_frontend_passkey = ~/.ssh/id_rsa
setup_bor_path =
```

Running the job: HTML results



Search and expansion

#Cluster	Rotation and Model Refinement					Translation Function					Packing					
	#Rots.	Top LLG	Mean LLG	Top Zscore	Mean Zscore	#Trans.	Top LLG	Mean LLG	Top Zscore	Mean Zscore	#Sol.	Top LLG	Mean LLG	Top Zscore	Mean Zscore	#Sol.
1	459	243.00	133.97	4.16	2.83	100	347.04	188.48	13.97	6.40	100	347.04	188.48	13.97	6.40	100
5	446	229.00	131.85	4.23	2.77	100	252.01	156.42	8.64	4.78	100	252.01	156.42	8.64	4.78	100
0	447	244.00	134.25	4.18	2.79	100	336.31	181.62	12.68	6.08	100	336.31	181.62	12.68	6.08	100
6	438	242.00	132.01	4.23	2.75	100	349.79	156.82	13.40	4.75	100	349.79	156.82	13.40	4.75	100

Backtracking

The current best solution is: frag41_0_0.pdb with FINALCC: 54.48 and n. residues traced 367
file is: /ccp4online/ARCIMBOLDO/iracaballero/1552843156/ARCIMBOLDO_BORGES/./11_EXP_alixe/1/5/0/frag41_0_0_rbr_1_ref.pdb

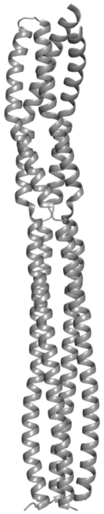
- REFINEMENT ROTATION AND MODEL
- EXPANSION
  Cycle 5:
  Final CC: **54.48%** N. Residues Traced: **367.00**

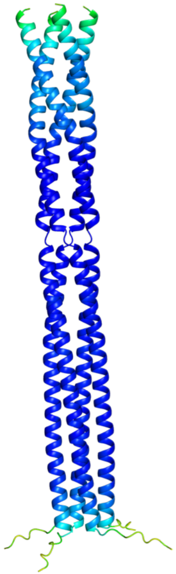
Verification

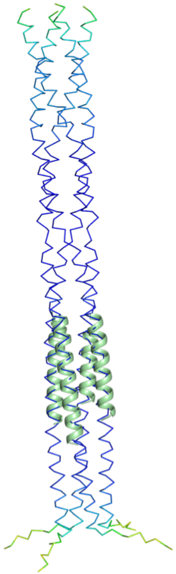
**Verification has determined that the structure is solved.
The best solution is clearly distinguised from a random one.**

Solved!

PDB entry: 8B8D
Resolution: 2.4Å
4 helices of 115 residues


Deposited PDB structure


AF3 model


Shredder fragment


Final solution

SUMMARY - ARCIMBOLDO

- ARCIMBOLDO_SHREDDER has a **`predicted_model mode`**:
  - No need to preprocess models.
  - If solution is straightforward, it will systematically eliminate model bias.
  - This feature can be run as standalone from a MR solution.

- The **`coiled_coil mode`** is available in ARCIMBOLDO_LITE and ARCIMBOLDO_SHREDDER:
  - Works at resolution up to 4Å.
  - Use ARCIMBOLDO_LITE with ideal helices and ARCIMBOLDO_SHREDDER with multimeric predictions.
  - It includes a verification step, key when going to lower resolutions.

SUMMARY - VERIFICATION

The core idea of ARCIMBOLDO is that correct solutions are identified when the minimal starting hypothesis reveals previously unknown structural features, while incorrect ones do not progress into meaningful solutions.

- **Predicted models:** the starting hypothesis leaves little to infer as fragments are now more accurate but this also introduce model bias → model-free verification that will systematically eliminate model bias.
- **Coiled coils:** figures of merit become unreliable → dedicated verification step to disprove the final solution against competing alternatives.
- **Low resolution in globular proteins:** the emergence of new features in unmodeled map regions confirmed partial model correctness.

ACKNOWLEDGEMENTS

ARCIMBOLDO TEAM: **Isabel Usón**, Josep Triviño, Elisabet Jiménez, Mari Carmen Martín & Iain Visa

~~Old~~ Always members of the ARCIMBOLDO TEAM: Claudia Millán, Massimo Sammito, Albert Castellvi, Ana Medina, Nicolas Soler & Rafael Borges

PHASER group: Airlie McCoy & Randy Read

Kay Diederichs & George Sheldrick

ibmb
Institut de Biologia Molecular de Barcelona
Molecular Biology Institute of Barcelona CSIC

CCP4