Structure Solved?
Validation and Deposition.

Eugene Krissinel

CCP4, STFC Research Complex at Harwell
Didcot, United Kingdom

eugene.krissinel@stfc.ac.uk

CCP4/DLS Data Collection and Analysis Workshop
DLS, Harwell, UK, 27 November - 5 December 2023

Research Complex at Harwell

Do we really know what we do?

- We barely have sufficient information for structure solution in MX

in a good case of having 30,000 reflections:

Target	*No. of atoms*	*No. of degress of freedom*	*Information content*
Small molecules	10 - 300	30 - 900	30 -1000
Macromolecules, polymers	3,000 - 30,000	9,000 - 90,000	**0.3 - 3**

and even worse because of noise in data

- Structure solution in MX only possible because we bring additional data/ information from other sources, primarily chemistry
- While we do so, bias or errors are possible
- Therefore, VALIDATION is an integral part of structure solution as any other stage of the whole process

Structure Validation

- Structure validation is a natural step in structure solution to establish that:
  - structure is in compliance with experimental data (e.g. Space Group is correct)
  - structure is chemically sensible
  - structure does not contain features not supported by electron density
  - all structural components are named and described correctly
- An ideal structure does not raise red flags in Coot — this is not achieved in many practical cases though; one need to perform:
  - electron density fit analysis (RSR-factor, atom map density correlation)
  - atom clash score
  - rotamer analysis (ideally no unusual conformations)
  - Ramachandran plot analysis (ideally no outliers)
  - difference map analysis (ideally no green blobs)
  - restraints rms checks
  - Mogul validation for ligands
- PDB-REDO is a powerful validation and structure completion tool — use it

Structure Validation

- Key validation criteria guidance figures

Validation criterion	*Ideal score*	*Median for 1.5 / 3Å structures*
R_{free}	Undefined	0.21 / 0.28
Real-space residual (% RSRZ>2)	Undefined	2.7 (resolution independent)
Clashscore (clashes per 1000 atoms, including H)	<5	8.8 / 3.9
Under-packing	1	1.2 / 2.2
Ramachandran score (% outliers)	0.05	0 / 1.7
Rotamer score (% poor)	0.5	1.7 / 9.6
Buried H-bonds (fraction unsatisfied)	0.02	0.025 / 0.08
RNA ribose puckers (%poor)	0.5	0 / 2.7

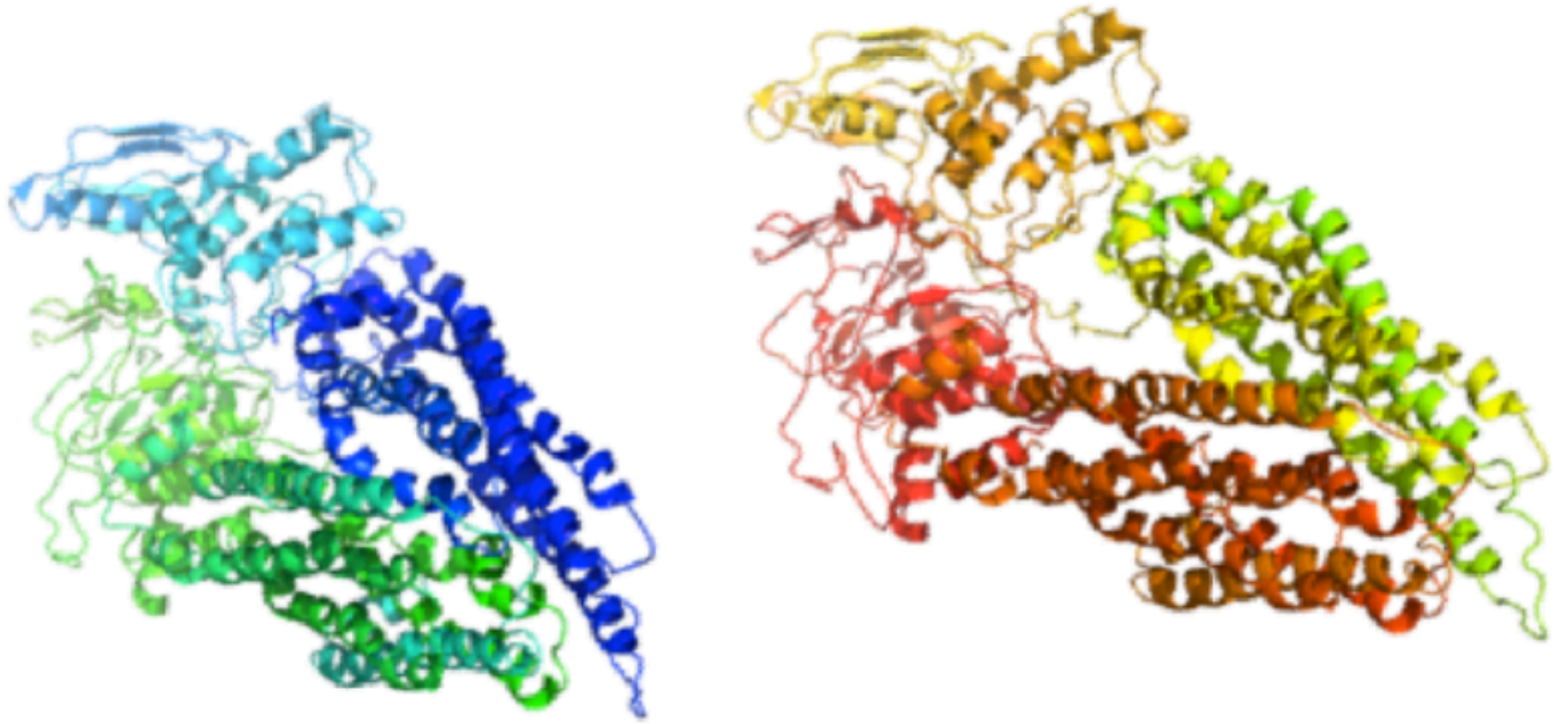
RETRACTED: Structure of MsbA from *Vibrio cholera*: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation

Geoffrey Chang[a],

[a]Department of Molecular Biology, CB-105, The Scripps Research Institute, La Jolla, CA 92037, USA

Edited by D. Rees. Available online 25 June 2003.

"were incorrect in both the hand of the structure and the topology. Thus, the biological interpretations based on the inverted models for MsbA are invalid."



RETRACTED: Structure of MsbA from *Vibrio cholera*: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation

Geoffrey Chang[a],

[a]Department of Molecular Biology, CB-105, The Scripps Research Institute, La Jolla, CA 92037, USA

Edited by D. Rees. Available online 25 June 2003.

"were incorrect in both the hand of the structure and the topology. Thus, the biological interpretations based on the inverted models for MsbA are invalid."

The following papers were retracted in 2007:[4][10]

1. Chang G, Roth CB. (2001) Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* 293(5536):1793-800. PMID 11546864
2. Pornillos O, Chen YJ, Chen AP, Chang G. (2005) X-ray structure of the EmrE multidrug transporter in complex with a substrate. *Science* 310(5756):1950-3. PMID 16373573
3. Reyes CL, Chang G. (2005) Structure of the ABC transporter MsbA in complex with ADP.vanadate and lipopolysaccharide. *Science* 308(5724):1028-31. PMID 15890884
4. Chang G. (2003). Structure of MsbA from Vibrio cholera: a multidrug resistance ABC transporter homolog in a closed conformation. *J Mol Biol* 330(2):419-30. PMID 12823979
5. Ma C, Chang G. (2004). Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli. *Proc Natl Acad Sci USA* 101(9):2852-7. PMID 14970332

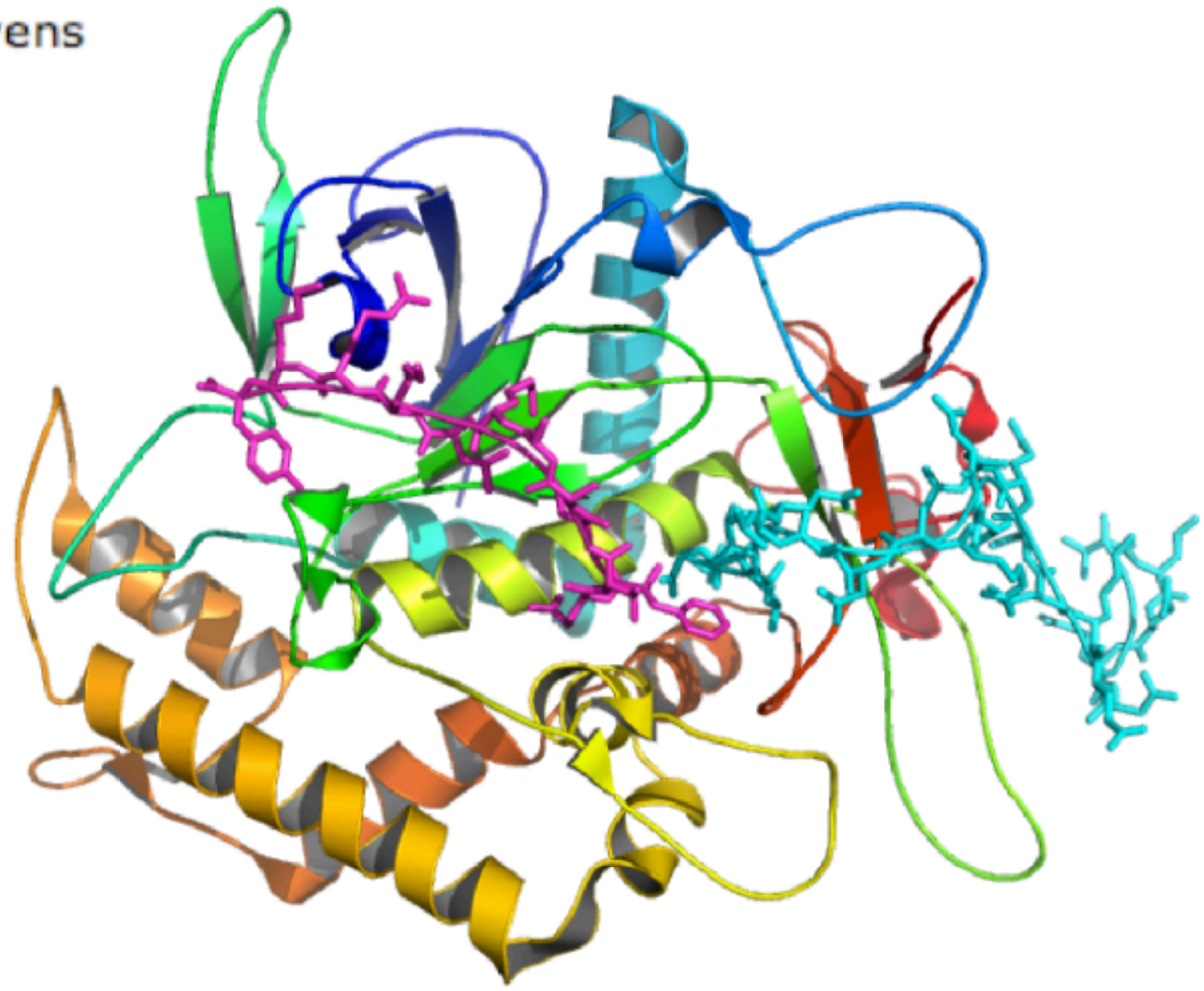
Nature Structural & Molecular Biology **16**, 795 (2009)
doi:10.1038/nsmb0709-795

Retraction: Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution

Michael A Hanson & Raymond C Stevens

"However, because of the lack of clear and continuous electron density for the peptide in the complex structure, the paper is being retracted."

1F83



Structure Validation Advanced

X-ray Validation Task Force

Method-specific Validation Task Forces have been convened to collect recommendations and develop consensus on additional validation that should be performed, and to identify software applications to perform validation tasks.

The recommendations of the X-ray Validation Task Force have been published:

A New Generation of Crystallographic Validation Tools for the Protein Data Bank
Randy J. Read, Paul D. Adams, W. Bryan Arendall III, Axel T. Brunger, Paul Emsley, Robbie P. Joosten, Gerard J. Kleywegt, Eugene B. Krissinel, Thomas Lütteke, Zbyszek Otwinowski, Anastassis Perrakis, Jane S. Richardson, William H. Sheffler, Janet L. Smith, Ian J. Tickle, Gert Vriend and Peter H. Zwart.
(2011) Structure 19: 1395-1412.

These recommendations are the basis for the validation reports produced by the wwPDB.

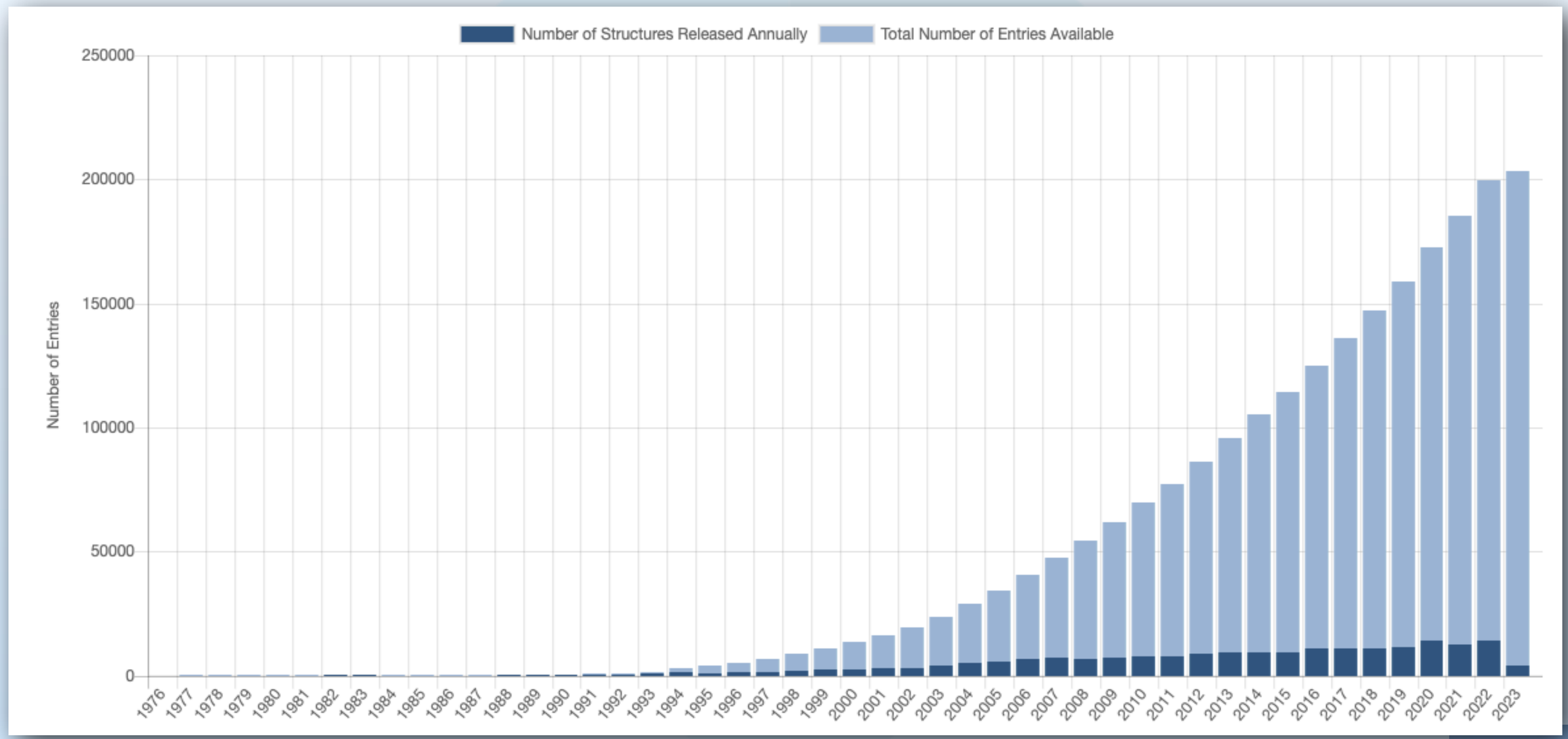
The wwPDB X-ray VTF was convened in 2008 at a workshop organized by Randy Read (Cambridge University), and sponsored by the RCSB PDB & PDBe. Detailed information about the workshop is **available**.

The wwPDB greatly appreciates the efforts of its Task Forces.

Questions? **info@wwpdb.org**.

What is the PDB

- The probably oldest freely available biological database



What is PDB contents?

- The PDB contains 3-dimensional structures of biological macromolecules, obtained by

  - X-ray Diffraction (*2019: 83.7%,* now: **85.6%**)
  - Nuclear Magnetic Resonance technique (*2019: 7.4%,* now: **6.8%**)
  - Electron Microscopy (*2019: 2%,* now: **7.3%**)
  - Neutron Diffraction (few)
  - Electron Diffraction (few)
  - Solid-State NMR (few)
  - Fiber Diffraction (few)

What is PDB Deposition?

- In brief, PDB Deposition is uploading files and filling in web-forms!
- In reality, PDB Deposition is a relatively laborious multi-stage process:
  1. *Structure validation*
  2. *Preparation of deposition files*
  3. *wwPDB validation*
  4. *Form filling*
  5. *Annotation — including more validation*
- Focus on validation and making sure that annotation is correct
  - *Deposition **is** a publication, take it very seriously*

2. Preparation of Deposition Files

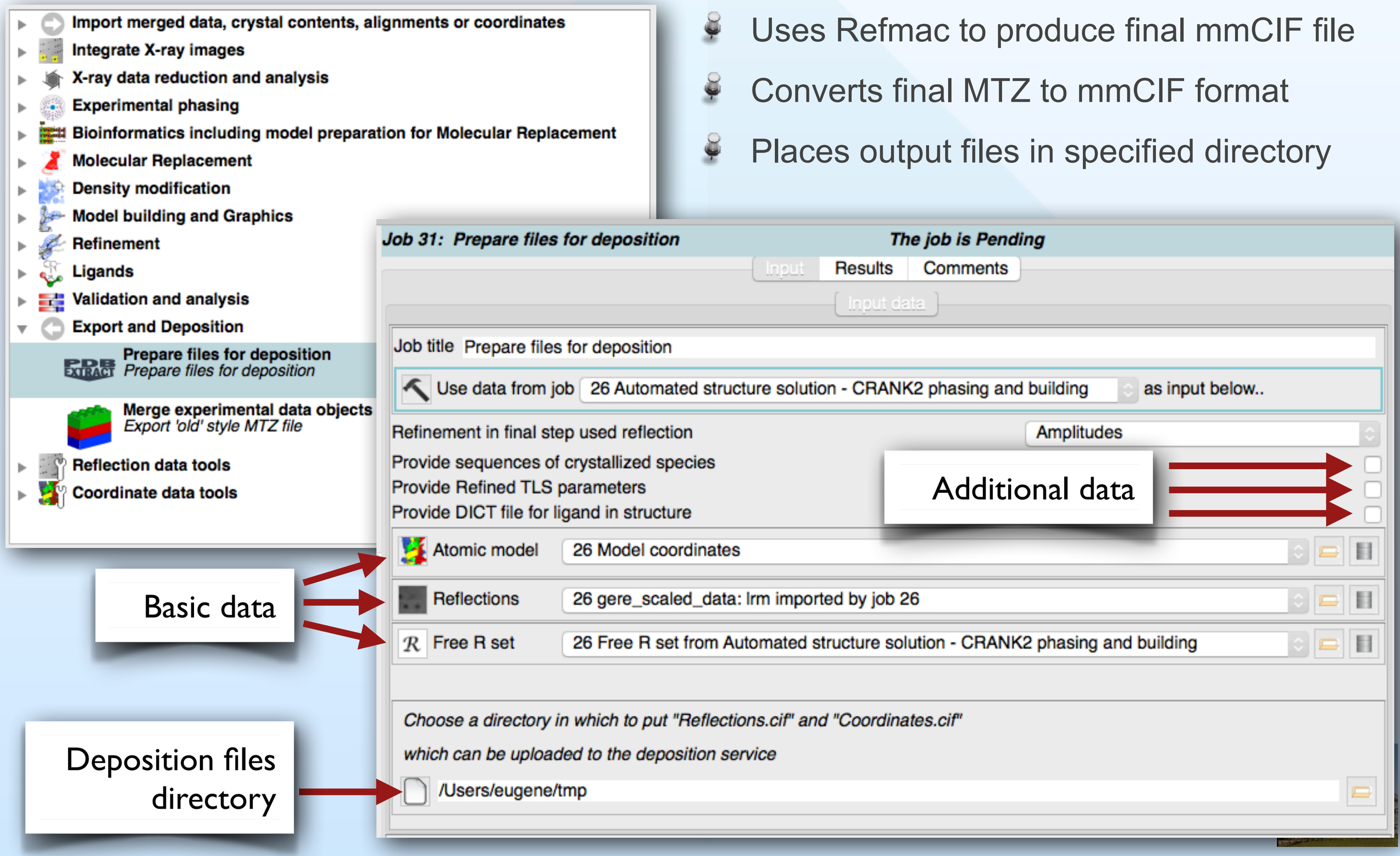
- The PDB requires the following basic information:
  1. Final model coordinates **in mmCIF format** (PDB format is now deprecated)
  2. mmCIF dictionaries for all ligands and non-standard/modified residues
  3. Structure factors **in mmCIF format**
  4. Relevant (protein/dna/rna) sequences

- In addition, you are expected to provide
  - data processing statistics (usually from XDS/DIALS or Xia-2)
  - data scaling and merging statistics (usually from Aimless in CCP4
  - final refinement scores (usually from Refmac in CCP4)
  - information on quaternary structure or oligomeric state (experimental or from PISA)
  - other annotation details (names, beamline info, etc etc)

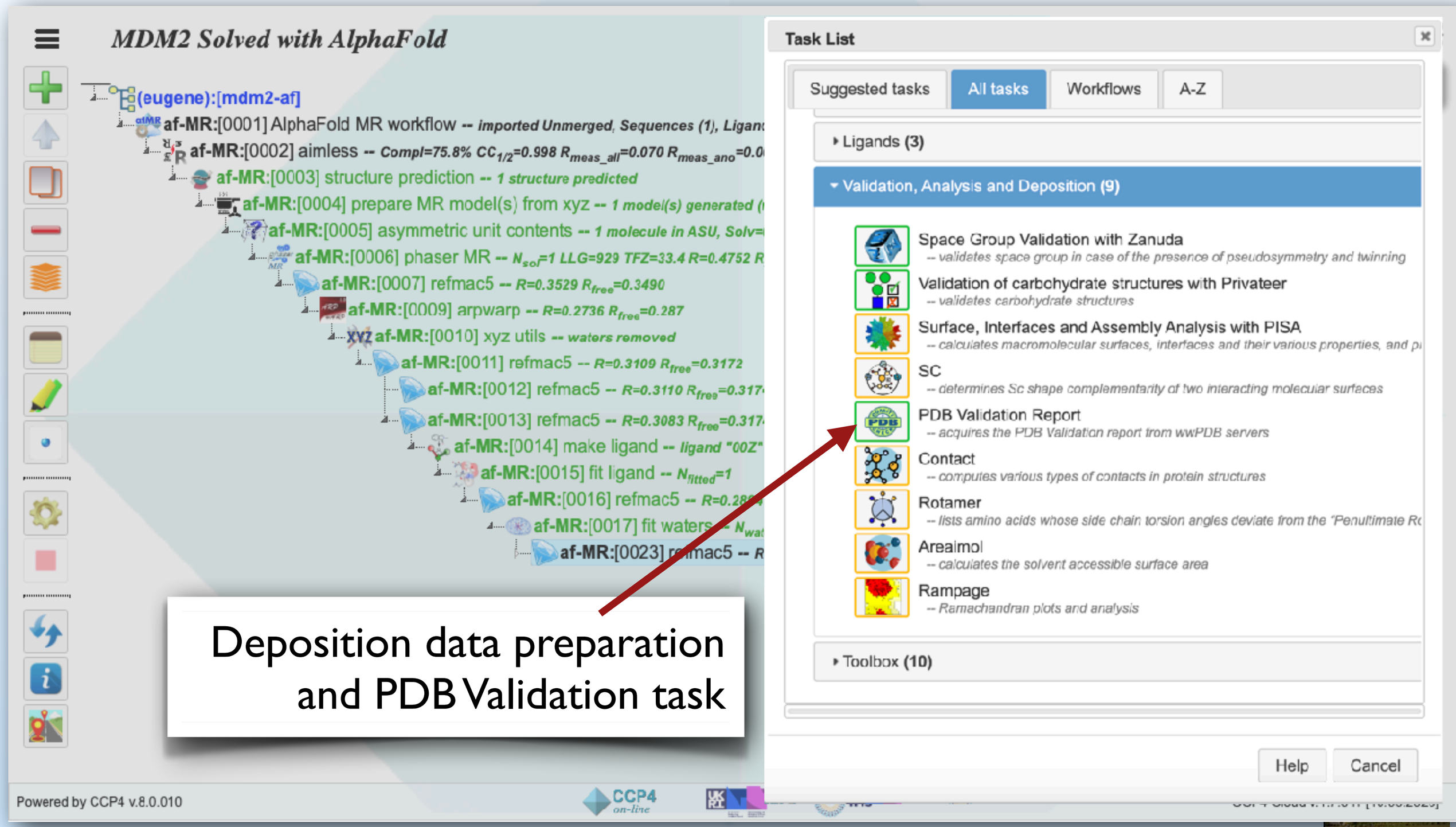
- Good news: much of this is prepared for you by CCP4 interfaces

Preparation of Deposition Files in CCP4i2

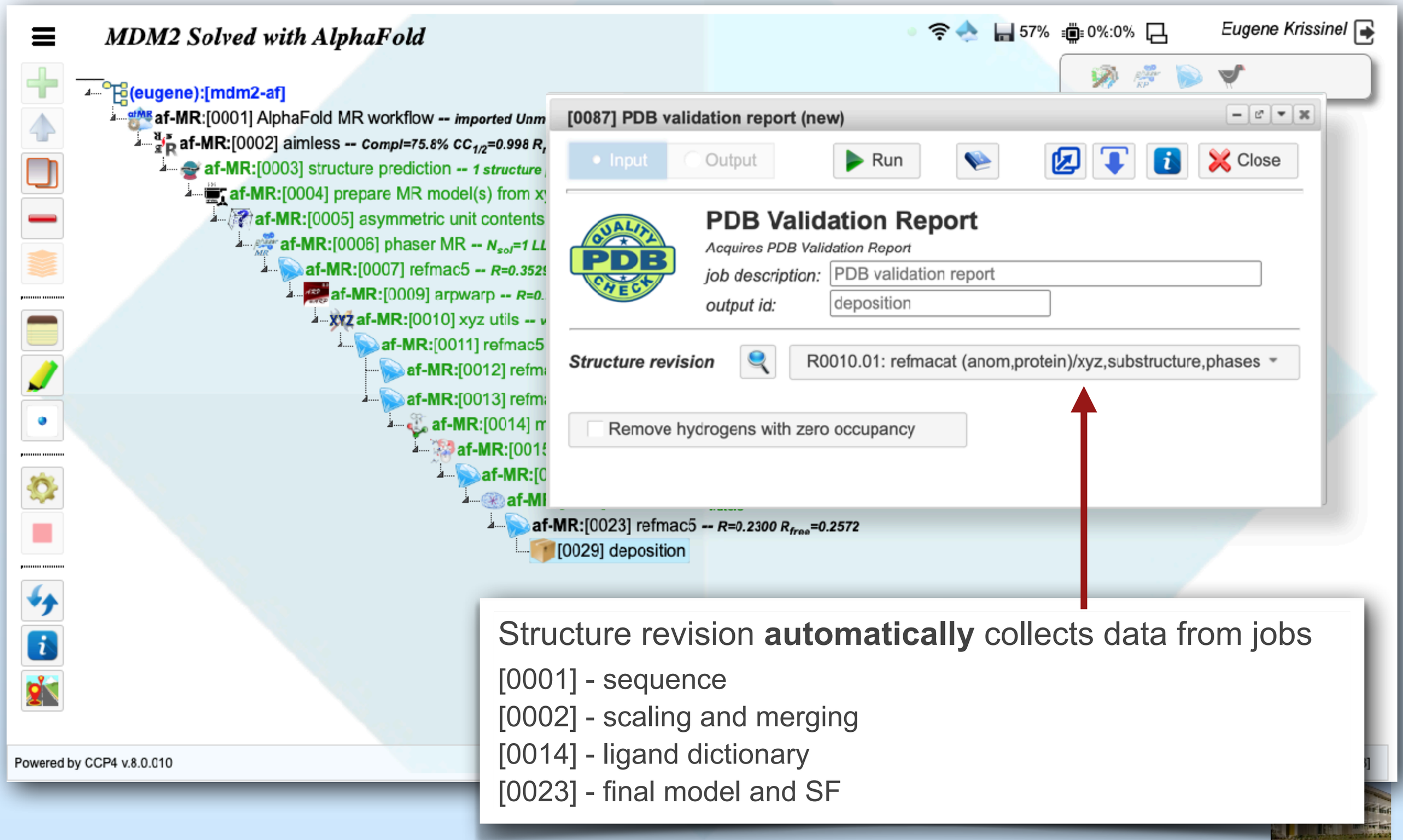
- Uses Refmac to produce final mmCIF file
- Converts final MTZ to mmCIF format
- Places output files in specified directory



Preparation of Deposition Files in CCP4 Cloud



Preparation of Deposition Files in CCP4 Cloud



Structure revision **automatically** collects data from jobs

[0001] - sequence
[0002] - scaling and merging
[0014] - ligand dictionary
[0023] - final model and SF

Preparation of Deposition Files in CCP4 Cloud

[0087] PDB validation report – finished

Input • Output | Close

Report | Main Log | Service Log | Errors

[0087] PDB Validation Report

CCP4 v.8.0.016; CCP4 Cloud v.unknown
Started: 2023-12-04 17:06:17
Finished: 2023-12-04 17:10:04
CPU: 00.149, Disk: 24.64M

Prepare deposition files

PDB id: not... 1 / 9 77%

Link to PDB deposition files

WORLDWIDE PDB PROTEIN DATA BANK

Preliminary Full wwPDB X-ray Structure Validation Report (i)

Dec 4, 2023 – 12:08 PM EST

This wwPDB validation report is NOT for manuscript review

This is a Preliminary Full wwPDB X-ray Structure Validation Report.

This report is produced by the standalone wwPDB validation server.
The structure in question has not been deposited to the wwPDB.
This report should not be submitted to journals.

We welcome your comments at *validation@mail.wwpdb.org*
A user guide is available at
https://www.wwpdb.org/validation/2017/XrayValidationReportHelp

Not Reviewed

PDB Validation Report

Preparation of Deposition Files in CCP4 Cloud

[0088] PDB deposition files -- finished

Input • Output | Close

Report | Main Log | Service Log | Errors

CCP4 v.8.0.016; CCP4 Cloud v.unknown
Started: 2023-12-04 17:13:20
Finished: 2023-12-04 17:13:20
CPU: 00.000, Disk: 0.04M

[0088] PDB Deposition Files

PDB Deposition:

a) Download the following 2 files in mmCIF format:

1. *Final atomic coordinates:* download
2. *Reflection data*
   either *merged and unmerged:* download *(recommended)*
   or *merged only:* download *(only if merged+unmerged*

b) Start new deposition session at wwPDB Deposition Site *(link opens in new tab/window)*

c) Follow instructions in the wwPDB deposition site and uploadthe files downloaded when prompted.

d) Consider archiving your Project (find more information here)

Legacy output. Do NOT deposit this file to the PDB, special use only

Final atomic coordinates in PDB format: download

Deposition files:

1. Coordinates+Sequences+Dictionaries+Data Statistics
2. Structure Factors

Link to PDB deposition

[0029] deposition -- finished

Input • Output Close

Report | Main Log | Service Log | Errors

[0029] References

Your results were obtained using developments from publications listed below. Please cite them when publishing:

- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007) *Clustal W and Clustal X version 2.0*. Bioinformatics **23**: 2947-2948; doi:10.1093/bioinformatics/btm404
- Vagin, A., and Teplyakov, A. (1997) *MOLREP: an automated program for molecular replacement*. J. Appl. Cryst. **30**: 1022-1025; doi:10.1107/S0021889897006766
- McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., Read R.J. (2007) *Phaser Crystallographic Software*. J. Appl. Cryst. **40**: 658-674; doi:10.1107/S0021889807021206
- Murshudov, G.N., Skubak, P., Lebedev, A.A., Pannu, N.S., Steiner, R.A., Nicholls, R.A., Winn, M.D., Long, F., and Vagin, A.A. (2011) *REFMAC5 for the refinement of macromolecular crystal structures*. Acta Cryst. **D67**: 355-367; doi:10.1107/S0907444911001314
- Tickle, I.J., Laskowski, R.A. and Moss, D.S. (1998) *Error Estimates of Protein Structure Coordinates and Deviations from Standard Geometry by Full-Matrix Refinement of γB- and BB2-Crystallin*. Acta Cryst. **D54**: 243-252; doi:10.1107/S090744499701041X
- Williams C. J., Hintze B. J., Headd J. J., Moriarty N. W., Chen V. B., Jain S., Prisant M. G., Lewis S. M., Videau L. L. , Keedy D. A., Deis L. N., Arendall W. B. III, Verma V., Snoeyink J. S., Adams P. D., Lovell S. C., Richardson J. S., Richardson D. C. (2018) *MolProbity: More and better reference data for improved all-atom structure validation*. Protein Science **27**: 293-315; doi:10.1002/pro.3330
- Grosse-Kunstleve R. W., Sauter N. K., Moriarty N. W., Adams P. D. (2002) *The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework*. J. Appl. Cryst. **35**: 126-136; doi:10.1107/S0021889801017824
- Perrakis, A., Morris, R., Lamzin, V.S. (1999) *Automated protein model building combined with iterative structure refinement.*. Nature Struct. Biol. **6**: 458-463; doi:10.1038/8263
- Perrakis, A., Harkiolaki, M., Wilson, K.S., Lamzin, V.S. (2001) *ARP/wARP and molecular replacement.*. Acta Cryst. **57**: 1445-1450; doi:10.1107/S0907444901014007
- Langer G., Cohen S.X., Lamzin V.S., Perrakis A. (2008) *Automated macromolecular model building for x-ray crystallography using ARP/wARP version 7.*. Nat. Protoc. **3**: 1171-1179; doi:10.1038/nprot.2008.91
- Long, F., Nicholls, R.A., Emsley, P., Grazulis, S., Merkys, A., Vaitkusb, A., Murshudov, G.N. (2017) *AceDRG: a stereochemical description generator for ligands*. Acta Cryst. **D73**: 112-122; doi:10.1107/S2059798317000067
- Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, K. (2010) *Features and development of Coot*. Acta Cryst. **D66**: 486-501; doi:10.1107/S0907444910007493
- Wojdyr, M. (2022) *GEMMI: A library for structural biology*. J. Open Source Softw. **7(73)**: 4200; doi:10.21105/joss.04200
- Collaborative Computational Project, Number 4 (2011) *Overview of the CCP4 suite and current developments*. Acta Cryst. **D67**: 235-242; doi:10.1107/S0907444910045749
- Krissinel, E., Uski, V., Lebedev, A., Winn, M., Ballard, C. (2018) *Distributed computing for macromolecular crystallography*. Acta Cryst. **D74**: 143-151; doi:10.1107/S2059798317014565

Throughout the Project, you may have used the following graphical viewers, please cite them at your discretion:

- ***UglyMol***: (javascript molecular graphics)

Citing Works

Impress yourself with how many people worked hard to help you solving your structure -- and think how you can pay them credits.

Whatever you do in crystallography, the citation list will go long.

3. PDB Validation Report

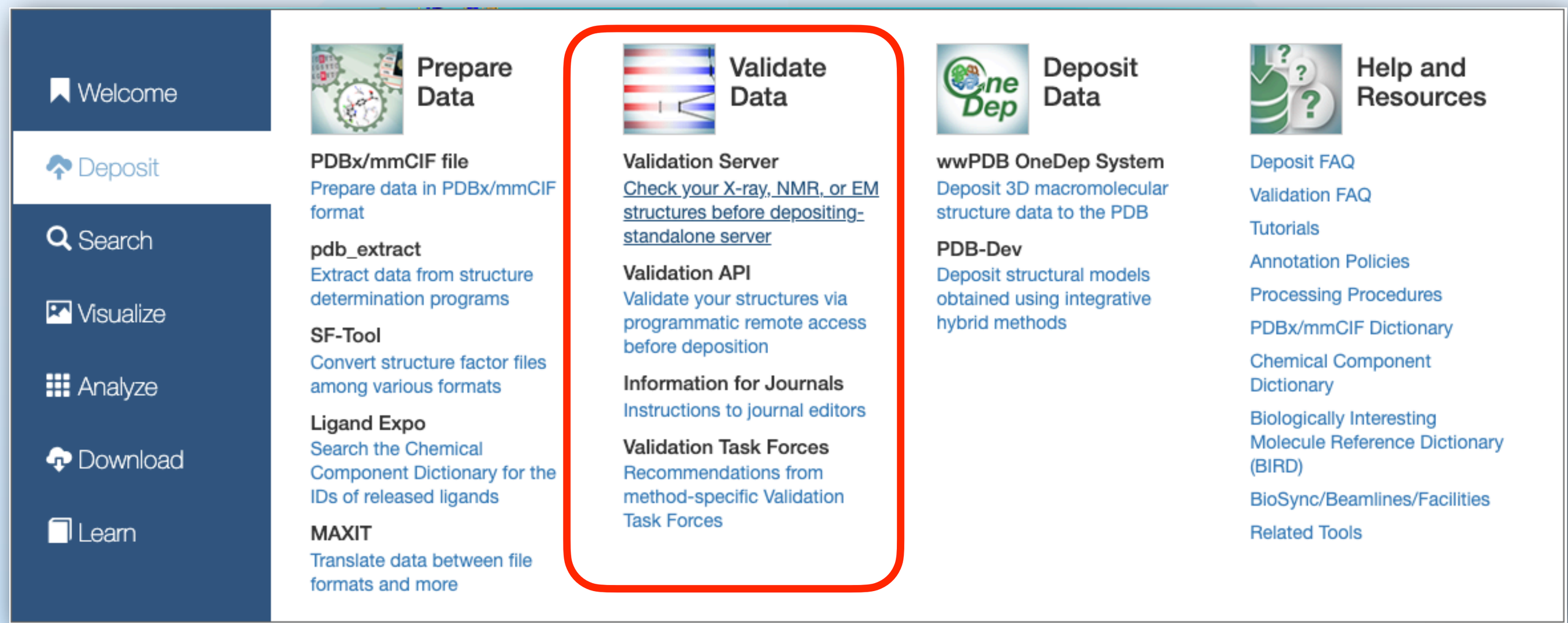
- Must be obtained and inspected before deposition
  - *PDB Validation Reports are not to be uploaded/deposited*
  - *PDB-VR will be generated by wwPDB as part of deposition process*
  - *checking and fixing all structure defects highlighted in PDB-VR will save you a fair bit of communication with the PDB, who will insist, in first turn, that obvious defects are rectified*
  - *you will have to justify major irregularities highlighted by PDB-VR if you think that they are structure feature rather than defects*
- PDB Validation Reports are automatically obtained by the Deposition Task in CCP4 Cloud
- If you are using CCP4i2, obtain PDB Validation Report by uploading mmCIF files, generated by the Deposition Task in CCP4i2, at

https://validate.wwpdb.org/

What is the PDB Validation Report

- *automatic service created as a result of the "Validation Task Force" initiative*
- *assesses quality of structure and correspondence with the experimental data*
- *checks for many obvious flaws; extremely useful*

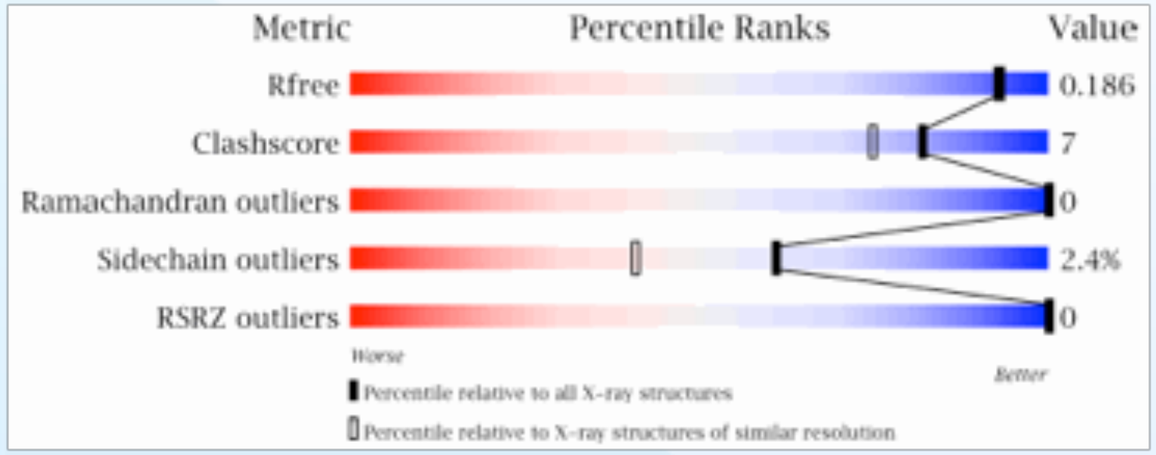
https://rcsb.org/#Category-deposit



PDB Validation Report

(excerpts)

1. Overall quality
   - *data for each slider is detailed in individual report sections*



2. Chain quality
   - *coloured representation of outliers*

Mol	Chain	Length	Quality of chain
1	A	207	
1	B	207	

3. Compound fit quality
   - *wrong geometry or poor fit in density*

Mol	Type	Chain	Res	Geometry	Electron density
6	GOL	A	381	-	X
4	SO4	A	391	-	X
5	OGA	A	370	X	-

4. Extended outlier statistics
   - *bond length, angles, chirality, planarity*
   - *close contacts*
   - *protein backbone (Ramachandran)*
   - *backbone torsions*
   - *sugar packing*

PDB Validation Report

1. Overall quality
   - *summary of key quality indicators to spot possible problems quickly*
2. Entry composition
   - *summary of the number of unique molecules and how they have been modelled*
3. Residue-property plots
   - *per-residue summary plots of quality information for protein, RNA and DNA molecules*
4. Data and refinement statistics
   - *Space Group, Cell, Completeness, R_{merge}, R_{free}, Resolution, Twinning tests etc etc*
5. Model quality
   - *summary on geometrical quality, chirality, close contacts etc*
6. Fit of model and data
   - *analysis of the fit of the molecule in the entry to experimental data (based on Real Space R-value)*

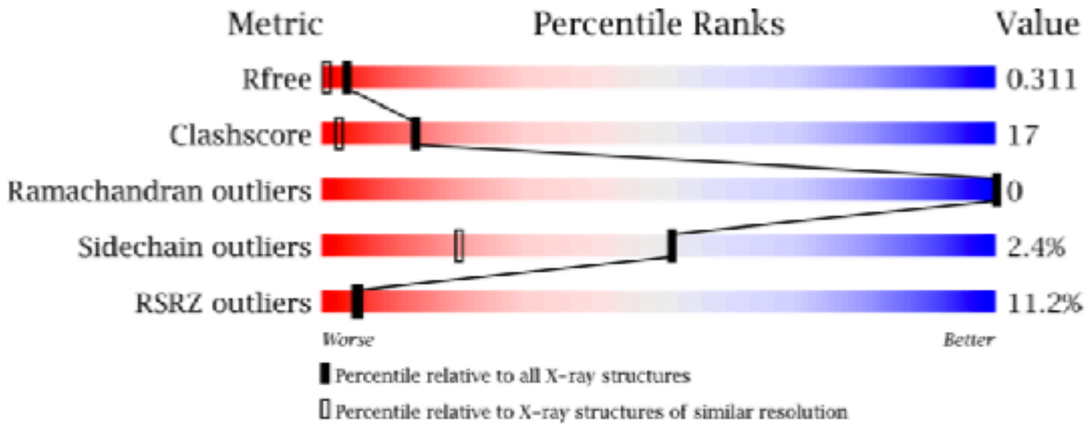
R. J. Read, P. D. Adams, W. B. Arendall III, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend and P.r H. Zwart. (2011) Structure 19: 1395-1412.

1 Overall quality at a glance

The following experimental techniques were used to determine the structure:
X-RAY DIFFRACTION

The reported resolution of this entry is 1.50 Å.

Percentile scores (ranging between 0-100) for global validation metrics of the entry are shown in the following graphic. The table shows the number of entries on which the scores are based.



Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	111664	2534 (1.50-1.50)
Clashscore	122126	2727 (1.50-1.50)
Ramachandran outliers	120053	2661 (1.50-1.50)
Sidechain outliers	120020	2659 (1.50-1.50)
RSRZ outliers	108989	2481 (1.50-1.50)

The table below summarises the geometric issues observed across the polymeric chains and their fit to the electron density. The red, orange, yellow and green segments on the lower bar indicate the fraction of residues that contain outliers for >=3, 2, 1 and 0 types of geometric quality criteria. A grey segment represents the fraction of residues that are not modelled. The numeric value for each fraction is indicated below the corresponding segment, with a dot representing fractions <=5% The upper red bar (where present) indicates the fraction of residues that have poor fit to the electron density. The numeric value is given above the bar.

Mol	Chain	Length	Quality of chain
1	AAA	97	10% 89% 5%

PDB Validation Report

1. Overall quality

PDB Validation Report

2. Entry composition

2 Entry composition

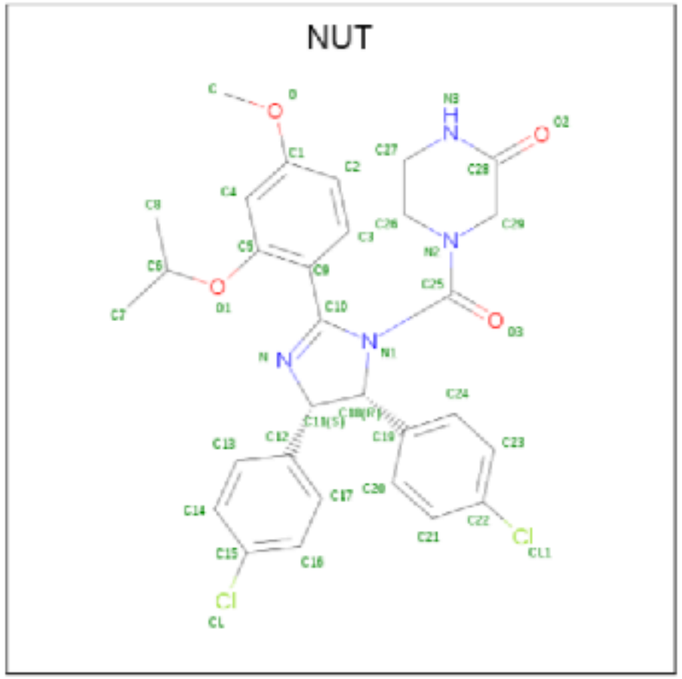
There are 2 unique types of molecules in this entry. The entry contains 789 atoms, of which 0 are hydrogens and 0 are deuteriums.

In the tables below, the ZeroOcc column contains the number of atoms modelled with zero occupancy, the AltConf column contains the number of residues with at least one atom in alternate conformation and the Trace column contains the number of residues modelled with at most 2 atoms.

- Molecule 1 is a protein.

Mol	Chain	Residues	Atoms						ZeroOcc	AltConf	Trace
			Total	C	N	O	S	Se			
1	AAA	92	749	487	124	134	1	3	0	0	0

- Molecule 2 is 4-({(4S,5R)-4,5-bis(4-chlorophenyl)-2-[4-methoxy-2-(propan-2-yloxy)phenyl]-4,5-dihydro-1H-imidazol-1-yl}carbonyl)piperazin-2-one (three-letter code: NUT) (formula: $C_{30}H_{30}Cl_2N_4O_4$).



Mol	Chain	Residues	Atoms					ZeroOcc	AltConf
			Total	C	Cl	N	O		
2	AaA	1	40	30	2	4	4	0	0

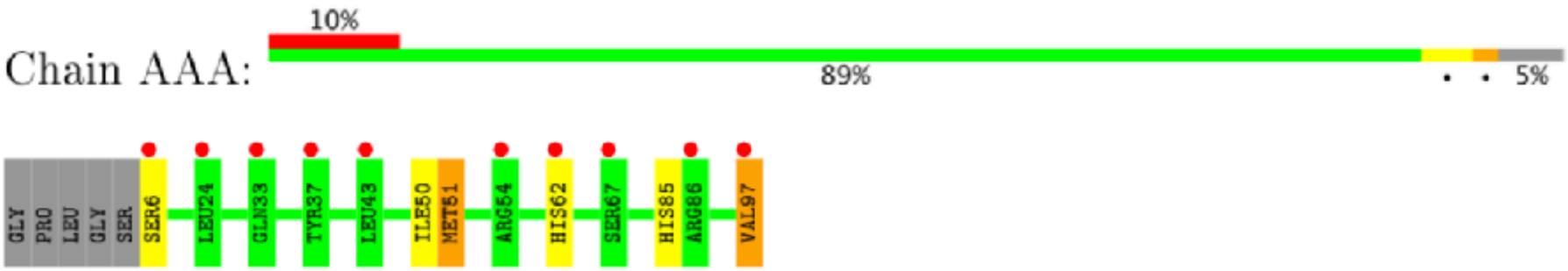
PDB Validation Report

3. Residue-property plots

3 Residue-property plots (i)

These plots are drawn for all protein, RNA and DNA chains in the entry. The first graphic for a chain summarises the proportions of the various outlier classes displayed in the second graphic. The second graphic shows the sequence view annotated by issues in geometryand electron density. Residues are color-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. A red dot above a residue indicates a poor fit to the electron density (RSRZ > 2). Stretches of 2 or more consecutive residues without any outlier are shown as a green connector. Residues present in the sample, but not in the model, are shown in grey.

- Molecule 1:



4 Data and refinement statistics

Property	Value	Source
Space group	P 65 2 2	Depositor
Cell constants a, b, c, α, β, γ	71.50Å 71.50Å 104.28Å 90.00° 90.00° 120.00°	Depositor
Resolution (Å)	61.92 – 1.50 61.92 – 1.50	Depositor EDS
% Data completeness (in resolution range)	98.7 (61.92-1.50) 98.7 (61.92-1.50)	Depositor EDS
R_{merge}	0.04	Depositor
R_{sym}	(Not available)	Depositor
$< I/\sigma(I) >$ [1]	1.50 (at 1.50Å)	Xtriage
Refinement program	refmac 5.8.0238 2018/15/10	Depositor
R, R_{free}	0.288 , 0.311 0.289 , 0.311	Depositor DCC
R_{free} test set	1236 reflections (4.84%)	wwPDB-VP
Wilson B-factor (Å^2)	18.2	Xtriage
Anisotropy	0.251	Xtriage
Bulk solvent k_{sol}(e/Å^3), B_{sol}(Å^2)	0.43 , 41.7	EDS
L-test for twinning[2]	$< \|L\| > = 0.50$, $< L^2 > = 0.33$	Xtriage
Estimated twinning fraction	No twinning to report.	Xtriage
F_o,F_c correlation	0.91	EDS
Total number of atoms	789	wwPDB-VP
Average B, all atoms (Å^2)	20.0	wwPDB-VP

Xtriage's analysis on translational NCS is as follows: *The largest off-origin peak in the Patterson function is 5.47% of the height of the origin peak. No significant pseudotranslation is detected.*

PDB Validation Report

5. Model quality

5.1 Standard geometry: *bond length, angles, chirality and planarity outliers*

5.2 Too close contacts: *atom clashes within ASU and between symmetry mates*

5.3 Torsion angles: *Ramachandran outliers*

5.4 Non-standard residues in protein, DNA and RNA chains: *are they correctly annotated?*

5.5 Carbohydrates: *any geometry and annotation issues*

5.6 Ligand geometry: *bond length, angles, chirality and planarity outliers*

5.7 Other polymers: *any geometry and annotation issues*

5.8 Polymer linkage issues: *are there linkage breaks?*

6. Fit of model and data

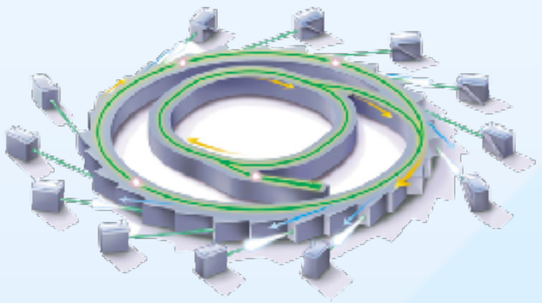
- *analysis of the fit of the molecule in the entry to experimental data (based on Real Space R-value)*

Reference:

R. J. Read, P. D. Adams, W. B. Arendall III, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend and P.r H. Zwart. (2011) Structure 19: 1395-1412.

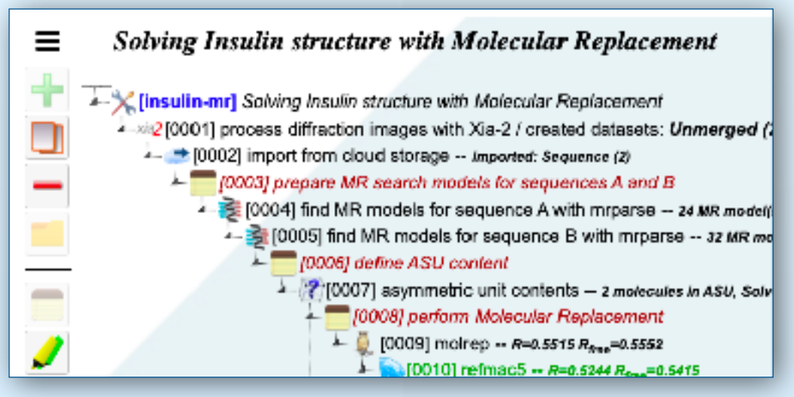
Final Step: Archive your Project in CCP4 Cloud

Experiment



evidence

Structure Solution Project



interpretation

Project Archive

Make your work available to reviewers

Share your knowledge

Release your storage quota

Archive ID: CCP4-1XYZ

Repository

PDB Code: 1XYZ

structures

Research

Final Step: Archive your Project in CCP4 Cloud

CCP4 Cloud Archive ID: **CCP4-RNASE** (could be published in a paper on web-site etc.)

Access URL:

https://cloud.ccp4.ac.uk/archive/access.html?id=CCP4-RNASE

Consider archiving completed projects in CCP4 Cloud when publishing your work