Do we treat Macromolecular Crystallography data FAIR?

L.M.J. Kroon-Batenburg
Utrecht University
The Netherlands

Universiteit Utrecht

Day-0 Workshop on MX raw image data formats, metadata and validation, 22-08-2020

Data publishing and management workflow

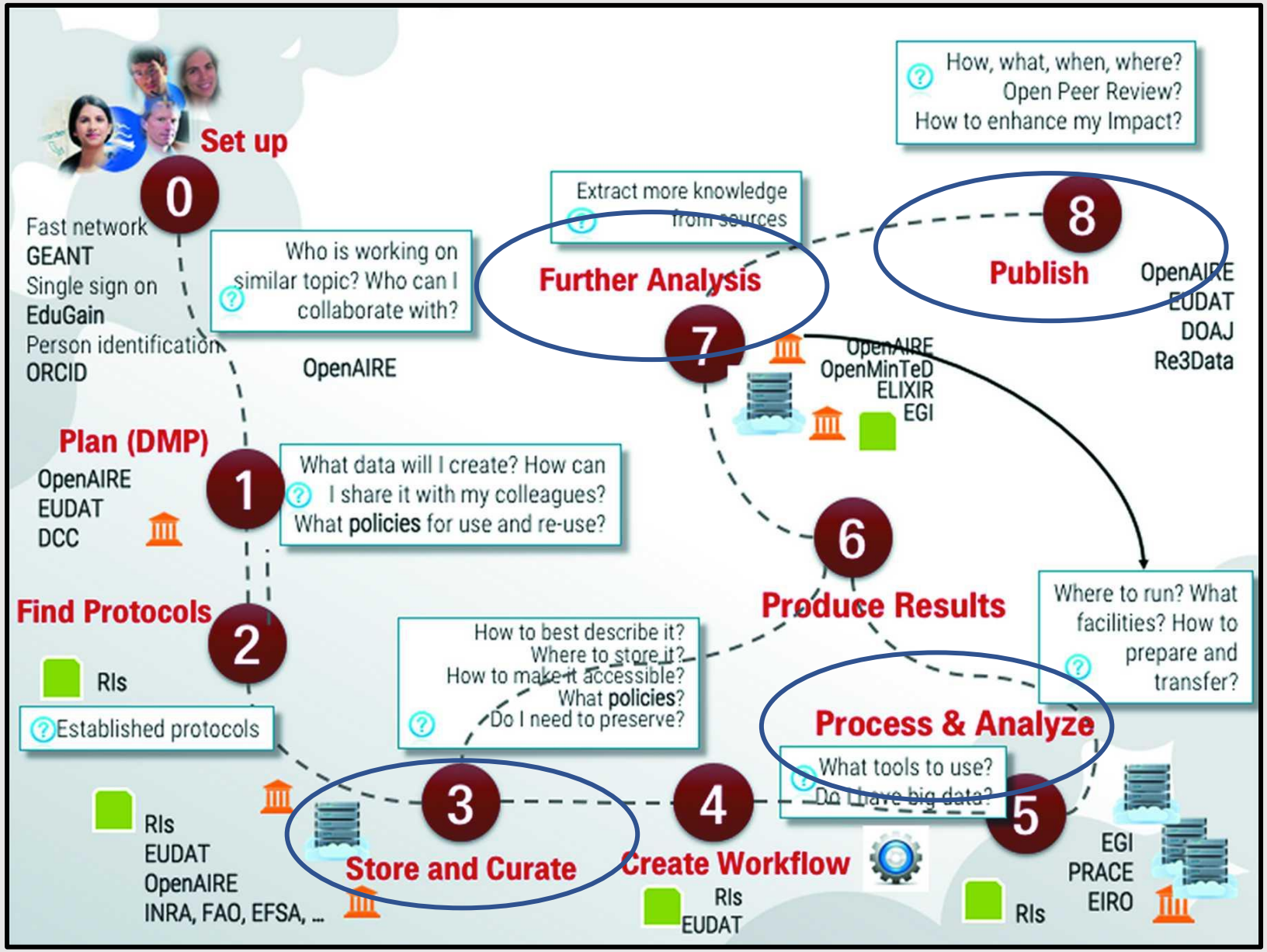


Illustration courtesy of Natalia Manova for the European OpenAIRE project

DDDWG recommendations

IUCr DDDWG Recommendations (top two)

•Authors should provide a permanent and prominent link from their article to the raw data sets which underpin their journal publication and associated database deposition of processed diffraction data (*e.g.* structure factor amplitudes and intensities) and coordinates, and which should obey the 'FAIR' principles, that their raw diffraction data sets should be Findable, Accessible, Interoperable and Re-usable (https://www.force11.org/group/fairgroup/fairprinciples).

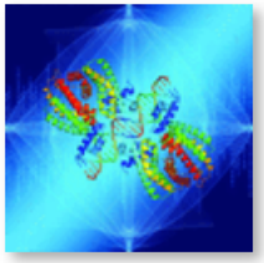
•A registered Digital Object Identifier (doi) should be the persistent identifier of choice (rather than a Uniform Resource Locator, url) as the most sustainable way to identify and locate a raw diffraction data set.

IUCr journals

D EDITORIAL

Acta Cryst. (2019). D**75**, 455-457
https://doi.org/10.1107/S2059798319004844

Cited by 1



Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography

J. R. Helliwell, W. Minor, M. S. Weiss, E. F. Garman, R. J. Read, J. Newman, M. J. van Raaij, J. Hajdu and E. N. Baker

The policy of IUCr Journals on diffraction data is defined.

Keywords: **FAIR**; **diffraction data**; **IUCr policy**.

Read article | Similar articles

"IUCr Journals are now taking the lead by encouraging authors to provide a doi for their deposited original raw diffraction data when they submit an article describing a new structure or a new method tested on unpublished diffraction data. "

FAIR

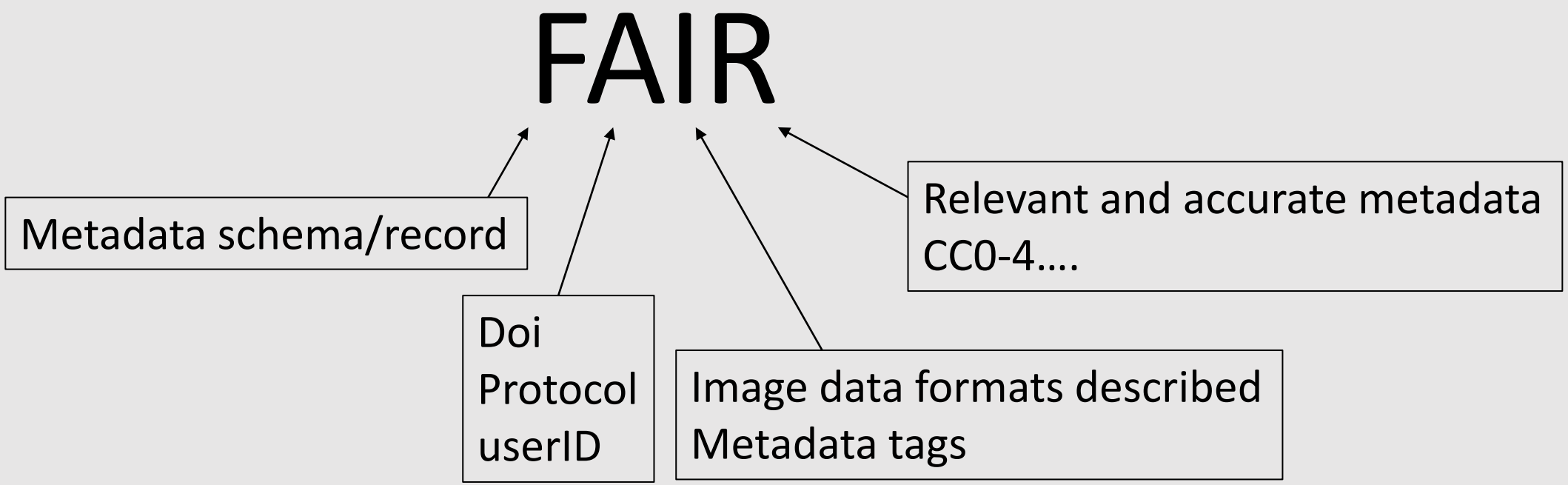
- **Findable:** easy to identify and find for both humans and computers, with metadata that **facilitate searching** for specific datasets,

- **Accessible:** stored for long term so that they can **easily be accessed** and/or downloaded with well-defined **access conditions**, whether at the level of metadata, or at the level of the actual data,

- **Interoperable:** ready to be combined with other datasets by humans or computers, without ambiguities in the **meanings of terms and values**,

- **Reusable:** ready to be used for future research and to be further processed using computational methods. This requires **adequate information about how the data** were obtained and processed (provenance) and an **appropriate license**

FAIR for raw data in MX



Raw data in MX

- What are the possibilities of raw data archiving?
- Can we adhere to the FAIR principles?
- Do we reuse the data?

Findable and accessible

Data:
- OpenAire
- DataCite

Repositories of databases:
- Re3data.org
- Fairsharing.org

Discipline specific repositories:
- SBGrid
- IRRMC
- CXI

General repositories:
- Zenodo
- Figshare
- Dryad
- Research gate
- ArXiv.org
- Mendeley

Universities, National, EUDAT

Synchrotron, Neutron Facilities and XFEL:
- ESRF
- DLS
- STFC ISIS
- Store. Synchrotron
- XFELs
- SciCat (ESS)
- ILL portal

Data Policies

DataCite
FIND, ACCESS, AND REUSE DATA

provides persistent identifiers (DOIs) for research data and other research outputs

DataCite: “x-ray diffraction” 19488 works

➔ **Raw images data, powder data, processed data or papers**

Raw data mostly:

- SBGrid
- IRRMC
- Zenodo
- CXI
- Ceon RepOD

- Figshare
- Dryad
- Mendeley
- DataShare Edinburgh
- Universities of Manchester, Leeds, Bath, Aberdeen, Cambridge, Strathclyde, Bristol, Cardiff, Utah
- Geological data

DataCite
FIND, ACCESS, AND REUSE DATA

DataCite Search

Works People Repositories Members Support Sign in

diffraction AND "SARS-CoV-2" Search

184 Works

X-Ray Diffraction data from SARS-CoV-2 Nucleocapsid N2b domain, source of 6WZO structure

Qiaozhen Ye
X Ray Diffraction published via SBGrid Data Bank
Native dataset for SARS-CoV-2 Nucleocapsid (N) dimerization domain, P1 form

No citations were reported. No usage information was reported.

https://doi.org/10.15785/sbgrid/785 Cite

Registration Year

- [ ] 2020 184

Resource Types

- [x] Dataset 184

Findable?

DataCite
FIND, ACCESS, AND REUSE DATA

diffraction AND "SARS-CoV-2" NOT zenodo NOT SBGrid → 24

NIH National Institutes of Health
Office of the Director
Data Science at NIH

Integrated Resource for Reproducibility in Macromolecular Crystallography

SARS-CoV-2 → 29
COVID-19 → 13

Metadata!

zenodo

→ “X-ray diffraction”: 1462 + Dataset: 117 Mostly Macromolecular crystallography raw data

→ Community: macromolecules AND diffraction: 151

→ Diffraction AND Covid-19: 78 entries
→ Diffraction AND “SARS-Cov-2”: 79 entries
→ Diffraction AND protein: 79 entries

different

Metadata:
protein sample
reference to pdb

Metadata:
Description of experiment
image headers or Nexus/HDF5

March 30, 2020

Dataset Open Access

Raw diffraction data for structure of SARS-CoV-2 main protease with Z2737076969 (ID: mpro-x0350 / PDB: 5RE8)

Aragao, David; Brandao-Neto, Jose; Carbery, Anna; Crawshaw, Adam; Dias, Alexandre; Douangamath, Alice; Dunnett, Louise; Fearon, Daren; Flaig, Ralf; Gehrtz, Paul; Hall, Dave; Krojer, Tobias; London, Nir; Lukacik, Petra; Mazzorana, Marco; McAuley, Katherine; Owen, David; Powell, Ailsa; Reddi, Rambabu; Resnick, Efrat; Skyner, Rachael; Snee, Matt; Strain-Damerell, Claire; Stuart, Dave; von Delft, Frank; Walsh, Martin; Wild, Conor; Williams, Mark; Winter, Graeme

Raw diffraction data for mpro-x0350 / PDB ID 5RE8 (see: https://www.ebi.ac.uk/pdbe/entry/pdb/5RE8) - SARS-CoV-2 main protease in complex with Z2737076969 (SMILES:FC=1C=CC=C(CNCC2=CC=CO2)C1) collected as part of an XChem crystallographic fragment screening campaign on beamline i04-1 at Diamond Light Source. The deposited structure was automatically processed with standard Diamond tools and PanDDA, however the raw data are being made available to allow reanalysis by any interested party. For more details see: https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html

Preview

mpro-x0350.zip

Mpro-x0350.run	4 Bytes
Mpro-x0350_1_0001.cbf	6.2 MB
Mpro-x0350_1_0002.cbf	6.2 MB
Mpro-x0350_1_0003.cbf	6.2 MB

161 views

3 downloads

See more details...

Indexed in

OpenAIRE

Publication date:
March 30, 2020

DOI:
DOI 10.5281/zenodo.3730547

Keyword(s):
COVID-19 | SARS-CoV-2 main protease | automated upload | PDB:5RE8 | Diamond Light Source / MX / XChem

Communities:

SBGrid Data Bank

SBGrid Data Bank

Publication Guidelines

Data ▾ About ▾ Get Help ▾ For Depositors ▾

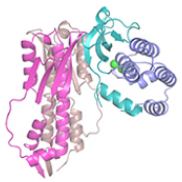
10.15785/SBGRID/ Data ID Go

PDB Code Go

Lab, Contributor, Biological Sample, Collection Facility, or PDB

X-Ray Diffraction data from LapD output domain in complex with LapG, source of 4U65 structure

610 datasets



VISUALIZE IN 3D >>

Data DOI: 10.15785/SBGRID/94 | ID: 94
Publication DOI: 10.7554/eLife.03650
4U65 Coordinates: Viewer, PDB (RCSB) (PDBe), MMDB
Sondermann Laboratory, Cornell University
Release Date: May 19, 2015

■ Reprocessing Instructions

beam center x=99.3 , y=100.0 indexed in P21 with HKL2000 DIstance is 217 (not 220 as indicated in the header)

■ Reprocessing Data

About Reprocessing

xia2 -2d | xia2 -3d | xia2 -3dii | xia2 -dials

Sufficient/Valid Metadata?

Mosflm, XDS, Dials via xia2

National Institutes of Health
Office of the Director
Data Science at NIH

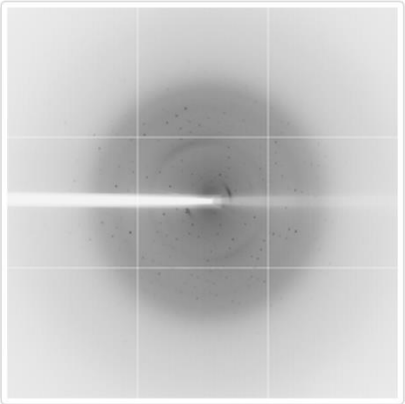
Integrated Resource for Reproducibility in Macromolecular Crystallography

Search examples
Find data related to a disease: COVID-19
Find a specific PDB ID: 4K6A
Free format search: potential drug target
Combining searches: drug AND cholera
Specific beamline: beamline = 21-ID-G
Resolution limit (Angstroms): resolution <1.25
Search by tag: workshop

Currently indexed projects: **5477**

Currently indexed datasets: **8839**

Diffraction project datasets IDP01325_3lus



Method: Molecular Replacement
Resolution: 1.96 Å
Space group: P 21 21 21

Download all images (1.1 GB)

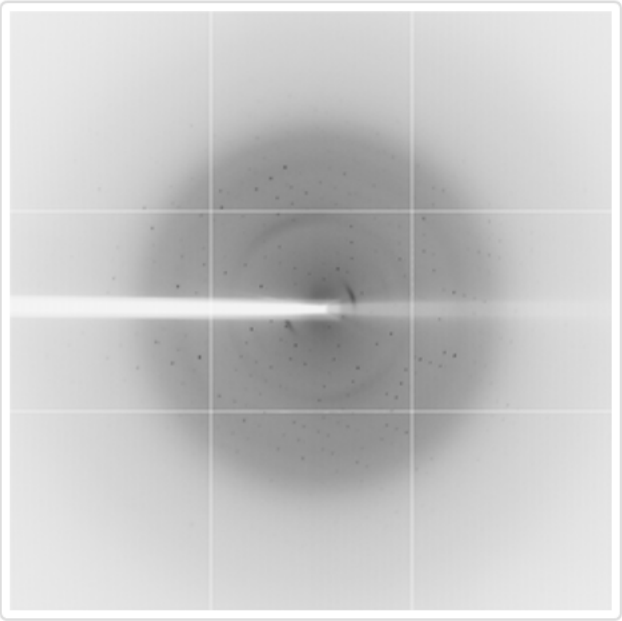
PDB website for 3LUS

doi:10.18430/M33LUS

Project details

Title	Crystal structure of a putative organic hydroperoxide resistance protein with molecule of captopril bound in one of the active sites from Vibrio cholerae O1 biovar eltor str. N16961
Authors	Nocek, B., Maltseva, N., Makowska-Grzyska, N., Kwon, K., Anderson, W., Joachimiak, A., NIAID
R / R_{free}	0.17 / 0.23
Unit cell edges [Å]	38.20 x 76.20 x 79.40
Unit cell angles [°]	90.0, 90.0, 90.0

Dataset 1325-capto-x1.####.img details



Number of frames	180 (1 - 180)
Distance [mm]	292.1
Oscillation width [°]	1.00
Omega [°]	-120.0
Wavelength [Å]	0.97929
Experiment Date	2009-11-21
Equipment	19-ID at APS (Advanced Photon Source)

Raw data link in PDBe

EMBL-EBI | Services | Research | Training | About us

Protein Data Bank in Europe
Bringing Structure to Biology

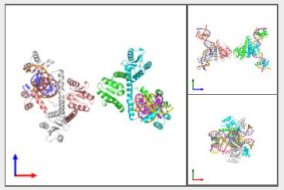
Search
Examples: hemoglobin, BRCA1_HUMAN
Advanced search
Feedback

PDBe › 5cy2

X-ray diffraction
4Å resolution

Tn3 resolvase - site III complex crystal form II
Source organism: *Escherichia coli*
Entry authors: Montano PS, Rice PA

Released: 11 Jan 2017
Model geometry
Fit model/data



Quick links

- 5cy2 overview
- Citations
- Structure analysis
- Function and Biology
- Ligands and Environments
- Experiments and Validation

Function and Biology

Details

Biochemical function: DNA binding
Biological process: DNA recombination
Cellular component: not assigned

Sequence domains:
- Resolvase, N-terminal catalytic domain
- Recombinase, conserved site
- Resolvase-like, N-terminal catalytic domain superfamily
- Homeobox-like domain superfamily
- Resolvase, HTH domain

Ligands and Environments

No bound ligands

Experiments and Validation

Details

X-ray source: APS BEAMLINE 19-ID
Spacegroup: C2
Unit cell: a: 144.9Å b: 151.92Å c: 106.14Å α: 90° β: 99.85° γ: 90°
R-values: R 0.209 | R_{work} 0.207 | R_{free} 0.25

Expression systems:
- *Escherichia coli*
- Not provided

Structure analysis

Details

Assembly composition: hetero tetramer (preferred)
Entry contents: 1 distinct polypeptide molecule
2 distinct DNA molecules

Macromolecules (3 distinct):
Transposon Tn3 resolvase

Molecule details ›
Chains: A, B, E, F
Length: 192 amino acids
Theoretical weight: 21.55 KDa
Source organism: *Escherichia coli*
Expression system: *Escherichia coli*
UniProt:

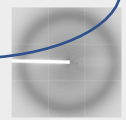
Archive	Experimental Data Available
Biological Magnetic Resonance Bank (BMRB)	Nuclear magnetic resonance (NMR) spectroscopic data
SBGrid Databank	X-ray diffraction image data
Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC)	X-ray diffraction image data
Electron Microscopy Public Image Archive (EMPIAR)	Raw electron microscopy image datasets

Experimental raw data

Links to raw experimental data available for this entry are listed below

Diffraction data related to PDB entry 5cy2 found at the SBGrid Data Bank

Data DOI:
10.15785/SBGRID/683
Total size: 3.2Gb



John Berrisford:
out of the 9665 X-ray entries that were released in 2019 we have DOI's for raw images in 205 of these entries.

Reuse of processed data

PDBe-KB COVID-19 Data Portal

Protein Data Bank in Europe - Knowledge Base

PDBe-KB COVID-19 Data Portal

An unprecedented number of scientific efforts are taking place worldwide in order to help combat the new coronavirus epidemic (COVID-19). One of the biggest challenges in this fast-moving situation is to share data and

P0DTC2 - Spike glycoprotein (S glycoprotein)

Spike protein S1 attaches the virion to the cell membrane by interacting with host receptor, initiating the infection. Binding to human ACE2 and CLEC4M/DC-SIGNR receptors and internalization of the virus into the endosomes of the host cell induces conformational changes in the S glycoprotein. Proteolysis by cathepsin CTSL may unmask the fusion peptide of S2 and activate membranes fusion within endosomes. Spike protein S2 mediates fusion of the virion and cellular membranes by acting as a class I viral fusion protein. Under the current model, the protein has at least three conformational states: pre-fusion native state, pre-hairpin intermediate state, and post-fusion hairpin state. During viral and target cell membrane fusion, the coiled coil regions

74 Structures

19 Ligands

7 Interactions

0 Functional Annotations

17 Similar Proteins

PDBe COVID-19 tweets

Curated Tweets by @PDBeurope

A collection of tweets relating to PDBe and the COVID-19 pandemic.

Protein Data Bank
@PDBeurope
This piece @Structure_CP from @annotated_sci looks at how the structural biology community have responded during the #COVID19 pandemic. This includes how the wwPDB (including @PDBeurope) & @EMDB_EMPIAR have increased efforts to

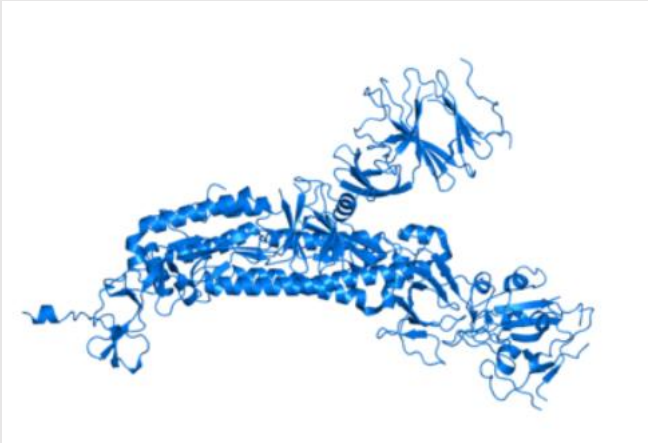
PDBe-KB › Spike glycoprotein

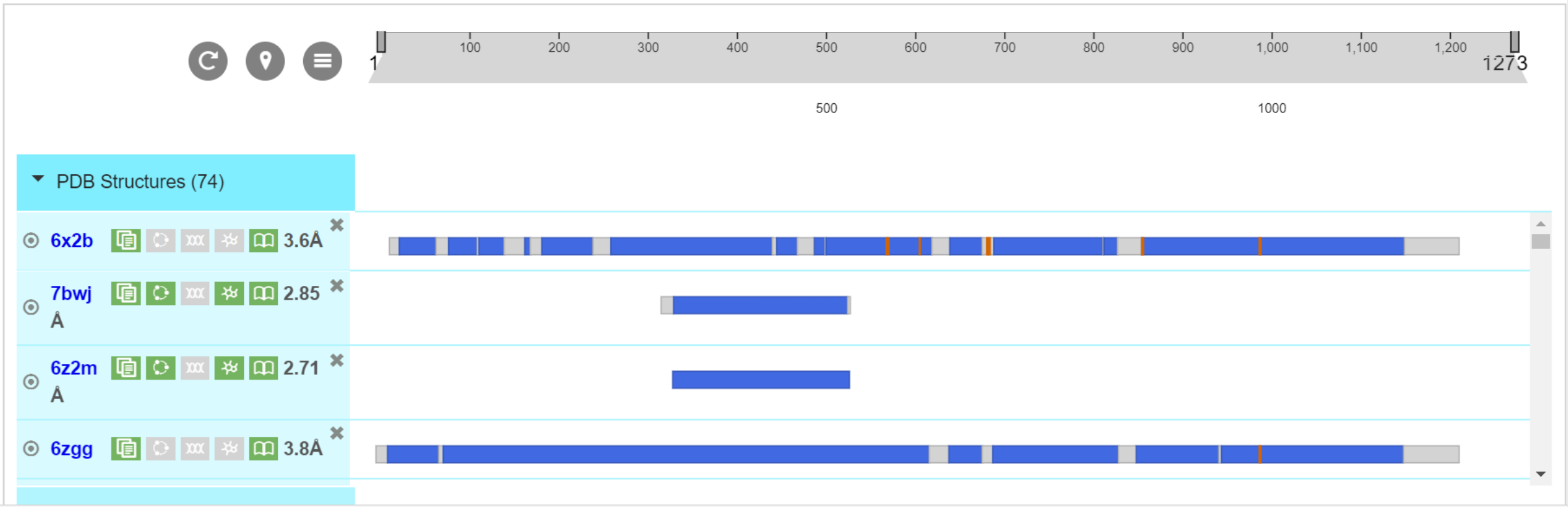
Gene: S

Organism: *Severe acute respiratory syndrome coronavirus 2 (2019-nCoV)*

Uniprot: P0DTC2 [go to UniProt]

Biological function: May down-regulate host tetherin (BST2) by lysosomal degradation, thereby counteracting its antiviral activity [go to UniProt]





Reuse of processed data

The FEBS Journal

FEBS PRESS science publishing by scientists

Ligand-centered assessment of SARS-CoV-2 drug target models in the Protein Data Bank

Alexander Wlodawer[1], Zbigniew Dauter[2], Ivan G. Shabalin[3], Miroslaw Gilski[4,5], Dariusz Brzezinski[3,5,6], Marcin Kowiel[5], Wladek Minor[3], Bernhard Rupp[7,8] and Mariusz Jaskolski[4,5]

https://covid-19.bioreproducibility.org/ 347

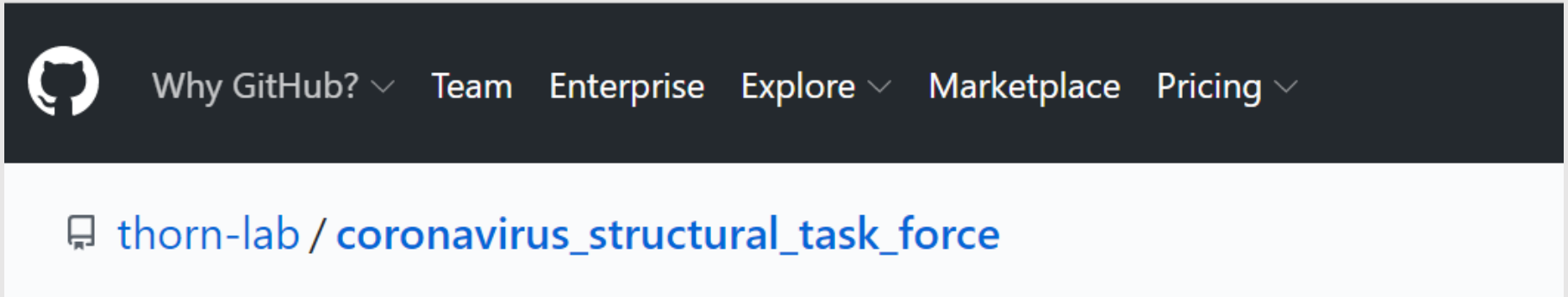
Search: (matched 232 out of 347 total records) Copy Excel CSV PDF Print

PDB	Resol.	Released	Title	Method	P_{Q1}	Issues		Re-refined?	Raw data	Ref.
6VYO	1.70 Å	2020-03-11	*Crystal structure of R...*	X-ray	78.6	-		Yes		-
6W9C	2.70 Å	2020-04-01	*The crystal structure ...*	X-ray	21.8	moderate	No	Yes		-
6W75	1.95 Å	2020-03-25	*1.95 Angstrom Resol...*	X-ray	61.6	-		No		-
6W61	2.00 Å	2020-03-25	*Crystal Structure of t...*	X-ray	61.8	-		No		-
6W4H	1.80 Å	2020-03-18	*1.80 Angstrom Resol...*	X-ray	63.8	-		No		-
6W4B	2.95 Å	2020-03-18	*The crystal structure ...*	X-ray	7.4	-		No		-
6W02	1.50 Å	2020-03-11	*Crystal Structure of A...*	X-ray	87.4	-		No		-

Whenever necessary and possible, we intend to start our re-analysis from reprocessing of the original diffraction images [28]. However, quite often we were unable to obtain the diffraction data despite the IUCr recommendation [29] and an earnest appeal from the community to make diffraction data related to CoV-2 public

Reuse of processed data

Andrea Thorn



Re-refinement: Isolde
Validation: Auspex, Molprobity
Links to: PDB-Redo and Buster(Global Phasing)

→ 83 links to raw diffraction data:
74 Zenodo: main protease (PANDDA project)
9 IRRMC
(2 in SBGrid are not mentioned)

Reuse of raw data

GΦL

Clemens Vonrhein & Gérard Bricogne

https://www.globalphasing.com/autoproc/wiki/index.cgi

(Re)processing available raw image data - Take-1

20200421	Looking at SARS-CoV-2 papain-like protease
	Looking at SARS-CoV-2 NSP16/NSP10 structures
	Looking at SARS-CoV-2 NSP3 structures
	Looking at SARS-CoV-2 NSP15 Endoribonuclease
	Looking at SARS-CoV-2 Nsp9 RNA binding protein

In total 10 data sets (from proteindiffraction.org) are being re-processsed

Re-refinement of the processed data is ongoing

Interoperable

Image data formats:
Mar345, MarCCD, ADSC, Raxis, Oxford, CMOS RDI, Pilatus (imgCIF/cbf), Eiger (HDF5) ….CSPAD,AGIPD...

Software packages can deal with most image formats:
HKL3000/XDS/d*Trek/Mosflm/Dials/EVAL

Vocabulary: metadata tags:
Plethora of Ascii key-words , imgCIF, Nexus

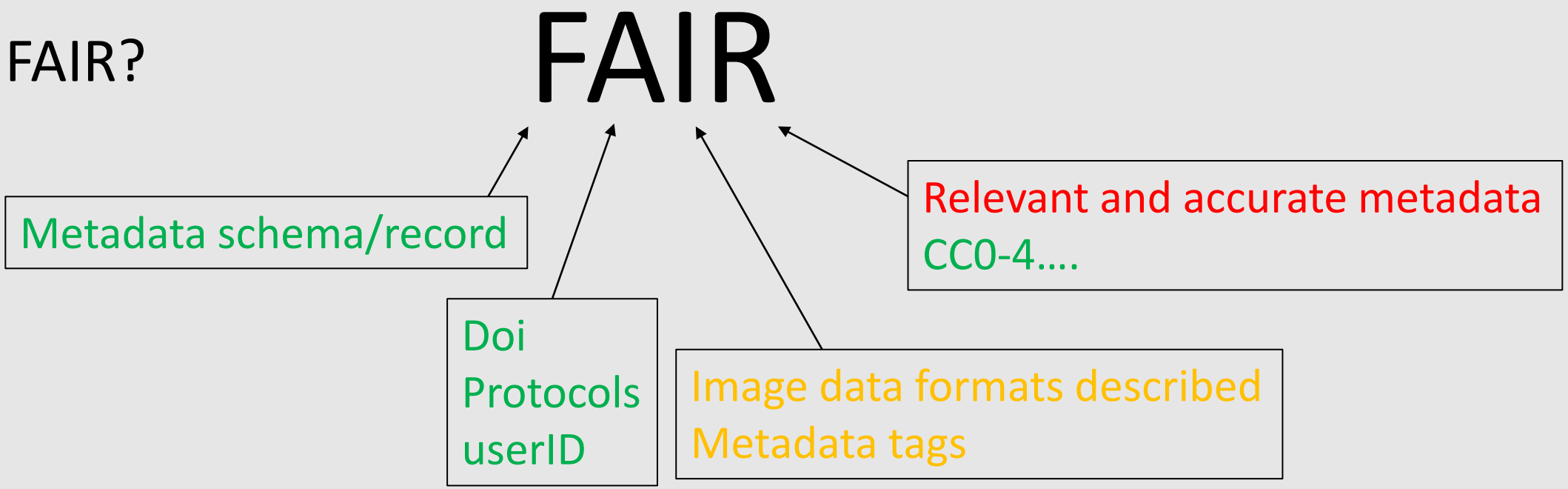
Raw data re-use

Reasons for reprocessing:

- Multiple lattices: % overlap (if we can go to $CC_{1/2}$ 0.14 this should matter)
- TDS/background (not solved in integration; also streaks not accounted for)
- Resolution cut-off
- Anisotropic data
- Unsolved structure
- Diffuse scattering (packing disorder or internal mobility)
- Incommensurate modulation

Conclusions

- FAIR?



- IUCr: imgCIF dictionary
- HDRMX NXmx Gold Standard using Nexus/HDF5