The DRI Project and forthcoming activities to support metadata-rich depositions

Genevieve Evans

Structural Biologist
& wwPDB data curator

Email: gle@ebi.ac.uk

PDBe
Protein Data Bank in Europe

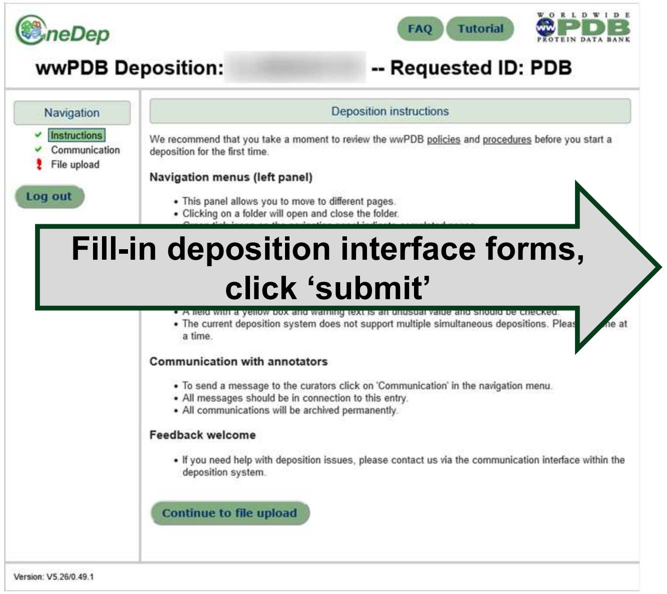
EMBL-EBI

Submitting data to the PDB… is hard work

Model file =
Coordinates (mandatory)
+
Metadata (optional)

Experimental data file =
Experimental data (mandatory)
+
Metadata (tiny bit)

Upload



OneDep

FAQ Tutorial

WORLDWIDE PDB PROTEIN DATA BANK

wwPDB Deposition: [redacted] **-- Requested ID: PDB**

Deposition instructions

We recommend that you take a moment to review the wwPDB policies and procedures before you start a deposition for the first time.

Navigation menus (left panel)

- This panel allows you to move to different pages.
- Clicking on a folder will open and close the folder.
- A field with a yellow box and warning text is an unusual value and should be checked.
- The current deposition system does not support multiple simultaneous depositions. Pleas ne at a time.

Communication with annotators

- To send a message to the curators click on 'Communication' in the navigation menu.
- All messages should be in connection to this entry.
- All communications will be archived permanently.

Feedback welcome

- If you need help with deposition issues, please contact us via the communication interface within the deposition system.

Continue to file upload

Version: V5.26/0.49.1

Fill-in deposition interface forms, click 'submit'

OneDep

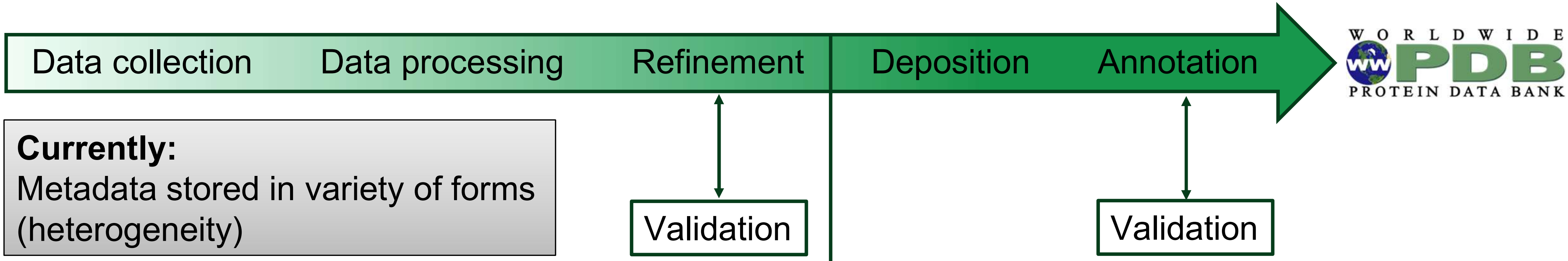
Annotation

Experimental data file
+
Annotated model file =
Coordinates
+
Metadata ("full")

The richer in metatada in the uploaded file(s), the less forms to fill-in the deposition interface.

The road to retain metadata

Each step must
- generate its own metadata
- retain the previous step's metadata



Data collection Data processing Refinement Deposition Annotation

WORLDWIDE PDB PROTEIN DATA BANK

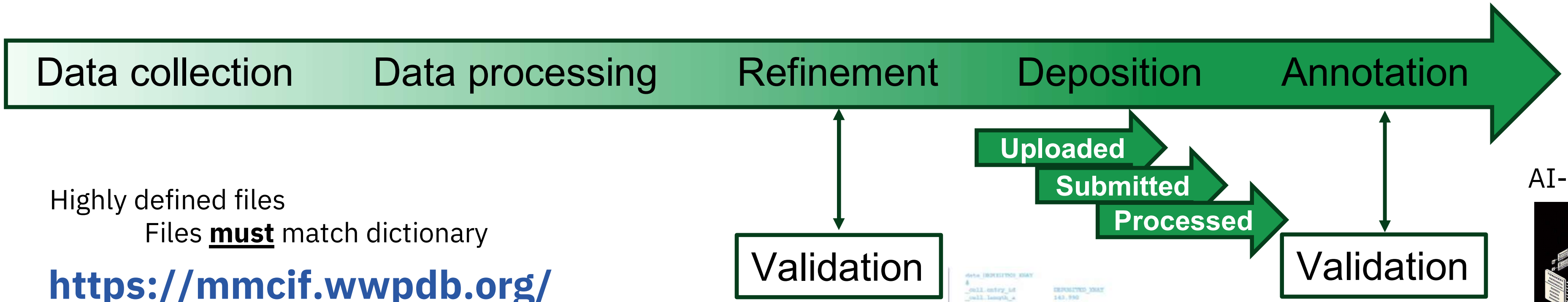
Validation

Validation

Currently:
Metadata stored in variety of forms (heterogeneity)

Goal: the file outputted by refinement software is as complete as possible to minimize the need to enter metadata manually in the deposition interface.

Metadata → Database


Data collection

Data processing

Refinement

Deposition

Annotation

Uploaded

Submitted

Processed

Validation

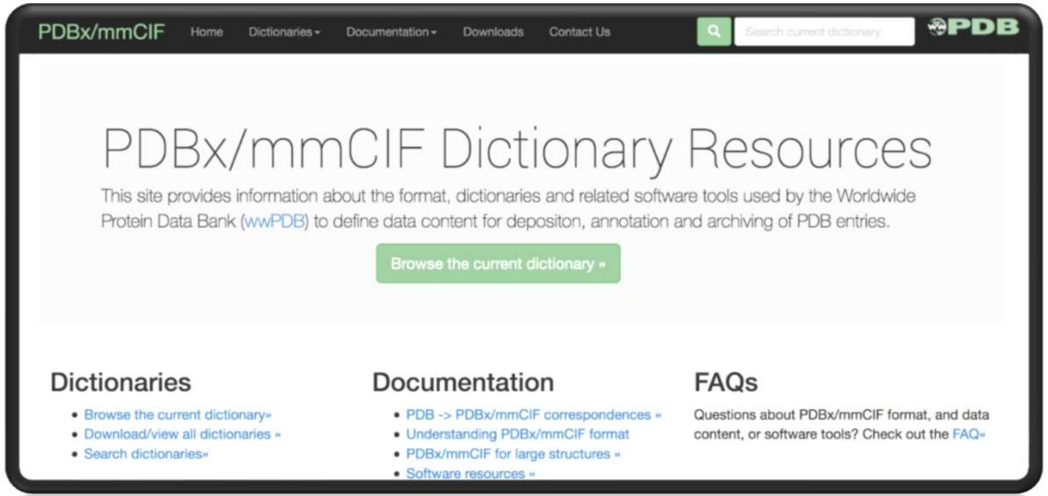
Validation

WORLDWIDE PDB PROTEIN DATA BANK

Highly defined files

Files **must** match dictionary

https://mmcif.wwpdb.org/



PDBx/mmCIF Dictionary Resources

This site provides information about the format, dictionaries and related software tools used by the Worldwide Protein Data Bank (wwPDB) to define data content for depositon, annotation and archiving of PDB entries.

Browse the current dictionary »

Dictionaries

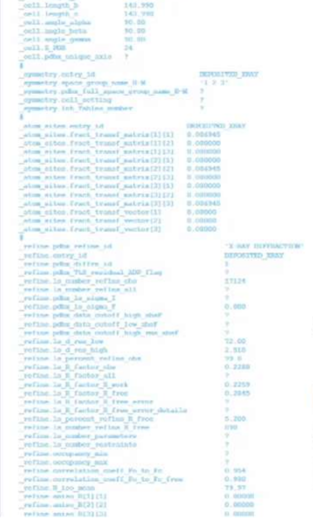
- Browse the current dictionary»
- Download/view all dictionaries »
- Search dictionaries»

Documentation

- PDB -> PDBx/mmCIF correspondences »
- Understanding PDBx/mmCIF format
- PDBx/mmCIF for large structures »
- Software resources »

FAQs

Questions about PDBx/mmCIF format, and data content, or software tools? Check out the FAQ»



```
data_DEPOSITED_XRAY
#
_cell.entry_id           DEPOSITED_XRAY
_cell.length_a           143.990
_cell.length_b           143.990
_cell.length_c           143.990
_cell.angle_alpha        90.00
_cell.angle_beta         90.00
_cell.angle_gamma        90.00
_cell.Z_PDB              24
_cell.pdbx_unique_axis   ?
#
_symmetry.entry_id                         DEPOSITED_XRAY
_symmetry.space_group_name_H-M             'I 2 3'
_symmetry.pdbx_full_space_group_name_H-M   ?
_symmetry.cell_setting                     ?
_symmetry.Int_Tables_number                ?
#
_atom_sites.entry_id                     DEPOSITED_XRAY
_atom_sites.fract_transf_matrix[1][1]    0.006945
_atom_sites.fract_transf_matrix[1][2]    0.000000
_atom_sites.fract_transf_matrix[1][3]    0.000000
_atom_sites.fract_transf_matrix[2][1]    0.000000
_atom_sites.fract_transf_matrix[2][2]    0.006945
_atom_sites.fract_transf_matrix[2][3]    0.000000
_atom_sites.fract_transf_matrix[3][1]    0.000000
_atom_sites.fract_transf_matrix[3][2]    0.000000
_atom_sites.fract_transf_matrix[3][3]    0.006945
_atom_sites.fract_transf_vector[1]       0.00000
_atom_sites.fract_transf_vector[2]       0.00000
_atom_sites.fract_transf_vector[3]       0.00000
#
_refine.pdbx_refine_id                           'X-RAY DIFFRACTION'
_refine.entry_id                                 DEPOSITED_XRAY
_refine.pdbx_diffrn_id                           1
_refine.pdbx_TLS_residual_ADP_flag               ?
_refine.ls_number_reflns_obs                     17124
_refine.ls_number_reflns_all                     ?
_refine.pdbx_ls_sigma_I                          ?
_refine.pdbx_ls_sigma_F                          0.000
_refine.pdbx_data_cutoff_high_absF               ?
_refine.pdbx_data_cutoff_low_absF                ?
_refine.pdbx_data_cutoff_high_rms_absF           ?
_refine.ls_d_res_low                             72.00
_refine.ls_d_res_high                            2.510
_refine.ls_percent_reflns_obs                    99.8
_refine.ls_R_factor_obs                          0.2288
_refine.ls_R_factor_all                          ?
_refine.ls_R_factor_R_work                       0.2259
_refine.ls_R_factor_R_free                       0.2845
_refine.ls_R_factor_R_free_error                 ?
_refine.ls_R_factor_R_free_error_details         ?
_refine.ls_percent_reflns_R_free                 5.200
_refine.ls_number_reflns_R_free                  890
_refine.ls_number_parameters                     ?
_refine.ls_number_restraints                     ?
_refine.occupancy_min                            ?
_refine.occupancy_max                            ?
_refine.correlation_coeff_Fo_to_Fc               0.954
_refine.correlation_coeff_Fo_to_Fc_free          0.930
_refine.B_iso_mean                               79.97
_refine.aniso_B[1][1]                            0.00000
_refine.aniso_B[2][2]                            0.00000
_refine.aniso_B[3][3]                            0.00000
```

Lines

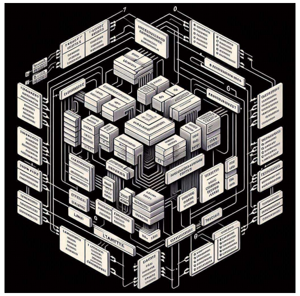
Uploaded: 2,729

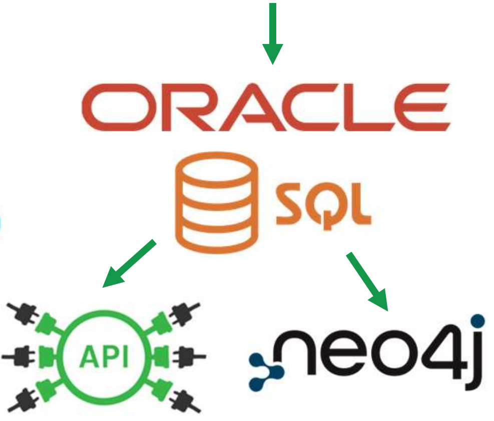
Submitted:

Processed:

mmCIF coordinate file
(small structure; 1 chain)

AI-generated image





endpoints

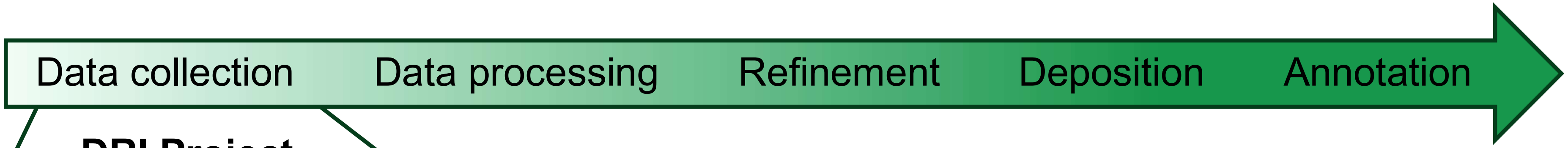
PDBe Protein Data Bank in Europe

PDBe-KB Protein Data Bank in Europe - Knowledge Base

AlphaFold Protein Structure Database

WORLDWIDE PDB PROTEIN DATA BANK

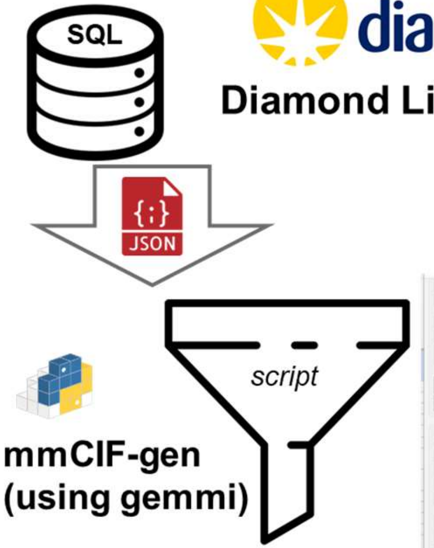
EMBL-EBI

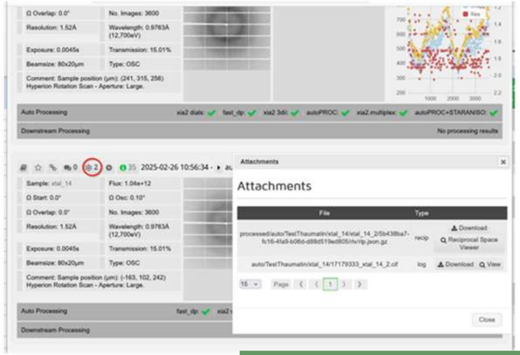


DRI Project

ISPyB/SynchWeb

Diamond Light Source





Ω Overlap: 0.0° | No. Images: 3600
Resolution: 1.52Å | Wavelength: 0.9763Å (12,700eV)
Exposure: 0.0045s | Transmission: 15.01%
Beamsize: 80x20µm | Type: OSC
Comment: Sample position (µm): (241, 315, 256) Hyperion Rotation Scan - Aperture: Large.

Auto Processing xia2 dials: ✓ fast_dp: ✓ xia2 3dii: ✓ autoPROC: ✓ xia2.multiplex: ✓ autoPROC+STARANISO: ✓

Downstream Processing No processing results

0 2 35 2025-02-26 10:56:34 -

Sample: xtal_14 | Flux: 1.04e+12
Ω Start: 0.0° | Ω Osc: 0.10°
Ω Overlap: 0.0° | No. Images: 3600
Resolution: 1.52Å | Wavelength: 0.9763Å (12,700eV)
Exposure: 0.0045s | Transmission: 15.01%
Beamsize: 80x20µm | Type: OSC
Comment: Sample position (µm): (-163, 102, 242) Hyperion Rotation Scan - Aperture: Large.

Auto Processing fast_dp: ✓

Downstream Processing

Attachments

Attachments

File	Type	
processed/auto/TestThaumatin/xtal_14/xtal_14_2/5b438ba7-fc16-4fa9-b06d-d88d519ed805/rlv/rlp.json.gz	recip	Download Reciprocal Space Viewer
auto/TestThaumatin/xtal_14/17179333_xtal_14_2.cif	log	Download View

- Experimental method (X-ray, NMR, EM, EC…)
- Data collection date
- Wavelength
- Synchrotron name
- Beamline name
- Detector type
- Detector name
- Serial crystallography experiment Y/N

- Temperature (not yet in SynchWeb / DLS ISPyB)
- Multiple datasets
- **Oscillation angle per image**
- **Total oscillation angle**
- **Beam transmission**

https://pypi.org/project/mmcif-gen/

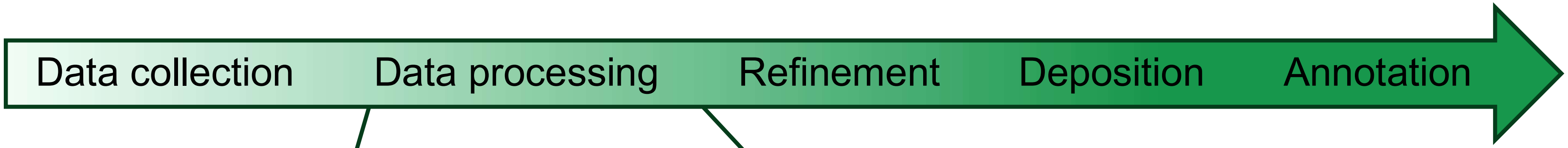
metadata
mmCIF file

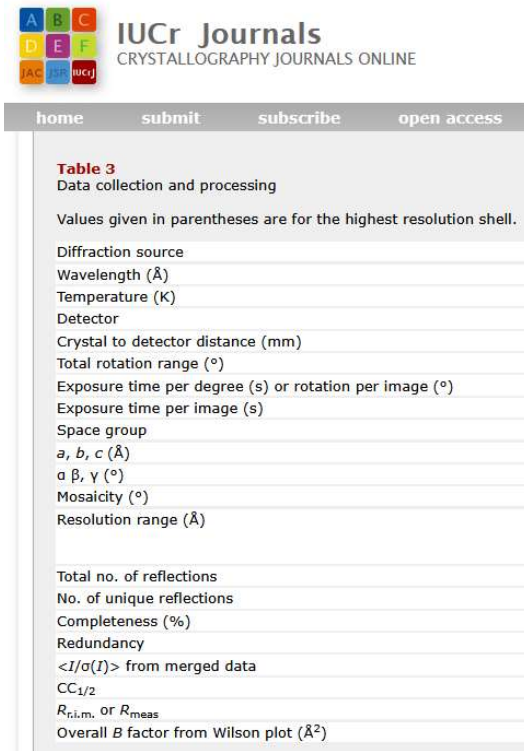
Global Phasing Limited

embedded
in **mtz file**

MTZ appendix - status (Global Phasing Ltd)

- ❑ The idea is to **append** (additional) **metadata** in some standard format to the end of an MTZ file:
  - This way metadata travel together with the most popular format for reflection data (MTZ), avoiding risks of de-synchronisation and loss of information along the path from experiment to deposition.
  - That appendix does not impact on existing MTZ-reading programs as far as we know (tested: CCP4, Phenix, GPhL, Gemmi and other programs/packages using the CCP4 library)
- ❑ Things to consider:
  - How to append?
  - Define start and end of appendix.
  - Define content of appendix.
  - Associate metadata with rest of MTZ file (i.e. reflection data and existing MTZ header).
  - How to extract?




IUCr Journals
CRYSTALLOGRAPHY JOURNALS ONLINE

home submit subscribe open access

Table 3
Data collection and processing

Values given in parentheses are for the highest resolution shell.

Diffraction source
Wavelength (Å)
Temperature (K)
Detector
Crystal to detector distance (mm)
Total rotation range (°)
Exposure time per degree (s) or rotation per image (°)
Exposure time per image (s)
Space group
a, *b*, *c* (Å)
α β, γ (°)
Mosaicity (°)
Resolution range (Å)

Total no. of reflections
No. of unique reflections
Completeness (%)
Redundancy
$\langle I/\sigma(I)\rangle$ from merged data
$CC_{1/2}$
$R_{\rm r.i.m.}$ or $R_{\rm meas}$
Overall *B* factor from Wilson plot (Å^2)

Example appendices extraction:

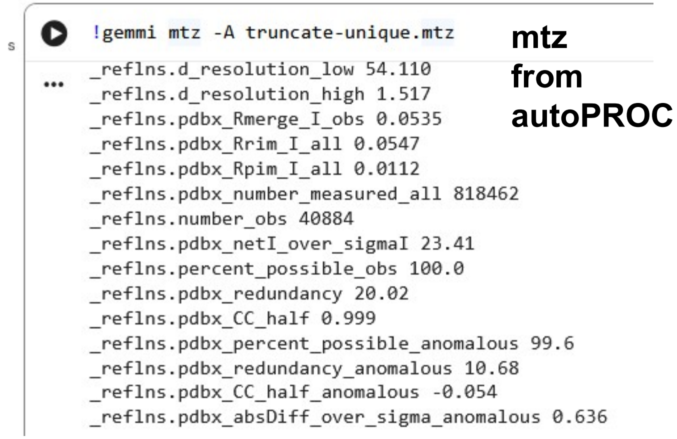
```
gemmi mtz -A truncate-unique.mtz
```

```
#MTZAPPENDIX-START\n
#MTZAPPENDIX-ITEM CIF DatasetID=1\n
…
#MTZAPPENDIX-END\n
```

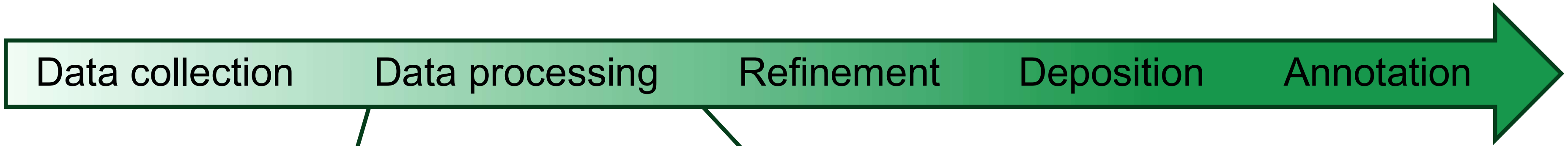
autoPROC tool:
aP_mtz_appendix –h

https://colab.research.google.com/drive/1L0FhSWeqqp50LBtNJBK7LOmEP0Vz4YnR?usp=sharing

https://github.com/glevans/PDB_Notebooks/tree/main/GemmiRecipes


```
!gemmi mtz -A truncate-unique.mtz
_reflns.d_resolution_low 54.110
_reflns.d_resolution_high 1.517
_reflns.pdbx_Rmerge_I_obs 0.0535
_reflns.pdbx_Rrim_I_all 0.0547
_reflns.pdbx_Rpim_I_all 0.0112
_reflns.pdbx_number_measured_all 818462
_reflns.number_obs 40884
_reflns.pdbx_netI_over_sigmaI 23.41
_reflns.percent_possible_obs 100.0
_reflns.pdbx_redundancy 20.02
_reflns.pdbx_CC_half 0.999
_reflns.pdbx_percent_possible_anomalous 99.6
_reflns.pdbx_redundancy_anomalous 10.68
_reflns.pdbx_CC_half_anomalous -0.054
_reflns.pdbx_absDiff_over_sigma_anomalous 0.636
```

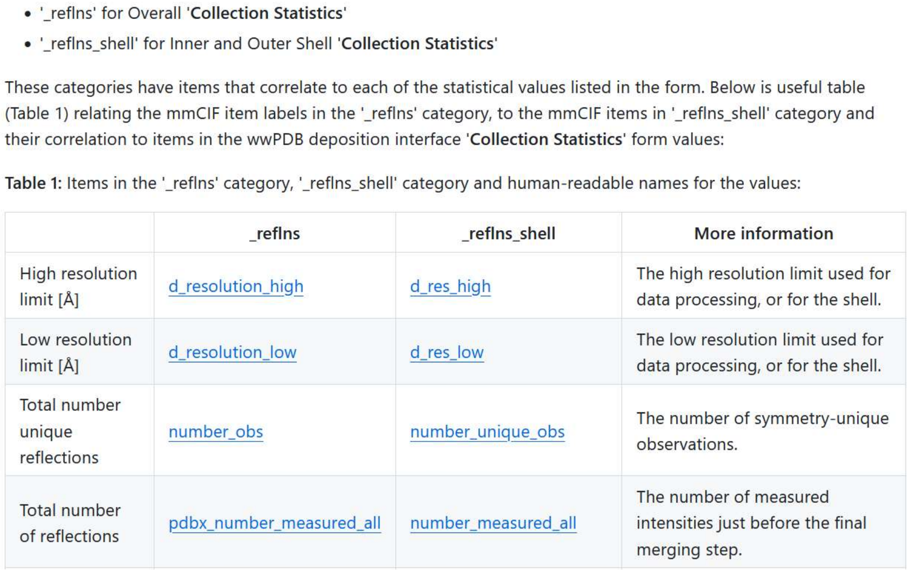
mtz from autoPROC



Working together towards metadata capture:

Metadata values & corresponding wwPDB capturing (mmCIF categories items)

https://github.com/glevans/automating_data_collection_stats (private)



- '_reflns' for Overall '**Collection Statistics**'
- '_reflns_shell' for Inner and Outer Shell '**Collection Statistics**'

These categories have items that correlate to each of the statistical values listed in the form. Below is useful table (Table 1) relating the mmCIF item labels in the '_reflns' category, to the mmCIF items in '_reflns_shell' category and their correlation to items in the wwPDB deposition interface '**Collection Statistics**' form values:

Table 1: Items in the '_reflns' category, '_reflns_shell' category and human-readable names for the values:

	_reflns	_reflns_shell	More information
High resolution limit [Å]	d_resolution_high	d_res_high	The high resolution limit used for data processing, or for the shell.
Low resolution limit [Å]	d_resolution_low	d_res_low	The low resolution limit used for data processing, or for the shell.
Total number unique reflections	number_obs	number_unique_obs	The number of symmetry-unique observations.
Total number of reflections	pdbx_number_measured_all	number_measured_all	The number of measured intensities just before the final merging step.

Interested?

Email: gle@ebi.ac.uk
with GitHub handle

Overall Wilson B factor is often NOT listed in logs (or mmCIF) with processing stats from autoprocessing pipelines.

PDBe Protein Data Bank in Europe PDBe-KB Protein Data Bank in Europe - Knowledge Base AlphaFold Protein Structure Database WORLDWIDE PDB PROTEIN DATA BANK EMBL-EBI

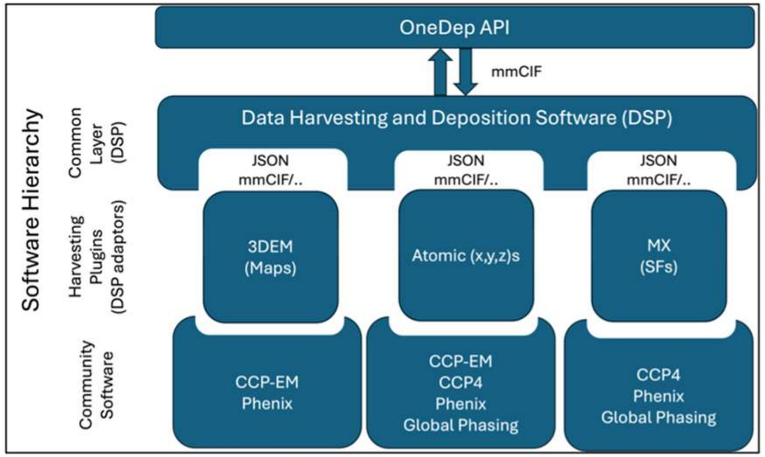
Biotechnology and Biological Sciences Research Council

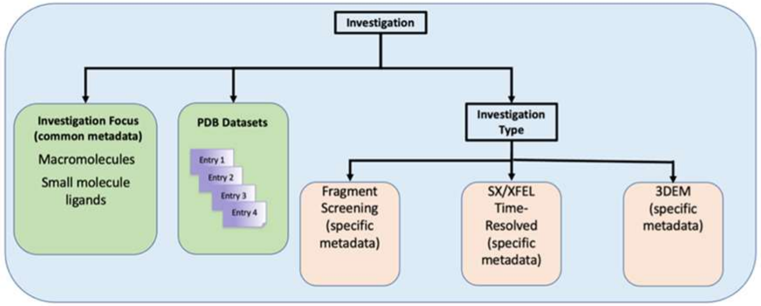
Data collection Data processing Refinement Deposition Annotation

Project awarded: Streamlining Data Deposition

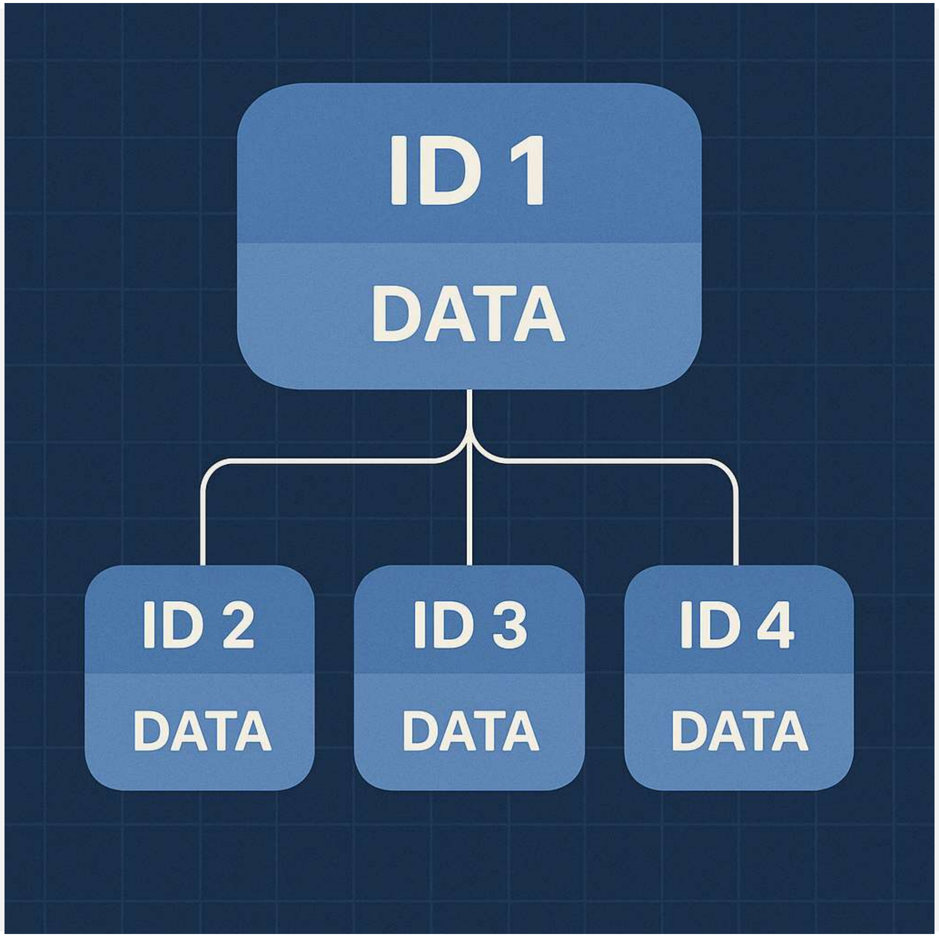
3 years project: November 2025 to October 2028

- Make Deposition API available for depositors
- Improve metadata copy based on previous depositions
- New OneDep APIs to:
  - Access ligand curation & validation
  - Add sequence annotation
  - Add assembly annotation
  - Access wwPDB validation
- Enable batch depositions and processing





Batch metadata → INVESTIGATION files



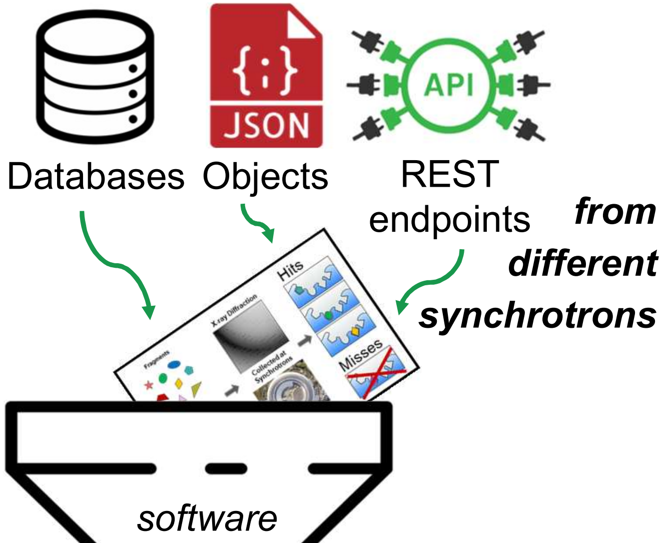
Concept:

- Umbrella file (ID 1) with defined data model to connect multiple entries (ID 2, ID 3 &ID 4).

Key features:

- **Value-adding, enriching metadata**
- Define types of umbrella files for different use case, *e.g.*:
  - Fragment screening
  - Time-resolved
  - EMDB specific cases – composite maps
- Only contains a subset of information found in a coordinate mmCIF file (not coordinates)
- Links to non-wwPDB databases

X-ray-based Fragment-based drug discovery (FBDD) data

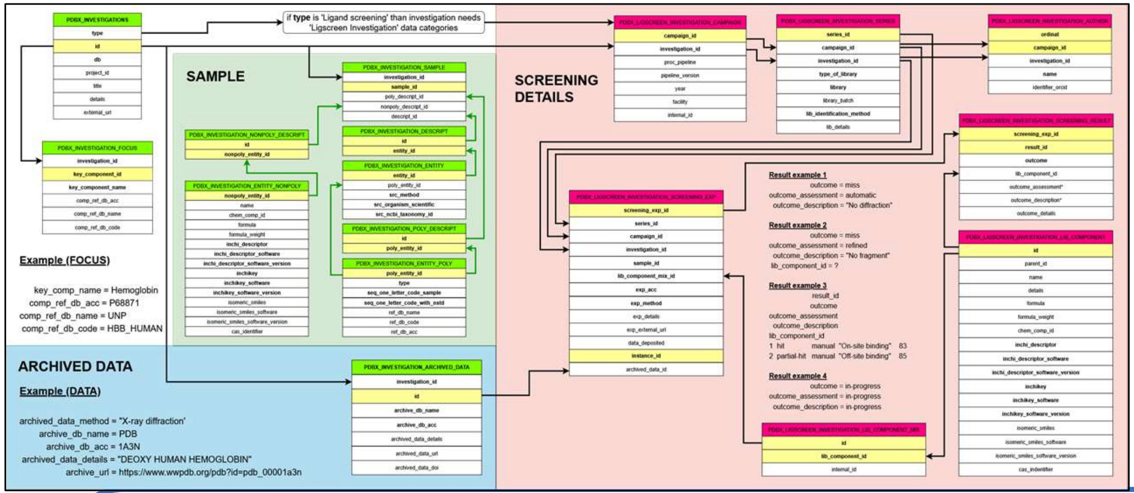


Investigation mmCIF Files

Software for generating umbrella FBDD mmCIF file
https://pypi.org/project/mmcif-gen/

- utilized for pilot repository
- ongoing development for synchrotron facilities

Standardizing data schema for FBDD data
https://github.com/PDBeurope/InvestigationCIF



PDBX_INVESTIGATIONS: type, id, db, project_id, title, details, external_url

if **type** is 'Ligand screening' than investigation needs 'Ligscreen Investigation' data categories

PDBX_INVESTIGATION_FOCUS: investigation_id, key_component_id, key_component_name, comp_ref_db_acc, comp_ref_db_name, comp_ref_db_code

Example (FOCUS)

key_comp_name = Hemoglobin
comp_ref_db_acc = P68871
comp_ref_db_name = UNP
comp_ref_db_code = HBB_HUMAN

SAMPLE

PDBX_INVESTIGATION_SAMPLE: investigation_id, sample_id, poly_descript_id, nonpoly_descript_id, descript_id

PDBX_INVESTIGATION_NONPOLY_DESCRIPT: id, nonpoly_entity_id

PDBX_INVESTIGATION_ENTITY_NONPOLY: nonpoly_entity_id, name, chem_comp_id, formula, formula_weight, inchi_descriptor, inchi_descriptor_software, inchi_descriptor_software_version, inchikey, inchikey_software, inchikey_software_version, isomeric_smiles, isomeric_smiles_software, isomeric_smiles_software_version, cas_identifier

PDBX_INVESTIGATION_DESCRIPT: id, entity_id

PDBX_INVESTIGATION_ENTITY: entity_id, poly_entity_id, src_method, src_organism_scientific, src_ncbi_taxonomy_id

PDBX_INVESTIGATION_POLY_DESCRIPT: id, poly_entity_id

PDBX_INVESTIGATION_ENTITY_POLY: poly_entity_id, type, seq_one_letter_code_sample, seq_one_letter_code_with_nstd, ref_db_name, ref_db_code, ref_db_acc

ARCHIVED DATA

Example (DATA)

archived_data_method = "X-ray diffraction'
archive_db_name = PDB
archive_db_acc = 1A3N
archived_data_details = "DEOXY HUMAN HEMOGLOBIN"
archive_url = https://www.wwpdb.org/pdb?id=pdb_00001a3n

PDBX_INVESTIGATION_ARCHIVED_DATA: investigation_id, id, archive_db_name, archive_db_acc, archived_data_details, archived_data_url, archived_data_doi

SCREENING DETAILS

PDBX_LIGSCREEN_INVESTIGATION_CAMPAIGN: campaign_id, investigation_id, proc_pipeline, pipeline_version, year, facility, internal_id

PDBX_LIGSCREEN_INVESTIGATION_SERIES: series_id, campaign_id, investigation_id, type_of_library, library, library_batch, lib_identification_method, lib_details

PDBX_LIGSCREEN_INVESTIGATION_AUTHOR: ordinal, campaign_id, investigation_id, name, identifier_orcid

PDBX_LIGSCREEN_INVESTIGATION_SCREENING_EXP: screening_exp_id, series_id, campaign_id, investigation_id, sample_id, lib_component_mix_id, exp_acc, exp_method, exp_details, exp_external_url, data_deposited, instance_id, archived_data_id

PDBX_LIGSCREEN_INVESTIGATION_SCREENING_RESULT: screening_exp_id, result_id, outcome, lib_component_id, outcome_assessment*, outcome_description*, outcome_details

PDBX_LIGSCREEN_INVESTIGATION_LIB_COMPONENT: id, parent_id, name, details, formula, formula_weight, chem_comp_id, inchi_descriptor, inchi_descriptor_software, inchi_descriptor_software_version, inchikey, inchikey_software, inchikey_software_version, isomeric_smiles, isomeric_smiles_software, isomeric_smiles_software_version, cas_indentifier

PDBX_LIGSCREEN_INVESTIGATION_LIB_COMPONENT_MIX: id, lib_component_id, internal_id

Result example 1

outcome = miss
outcome_assessment = automatic
outcome_description = "No diffraction"

Result example 2

outcome = miss
outcome_assessment = refined
outcome_description = "No fragment"
lib_component_id = ?

Result example 3

result_id	outcome	outcome_assessment	outcome_description	lib_component_id
1	hit	manual	"On-site binding"	83
2	partial-hit	manual	"Off-site binding"	85

Result example 4

outcome = in-progress
outcome_assessment = in-progress
outcome_description = in-progress

Synchrotrons and associated facilities involved in developing this data model:

- ○ **The Crystallisation Facility** at the **European Molecular Biology Laboratory (EMBL) Grenoble** and **European Synchrotron Radiation Facility (ESFR)** in France
- ○ **XChem: Diamond Fragment Screening** at **Diamond Light Source (DLS)** in the United Kingdom
- ○ **Fragment Screening Facility** at **Berlin synchrotron BESSY-MX and Helmholtz-Zentrum Berlin/HZB** in Germany
- ○ **FragMAX** at Swedish synchrotron **MAX IV** in Sweden

Acknowledgements

Resources

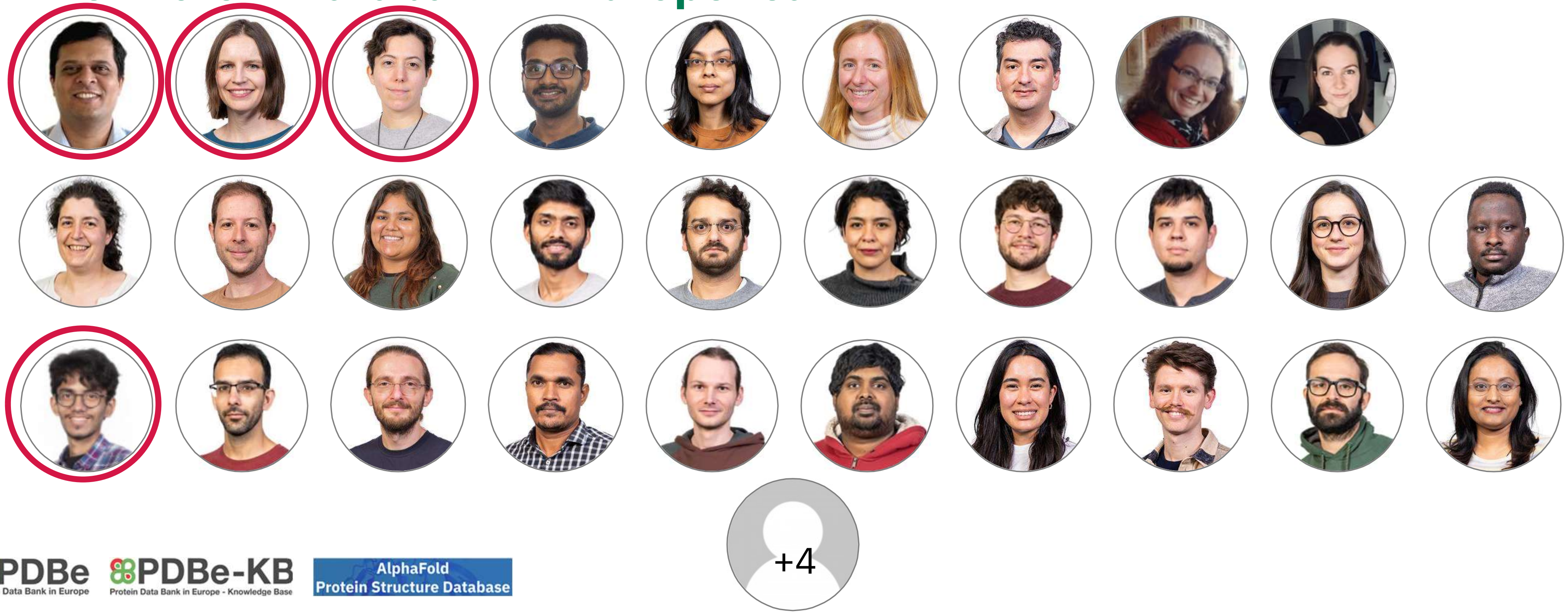
PDBe
Protein Data Bank in Europe

PDBe-KB
Protein Data Bank in Europe - Knowledge Base

AlphaFold
Protein Structure Database

in PDBe
𝕏 @PDBeurope
🦋 @PDBeurope
f proteindatabank
▶ proteindatabank
◎ pdbeurope

The Protein Databank in Europe team



Acknowledgements

Funding

EMBL

UK Research and Innovation

wellcome

Funded by the European Union

Google DeepMind

Collaborators

WORLDWIDE PDB PROTEIN DATA BANK

PDBe-KB Protein Data Bank in Europe - Knowledge Base

3D-Beacons

CEITEC Central European Institute of Technology BRNO | CZECH REPUBLIC

3D-BioInfo

Google DeepMind

Thank you, feedback welcome!

DLS

- I03 Beamline
  - Dave Hall
  - Mark Williams
  - Neil Paterson
- ISPYB
  - Irakli Sikharulidze
- XChem / I04-1 Beamline
  - Frank von Delft
  - Daren Fearon
  - Warren Thompson
  - Jasmin Aschenbrenner

diamond

Global Phasing Ltd

- Software developers
  - Gerard Bricogne
  - Clemens Vonrhein
  - Marcin Wojdyr

GΦL
Global Phasing Limited

MaxIV

- FragMax
  - Tobias Krojer

MAX IV

FragMAX
MAX IV Laboratory

BESSY

- MX Team
  - Manfred Weiss

HZB Helmholtz Zentrum Berlin

CCP4

- Software developers
  - Eugene Krissinel

CCP4

EMBL Grenoble

ESRF
The European Synchrotron

- CRIMS at ESRF (Crystallographic Information Management System / CRIMS)
  - José A. Marquez
  - Raphaël Bourgeas

PDBe
Protein Data Bank in Europe

pdbhelp@ebi.ac.uk

PDBe

EMBL-EBI