Experiences with MX data reuse at Diamond

David Aragão
david.aragao@diamond.ac.uk
Linkedin: https://www.linkedin.com/in/davidaragao
orcID: https://orcid.org/0000-0002-6551-4657
Mastodon: @aragaod@aus.social

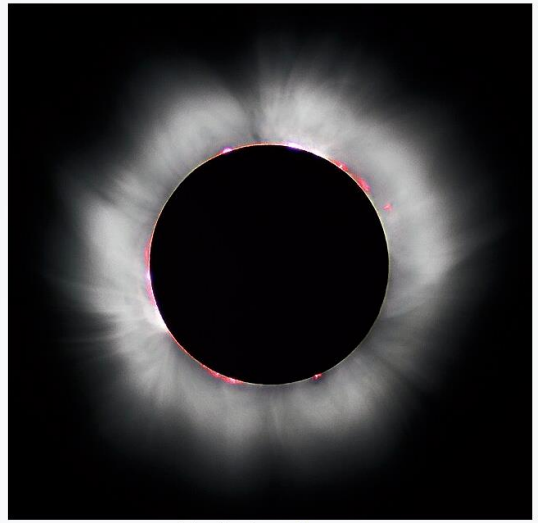
Vivian Li
Steve Collins
Graeme Winter
Richard Gildea
Elliot Nelson
Ralf Flaig

Remembering Raimond Ravelli (1968-2023)

Néron

View of the Néron from the Esclangon Bridge, to the south, crossing the Drac between Grenoble and Fontaine.

Solar eclipse of August 11, 1999



Totality from France

https://www.embl.org/about/info/alumni/community/obituaries/raimond-ravelli/

8-inch, 5¼-inch, and 3½-inch floppy disks

Harwell Campus

- **The UK's national synchrotron facility, open from 2007 to researchers in universities and industry, and co-owned and funded by the UK government through STFC and the Wellcome Trust (86:14 ratio).**

- **A total of 12,000 registered users, and more than 100 companies paying for proprietary use.**
- **33 operational beamlines**
- **First scientific users from 2007**

Diamond Beamlines and Science Groups

Cryo-TXM B24
Microfocus MX I24
Circular Dichroism B23
Long Wavelength MX I23
MIRIAM: IR Microspectroscopy B22
Small Angle Scattering and Diffraction I22
High Throughput SAXS B21
Inelastic X-ray Scattering I21
LOLA: Versatile X-ray Spectroscopy I20
Small-Molecule Single-Crystal Diffraction I19
Core XAS B18
Microfocus Spectroscopy I18
Test Beamline B16
Materials and Magnetism I16
XPDF: X-ray Pair Distribution Function I15-1
Extreme Conditions I15
Hard X-ray Nanoprobe I14
I02-2 VMXi
I02-1 VMXm
I03 MX
I04-1 MX and XChem
I04 Microfocus MX
I05 ARPES
I06 Nanoscience
I07 Surface and Interface Diffraction
B07 VERSOX: Versatile Soft X-ray
I08 Scanning X-ray Microscopy
I09 SISA: Surface and Interface Structural Analysis
I10 BLADE: X-ray Dichroism and Scattering
I11 High Resolution Powder Diffraction
Long Duration Experiments (LDE)
DIAD: Dual Imaging and Diffraction (K11)
I12 JEEP: Joint Engineering, Environmental and Processing
I13 X-ray Imaging and Coherence
Krios I
Scios
Talos Arctica
Krios II
Krios III
Krios IV
ARM 300F
ARM 200F

Macromolecular Crystallography
Crystallography
Structures and Surfaces
Biological Cryo-Imaging
Magnetic Materials
Spectroscopy
Imaging and Microscopy
Soft Condensed Matter

Large data generating MX beamlines

- I04-1 (XCHEM)
- I04 (micro, UDC)
- I03 (UDC)
- I24 (micro, SSX)

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

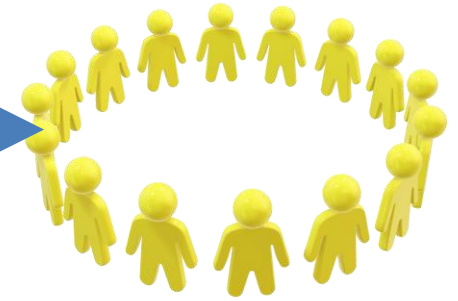
FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

Data storage



Spinning disk / SSD

Long term storage (Tape)

- What do we need to analyse data later?
- What do need to find useful data?

diamond

research papers

IUCrJ
ISSN: 2052-2525
BIOLOGY | MEDICINE

Volume 7 | Part 5 | September 2020 | Pages 784-792
https://doi.org/10.1107/S2052252520008672
OPEN ACCESS
Cited by 8

Gold Standard for macromolecular crystallography diffraction data

Herbert J. Bernstein,[a*] Andreas Förster,[b] Asmit Bhowmick,[c] Aaron S. Brewster,[c] Sandor Brockhauser,[d,e,f] Luca Gelisio,[g] David R. Hall,[h] Filip Leonarski,[i] Valerio Mariani,[g] Gianluca Santoni,[j] Clemens Vonrhein[k] and Graeme Winter[h]

[a]Ronin Institute for Independent Scholarship, c/o NSLS II, Brookhaven National Laboratory, Upton, New York, USA, [b]DECTRIS Ltd, Täfernweg 1, 5405 Baden-Dättwil, Switzerland, [c]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA, [d]European XFEL GmbH, Holzkoppel 4, 22869 Schenefeld, Germany, [e]Biological Research Centre Szeged (BRC), Temesvári krt. 62, 6726 Szeged, Hungary, [f]University of Szeged, Arpad ter 2, 6720 Szeged, Hungary, [g]Center for Free-Electron Laser Science, Notkestrasse 85, 22607 Hamburg, Germany, [h]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot OX11 0DE, United Kingdom, [i]Swiss Light Source, Paul Scherrer Institut, Forschungsstrasse 111, 5232 Villigen PSI, Switzerland, [j]Structural Biology Group, European Synchrotron Radiation Facility, 71 Avenue des Martyrs, 38000 Grenoble, France, and [k]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, United Kingdom
[*]Correspondence e-mail: yayahjb@gmail.com

Edited by M. Takata, SPring-8, Japan (Received 6 February 2020; accepted 26 June 2020; online 10 July 2020)

https://doi.org/10.1107/S2052252520008672

Multi-crystal cubic insulin example data set recorded on i24

Winter, Graeme

Data collector(s)

Winter, Graeme

Researcher(s)

Bertram, Felicity

Data collected as part of routine commissioning on Diamond Light Source beamline i24 15th August 2022 with samples prepared by Felicity Bertram at Diamond following standard techniques. Data are individually incomplete but combined make for a reasonably complete and reasonable data set.

Purpose of the data upload is to make data available for tutorials using the DIALS toolchain (see e.g. examples at https://github.com/graeme-winter/dials_tutorials) however data are available for all purposes without limitation.

Files (6.5 GB)

Name	Size	
ins11_b200_1.nxs	55.7 kB	Download
md5:6780fcea0a65d3ee1d30c032ce1e7278		
ins11_b200_1_000001.h5	184.4 MB	Download
md5:1627e485b09717cf3767608a4a604cfb		
ins11_b200_1_master.h5	55.7 kB	Download
md5:6780fcea0a65d3ee1d30c032ce1e7278		

https://zenodo.org/record/7085897

August 4, 2023

Dataset Open Access

Crystal structures of SARS-CoV-2 main protease screened against COVID Moonshot compounds by X-ray Crystallography at the XChem facility of Diamond Light Source

Fearon, D.; Aimon, A.; Aschenbrenner, J.C.; Balcomb, B.H.; Barker, I.A.; Bertram, F.K.R.; Brandao-Neto, J.; Dias, A.; Douangamath, A.; Dunnett, L.; Godoy, A.S.; Gorrie-Stone, T.J.; Koekemoer, L.; Krojer, T.; Lithgo, R.M.; Lukacik, P.; Marples, P. G.; Mikolajek, H.; Nelson, E.; Nidamarthi, K.H.V; Owen, C. D.; Powell, A.J.; Rangel, V. L.; Skyner, R.; Strain-Damerell, C.M.; Thompson, W.; Tomlinson, C. W. E.; Wild, C.; Walsh, M.A.; von Delft, F.

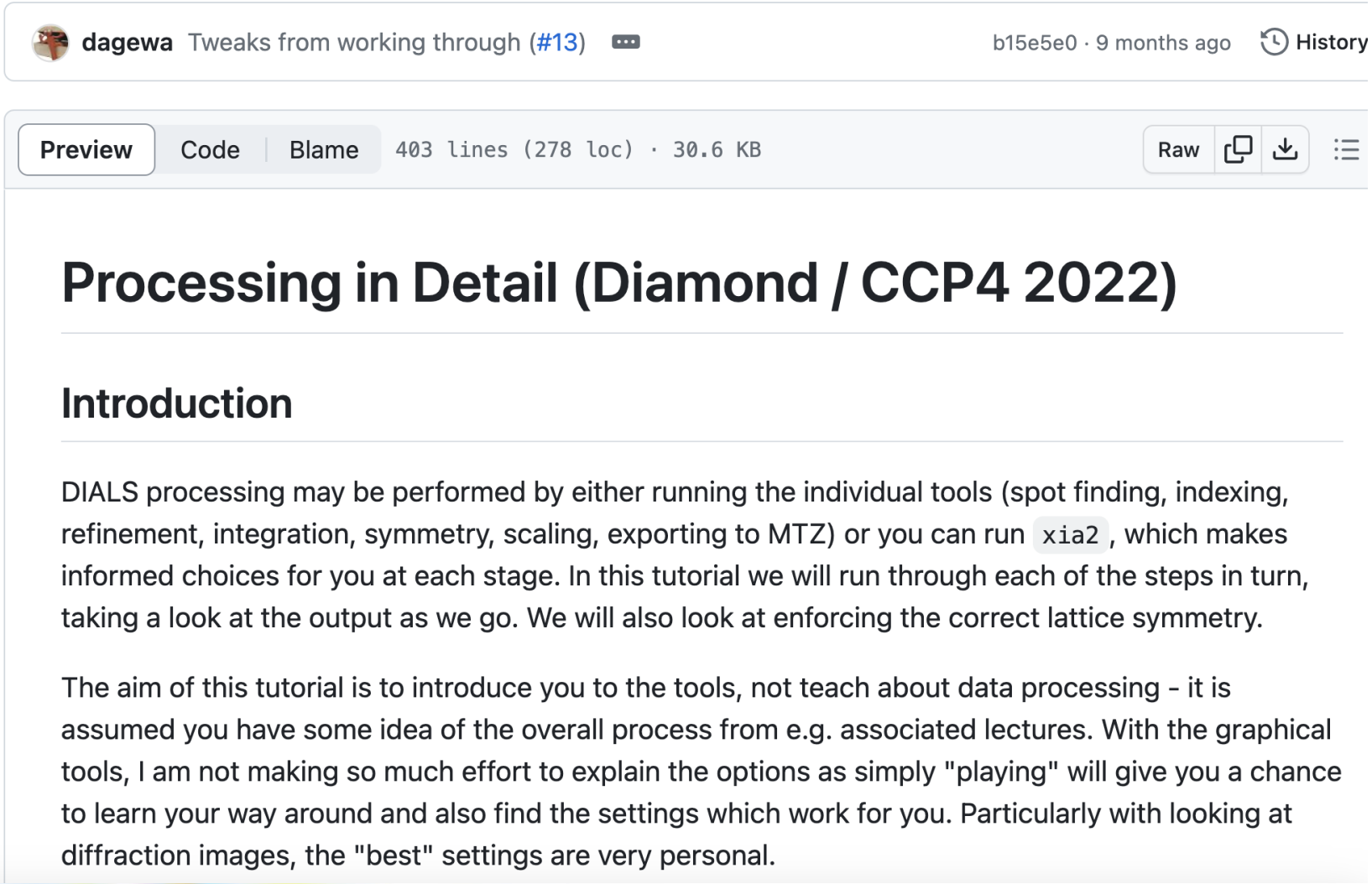
Bulk repositiory of structures of SARS-CoV-2 main protease in complex with COVID Moonshot inhibtor compounds

Files (814.8 MB)

https://zenodo.org/record/8214528

diamond

Training



https://github.com/graeme-winter/dials_tutorials/blob/main/insulin/processing_in_detail.md

diamond

Current DLS situation

- DLS has changed its policy from April 2019
  - Academic funded data collected thereafter would be made public from March 2022
  - Data owners being told 6 months in advance of such
  - DLS has not yet released any data but still has the intention to do it
  - Data prior to April 2019, DLS has not yet deleted any data, but such data would be only available at the explicitly request of the data owner

Acta Crystallogr D Biol Crystallogr. 2014 Oct 1; 70(Pt 10): 2510–2519.
Published online 2014 Sep 30. doi: 10.1107/S1399004714016174
PMCID: PMC4187999
PMID: 25286837

Operation of the Australian Store.Synchrotron for macromolecular crystallography

Grischa R. Meyer,[a] David Aragão,[b] Nathan J. Mudie,[b] Tom T. Caradoc-Davies,[b] Sheena McGowan,[c] Philip J. Bertling,[d] David Groenewegen,[d] Stevan M. Quenette,[a] Charles S. Bond,[e] Ashley M. Buckle,[c] and Steve Androulakis[f,*]

Author information Article notes Copyright and License information PMC Disclaimer

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4187999/

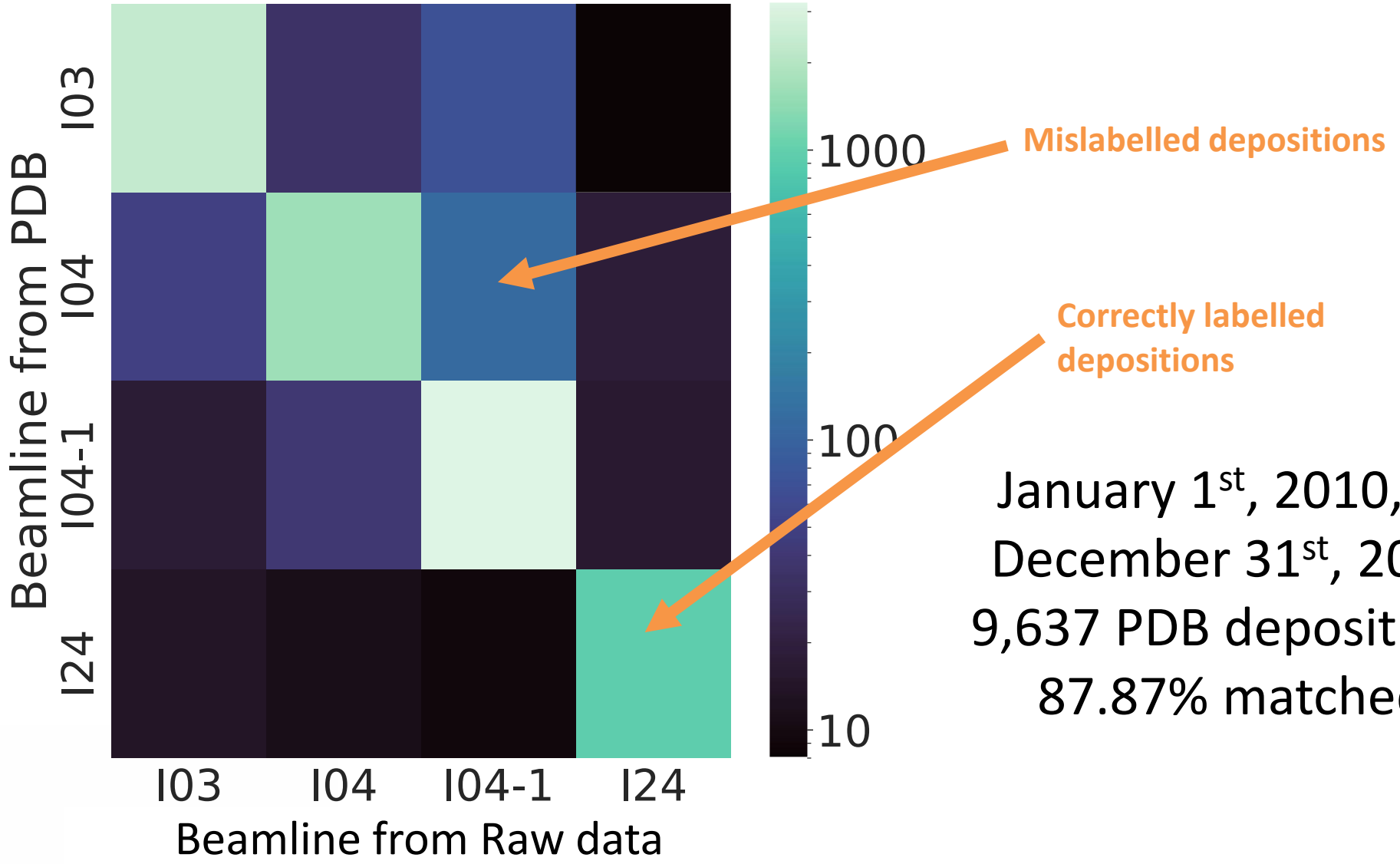
- How useful would be to have that raw data available?
  - 2021 submitted 3 Months summer student project to investigate this
  - Student started worked June to September 2022

Can we investigate DLS data and see usefulness?

- Are our instruments getting better?
- Are we collecting more but also efficiently ?
- How many collections until the publishable final dataset ?
- What is the phasing vs MR trends?
- Can it inform where we should direct software and hardware development on our I04 beamline and MX in general?
- Calculate how many datasets of the same protein are required before a structure is deposited for that structure
- Identify (prototyping on I04 beamline) datasets that have been solved prior to the dataset used in the PDB deposition
- Improve our Unattended Data collections recipes and strategies

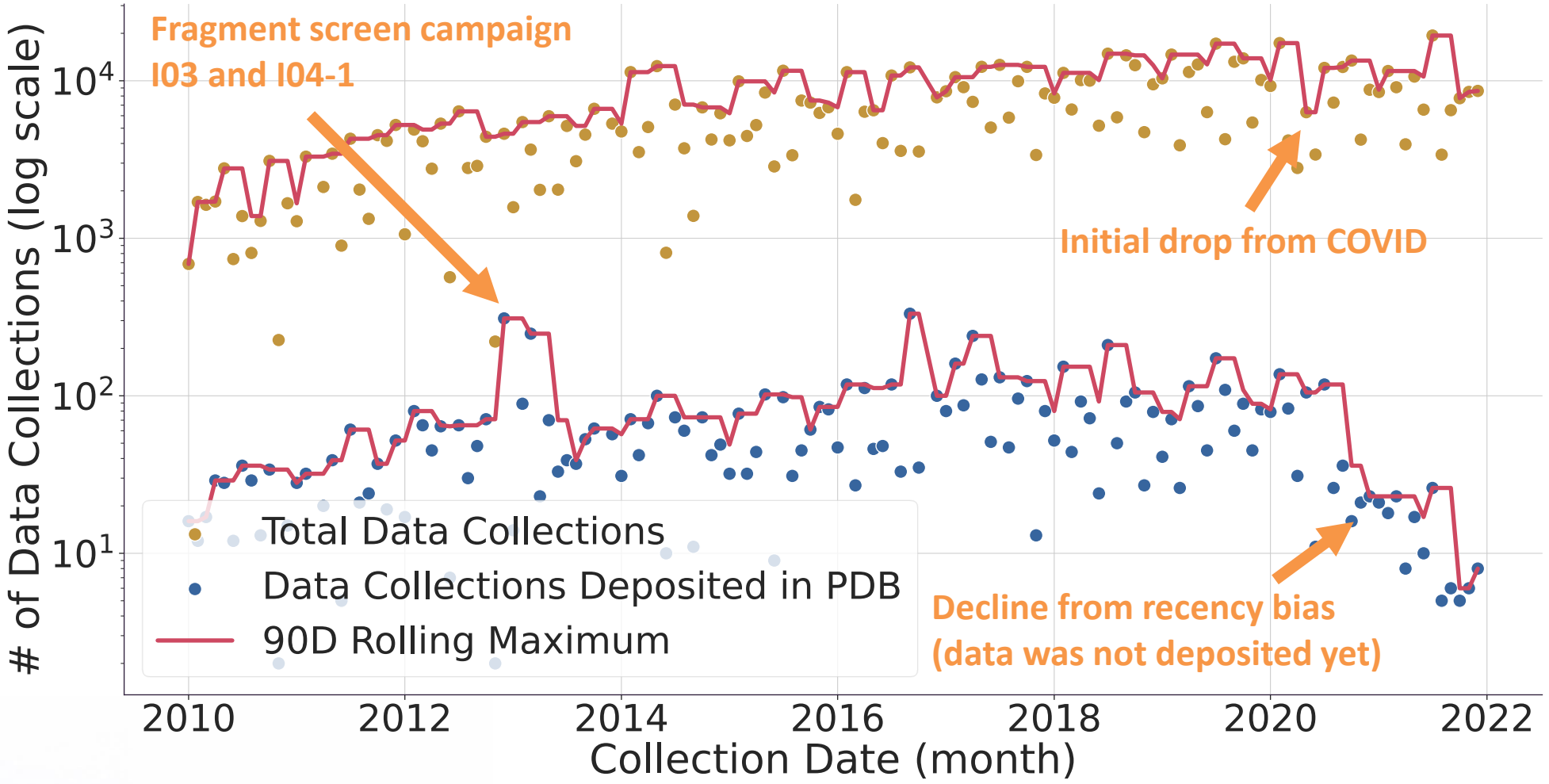
Vivian Li
Elliot Nelson
Richard Gildea
Ralf Flaig

Do this mistakes matter?

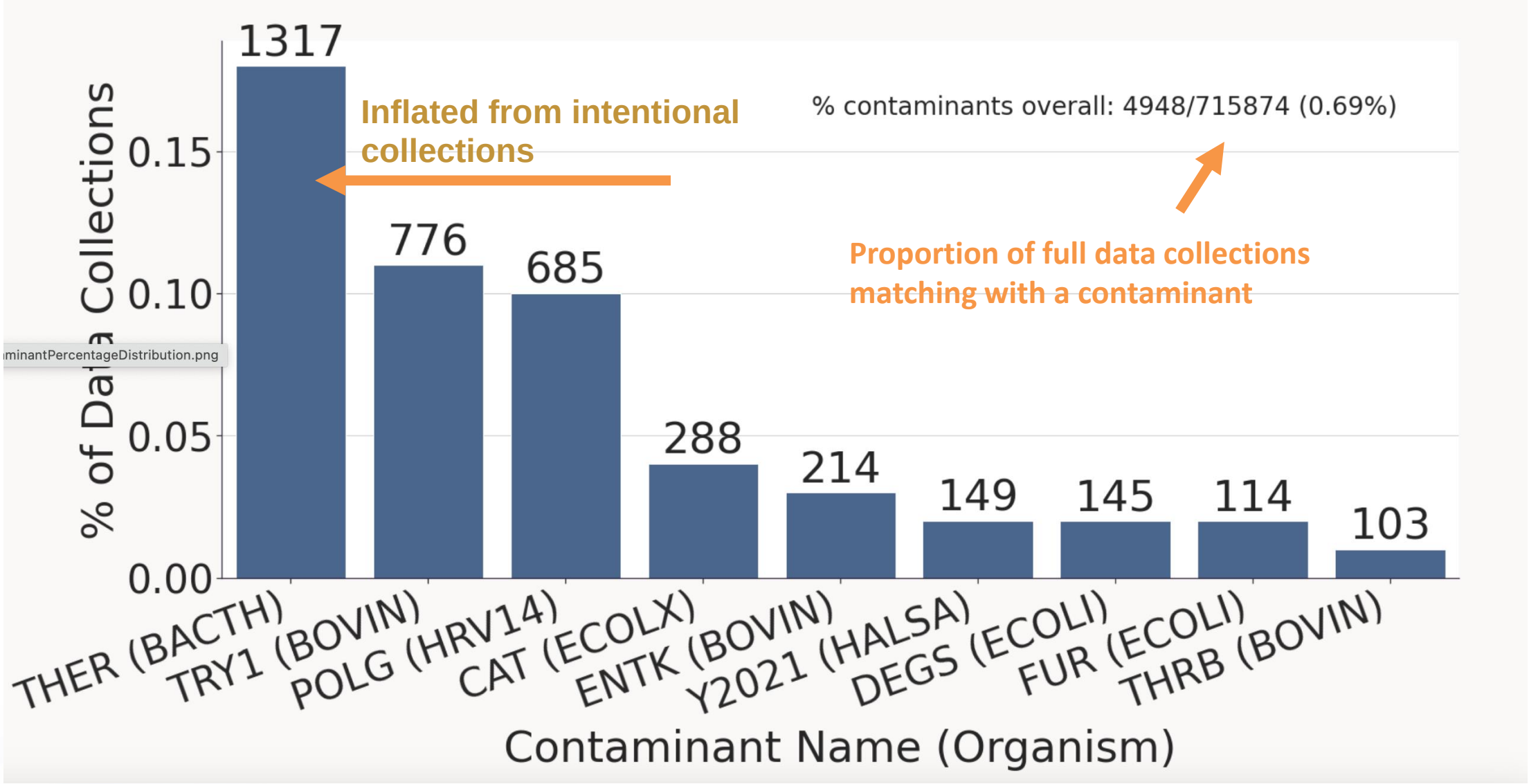


January 1st, 2010, to December 31st, 2020
9,637 PDB depositions
87.87% matched

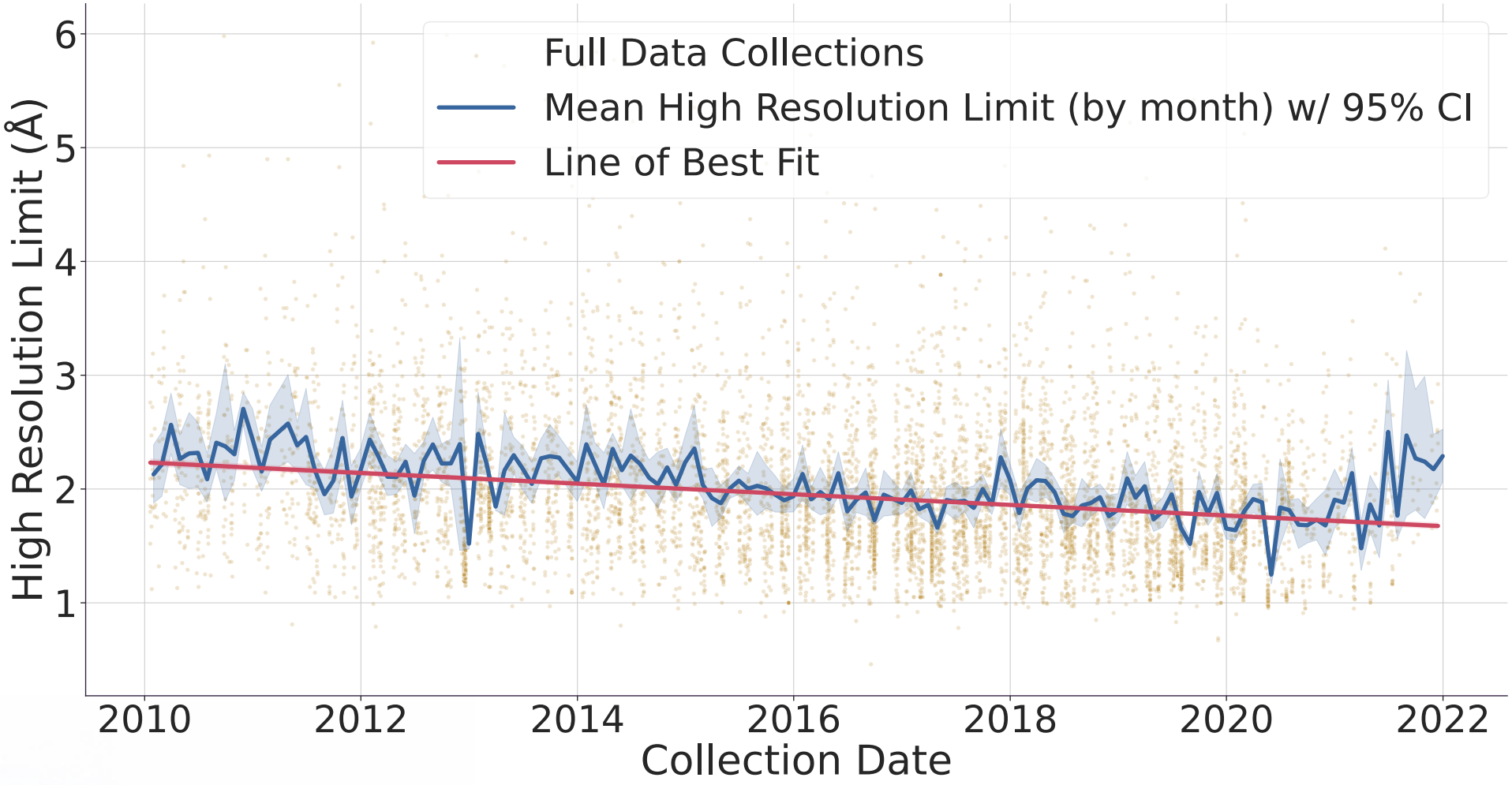
Differences between information in PDB and Data on Tape

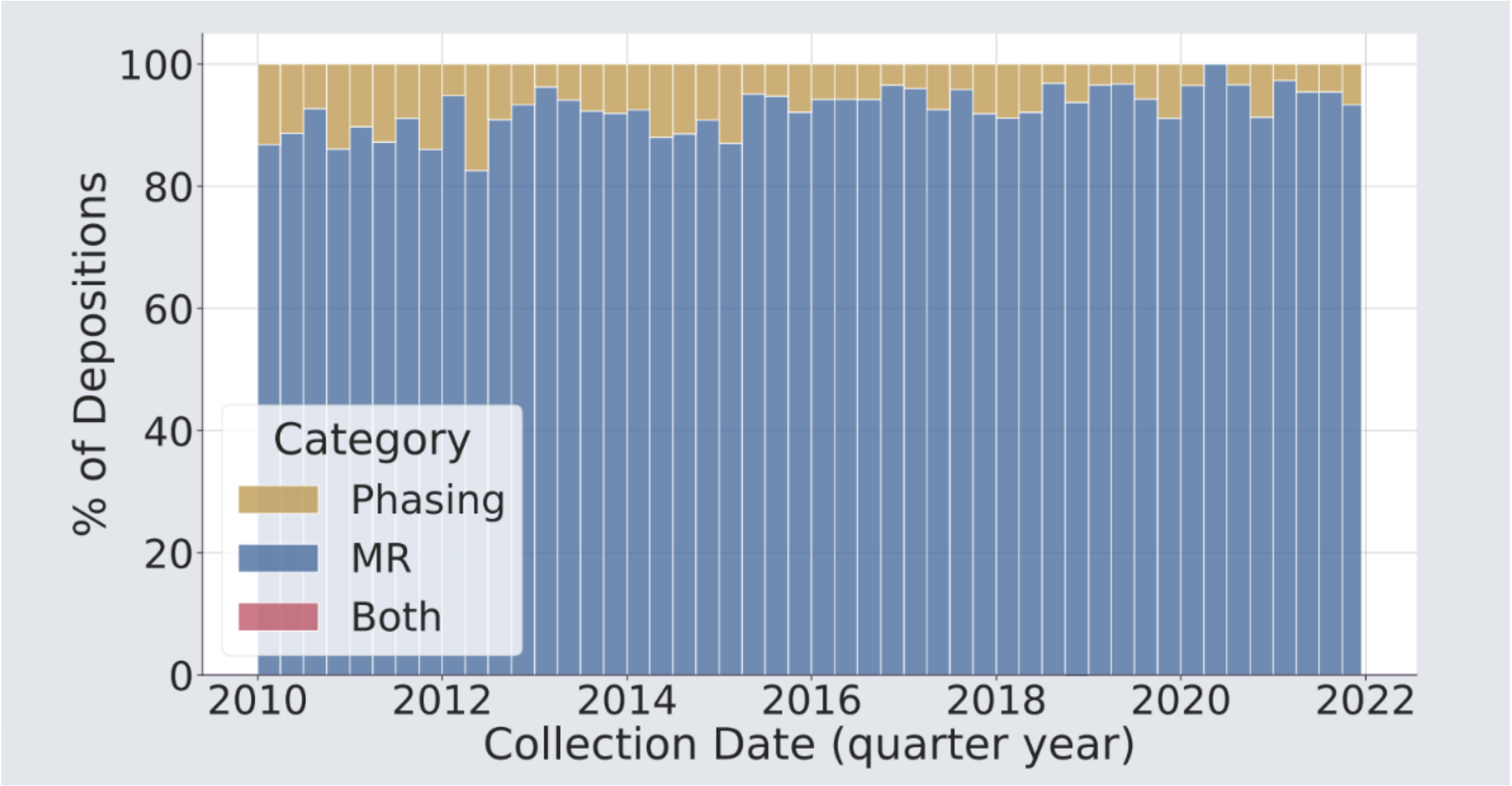


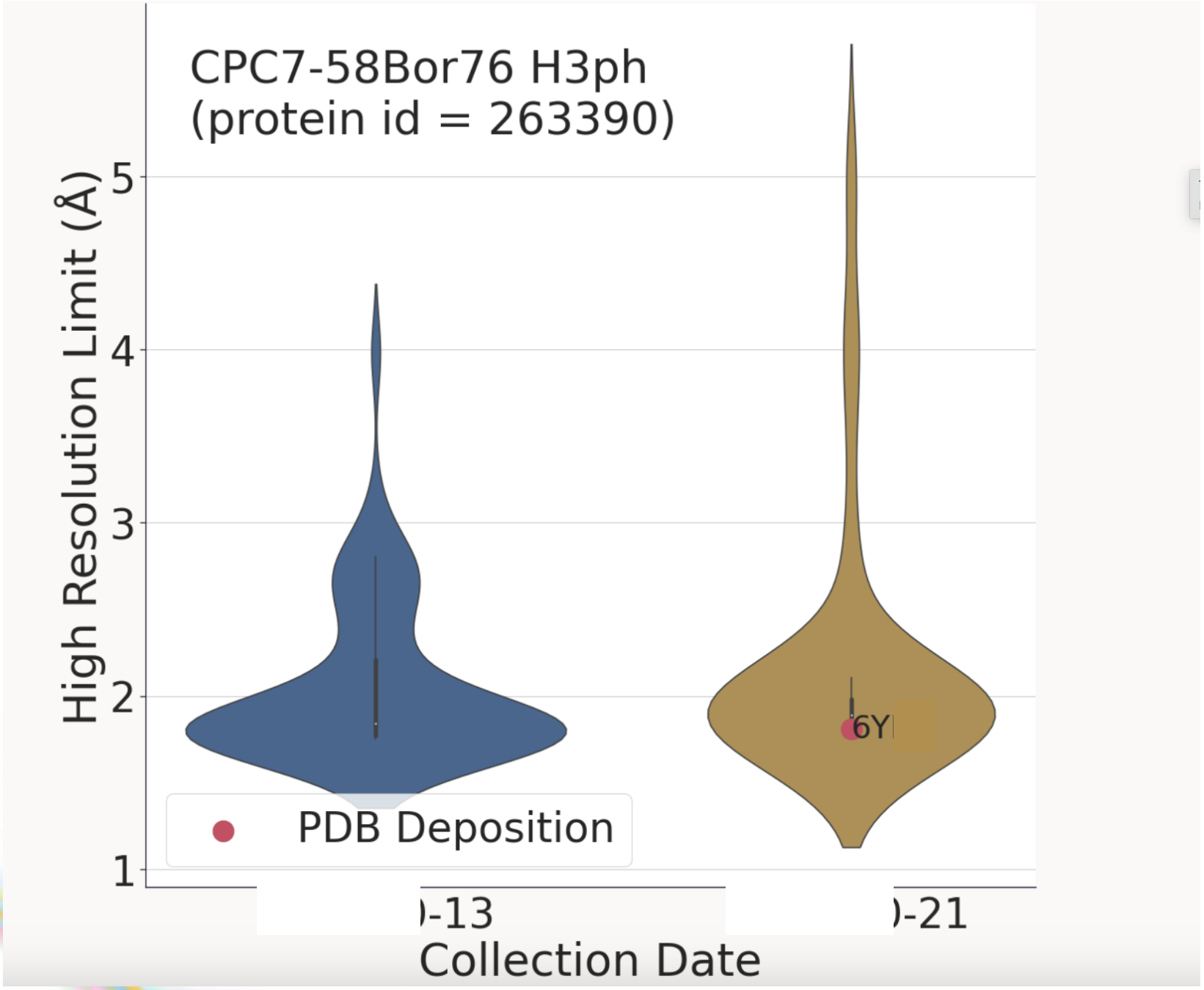
Full data collections collected and deposited over time

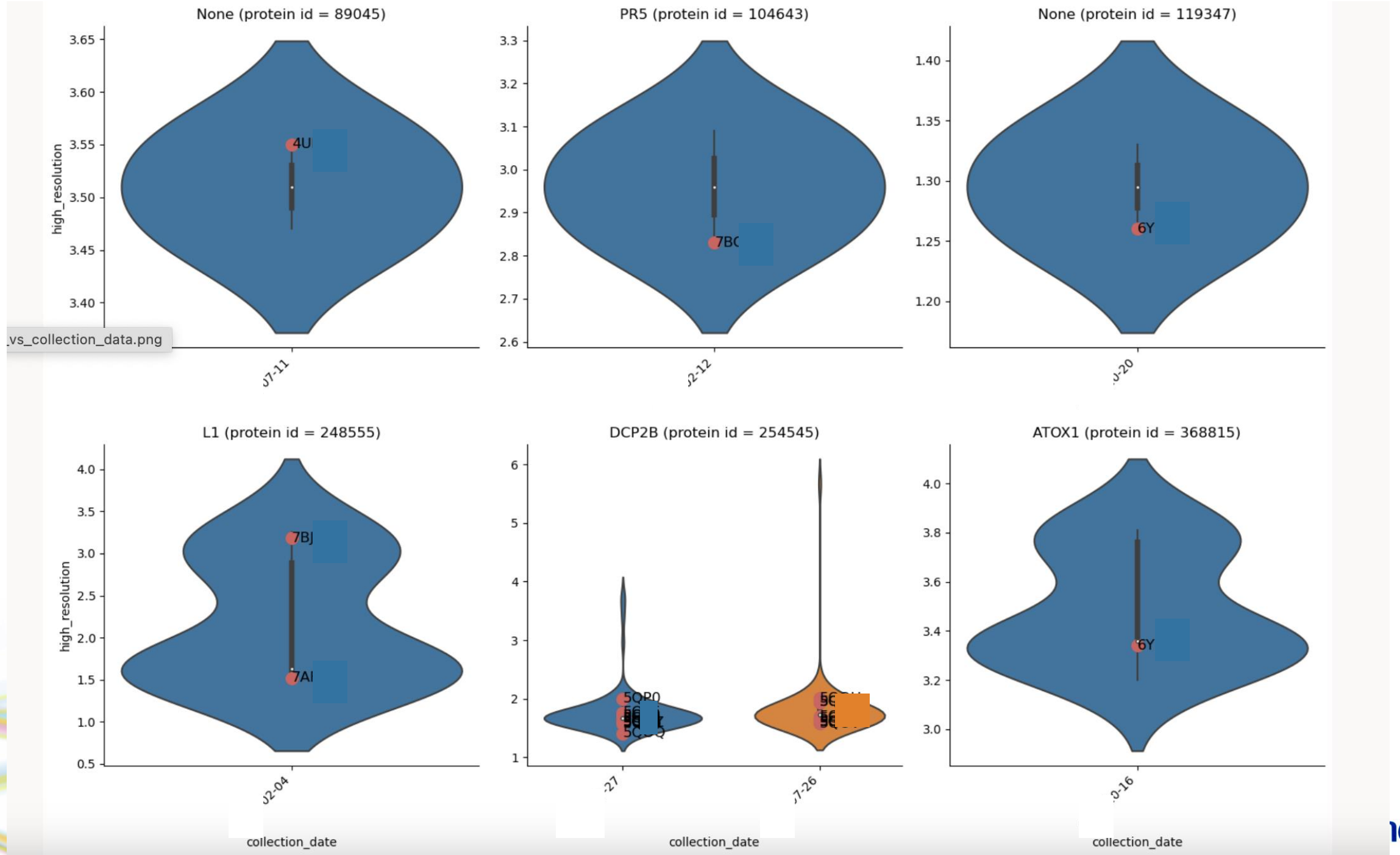


Distribution of contaminant proteins in full data collections from MX beamlines (I03, I04, I04-1, I24) collected between January 1, 2010, and January 1, 2022.
Full names from left to right are Thermolysin, Cationic trypsin, Genome polyprotein, Chloramphenicol acetyltransferase, Enteropeptidase, Putative heme-binding protein VNG_2021C, Serine endoprotease DegS, Ferric uptake regulation protein, and Prothrombin


Full Data Collections
Mean High Resolution Limit (by month) w/ 95% CI
Line of Best Fit
High Resolution Limit (Å)
1
2
3
4
5
6
2010
2012
2014
2016
2018
2020
2022
Collection Date


100
80
60
40
20
0
% of Depositions
Category
Phasing
MR
Both
2010
2012
2014
2016
2018
2020
2022
Collection Date (quarter year)


CPC7-58Bor76 H3ph
(protein id = 263390)
High Resolution Limit (Å)
5
4
3
2
1
6Y
PDB Deposition
-13
-21
Collection Date


None (protein id = 89045)
PR5 (protein id = 104643)
None (protein id = 119347)
L1 (protein id = 248555)
DCP2B (protein id = 254545)
ATOX1 (protein id = 368815)
high_resolution
collection_date
4U
7BC
6Y
7BJ
7AI
5QP0
_vs_collection_data.png

Acknowledgements

Project team

Vivian Li, Elliot Nelson, Richard Gildea, Ralf Flaig

Technical help with SQL/maria DB:

Karl Levik

Diamond,

For the funding