Raw data opportunities for biological crystallography publishing

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

Universiteit Utrecht

Workshop: Data Science Skills in Publishing: for authors, editors and referees.
Vienna, 18-08-2019

Data publishing and management workflow

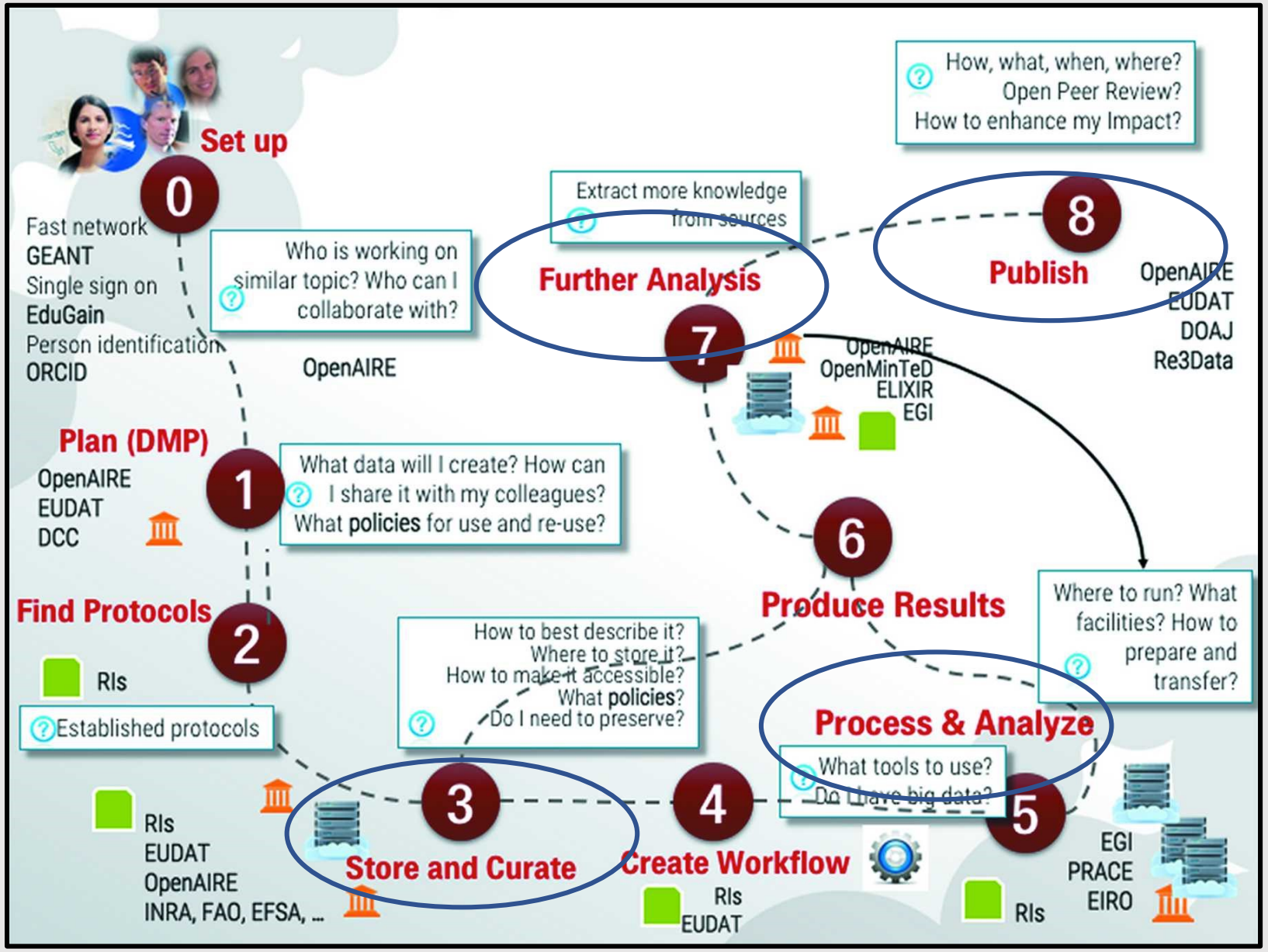


Illustration courtesy of Natalia Manova for the European OpenAIRE project

DDDWG recommendations

IUCr DDDWG Recommendations (top two)

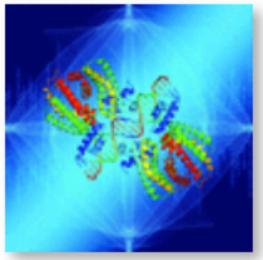
•Authors should provide a permanent and prominent link from their article to the raw data sets which underpin their journal publication and associated database deposition of processed diffraction data (*e.g.* structure factor amplitudes and intensities) and coordinates, and which should obey the 'FAIR' 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.

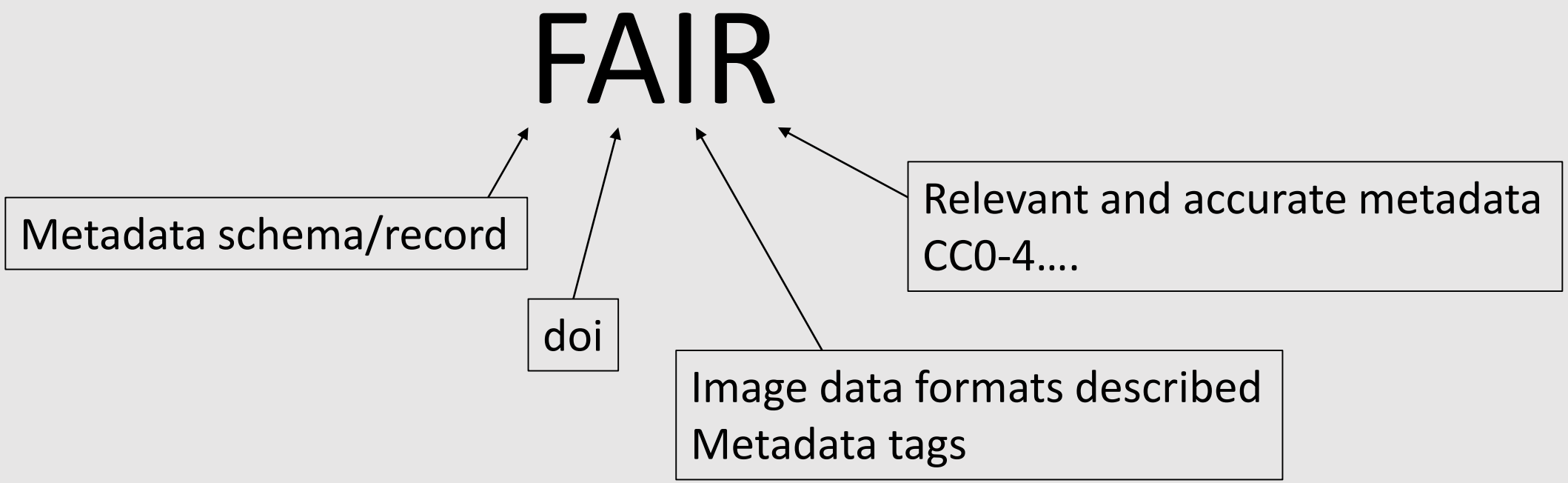
Read article Similar articles

"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. "

Raw data opportunities

- What are the possibilities of raw data archiving?
- Can we adhere to the FAIR principles?
- In the refereeing process can we validate the quality and analysis of the raw data?
- Can we do new science?

FAIR for raw data in Crystallography



Where to archive?

Institutional

Social sciences

Repository Name	Information on fees/costs	Size limits	Integrated with *Scientific Data*'s manuscript submission system	Re3data / FAIRSharing entry
Dryad Digital Repository	$120 USD for first 20 GB, and $50 USD for each additional 10 GB	None stated	**Yes ✓**	view FAIRsharing entry
figshare	100 GB free per *Scientific Data* manuscript. Additional fees apply for larger datasets	1 TB per dataset	**Yes ✓** - To qualify for the 100 GB of free storage, data must be uploaded to figshare via our submission system. Download instructions.	view FAIRsharing entry
Harvard Dataverse	Contact repository for datasets over 1 TB	2.5 GB per file, 10 GB per dataset	No	view re3data entry
Open Science Framework	Free of charge	5 GB per file, multiple files can be uploaded	No	view FAIRsharing entry
Zenodo	Donations towards sustainability encouraged	50 GB per dataset	No	view re3data entry
Mendeley Data	Contact repository for datasets over 10 GB	10 GB per dataset	No	view FAIRsharing entry

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

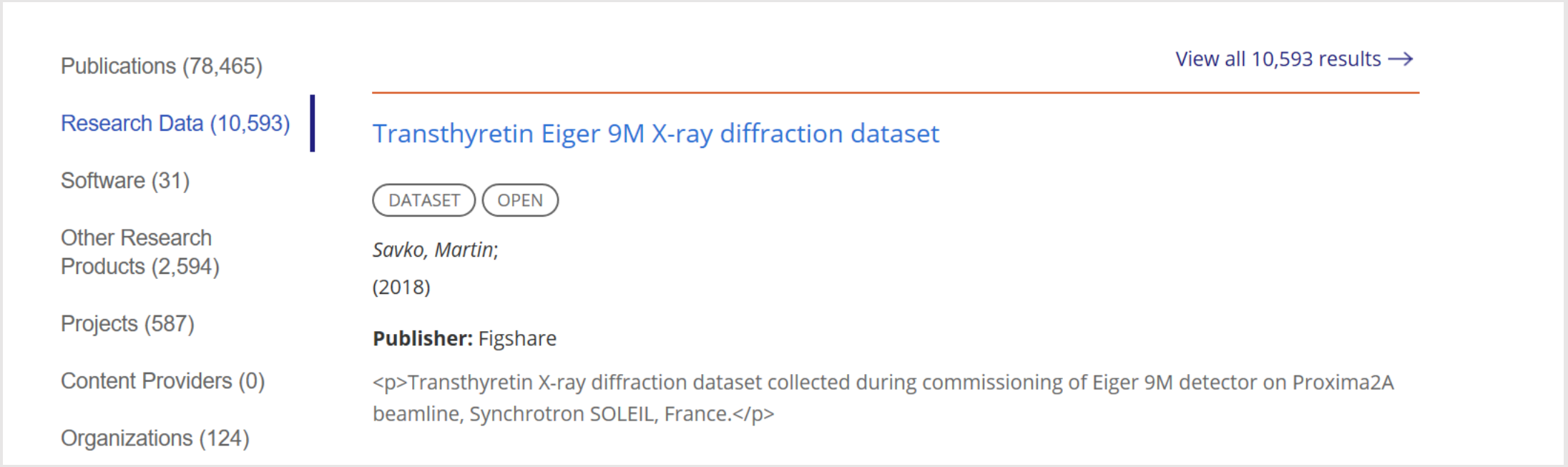
Synchrotron, Neutron Facilities and XFEL:
- ESRF
- DLS
- STFC ISIS
- Store. Synchrotron
- XFELs
- SciCat (ESS)
- ILL portal

DOIs restricted

OpenAIRE | EXPLORE

X-ray diffraction : 10,593

Mostly Zenodo and Figshare



Not SBGrid and IRRMC

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

DataCite: “x-ray diffraction” 15931 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 DOI metadata schema (Json record)


```
{
 "id": "https://doi.org/10.15127/1.219240",
 "doi": "10.15127/1.219240",
 "url": "https://www.escholar.manchester.ac.uk/uk-ac-man-scw:219240",
 "types": {
  "ris": "GEN",
  "bibtex": "misc",
  "citeproc": "article",
  "schemaOrg": "CreativeWork"
 },
 "creators": [
  {
   "name": "Tanley, Simon",
   "nameType": "Personal",
   "givenName": "Simon",
   "familyName": "Tanley",
   "affiliation": []
  }
 ],
 "titles": [
  {
   "title": "HEWL_carboplatin_aq_glycerol"
  },
 {
   "title": "4dd2",
   "titleType": "Subtitle"
  }
 ],
 "publisher": "The University of Manchester",
 "container": {},
 "subjects": [],
 "contributors": [],
 "dates": [
  {
   "date": "2014",
   "dateType": "Issued"
  }
 ],
 "publicationYear": 2014,
 "identifiers": [
  {
   "identifier": "https://doi.org/10.15127/1.219240",
   "identifierType": "DOI"
  }
 ],
 "sizes": [],
 "formats": [],
 "rightsList": [],
 "descriptions": [
  {
   "description": "The International Union of Crystallography has for many years been advocating archiving of raw data to accompany structural papers. Recently, it initiated the formation of the Diffraction Data Deposition Working Group with the aim of developing standards for the representation of these data. A means of studying this issue is to submit exemplar publications with associated raw data and metadata. A recent study on the effects of dimethyl sulfoxide on the binding of cisplatin and carboplatin to histidine in 11 different lysozyme crystals from two diffractometers led to an investigation of the possible effects of the equipment and X-ray diffraction data processing software on the calculated occupancies and B factors of the bound Pt compounds. 35.3 Gb of data were transferred from Manchester to Utrecht to be processed with EVAL. A systematic comparison shows that the largest differences in the occupancies and B factors of the bound Pt compounds are due to the software, but the equipment also has a noticeable effect. A detailed description of and discussion on the availability of metadata is given. By making these raw diffraction data sets available via a local depository, it is possible for the diffraction community to make their own evaluation as they may wish.",
   "descriptionType": "Abstract"
  }
 ],
 "geoLocations": [],
 "fundingReferences": [],
 "relatedIdentifiers": [],
 "providerId": "bl",
 "clientId": "bl.mchester",
 "state": "findable"
}
```

DataCite
FIND, ACCESS, AND REUSE DATA

“X-ray diffraction”: 15931
Restriction “Dataset” : 8804

“Raw diffraction”: 84

“X-ray diffraction“ AND Tanley: 15
Restriction “Dataset” : none

“Diffraction images”: 302

“X-ray diffraction” AND “crystal”: no Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

(“X-ray-diffraction" OR "raw data" OR "diffraction images") AND (macromolecule OR protein) : 446

no Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

(“X-ray-diffraction" OR "raw data" OR "diffraction images"): 34568

yes Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

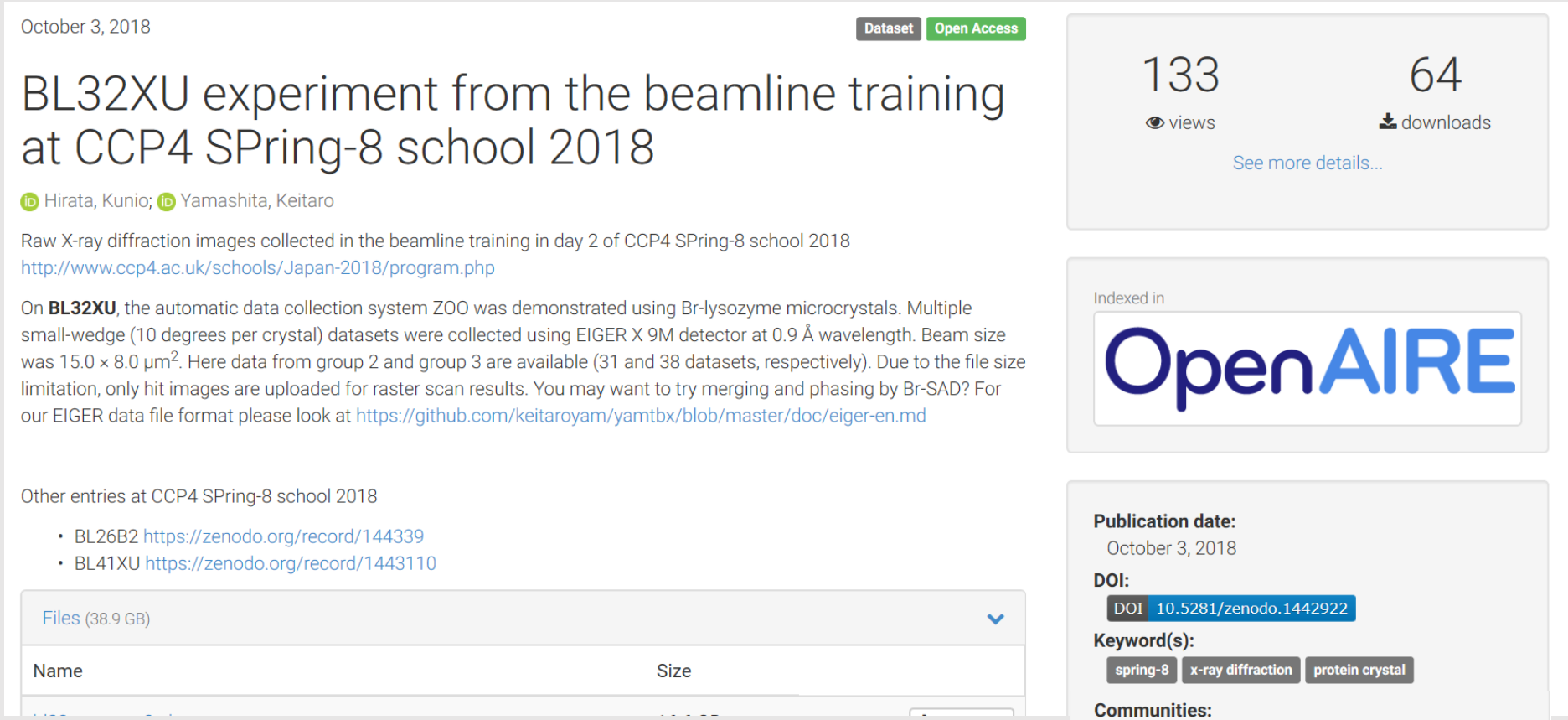
Helliwell: yes Raw diffraction images for pgp3 IUCrJ 2018 N. E. Chayen and J.R. Helliwell

“X-ray diffraction”: 1292

+ Dataset: 109

Mostly Macromolecular crystallography raw data

metadata



October 3, 2018

Dataset Open Access

BL32XU experiment from the beamline training at CCP4 SPring-8 school 2018

Hirata, Kunio; Yamashita, Keitaro

Raw X-ray diffraction images collected in the beamline training in day 2 of CCP4 SPring-8 school 2018
http://www.ccp4.ac.uk/schools/Japan-2018/program.php

On **BL32XU**, the automatic data collection system ZOO was demonstrated using Br-lysozyme microcrystals. Multiple small-wedge (10 degrees per crystal) datasets were collected using EIGER X 9M detector at 0.9 Å wavelength. Beam size was 15.0 × 8.0 μm^2. Here data from group 2 and group 3 are available (31 and 38 datasets, respectively). Due to the file size limitation, only hit images are uploaded for raster scan results. You may want to try merging and phasing by Br-SAD? For our EIGER data file format please look at https://github.com/keitaroyam/yamtbx/blob/master/doc/eiger-en.md

Other entries at CCP4 SPring-8 school 2018

- BL26B2 https://zenodo.org/record/144339
- BL41XU https://zenodo.org/record/1443110

Files (38.9 GB)

Name	Size

133 views

64 downloads

See more details...

Indexed in OpenAIRE

Publication date:
October 3, 2018

DOI:
DOI 10.5281/zenodo.1442922

Keyword(s):
spring-8 x-ray diffraction protein crystal

Communities:
Macromolecular Crystallography

License (for files):
Creative Commons Attribution 4.0 International

“X-ray diffraction” AND Helliwell: none

Helliwell: 18

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

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

526 datasets

■ 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

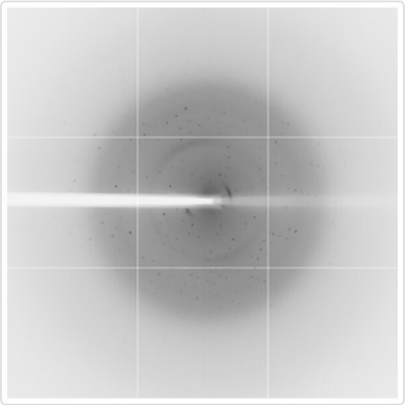
Integrated Resource for Reproducibility in Macromolecular Crystallography

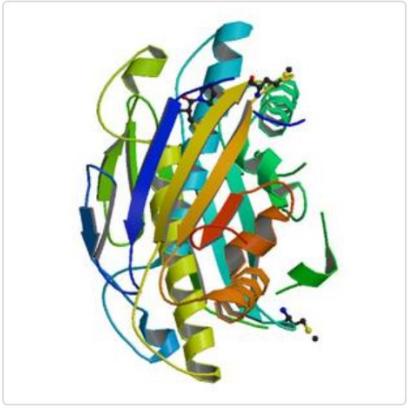
Search examples
Find a specific PDB ID: 4K6A
Free format search: potential drug target
Combining searches: drug AND cholera
Specific beamline: beamline = 21-ID-G
Resolution limit (Angstroms): resolution <1.25
Search by tag: workshop

Currently indexed projects: **5265**

Currently indexed datasets: **8552**

Diffraction project datasets IDP01325_3lus





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

Download all images (1.1 GB)

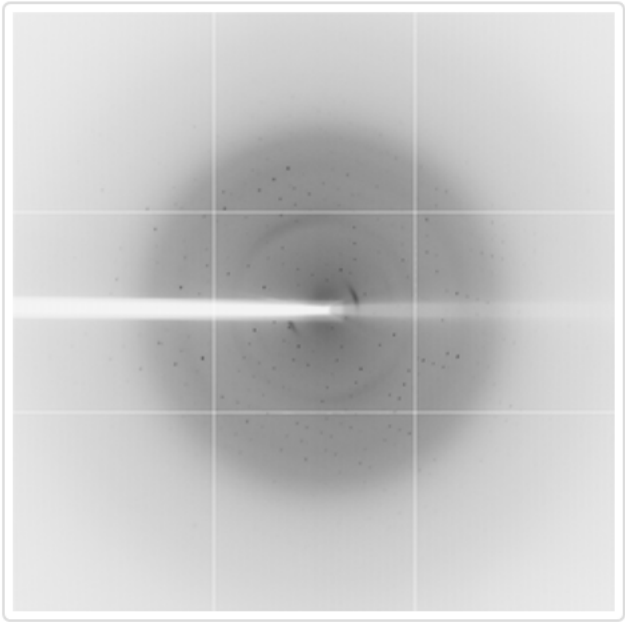
PDB website for 3LUS

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

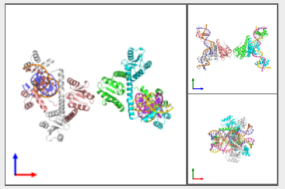
Tn3 resolvase - site III complex crystal form II

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

View
Downloads
3D Visualisation

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

Structure analysis

Details

Assembly composition: hetero tetramer (preferred)

Entry contents: 1 distinct polypeptide molecule
2 distinct DNA molecules

Macromolecules (3 distinct):

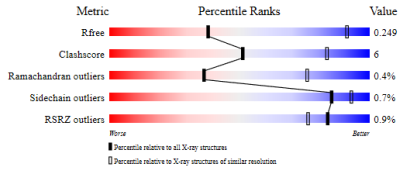
Transposon Tn3 resolvase

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:

Experiments and Validation

Details



X-ray source: APS BEAMLINE 19-ID

Spacegroup: *C*2

Unit cell: a: 144.9Å b: 151.92Å c: 106.14Å
α: 90° β: 99.85° γ: 90°

R-values:

R	R_{work}	R_{free}
0.209	0.207	0.25

Expression systems:
- *Escherichia coli*
- Not provided

PDB-REDO

The sliders below show the change in model quality between original PDB entry and the PDB-REDO entry

Model Geometry
Fit model/data

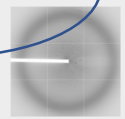
PDB-REDO

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



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

Re-useable: core metadata

Minimal Metadata	imgCIF tags
• Data binary format	_array_structure_byte_order,_array_structure _compression_type
• Number of pixels, pixel size (binning mode)	_array_structure_list.index;
• Beam Center (mm, pixels)	_array_structure_list.dimensions
• Origin of data frame	_array_element_size.size
• Wavelength	_diffrn_detector_element.center[1]
• Rotation axis	_diffrn_detector_element.center[2]
• Rotation range per frame	_diffraction_radiation.wavelength.wavelength
• Axes and offsets	_diffrn_scan_axis.axis_id, _diffrn_scan_axis.displacement_start
• Detector-to-sample distance	_diffrn_scan_axis.displacement.increment
	_axis.id, _axis.vector[1].., _ _axis.offset[1]..

continued…

Implicitly assumed (Expert knowledge)	Advanced
• Orientation of rotation axis • Rotation direction • Dectector swing angle (0°) • Polarization • Detector type	• Sensor thickness • Baseline offset • Overflow level • Polarization • Gain • Detector swing • Multi axis goniometer • Exposure time • Bad pixels • Time stamp

COMCIFS/CommDat: imgCIF metadata and CheckCIF

HDRMX and Nexus

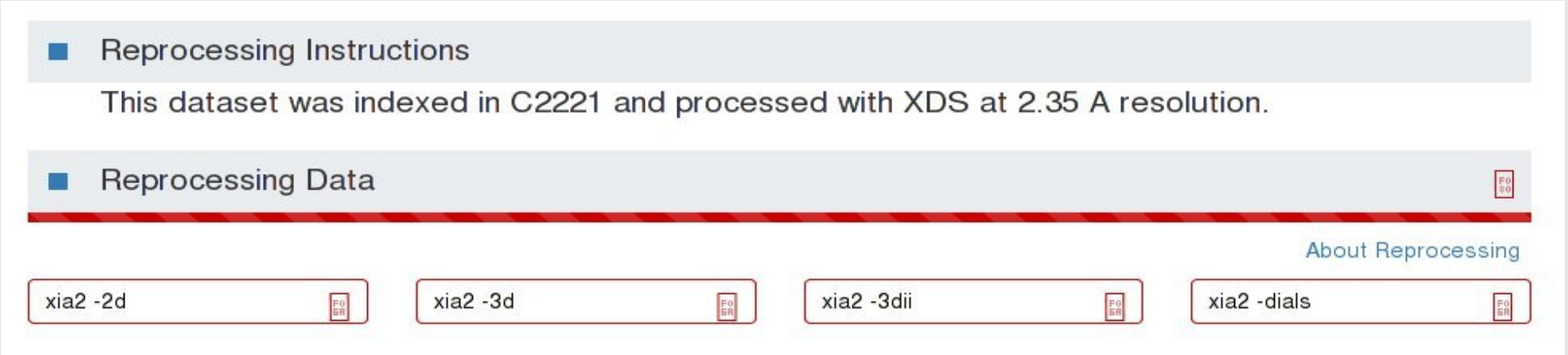
Raw data re-use

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
- Unsolved structure
- Diffuse scattering (packing disorder or internal mobility)
- Incommensurate modulation

383

5HV4	2014	Resolution 2.35 Å	XDS



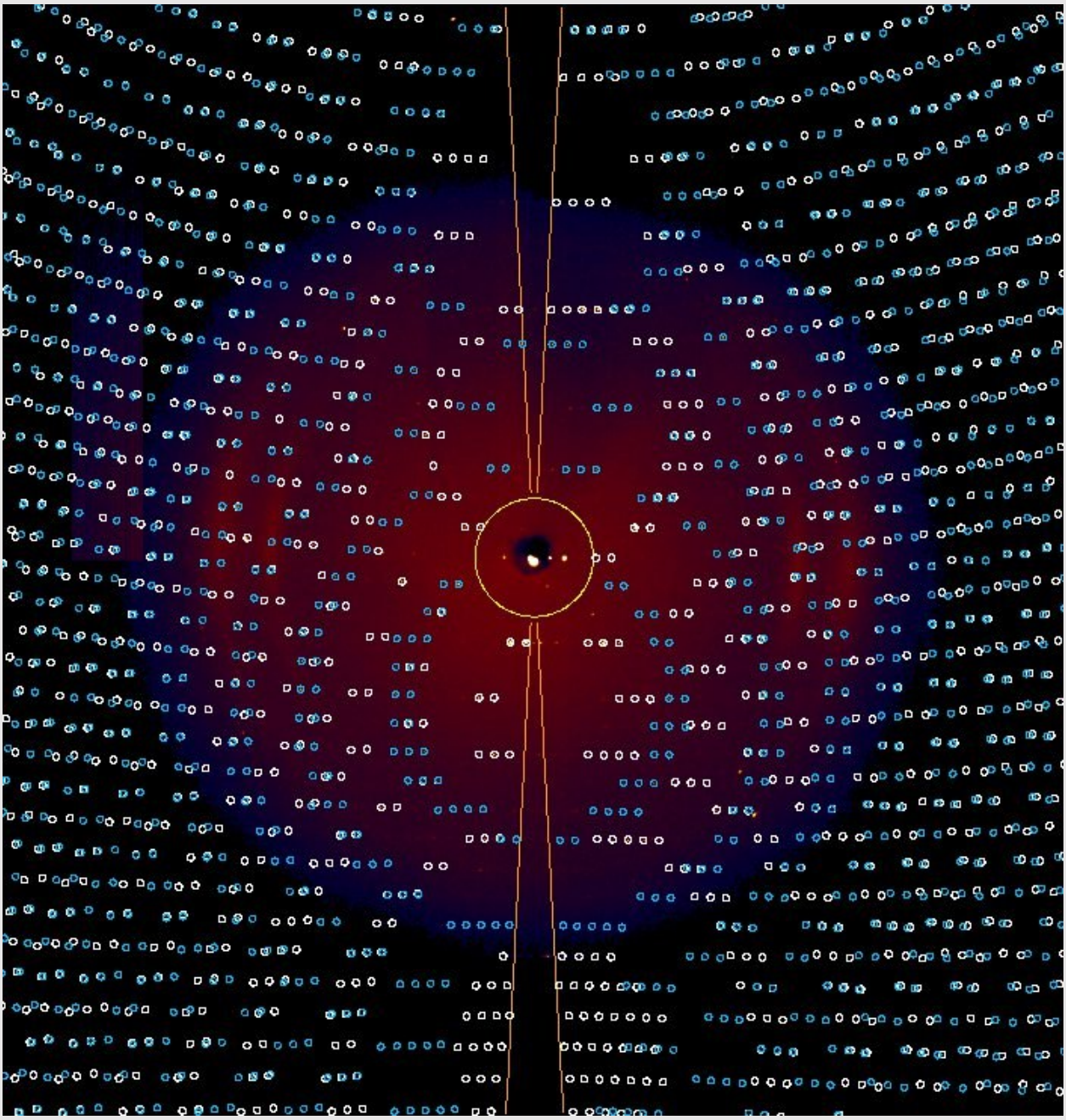
Monoclinic: 42.6 75.2 76.5 90 106.2 90
-> C-centered orthorhombic 42.6 146.4 75.2 90 90 90

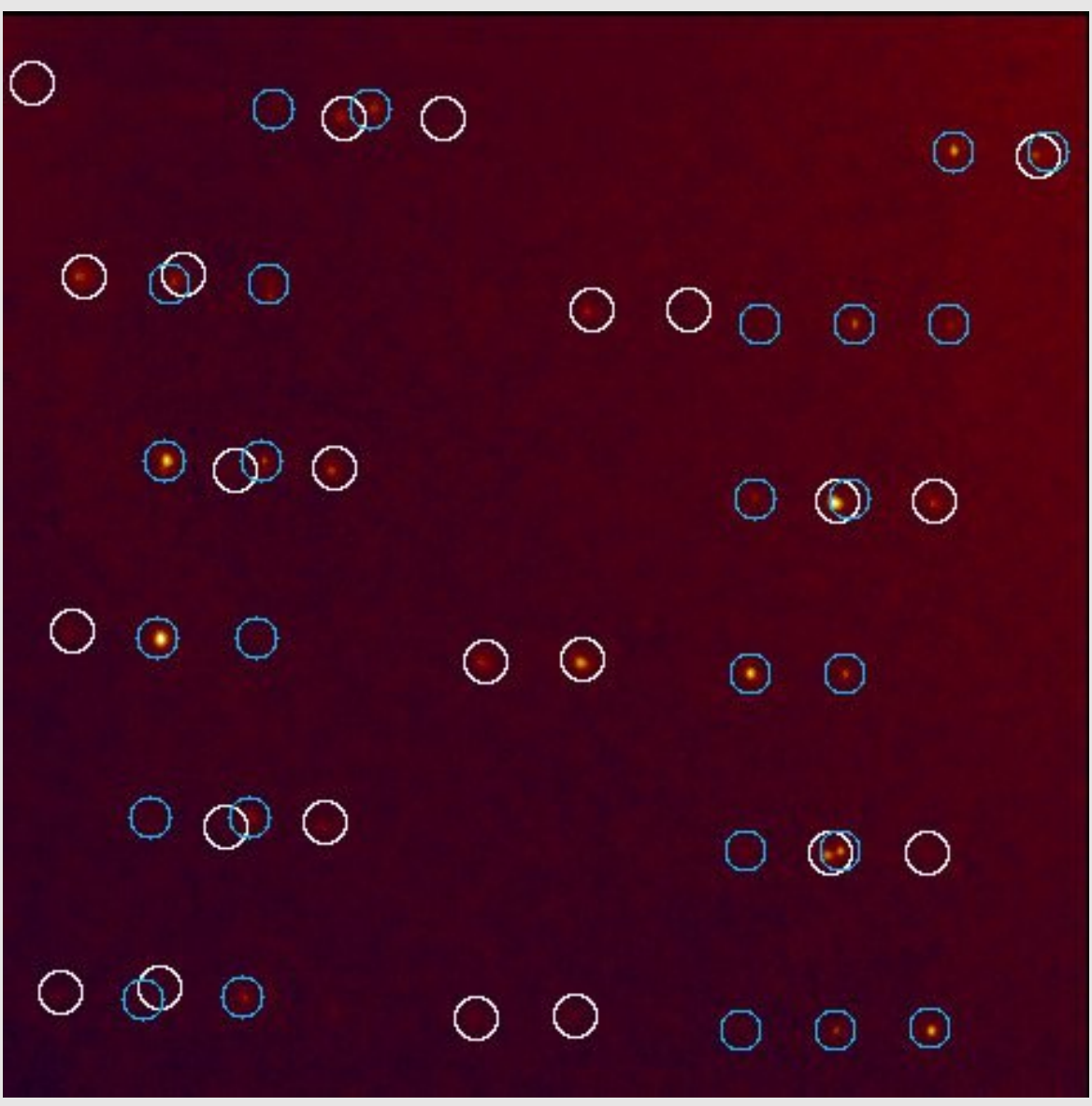
Dirax finds two matrices, Data processed with EVAL

Problems with reprocessing:
1) offset of omega-axis (~1°)
2) twin 177.6° rotation around (0,1,-2) in orthorhombic lattice

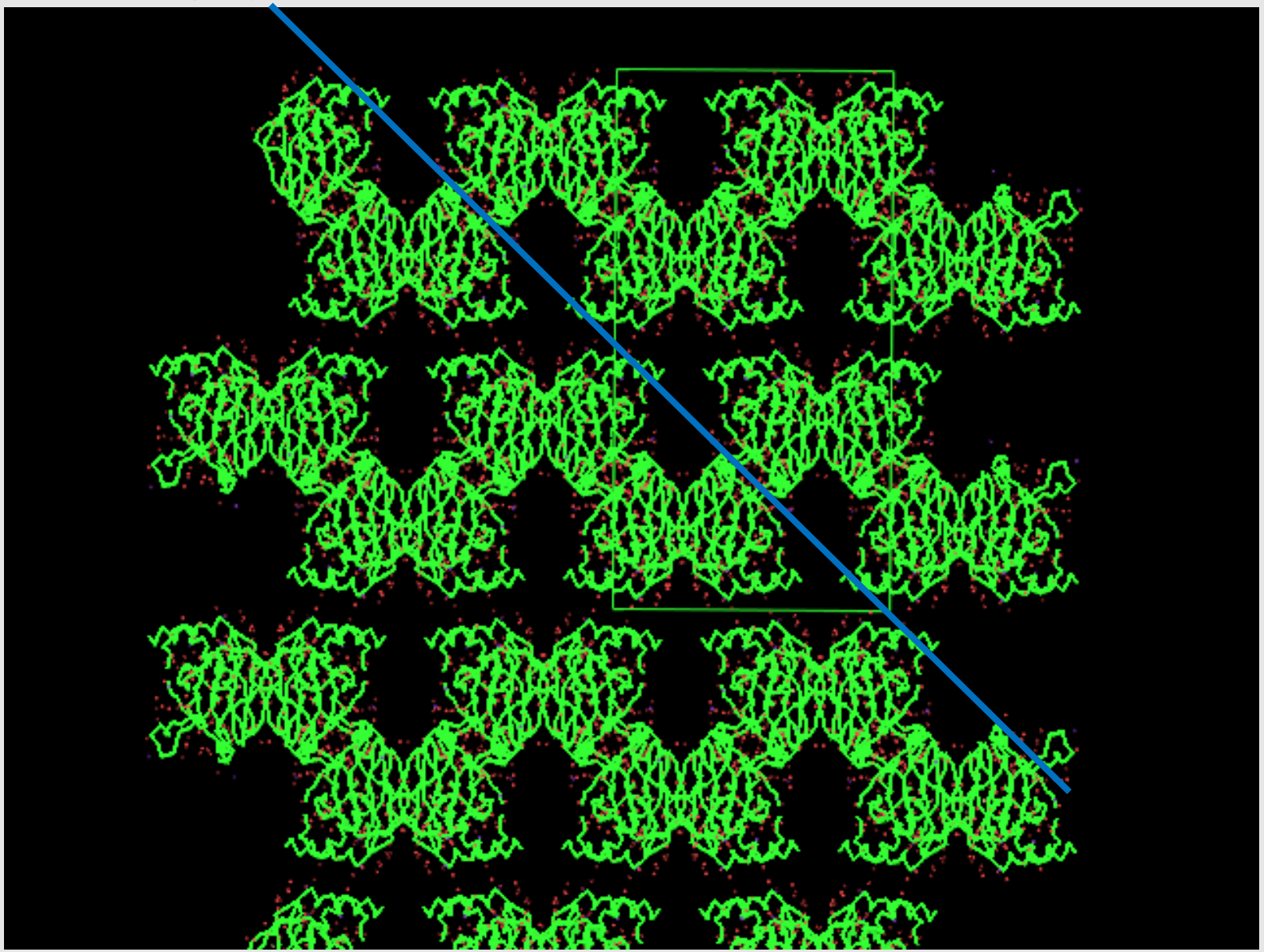
23% deconvoluted 15% overlapping

Beamline: 4.2.2 ALS
Detector: CMOS_8M RDI
Format: img
Header: ADSC





Twinning operation

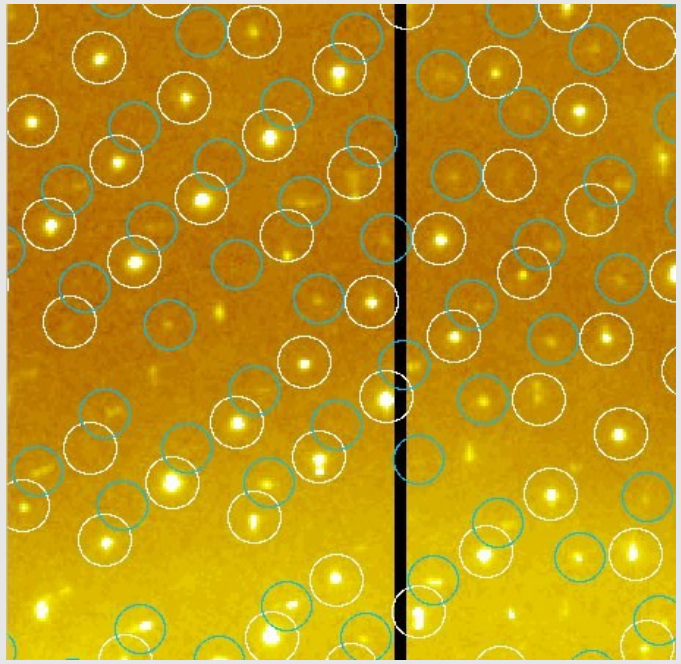


BJP1_3m8t

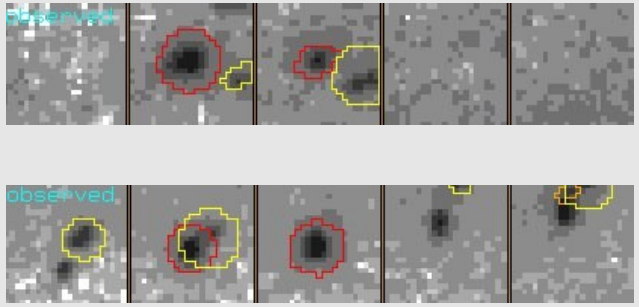
3M8T	2008	Resolution 1.33 Å	MOSFLM

unitcell: 47.2 45.2 77.4 100.9 89.8 118.1

Symmetry P1, two independent molecules



Second fragment 3.8° rotation, arbitrary axis
5% overlap with first matrix



EVAL15 box

Multiple crystal forms

E. coli enzyme N-acetylneuraminic lyase

SCIENTIFIC REPORTS

OPEN

Received: 23 January 2018
Accepted: 19 September 2018
Published online: 05 October 2018

Pathological macromolecular crystallographic data affected by twinning, partial-disorder and exhibiting multiple lattices for testing of data processing and refinement tools

Ivan Campeotto[1,2,3], Andrey Lebedev[4], Antoine M. M. Schreurs[5], Loes M. J. Kroon-Batenburg[5], Edward Lowe[2], Simon E. V. Phillips[1,4], Garib N. Murshudov[6] & Arwen R. Pearson[1,7]

Data Records

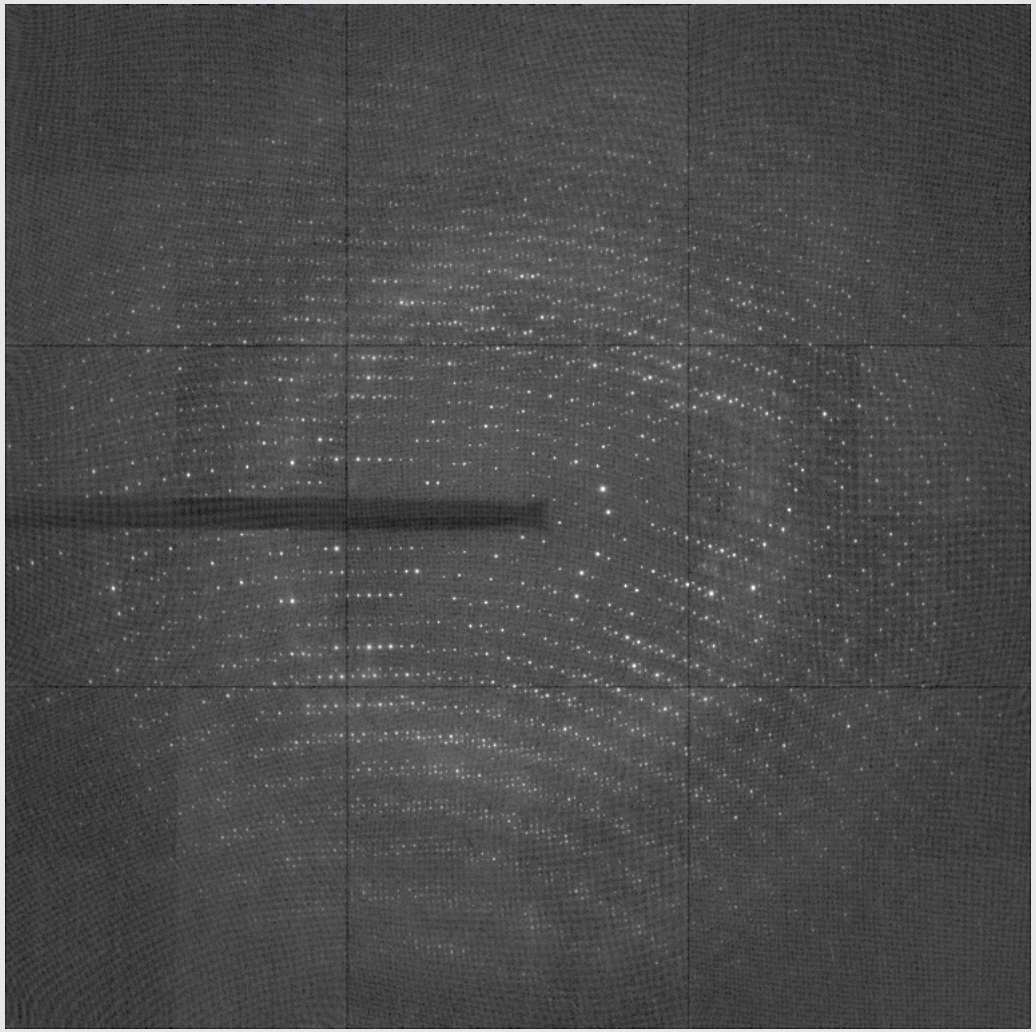
The datasets (raw diffraction images) discussed in this manuscript have been deposited in the publicly available database zenodo at, https://doi.org/ 10.5281/zenodo.54568 and 10.5281/zenodo.1240503. Structural models and processed structure factor data deposited in the PDB are available under the accession codes given in Table 1, with the exception of dataset Y137A, as the R factor indices were not satisfactory for PDB deposition.

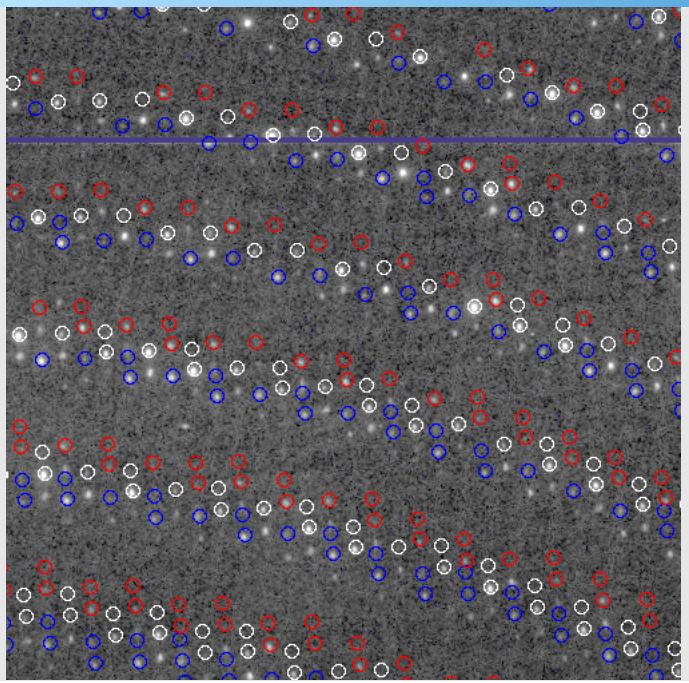
Datasets	Res	Crystal form and cell parameters	Obliquity* (ω)	Twinning fraction*	Twin on	Twin off	PDB code	Diamond Station
Wild type apo	2.20 Å	$P2_1$ crystal form I a = 54.8 b = 142.2 c = 84.2 α = 90.00 β = 108.97 γ = 90.00	0.019	0.372	R_{factor} = 0.200 R_{free} = 0.267	R_{factor} = 0.251 R_{free} = 0.319	2WO5	I02
Wild type pyruvate complex	1.65 Å	$P2_1$ crystal form I a = 54.7 b = 142.5 c = 83.6 α = 90.00 β = 109.16 γ = 90.00	0.070	0.334	R_{factor} = 0.201 R_{freev} = 0.245	R_{factor} = 0.256 R_{freev} = 0.293	2WNN	I03
E192N apo	1.80 Å	$P2_1$ crystal form I a = 54.6 b = 142.8 c = 84.3 α = 90.00 β = 108.8 γ = 90.00	0.130	0.463	R_{factor} = 0.195 R_{free} = 0.244	R_{factor} = 0.272 R_{free} = 0.320	2WNQ	I04
E192N pyruvate complex	1.80 Å	$P2_1$ crystal form I a = 56.9 b = 143.0 c = 83.9 α = 90.00 β = 109.8 γ = 90.00	0.000	—	R_{factor} = 0.178 R_{free} = 0.209	R_{factor} = 0.187 R_{free} = 0.223	2WNZ	I02
E192N + pyruvate + THB**	2.05 Å	$P2_1$ crystal form I a = 57.0 b = 143.7 c = 84.3 α = 90.00 β = 109.9 γ = 90.00	0.130	—	R_{factor} = 0.192 R_{free} = 0.238	R_{factor} = 0.191 R_{free} = 0.242	2WPB	I03
Y137A pyruvate complex	1.80 Å	$P2_1$ crystal form I a = 54.7 b = 142.2 c = 83.6 α = 90.0 β = 109.0 γ = 90.0	0.119	0.149	R_{factor} = 0.287 R_{free} = 0.331	R_{factor} = 0.296 R_{free} = 0.357	n./a.***	I04
Y137A pyruvate, ManNAc and Neu5Ac complex	2.00 Å	$P2_1$ crystal form I a = 56.1 b = 143.5 c = 83.6 α = 90.0 β = 109.6 γ = 90.0	0.094	0.497	R_{factor} = 0.183 R_{free} = 0.236	R_{factor} = 0.265 R_{free} = 0.321	4BWL	I02
Wild type apo	1.90 Å	$P2_1$ crystal form II a = 84.3 b = 95.9 c = 91.4 α = 90.00 β = 115.33 γ = 90.00	2.10	—	R_{factor} = 0.198 R_{free} = 0.226	R_{factor} = 0.197 R_{free} = 0.225	2YGY	I02
E192N/Y137F pyruvate complex	1.80 Å	$P2_1$ crystal form III a = 78.0 b = 116.7 c = 83.7 α = 90.0 β = 118.06 γ = 90.00	0.290	0.328	R_{factor} = 0.156 R_{free} = 0.183	R_{factor} = 0.206 R_{free} = 0.228	2YGZ	I02
E192N + pyruvate complex	1.85 Å	$P2_1$ crystal form III a = 78.1 b = 116.5 c = 83.7 α = 90.00 β = 116.5 γ = 90.00	0.150	0.096	R_{factor} = 0.165 R_{free} = 0.186	R_{factor} = 0.174 R_{free} = 0.193	2XFW	I02
E192N + pyruvate	1.45 Å	$P2_12_12_1$ crystal form IV a = 78.3 b = 108 c = 148.3 α = β = γ = 90.00	0.000	—	R_{factor} = 0.191 R_{free} = 0.201	R_{factor} = 0.188 R_{free} = 0.205	2WKJ	I04

Four crystal form I structures have incommensurate modulation
q-vector: ~0.16 0.0 ~0.43
and twinning (-h,-k,h+l)

Incommensurate modulation

E. coli enzyme N-acetyl-neuraminic lyase





```
rmat from j1r.rmat
checking laue with one qvector
qvec setup for symmetry 2/m
RMAT 1 j1r
                RMAT                                        DMAT
  -0.0172030  -0.0026438  -0.0008334       -50.8232994 -20.1229591   6.3027682
  -0.0075459   0.0064414  -0.0017993       -53.9012222 131.3275757 -15.3532276
  -0.0041504  -0.0007531  -0.0124649        20.1790657  -1.2342228 -81.3962860
Determinant:  0.1606476E-05                              622480.2
cell from rmat: 55.02423 142.78653 83.86938 90.0003 109.1485 90.0002 V= 622480.25
pg constrained: 55.02423 142.78653 83.86938 90.0000 109.1485 90.0000 V= 622480.19
Sigma 0.0266 0.0627 0.0144 0.011 0.019 0.025 Volume 283.49
Bravais=P pg=2/m
Qvectype=2 MODULATED
ActiveQvecRadius=0.007152
QVEC     dh      dk      dl    Order
   1  0.1653  0.0000  0.4316    1
3 generated QV Combinations
```

Validation

Validation of the structural model:

- IUCr CheckCIF
- wwPDB validation report

} Structural model in the light of the processed data (hkl, I, $\sigma(I)$)

“CheckCIF” for raw images:

- Check for core metadata
- Validation of the analysis of raw data:
  - is everything in the diffraction pattern understood?
  - Referee or automatic validation?

Automatic validation of diffraction image analysis?

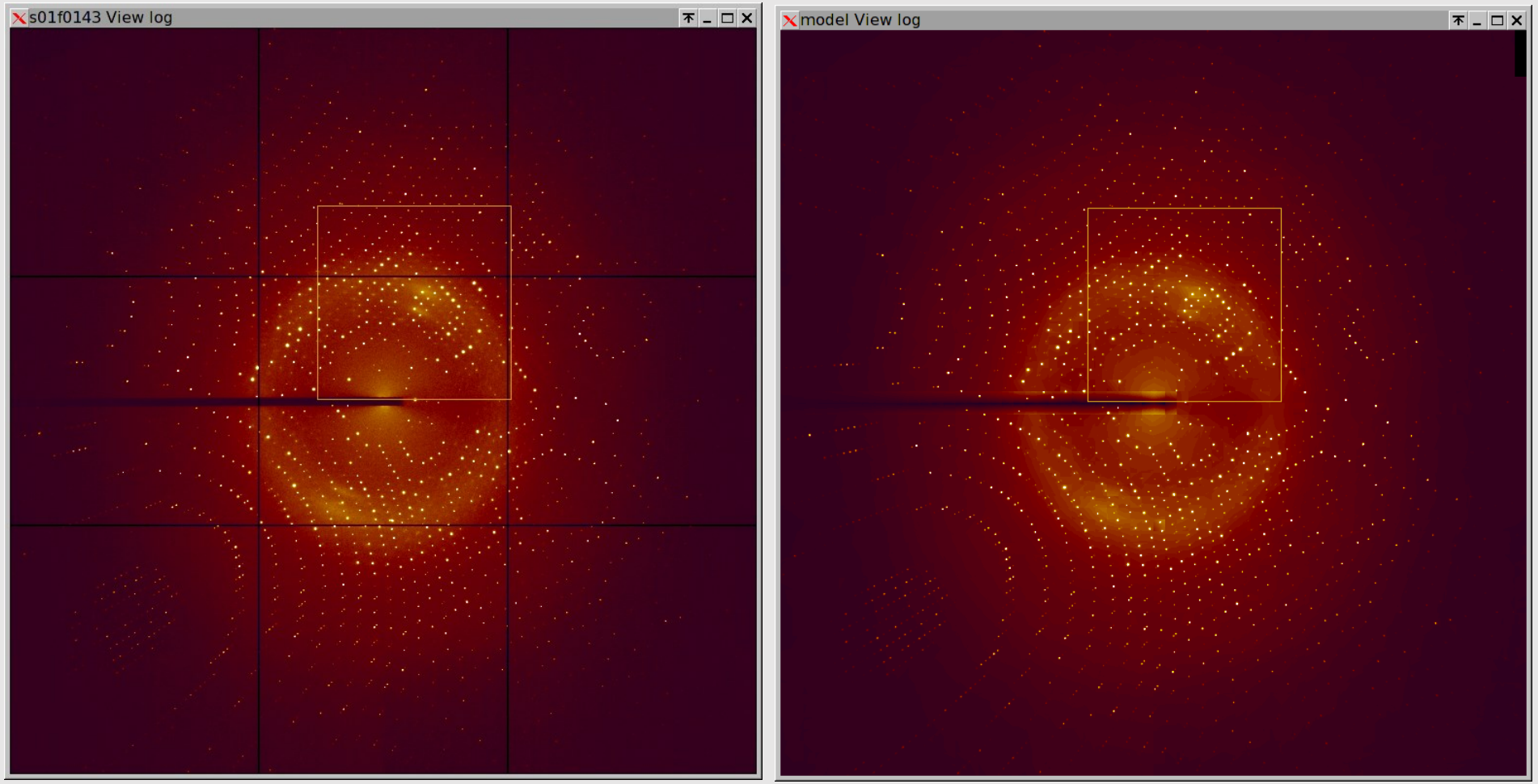
Thaumatin data from I03 Diamond Light Source Detector ADSC 3000x3000 pixels

Duncan Sneddon

ρ_{obs}

$\rho_{model} = JP_i + \sum_m^M J_m P_{im} + ax_i + by_i + c$

EVAL

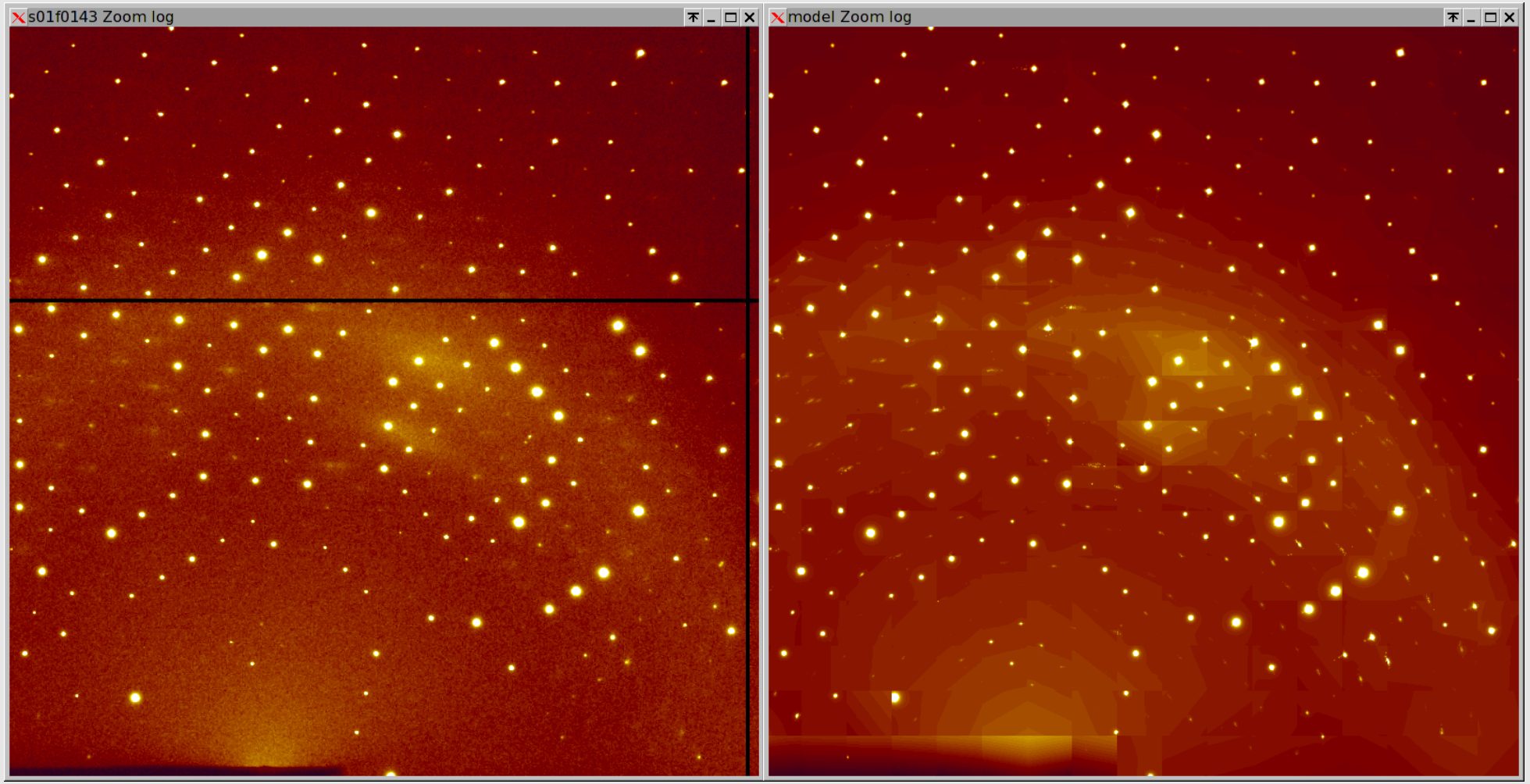


Reconstruction from modelling 5 images

Modelled image with EVAL

Iobs

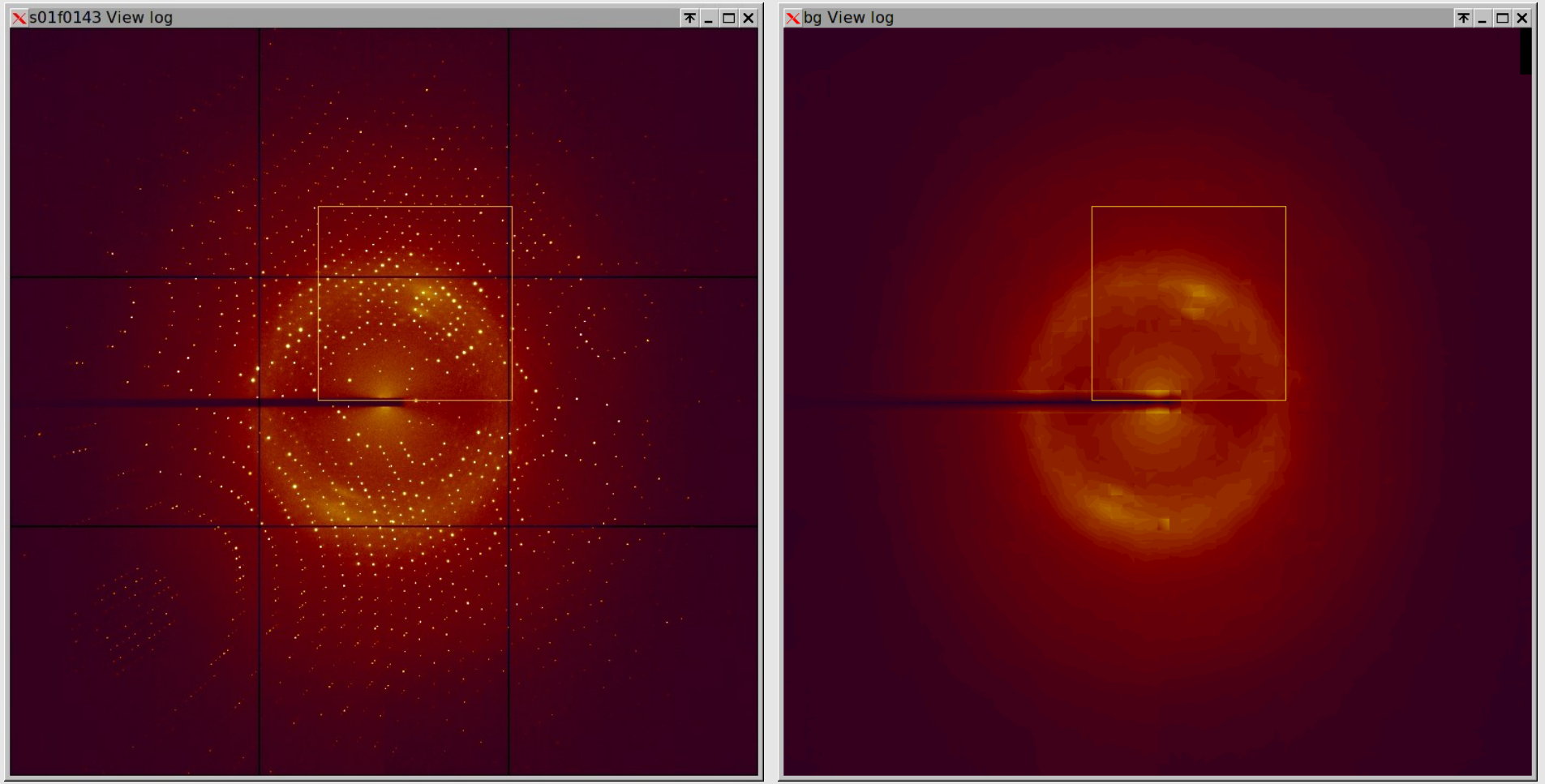
Imodel



Distribution of pixel values?

Iobs

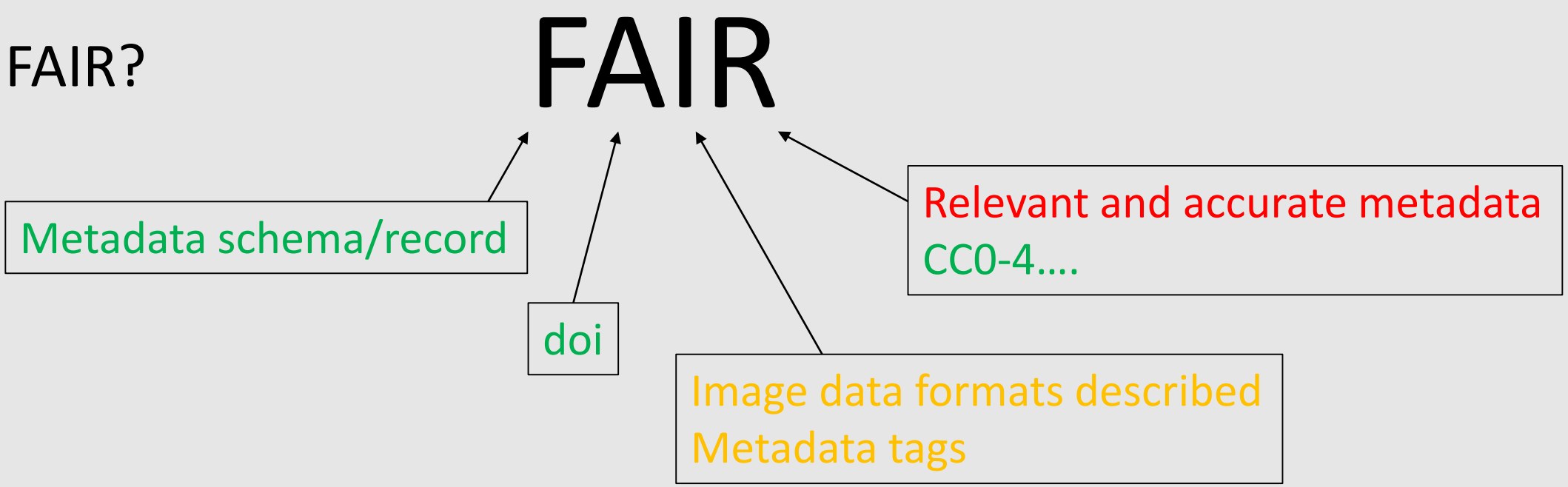
Ibg= $ax_i + by_i + c$



The diffuse background reconstructed from observations per box

Conclusions

- FAIR?



- COMCIFS/CommDat: imgCIF metadata and CheckCIF
- HDRMX and Nexus
- Referees can have a look at the images and the indexing
- Automatic validation of raw data interpretation?