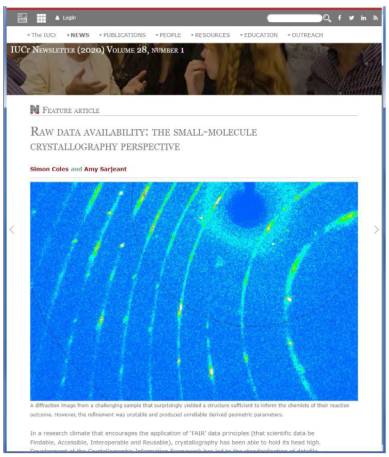
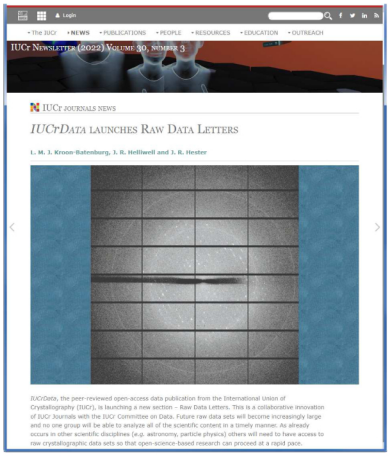
Raw diffraction data reuse: the good, the bad and the challenging
A Satellite Workshop to the XXVI IUCr Congress

Organised by Loes Kroon-Batenburg, Selina Storm, John R. Helliwell and Brian McMahon under the auspices of the IUCr Committee on Data

A major effort of the IUCr Diffraction Data Deposition Working Group (2011 to 2017) and now the IUCr Committee on Data since 2017 has been exploring the practicalities, the costs and benefits, and the opportunities for new crystallographic science arising from large-capacity data archives that have become available. We think it timely to hold a full-day workshop aimed at: (1) discussing current practices in raw data archival and sharing, (2) educating those who generate and deal in crystallographic data on best practices in data reuse in various categories of crystallographic science by leading experts, (3) offering a summing up, including the role of *IUCrData*'s new *Raw Data Letters*. We expect attendees to learn about the opportunities for raw data reuse, including the use of raw data as test data sets for machine learning, and to achieve an understanding of how to effectively archive their own raw data to maximise the potential for data sharing and reuse in the future. This workshop will explore in detail the successes and challenges in practice of raw data sharing and reuse. Being a full day, it will complement the microsymposia on 'Raw diffraction data reuse: warts and all' and 'Interoperability of Crystallographic Data and Databases' in the Congress itself; most importantly the microsymposia will allow for the usual submission of up to four abstracts from anywhere in the world whereas the workshop is principally made up of invited speakers. Furthermore, the microsymposia will highlight the importance of databases. Of the workshop and the microsymposia the keynote on the 'European Photon and Neutron Open Science Cloud' (by Andy Götz of ESRF) is the major highlight, making it the world leading effort of this consortium of more than ten European synchrotron and X-ray laser radiation sources with raw data management and sharing.



Included with this programme are:
[1] abstracts from the Keynote Lecture and Microsymposia sponsored by CommDat; and reprints of: [2] the *IUCr Newsletter* article by Simon Coles and Amy Sarjeant discussing the results of a community survey on the the desirability of archiving raw diffraction images in chemical crystallography structure determinations; [3] the *IUCrData* editorial announcing the new *Raw Data Letters* section; and [4] the first published *Raw Data Letters* articles.



Raw diffraction data reuse: the good, the bad and the challenging

A Satellite Workshop to the XXVI IUCr Congress

Timetable

Tuesday 22 August, Room 205, 8:20 am – 4:00 pm

Session 1: Facility and raw data providers. Part I

Chair: John R. Helliwell *Technical Co-chair:* Brian McMahon

8.20 am Opening remarks

8.30 am Scientific computing and data management at the Australian Synchrotron
Andreas Moll
Australian Synchrotron, Clayton, Victoria, Australia

8.50 am Managing and curating data flows at PETRA IV
Anton Barty
Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany

9.10 am DAPHNE4NFDI: DAta from PHoton and Neutron Experiments for NFDI
Bridget Murphy
Christian-Albrechts-Universität zu Kiel, Institut für Experimentelle und Angewandte Physik, Leibnitzstraße 19, 24118 Kiel, Germany

9.30 am Making the most of data from the ESRF
Andy Götz
ESRF, Grenoble, France

9.50 am Coffee break

Session 2: Facility and raw data providers. Part II

Chair: Selina Storm *Technical Co-chair:* Brian McMahon

10.10 am Scientific computing, data sharing and reuse at PSI
Alun Ashton
Paul Scherrer Institute, Villigen Switzerland

10.40 am X-tal Raw Data Archive (XRDa): A crystallographic raw diffraction image archive in Asia
Genji Kurisu
PDBj, Japan

11.00 am Handling of big data at the European XFEL
Fabio Dall'Antonia
European X-ray Free-electron Laser Facility GmbH, Schenefeld, Germany

11.20 am A subject specific repository for MX (proteindiffraction.org)
Wladek Minor
University of Virginia, Charlottesville, USA

11.40 am Processing data in serial crystallography on-the-fly: what kind of raw data do we want to store?
Alexandra Tolstikova
Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany

12.00 noon Lunch break

Session 3: Raw data reusers

3.1: Macromolecular crystallography

Chair: John R. Helliwell *Technical Co-chair:* James Hester

12.40 pm The raw, the cooked and the medium-rare: unmerged diffraction data as a rich source of opportunities for data re-use and improvements in methods and results
Gerard Bricogne
Global Phasing Limited, Cambridge, UK

1.00 pm Experiences with MX data reuse at Diamond
David Aragao
Diamond Light Source, UK

1.20 pm Raw data reuse: what it means for CCP4
Eugene Krissinel
CCP4, Rutherford Appleton Laboratory, UK

1.40 pm Reusing raw data for machine learning in MX
Melanie Vollmar
EMBL-EBI, Hinxton, UK

2.00 pm Tea break

3.2: Chemical crystallography

Chair: Loes Kroon-Batenburg *Technical Co-chair:* James Hester

2.20 pm Use of raw data for diffraction space visualization: What are we missing in an integrated HKL file?
Jim Britten
MAX Diffraction Facility, McMaster University, Hamilton, ON, Canada

2.40 pm The increasing diversity of small molecule data: can one size fit all?
Simon Coles
School of Chemistry, University of Southampton, UK

3.2: Powder diffraction

3.00 pm Powder diffraction raw data
Elena Boldyreva
Institute of Catalysis Russian Academy of Sciences, Novosibirsk, Russia

3.20 pm Powder diffraction data sharing and reuse: advantages and possible practical obstacles
Miguel A. G. Aranda
Universidad de Málaga, Spain

3.40 pm Summing up: the role of *IUCrData*'s new *Raw Data Letters* in serving all the above
Loes Kroon-Batenburg[1] / Selina Storm[2]
[1] *Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands*
[1] *Germany*

3.55 pm Close of Workshop

6.00 pm Congress Opening Ceremony

Abstracts

Scientific computing and data management at the Australian Synchrotron

A. Moll

ANSTO – Australian Synchrotron, 800 Blackburn Road, 3168, Clayton, Victoria, Australia
Email: andreasm@ansto.gov.au

The Australian Synchrotron is a division within ANSTO and one of Australia's premier research facilities. It produces powerful beams of light that are used to conduct research in many important areas including health and medical, food, environment, biotechnology, nanotechnology, energy, mining, agriculture, advanced materials and cultural heritage.

After 15 years of uninterrupted operation with the original ten experimental end stations, called beamlines, the Australian Synchrotron is currently entering an exciting new phase with the addition of eight new beamlines, including a new high-throughput Crystallography beamline. This created an opportunity for the Scientific Computing team to redesign the whole software stack from the ground up.

This presentation will take you on a journey of Scientific Computing at the Australian Synchrotron. You will learn how we employ modern, industry standard tools and architectures in a research environment in order to handle the large data throughput of modern detectors and provide the robustness our users expect from us. A particular focus will be on our use of cloud technologies, running on-premises, across our whole stack from hardware control to data processing on GPUs.

Keywords: Scientific computing; Data management; Australian Synchrotron

Managing and curating data flows at PETRA IV

Anton Barty

Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany
Email: anton.barty@desy.de

The upgrade of the existing PETRA-III synchrotron at DESY to a fourth-generation light source, PETRA-IV, includes not only an increase in brightness but also a new and expanded portfolio of instruments. A good fraction of the planned instruments will generate in excess of a petabyte of data per day during routine operation – a figure that already occurs today at some instruments. At such data rates, retaining all data on disk for 6 months and on tape for 10 years is no longer economically feasible. Instead, rapid analysis using validated pipelines will reduce archived data volumes while providing faster turnaround of results to users performing routine measurements.

The expectation that a majority of users will be experts in their own scientific fields but not necessarily experts in photon science data analysis highlights the need for the provision of high-level integrated data analysis and data management services to users. The PETRA-IV project envisages the provision of analytic services to the wider scientific community, wherein the timely provision of analysed data to users at conclusion of the measurement is essential. With data volumes exceeding the logistical capacity of most users, and especially non-expert users, these services must be provided by the facility or similar large scale research infrastructure. Provision must also be made for commercial measurement services on top of the same core infrastructure where data must be treated in confidence rather than being destined for open publication.

Integrated data analysis and data management services are required at the facility to support the full data life cycle from proposal through data taking and on to data analysis, publication, archiving. This includes collection of metadata such as persistent sample identifiers alongside the data, through to eventually making the data open for re-use by the wider community according to FAIR principles (FAIR data stands for Findable, Accessible, Interoperable and Reusable data). Already moving in this direction is the DAPHNE4NFDI national science research data infrastructure project – a cross-institutional project spanning 17 German research institutions currently addressing the topics of data management for photon and neutron science communities.

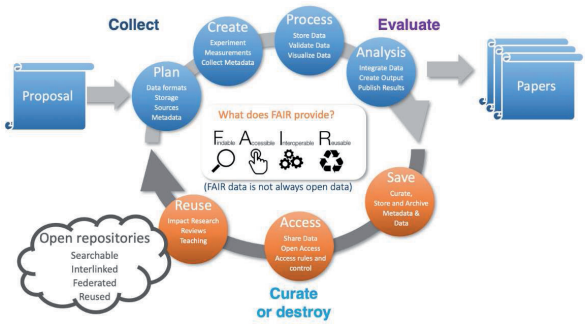


Fig. 1. The typical data life cycle at a photon source from proposal through experiment to analysis, curation and open data.

Thanks to Patrick Fuhrmann DESY for the figure, to the DAPHNE4NFDI collaboration, and to the PETRA-IV data team for contributions to the PETRA-IV data management plan.

DAPHNE4NFDI: DAta from PHoton and Neutron Experiments for NFDI

Anton Barty[1], Lisa Amelung[1], Christian Gutt[2], Astrid Schneidewind[3], Wiebke Lohstroh[4], Jan-Dierk Grunwaldt[5], Sebastian Busch[6], Tobias Unruh[7], Frank Schreiber[8] and Bridget Murphy[9]

[1] *DESY, Hamburg, Germany*
[2] *Universität Siegen, Siegen, Germany*
[3] *FZ Jülich, Jülich, Germany*
[4] *TU München, Munich, Germany*
[5] *Karlsruhe Institute of Technology, Karlsruhe, Germany*
[6] *Helmholtz-Zentrum Hereon, Geesthacht, Germany*
[7] *University Erlangen-Nurnberg, Nuremberg, Germany*
[8] *University of Tubingen, Tübingen, Germany*
[9] *Kiel University, Kiel, Germany*

The photon and neutron science community encompasses users from a broad range of scientific disciplines. With the advent of high-speed detectors and increasingly complex instrumentation, the user community faces a common need for high-level, rapid data analysis and the challenge of implementing research data management for increasingly large and complex datasets. The aim of DAPHNE4NFDI [1] is to create a comprehensive infrastructure to process research data from large scale photon and neutron infrastructures according to the FAIR principles (Findable, Accessible, Interoperable, Repeatable).

DAPHNE4NFDI brings together users representing key scientific application domains with the large-scale research facilities in photon and neutron science in order to advance the state of data management in the community. The overall goals of DAPHNE4NFDI are:

1. Improve metadata capture through consistent workflows supported by user-driven online logbooks that are linked to the data collection, thus enabling a richer capture of information about the experiments than is currently possible;

2. Establish a community repository of processed data, new reference databases and analysis code for published results, linked, where possible, to raw data sources, to sustainably improve access to research data and enable data and software re-use;

3. Develop, curate and deploy user-developed analysis software on facility computing infrastructure so that ordinary users can benefit from and repeat the analysis performed by leading power user groups through common data analysis portals.

This work is supported in the context of the work of the NFDI e.V. The consortium DAPHNE4NFDI is funded by the DFG – project number 460248799.

[1] Barty, Anton; Gutt, Christian; Lohstroh, Wiebke; Murphy, Bridget; Schneidewind, Astrid; Grunwaldt, Jan-Dierk; Schreiber, Frank; Busch, Sebastian; Unruh, Tobias *et al.* (2023). DAPHNE4NFDI – Consortium Proposal. *Zenodo*, DOI 10.5281/zenodo.8040605

Making the most of data from the ESRF

Andy Götz

European Synchrotron Radiation Facility (ESRF), 71 avenue des Martyrs, CS 40220, 38043 Grenoble Cedex 9, France

This talk will present the ESRF approach to data and metadata management for users and the community in general. The latest developments on laboratory information system to provide the right tools to structural biologists to make the most of their data will be presented. Recent developments include developing a flexible but powerful platform for processing data and providing the tools to view and evaluate the results efficiently. The ESRF data repository is being extended to include not only processed data for structural biology but also other domains for example tomography of human organs, fossils and materials science. The talk will conclude with a brief overview of how the ESRF data repository relates to the European PaN data commons and where the next big challenges lie to making data reuse a reality.

Scientific computing, data sharing and reuse at PSI

A. Ashton

Science IT Infrastructure and Services (AWI) Department, Paul Scherrer Institut, Switzerland
Email: alun.ashton@psi.ch

The Paul Scherrer Institute (PSI) develops, builds and operates complex large research facilities. The large scientific research facilities at PSI, such as the Swiss Light Source SLS, the free-electron X-ray laser SwissFEL, the SINQ neutron source, the SμS muon source and the Swiss research infrastructure for particle physics CHRISP, offer out-of-the-ordinary insights into the processes taking place in the interior of different substances and materials. These are the only such facilities within Switzerland, and some are the only ones in the world.

With new end-stations and detectors at SwissFEL and an upgrade to the SLS due for completion in 2025, the data volumes and computational requirements from photon experiments alone will undoubtedly exceed the current peak of a petabyte of raw data a week. Consequently, with support from initiatives including the ETH Domain program on Open Research Data, the Swiss Data Science Center (SDSC) and the Swiss National Supercomputing Centre (CSCS) and previously from the EU H2020 project ExPaNDS, PSI is already embarking on a holistic approach to handling data and the data lifecycle and finding novel ways to reduce, reuse and share experimental data at each stage of the lifecycle.

PSI recently expanded its focus areas and established a new research division: Scientific Computing, Theory and Data (SCD). In recognition of the importance and globally unique ensemble of large facilities at PSI, a keystone to the new division is supporting the operations and experiments with their increasing challenges and opportunities for a unique digital environment. This presentation will focus on the activities of SCD and its collaborators to deliver and improve scientific computing, data sharing and reuse at PSI.

X-tal Raw Data Archive (XRDa): A crystallographic raw diffraction image archive in Asia

G-J. Bekker[1] and G. Kurisu[1,2]

[1] *Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan*
[2] *Protein Research Foundation, Minoh, Osaka 565-8686, Japan*
Email: gkurisu@protein.osaka-u.ac.jp

The Protein Data Bank (PDB) is a public archive of atomic coordinates and crystallographic structure factors including professionally curated meta data. The PDB archive is maintained by the world-wide PDB (wwPDB), a global

organization founded in 2003 by RCSB PDB in the USA, PDBe in Europe, and PDBj in Japan, and later jointly managed with the Biological Magnetic Resonance Data Bank (BMRB) and the Electron Microscopy Data Bank (EMDB) as the wwPDB core members [1]. Quite recently, the wwPDB organization welcomed Protein Data Bank China (PDBc) as an Associate member of the wwPDB, and PDBc has started remote processing of some of the structures deposited to and allocated by PDBj. In addition to the PDB core archive, we, PDBj, collaborate with other wwPDB members to maintain the BMRB for experimental data from NMR experiments, and the EMDB for Coulomb potential maps from single-particle or sub-tomogram averaging in Cryo-Electron Microscopy. PDBj is the only wwPDB member who engages in the processing of data for all these three wwPDB core archives [2]. Although the above structural data in the core archives (PDB, BMRB, and EMDB) are actively collected and curated by the wwPDB, the raw image data that were the direct result of the primary experiments, and were used to determine the structures by macromolecular crystallography or cryo-EM microscopy are not collected by the wwPDB. For cryo-EM data, sub-members of the EMDB collect experimental raw micrographs or movies, and archive these in the Electron Microscopy Public Image Archive (EMPIAR). PDBj has been functioning as a local distributor of the EMPIAR archive since 2018, based on a bilateral agreement between EMBL-EBI and Institute for Protein Research, Osaka University. EMPIAR at PDBj (EMPIAR-PDBj) holds the exact same entries as EMPIAR at EMBL-EBI, and we have helped local depositors to transfer their large images/movies, thereby providing our own services (including deposition) through our original website for EMPIAR-PDBj (empiar.pdbj.org/).

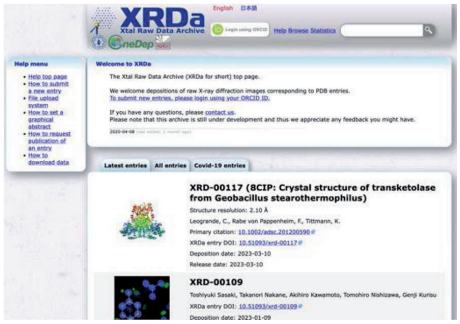


Fig. 1. The front page of XRDa as operated by PDBj (xrda.pdbj.org).

For macromolecular crystallography (MX) raw images, two major archives currently exist; Diamond Light Source in the UK, and the SBDB (SB Grid Data Bank), CXIDB (Coherent X-ray Image Data Bank) and IRRMC (Integrated Resource for Reproducibility in Macromolecular Crystallography) in the USA. However, up till now, no such archive for depositions from Asia has existed, and neither of the existing ones in the UK and the USA are wwPDB members. From 2020, PDBj has started our original diffraction archive named "X-tal Data Archive" (XRDa, xrda.pdbj.org) that securely stores the experimental diffraction images from Asian depositors. As a member of the wwPDB, we have streamlined deposition with the wwPDB's OneDep system. For depositors from Asia, after depositing their structural data to PDBj *via* wwPDB's OneDep system, their entry will be automatically linked to their ORCiD-ID in XRDa.

Depositors to XRDa can login using their ORCiD-ID, where any PDB IDs that have been registered in OneDep by them or their co-authors will be available. In addition, depositors can also submit their raw data before submitting their structures to the PDB (and link these afterwards), or submit raw data for structures not to be submitted to the PDB, *e.g.* for micro electron diffraction data of small molecules. Following login, users can easily deposit diffraction images *via* the "My entries" page. Once submitted, PDB-linked entries will enter a holding status and will be automatically co-released, while independent entries will be released immediately. Please feel free to deposit your diffraction images to PDBj.

XRDa is operated by PDBj and supported by the Platform Project for Supporting Drug Discovery and Life Science Research (BINDS) from AMED under Grant Number JP21am0101066.

[1] wwPDB consortium. (2019). *Nucleic Acids Research*, **47**, D520–D528.
[2] Bekker, G. J., Yokochi, M., Suzuki, H., Ikegawa, Y., Iwata, T., Kudou, T., Yura, K., Fujiwara, T., Kawabata, T. and Kurisu, G. (2022). *Protein Science* **31**, 173–186.

Keywords: PDB, BMRB, EMDB, EMPIAR, Xtal raw data

Handling of big data at the European XFEL

Fabio Dall'Antonia, Janusz Malka, Egor Sobolev, Philipp Schmidt, Krzysztof Wrona and Luca Gelisio

European X-ray Free-electron Laser Facility GmbH, Holzkoppel 4, 22869 Schenefeld, Germany

The European XFEL (EuXFEL) is a unique photon-source facility producing free-electron laser (FEL) pulses in the soft and hard X-ray regime, of extreme brightness and ultra-short duration. These are delivered at MHz repetition rate, enabling various experimental techniques and time-resolved setups. The seven scientific instruments mostly employ pixelized area detectors that can record up to 8,000 1-Mpx images per second.

These opportunities for research come at the cost of huge data volumes, which can reach a few PiB per beam-time, posing challenges for data storage and retention, as well as for data re-use purposes.

EuXFEL data collected with imagers requires facility services for the correction of pixel intensities. First steps of technique-specific data reduction such as azimuthal integration or crystallographic indexing are done by users remotely on facility resources as well, since download to local computers is not feasible. Currently we are in the process of updating the scientific data policy so as to account for data reduction prior to the long-term storage on disks, as well as for FAIR principles [1] of data management.

We are also developing facility services to apply specific data reduction techniques. For example, in case of serial femtosecond crystallography (SFX) [2] data can typically be reduced to only a few percent, sometimes even below 1%, of recorded frames since many FEL shots miss the sample crystals delivered by a liquid jet, leading to images without Bragg diffraction. In this case we are working on facility services that implement Bragg peak detection for automatic filtering procedures, either before data acquisition or at early stages of the offline correction and processing pipeline.

Concerning the re-use of open data, EuXFEL has got cloud-based services in the testing stage, which are initially employed for educational purposes but shall become a means for remote data analysis of selected and filtered data sets of each proposal after the embargo period.

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. *et al.* (2016). *Sci. Data* **3** 160018.
[2] Wiedorn, M. O., Oberthür, D., Bean, R. *et al.* (2018). *Nature Commun.* **9**, 4025.
Keywords: FEL, data reduction, SFX, FAIR data

A subject specific repository for MX (proteindiffraction.org)

W. Minor, M. Cymborowski and D. R. Cooper

Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA22903, USA

Preservation and public accessibility of primary experimental data are cornerstones necessary for the reproducibility of empirical sciences. We present the Integrated Resource for Reproducibility in Molecular Crystallography (IRRMC). In its first six years, several hundred crystallographers have deposited thousands of datasets representing more than 5,800 indexed diffraction experiments. We will present several examples of the crucial role that original diffraction data played in improving previously determined protein structures.
Keywords: reproducibility, data, drug Discovery

Processing data in serial crystallography on-the-fly: what kind of raw data do we want to store?

A. Tolstikova[1], T. A. White[1], T. Schoof[1], S. Yakubov[1], V. Mariani[2], A. Henkel[3], B. Klopprogge[3], A. Prester[3], S. De Graaf[3], M. Galchenkova[3], O. Yefanov[3], J. Meyer[1], G. Pompidor[1], J. Hannappel[1], D. Oberthuer[3], J. Hakanpää[1] and M. Gasthuber[1], A. Barty[1]

[1]*Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany*
[2]*Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, USA*
[3]*Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, Hamburg, Germany*
Email: alexandra.tolstikova@desy.de

Serial crystallography experiments involve collecting large amounts of diffraction patterns from individual crystals, resulting in terabytes or even petabytes of data. However, storing all this data has already become unsustainable,

and as facilities move to new detectors and faster acquisition rates, data rates continue to increase. One potential solution is to process data on-the-fly without writing it to disk. Recently, we have implemented a system for real-time data processing during serial crystallography experiments at the P11 beamline at PETRA III. Our pipeline, which uses CrystFEL software [1] and the ASAP::O data framework, can process frames from a 16-megapixel Dectris EIGER2 X detector at its maximum full-frame readout speed of 133 frames per second. The pipeline produces un-merged Bragg reflection intensities that can be directly scaled and merged for structure determination.

Processing serial crystallography data on-the-fly offers numerous advantages. It allows for real-time data quality control during the experiment, decreases the time spent between data collection and obtaining a final structure and can significantly reduce the amount of data that needs to be stored and managed. However, there are potential disadvantages and risks to not storing all raw data, such as losing the ability to revisit the data for reanalysis or to reproduce the results. Therefore, careful consideration is needed when deciding which data to store and which to discard. In this talk, we will discuss the challenges and opportunities of real-time data processing in serial crystallography and explore possible strategies for deciding which data to store.

[1] White T. A., Mariani V., Brehm W., Yefanov O., Barty A., Beyerlein K. R., Chervinskii F., Galli L., Gati C., Nakane T., Tolstikova A., Yamashita K., Yoon C. H., Diederichs K. & Chapman H. N. (2016). *J. Appl. Cryst.* **49**, 680–689.

Keywords: serial crystallography, real-time data processing, data reduction

The raw, the cooked and the medium-rare: unmerged diffraction data as a rich source of opportunities for data re-use and improvements in methods and results

G. Bricogne, C. Flensburg, R. H. Fogh, P. A. Keller, I. J. Tickle and C. Vonrhein

Global Phasing Limited, Sheraton House, Castle Park, Cambridge CB3 0AX, UK
Email: gb10@globalphasing.com

Deposition into the PDB of experimental diffraction data, in the form of merged intensities or of 'structure factor[amplitude]s', to accompany atomic models determined and/or refined from them, was made mandatory in 2008. This brought benefits that went well beyond the intended purpose of making the deposited models verifiable and correctable, in the form of an unanticipated 'virtuous circle' whereby deposited data fuelled improvements in refinement software that in turn enabled improvements to be made in the initially deposited models, from the same data. The need to manage the outcome of this continuous improvement process led to the introduction of a versioning mechanism into the archiving of atomic models by the PDB in 2017.

The creators of the Electron Density Server had already noted in 2004 that 'perhaps we should consider deposition of unmerged intensities or even raw diffraction images in the future' [1], without however anticipating the potential for a similar auto-catalytic cycle of simultaneous improvements in data reduction methods and in final structural results that could follow. This potential was later articulated in *e.g.* [2] in the following terms: 'those [merged] deposited X-ray data are only the best summary of sets of diffraction images according to the data-reduction programs and practices available at the time they were processed. Just like refinement software, those programs and practices are subject to continuing developments and improvements, especially in view of the current interest and efforts towards better understanding radiation damage during data collection and in taking it into account in the subsequent processing steps.'

Strong general support for the idea of archiving raw diffraction images, together with the recognition that this task was beyond the remit and resources of the PDB, has led over the past decade to the emergence of a delocalised infrastructure (whereby raw data storage and curation takes place at synchrotrons and other dedicated repositories while the PDB provides a capability to annotate entries with a DOI that points to the raw data storage location) that is a major topic in this Workshop.

Our own interest has been to document the scientific case for depositing and archiving suitably annotated unmerged diffraction data into the PDB, a goal achievable with modest storage requirements while already creating a standardised resource capable of feeding improvements in scaling and merging methods resulting in better refined models than those originally deposited. This goal is the focus of the current activities of the Subgroup on Data Collection and Processing of the PDBx/mmCIF Working Group of the wwPDB, in which we participate, to expand the mmCIF dictionary to support such extended deposition and archiving.

Crucially, unmerged data collected by the rotation method can preserve instrumental metadata about the image number and the detector position at which each diffraction spot was located and integrated, providing a broader decision-making scope over the way it is incorporated into the scaling and merging process. This opens a wide range of possibilities for improving any initially performed scaling/merging steps and for extraction of further data. We will present examples touching upon the following areas:

1. production of full validated data quality metrics that are often incomplete or inconsistent in deposited merged data;
2. detection of problematic images and image ranges, and remediation by their selective exclusion from scaling/merging;
3. anisotropic diffraction limit analysis (or re-analysis) with STARANISO, if not already performed;
4. extraction of previously unexploited anomalous signal and computation of anomalous difference Fourier maps;
5. 'reflection auditing' by tracing outliers detected at the refinement stage back to their unmerged contributors in terms of specific image numbers and detector positions, thus diagnosing ice rings, poor beamstop masks, angular overlaps, *etc.*;
6. detection of radiation damage *via* $F_{\rm early} - F_{\rm late}$ maps; adapting parametrisation to patterns of structural radiation damage.

We are grateful to the PDBx/mmCIF Subgroup on Data Collection and Processing, especially Aaron Brewster, Ezra Peisach, Stephen Burley and David Waterman, for a stimulating collaboration that provided a context for presenting these investigations.

[1] Kleywegt, G. J., Harris, M. R., Zou, J.-Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
[2] Joosten, R. P., Womack, T., Vriend, G. & Bricogne, G. (2009). *Acta Cryst.* D**65**, 176–185.

Keywords: raw diffraction data, unmerged diffraction data, scaling and merging methods

Note: the same abstract has been submitted to Microsymposium A118: Raw Diffraction Data Reuse: Warts and All (see pp. 19–25). Different aspects of the topic will be discussed across both sessions.

Experiences with MX data reuse at Diamond

D. Aragao, V. Li, S. Collins, G. Winter, R. Gildea, E. Nelson and R. Flaig

Diamond Light Source, Harwell Science and Innovation Campus, Chilton, Didcot, OX11 0DE, UK
Email: David.Aragao@diamond.ac.uk

Diamond Light Source (DLS) operates seven beamlines for macromolecular crystallography [1]. These instruments together with others located at similar facilities worldwide record tens of thousands of datasets every week. These are generally left accessible on disk for the industrial partners or academic researchers to analyse. Afterwards, the data are normally stored in a tape system or deleted altogether. With the advent of modern detectors all DLS MX beamlines write HDF5 compressed data and follow the NXmx gold standard [2]. The size of the files and availability of a standard mean these data sets are portable and can be reduced successfully without any extra information. These are critical for data usefulness in the decades to come. On the business end of things, DLS has changed its policy from April 2019 so that academic funded data collected thereafter would be made public from March 2022 making an effective maximum of 3 years embargo with data owners being told 6 months in advance of such. However, DLS has not yet released any data but still has the intention to do it. As for data prior to April 2019, DLS has not yet deleted any data, but such data would be only available at the explicit request of the data owner. To further explore ideas of what can be done with the availability of raw data we present an I04 beamline [3-4] 3-month summer student project where some analysis of the current data in storage, a comparison with the published structures in the PDB and a basic attempt to infer long-term data trends was made.

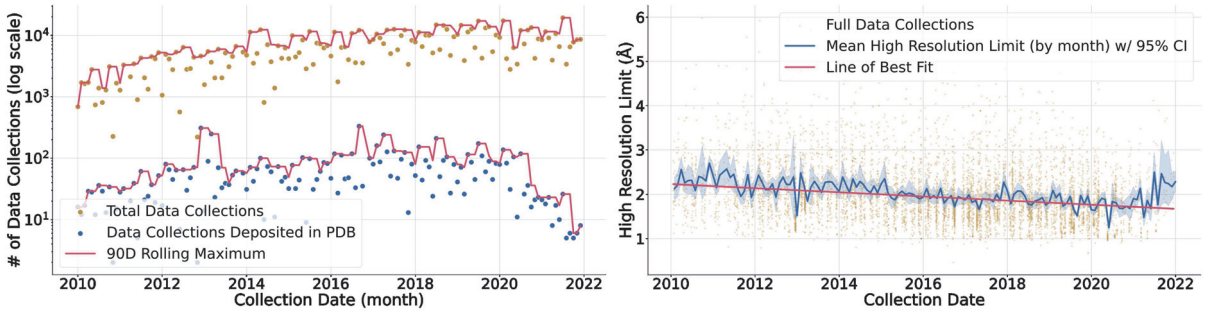


Fig. 1. (left) Evolution of the number of data collected (gold dots) and protein structure depositions (blue dots). Lines represent a 90-day rolling window to smooth the DLS shutdowns where no data is collected. The *y*-axis of the graph is logarithmic to allow to compare the 1% deposited structures *vs* the $> 10^3$ datasets collected in the same period. (right) Evolution of protein structure high-resolution limit over time with mean and best fit line. The faded gold dots represent individual full data collections, while the blue line is the average high-resolution limit for each month, surrounded by a 95% confidence interval. The red line is a linear regression line of best fit, and we can see that the high-resolution limit has been getting better over time, improving from around 2.2 ångströms to around 1.7 ångströms.

We thank Karl Levy for valuable input and suggestions on how to query the ISPYB database and Diamond Light Source undergraduate summer placement for funding a student at the I04 beamline.

[1] https://www.diamond.ac.uk/Instruments/Mx.html
[2] H.J. Bernstein *et al.* (2020). *IUCrJ*, **5**, 784–792
[3] R. Flaig *et al.* (2017). *Acta Cryst.* A**73**, a71
[4] https://www.diamond.ac.uk/Instruments/Mx/I04.html

Keywords: raw data, open data, data reuse

Raw data reuse: what it means for CCP4

Eugene Krissinel

CCP4, Rutherford Appleton Laboratory, UKRI STFC, Harwell Campus, Didcot, Oxfordshire, OX11 0QX UK
Email: eugene.krissinel@stfc.ac.uk

Collaborative Computational Project Number 4 in Macromolecular Crystallography (CCP4 UK) has a mission to distribute, develop and facilitate development of crystallographic software for all stages of the structure determination pipeline, from raw image processing to phasing, refinement, completion, validation and deposition. Over the 44-year history of the Project, crystallographic software underwent a series of evolutionary changes, caused by advances in theory, sample preparation techniques, quality and properties of raw data.

MX is often regarded as a technique with limited reproducibility, which suggests high importance of data retention in the field. For many years, the Protein Data Bank (PDB) collected only end results of interpretation of raw data, the atomic coordinates, leaving no scope for revisiting structure determination in future. The situation improved in 1999 and further in 2020 when, respectively, merged and unmerged data became available for deposition. Deposition of unprocessed, raw data is a natural next step in this direction, which is rather demanding on the storage side and is actively discussed in value for cost terms. We would like to bring the software effect into consideration.

Notably, there is no single established format for raw data in MX, and in addition, raw data should be processed with instrument/detector specifics (see Reference [1] as an example). Data processing software, such as *XDS* [2], *HKL* [3], *Mosflm* [4], *d*TREK* [5], *DIALS* [6], include extendable sets of routines or plugins for dealing with the variety of formats. Commitment to raw data retention and reusability effectively means commitment to maintaining data processing software, format plugins, and associated beamline metadata forever. This is a significant challenge as software ages faster than data and usually gets retired exactly for maintainability reasons. For example, *Mosflm* and *d*TREK* are effectively in sunset mode, and the newest development in the field, *DIALS*, is not supposed to work with all older formats. Most probably, even if raw data were kept for all PDB entries from day zero, we would not be able to use the oldest datasets today. A possible solution to this problem may be in introducing a “storage”

format, but in any case, reuse and storage of raw data cannot be detached from running the corresponding software project.

STFC and Diamond synchrotron show an example in maintaining raw data. In 40 days after collection, data are pushed from beamlines to long-term storage, from where they can be downloaded years later. CCP4 is in a good position to facilitate reuse of such data by setting links between data facilities at Diamond, STFC/SCD and CCP4 Cloud [7]. This matches well with introducing CCP4 Cloud Archive facility in January 2023, where completed structure determination projects can be deposited, so that not only the project data and metadata but also the way the structure was solved can be retained; archived projects can be revisited and revised in future.

Linking this facility with raw data storage and PDB entry would provide a fully accountable data line for MX. This bears obvious benefits for researchers and makes a foundation for the efficient reuse of collected data, also helping to maximise longevity and robustness of data-handling software. Works have begun in this direction, and much will depend on community take up and feedback.

CCP4 is funded by BBSRC UK (Grant BB/S006974/1) and industrial licencing.

[1] http://www.globalphasing.com/autoproc/wiki/index.cgi?BeamlineSettings
[2] Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
[3] Minor, W., Cymborowski, M., Otwinowski, Z. and Chruszcz, M. (2006). *Acta Cryst.* D**62**, 859–866.
[4] Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R. and Leslie, A.G.W. (2011). *Acta Cryst.* D**67**, 271–281.
[5] Pflugrath, J.W. (1999). *Acta Cryst.* D**55**, 1718–1725.
[6] Winter, G., Waterman, D.G., Parkhurst, J.M. *et al.* (2018). *Acta Cryst.* D**74**, 85–97.
[7] Krissinel, E., Lebedev, A., Uski, V., Ballard, C. *et al.* (2022). *Acta Cryst.* D**78**, 1079–1089.

Keywords: crystallographic computing, raw image data, structure determination

Reusing raw data for machine learning in MX

M. Vollmar[1] and G. Evans[2,3]

[1] *EMBL-EBI, Hinxton, United Kingdom*
[2] *Rosalind Franklin Institute, Harwell, United Kingdom*
[3] *Diamond Light Source Ltd, Harwell, United Kingdom*
Email: melaniev@ebi.ac.uk

Large quantities of raw diffraction data from protein crystals are collected at synchrotron facilities and in-house X-ray sources every day. The vast majority of this data never yields a protein structure and never leaves the local data storage. Over the last five years there has been a steady increase in interest in the development of machine learning and artificial intelligence models in structural biology. To train any predictive model for high-quality predictions, large quantities of data are required, preferably standardised, curated and labelled.

However, the closed state of data storage, *i.e.* the data is only found on local storage, makes accessing raw diffraction data challenging for anyone who wants to use such data for developing machine learning and artificial intelligence models. Additionally, if raw diffraction data has been made publicly available it usually only represents a certain type of data, namely the one that resulted in successful structure solution, while any diffraction data that did not yield an atomic model remains hidden. Contacting data holders to gain access also often brings the challenge of tracing and finding the raw data on local storage depending on how well data and file management are handled within a facility or research group.

Here, we provide a retrospective analysis and share our experiences when developing a machine learning model using raw diffraction data [1]. We describe the challenges in finding suitable data, difficulties accessing that data, efforts needed to trace data locally and, finally, how a well-defined set of raw diffraction data was used to train a machine learning model.

We thank Arnaud Baslé, Dominic Jaques, Garib Murshudov, James Parkhurst and David G. Waterman for their contributions and vivid discussions when developing a machine learning model as described in [1].

[1] Vollmar, M., Parkhurst, J., Jaques, D., Baslé, A., Murshudov, G., Waterman, D. and Evans, G. (2020). *IUCrJ*, **7**, 342–354.

Keywords: machine learning, raw diffraction data, open/closed access

Use of raw data for diffraction space visualization: What are we missing in an integrated HKL file?

Jim Britten

MAX Diffraction Facility, McMaster University, Hamilton, ON, Canada
Email: britten@mcmaster.ca

For many years chemists, biochemists and physicists have been using area detectors to collect 3D diffraction data from crystals. Engineers are collecting 3D data for texture and residual stress analyses. Various software packages analyse the frames and the derived data is used to solve the problem at hand.

In this presentation we will use 3D reciprocal space visualization software [1] to look carefully at examples of these types of data. Very often we can find evidence of unexpected features that have been ignored or missed by the data processing software. In some cases, we can do further processing of the data based on the new-found information. In other cases, we need to flag the data set as a candidate for future analysis when the appropriate software becomes available. It may even trigger the development of new software.

Diffraction space is more complicated than we often imagine and contains unmined information about our samples. Scattering from aperiodic crystals is an obvious example. The value of preserving raw 3D diffraction data cannot be underestimated.

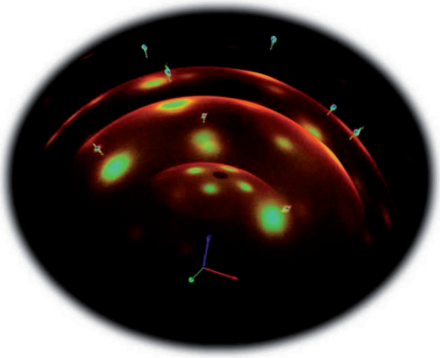


Fig. 1. Diffraction from a highly oriented nanowire film on a single crystal substrate.

[1] *MAX3D* reciprocal space visualization software package (https://rhpcs.mcmaster.ca/ guanw/max3d/MAX3D-5.0-2021.10.12.zip)

Keywords: reciprocal space, visualization, *MAX3D*

The increasing diversity of small molecule data: can one size fit all?

Simon J. Coles

School of Chemistry, Faculty of Engineering and Physical Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK
Email: S.J.Coles@soton.ac.uk

Today's accepted approaches to handling chemical crystallography data have largely been established in the 'boom period' of crystal structure analysis, that is the late 1990s and early 2000s as CCD area detectors took hold and data volumes increased significantly. Chemical crystallography is facing another change, with a range of alternative structure determination methods becoming viable and dynamic crystallography seeing more widespread use. Some examples of initiatives from our laboratory illustrate the nature of this imminent expansion in our field.

3D-Electron Diffraction (3D-ED) is set to significantly impact on small-molecule crystal structure analysis with the introduction of new dedicated instrumentation that will dramatically increase the volume of results generated and lift the technique from research in itself, to being generally applicable and providing a widespread service. 3D-ED is generating some truly amazing results, producing structures from nano crystallites traditionally considered as powders, on materials that normally would never have been applicable to single-crystal analysis. However, the nature of the experiment is challenging and invariably datasets from several crystallites must be merged to maximise

the completeness of data, with the result that structures do not meet the same quality standards as we have come to expect from X-ray single-crystal analysis. Other similarly emergent structure determination techniques applied to particular problems, such as NMR Crystallography and XFEL studies, present wonderful opportunities but come with the same data quality problem.

The Crystal Sponge technique enables the uncrystallisable to be crystallised. Compounds that do not crystallise well, or at all, or that can only be synthesised in minute amounts can be soaked into a crystalline porous material and the composite host+guest structure determined. This provides a molecular structure that can have great value for synthetic chemists for characterisation or confirmation of product. However, the experimental technique is variable in that molecules can arrange in the sponge in different ways significantly affected by soaking conditions and this leads to diffuse diffraction, disorder and lower quality results. Similarly dynamic crystallography, that is structures under change mediated by *e.g.* temperature, pressure, gas adsorption, electric current, also suffers from these effects.

These exciting advances are set against the backdrop of traditional X-ray crystal structure analysis, with $>$ 100 years of enhancing instrumentation, $>$ 50 years of collecting results into databases, 40+ years of trusted common refinement processes, 30 years of standards and 20+ years of validation tools. So, the established processes, metrics, *etc.* for small-molecule crystallography provide a well-established and trusted 'quality framework' for our results. This means that the small-molecule crystallography community now caters very well for the validation and quality control of relatively routine structures as part of the checking and publication process. However, this quality framework doesn't cater well for these exciting new frontiers of chemical crystallography in the sense that results are deemed to be of a lower quality. But the results of these experiments drive and underpin investigations in ways that could never have happened before and with a comparable accuracy to the gold standard of single-crystal X-ray analysis.

Being able to answer questions such as 'what is the compound I have made?', 'what is this reaction by-product?', 'how has my structure changed?', 'how does this material manifest these properties?', particularly for materials that are not ideally crystalline, can be crucial to further the progress of research.

Clearly these are strong examples that extend the community discussions around making raw chemical crystallography data available [1]. But how can we balance this current contrast between well established and emergent techniques? Firstly, it is necessary to consider extending the current quality framework and secondly it is imperative to make raw data available alongside the results from these new techniques. This talk will present the concept of 'structure grading' as an indicator of what claims can be made based on a particular result. These claims, and therefore the structure grading, should be backed up by the raw data – particularly in the case of emergent techniques where it is highly likely that methods will improve and so a better result can be derived in the future from the original data. The talk will therefore also consider how it can be shown that the best possible result has been obtained from the raw data, or indications can be provided that declare that there is room for future improvement.

[1] When should small molecule crystallographers publish raw diffraction data? (2021). Twenty-Fifth Congress and General Assembly of the International Union of Crystallography, https://www.iucr.org/resources/data/commdat/prague-workshop-cx.

Keywords: raw data, chemical crystallography, structural chemistry

Powder diffraction raw data

Nicola P.M. Casati[1] and Elena V. Boldyreva[2,3]

[1] *Paul Scherrer Institut, WLGA/229, Forschungsstrasse 111, 5232 Villigen PSI, Schweiz, E-Mail: nicola.casati@psi.ch*
[2] *Boreskov Institute of Catalysis Russian Academy of Sciences, Lavrentieva ave., 5, Novosibirsk, 630090, Russia, E-mail: eboldyreva@catalysis.ru*
[3] *Novosibirsk State University, Pirogova str., 2, Novosibirsk, 630090, Russia, E-mail: e.boldyreva@g.nsu.ru*

Powder Diffraction is a valuable tool when studying a statistical number of objects, *e.g.* in quantification or probing chemical reactions or phase transitions. It also provides information on many important compounds and materials that cannot be prepared as single crystals, or that, while starting as single crystals, become polycrystalline due to structural transformations on variations of temperature, pressure, irradiation *etc.*

Data analysis for powder diffraction is normally not only less straightforward then for single crystals, but is also more prone to ending in false minima and getting wrongly interpreted. A more delicate and possibly robust approach is needed to retrieve the correct information. It is therefore desirable, at times, to go back to original data and find

out the origin of specific features in the data themselves, whether coming from the setup, from mishandling or from the experiment itself. This helps not only avoiding mistakes but also extracting the maximum information, which at times can only be viewed *a posteriori* after more knowledge on the system has been acquired. In this context correct data handling is pivotal to being able to successfully look back at 'old' but still valuable patterns.

In the present contribution, we aim to define the important aspects in regulating powder diffraction raw data, including the level at which 'raw is raw' and the metadata that needs to be included, to make sure enough information is kept. This, to ensure it is possible to look back successfully at previous data and/or correct problematic behaviour. The discussion will involve both 1D and 2D detector datasets, including very large ones (several thousand patterns). As one of the examples, matching several aspects, we use a high-pressure study of phase transitions in L-Serine. We illustrate, how the raw data collected in 2006 [1] could be used almost 10 years later, in order to get new information from these data, viewed through the new experience obtained in 2015 [2].

EVB acknowledges funding from the Ministry of Science and Education RF (project AAAA-A21-121011390011-4).

[1] Boldyreva, E. V., Sowa, H., Seryotkin, Y. V., Drebushchak, T. N., Ahsbahs, H., Chernyshev, V. & Dmitriev, V. (2006). Pressure-induced phase transitions in crystalline L-serine studied by single-crystal and high-resolution powder X-ray diffraction. *Chemical Physics Letters*, **429**(4-6), 474-478.
[2] Fisch, M., Lanza, A., Boldyreva, E., Macchi, P. & Casati, N. (2015). Kinetic control of high-pressure solid-state phase transitions: A Case Study on L-serine. *The Journal of Physical Chemistry C*, **119**(32), 18611-18617.

Powder diffraction data sharing and reuse: advantages and possible practical obstacles

Miguel A. G. Aranda

Universidad de Málaga, 29071-Málaga, Spain
Email: g_aranda@uma.es

Scientific data in our crystallographic community can be classified, in broad terms, in three large categories: raw, reduced and derived data. On the one hand and for decades, IUCr has been and is being very active in promoting the sharing of reduced and derived data in independently verified databases. The final results, in narrative style, are also shared in the scientific journals. On the other hand, the need for raw data sharing is clearly increasing, being nowadays technically feasible and likely cost-effective.

Within the crystallography field, the powder diffraction (PD) community is a subgroup dealing with several goals, mainly (1) average crystal structure determination; (2) quantitative phase analyses; (3) microstructural analyses; and (4) local structure determination and quantitative analyses of nanocrystalline materials. It should be noted that many PD users are not directly associated with crystallography but with material science, solid-state chemistry and physics, *etc.*, some practices being different in different fields. For PD, derived data for objectives (2) and (3) and to a large extent (4) cannot be incorporated in current 'standard' (independently verified) databases. Therefore, and in my opinion, the need for sharing raw PD data is even more compelling than that of sharing raw single-crystal diffraction data.

In order to ensure that raw powder diffraction data sharing is useful, the methodology has to be robust. From the computing point of view, the shared data must be findable, accessible, interoperable and reusable — *i.e.* comply with FAIR standards. However, this is necessary but not sufficient. On the other hand, and from the involved scientific community point of view, the shared data must have sufficient quality and their quantitative reuse should be relatively easy.

Some possible benefits of sharing powder diffraction raw data were discussed in a previous publication (*J. Appl. Cryst.* (2018), **51**, 1739–1744). In this communication, I will further elaborate on the benefits but mainly on some practical obstacles to be addressed. For powder diffraction data from point detectors, the sharing seems to be straightforward. However, for powder diffraction data taken from 2D detectors, this is not the case. It is noted that both correction and integration steps have choices that need to be unified. This is a challenging task that needs to be undertaken.

Keywords: raw data, FAIR standard, data reduction, data correction, data integration

Abstracts and posters from other sessions

Keynote Lecture

Friday 25 August 2023, Room 203/204, 10:20 am – 11:10am

Europe's Photon and Neutron Open Science Clouds for Raw and Processed Data: Aims and Achievements to Date

Andy Götz

European Synchrotron Radiation Facility (ESRF), 71 avenue des Martyrs, CS 40220, 38043 Grenoble Cedex 9, France

The Photon and Neutron Open Science Cloud (PaNOSC) and the European Open Science Cloud Photon and Neutron Data Services (ExPaNDS) are two European Community financed projects comprised of 8 synchrotrons, 2 FELs, 3 laser and 4 neutron central facilities, which have been established in Europe to facilitate Open Science, an enhanced science methodology. Through such a coordination there can be: (i) an increased fraction of published research by release of raw data to the public after a 3 year embargo period; (ii) in cases of an agreed cooperation between specific research communities where measurements are now outpacing the capabilities of individual teams to analyse the raw data; (iii) for more conventional 'single research team' science publication can be underpinned by a single digital object identifier (DOI) to the appropriate dataset held in the facility data archive, without need for raw dataset transfer to the home university. The implementation of data archiving is facilitated by great increases in tape storage capabilities. There are 30 case studies that document the opportunities of these projects including open science and reproducibility of science. Crystallography, diffraction and scattering experiments certainly yield big data flows at the relevant ESRF EBS beamlines with X-ray imaging, EM and SSX beamlines reaching massive data flow levels. All these benefits must be balanced with carbon footprint of such a data archive in today's world. The carbon footprint of archiving petabytes of data is a balancing act between the financial cost, the impact on the environment and the value of the data compared with the cost of redoing the experiment. The example of the ESRF data archive which is based on 'cold storage' (*i.e.* tape storage) will be used to illustrate how the different costs can be calculated and compared.

Microsymposium A118: Raw Diffraction Data Reuse: Warts and All

Wednesday 23 August 2023, Room 216, 4:00 pm – 6:30 pm

Situations where small-molecule raw data should be made available

Simon J. Coles

School of Chemistry, Faculty of Engineering and Physical Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK

It is now *de facto* best practice to deposit structure factors when publishing and making small-molecule crystal structures available. This means that the small-molecule crystallography community now caters very well for the validation of 'routine structures' as part of the publication process. The clear benefits that we are now seeing arise from this approach are that journal articles are better evidenced and the crystallographic databases contain even better quality records. Increasing personal experience of the need/desire to assess structures in different ways not necessarily supported by the model published by the original authors illustrates the value of providing structure factors in promoting the appropriate reuse of crystal structures. Work that makes use of collections of structures from databases generally involves a lot of 'data massaging' depending on the goal of the research, *e.g.* re-refinement with more modern software approaches to improve accuracy or without restraints/constraints to explore geometry. A very compelling recent scenario is that one can now perform quantum crystallographic aspherical atom refinements with the original data made available on publication to greatly improve the structure and extract chemical or bonding information that wasn't possible at the time.

This fundamental change in the way we communicate our results has greatly enhanced validation and reuse of crystal structures; however this is generally only the case if all aspects of a raw image are fully and/or properly accounted for and the model is correct or appropriate. So, our community can clearly take more steps to better support reanalysis and reinterpretation of the data collected in single-crystal diffraction experiments.

It follows that in some cases raw data may no longer be required. If it can be conclusively shown, preferably by a trusted, automated and enduring mechanism, that all diffraction events are accounted for by the integration and processing stages of the analysis, that nothing further could be obtained from the raw data, then arguably there is no need to retain them. This would obviate the management and financial burdens of curation of a considerable proportion of the raw data generated.

However, there are a significant proportion of cases of results arising from raw data where thorough evidence and justification may be necessary or where it is highly likely that a better analysis may be performed in the future because of method and software innovations. Moreover, there are increasing pressures from bodies *e.g.* funders to make the data relating to research outputs Findable, Accessible, Interoperable and Reusable (FAIR) and it may be important to keep raw data for these reasons.

This talk will present a range of small-molecule crystallography cases where raw data publication would be key *e.g.* in underpinning advanced or dynamic crystallographic experiments, validating claims and quality, evidencing pathological samples and diffraction, supporting future development and to crowd source solutions. These cases are based on the outcomes of a significant community survey [1] on raw data management practices and workshop discussions [2]. There are a number of different approaches that could be implemented to address these cases and these will be outlined as well as showcasing the recent introduction of a new category of article in *IUCrData – Raw Data Letters* [3].

[1] Coles, S.J. & Sarjeant, A. (2018). *IUCr Newsletter*, volume **26**, number 2; Coles, S.J. & Sarjeant, A. (2020). *IUCr Newsletter*, volume **28**, number 1.
[2] When should small molecule crystallographers publish raw diffraction data (2021). Twenty-Fifth Congress and General Assembly of the International Union of Crystallography, https://www.iucr.org/resources/data/commdat/prague-workshop-cx.
[3] Kroon-Batenburg, L. M. J., Helliwell, J. R. & Hester, J. R. (2022). *IUCrData* **7**, x220821.

Keywords: raw data, chemical crystallography, structural chemistry

Compression and data reduction in serial crystallography

M. Galchenkova[1], A. Tolstikova[2], D. Oberthuer[1], J. Sprenger[1], W. Brehm[1], T. A. White[2], A. Barty[2], H.N. Chapman[1] and O.M. Yefanov[1]

[1]*Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany;*
[2] *Deutsches Elektronen-Synchrotron, Notkestrasse 85, Hamburg 22607, Germany*
Email: oleksandr.yefanov@cfel.de

Protein crystallography is one of the most successful methods for biological structure determination. This technique requires many diffraction snapshots to get 3D structural information of the studied protein. Even more patterns are needed for studying fast protein dynamics that can be achieved using serial crystallography (SX). Fortunately, new X-ray facilities such as 4th-generation synchrotrons and Free Electron Lasers (FELs) combined with newly developed X-ray detectors opened a way to carry out these experiments at a rate of more than 1000 images per second. The drawback of this increase in acquisition rate is the volume of collected data – up to 2 PB of data per experiment could be easily obtained. Therefore, new data reduction strategies have to be developed and deployed. Lossless data reduction methods will not change the data, but usually fail to achieve a high compression ratio. On the other hand, lossy compression methods can significantly reduce the amount of data, but they require careful evaluation of the resulting data quality.

We have tested different approaches for both lossless and lossy compression applied to SX data, proposed some new ways for lossy compression and demonstrated appropriate methods for data quality assessment. By checking the resulting statistics of compressed data (like CC^*/R_{split}, R_{free}/R_{work}) we have demonstrated that the volume of the measured data can be greatly reduced (10–100 times!) while the quality of the resulting data was kept almost constant. In addition, we tested lossy compression methods on the SAD dataset (thaumatin collected at 4.57 keV, measured at the SwissFEL) and demonstrated that even such very sensitive data can be successfully compressed. This allowed us to determine the limit of application for all considered lossy compressions. Some of the proposed compression strategies, tested on SX and MX datasets, can be used for other types of experiments, even with different sources (for example electron and neutron diffraction).

The authors are thankful to K.Nass and D.Ozerov for sharing the SAD, to V. Mariani for useful comments. Taking into account the practical impact of this work, starting from 2020 the authors shared the ideas described in this paper with the data scientists at eXFEL, LCLS, SwissFEL, ESRF, APS, Petra III as well as at different conferences and workshops to demonstrate the ways of data compression and quality checks for the SX data.

Keywords: serial crystallography, compression, data reduction

The raw, the cooked and the medium-rare: unmerged diffraction data as a rich source of opportunities for data re-use and improvements in methods and results

G. Bricogne, C. Flensburg, R. H. Fogh, P. A. Keller, I. J. Tickle and C. Vonrhein

Global Phasing Limited, Sheraton House, Castle Park, Cambridge CB3 0AX, UK
Email: gb10@globalphasing.com

Deposition into the PDB of experimental diffraction data, in the form of merged intensities or of 'structure factor[amplitude]s', to accompany atomic models determined and/or refined from them, was made mandatory in 2008. This brought benefits that went well beyond the intended purpose of making the deposited models verifiable and correctable, in the form of an unanticipated 'virtuous circle' whereby deposited data fuelled improvements in refinement software that in turn enabled improvements to be made in the initially deposited models, from the same data. The need to manage the outcome of this continuous improvement process led to the introduction of a versioning mechanism into the archiving of atomic models by the PDB in 2017.

The creators of the Electron Density Server had already noted in 2004 that 'perhaps we should consider deposition of unmerged intensities or even raw diffraction images in the future' [1], without however anticipating the potential for a similar auto-catalytic cycle of simultaneous improvements in data reduction methods and in final structural results that could follow. This potential was later articulated in *e.g.* [2] in the following terms: 'those [merged]

deposited X-ray data are only the best summary of sets of diffraction images according to the data-reduction programs and practices available at the time they were processed. Just like refinement software, those programs and practices are subject to continuing developments and improvements, especially in view of the current interest and efforts towards better understanding radiation damage during data collection and in taking it into account in the subsequent processing steps.'

Strong general support for the idea of archiving raw diffraction images, together with the recognition that this task was beyond the remit and resources of the PDB, has led over the past decade to the emergence of a delocalised infrastructure (whereby raw data storage and curation takes place at synchrotrons and other dedicated repositories while the PDB provides a capability to annotate entries with a DOI that points to the raw data storage location) that is a major topic in this Workshop.

Our own interest has been to document the scientific case for depositing and archiving suitably annotated unmerged diffraction data into the PDB, a goal achievable with modest storage requirements while already creating a standardised resource capable of feeding improvements in scaling and merging methods resulting in better refined models than those originally deposited. This goal is the focus of the current activities of the Subgroup on Data Collection and Processing of the PDBx/mmCIF Working Group of the wwPDB, in which we participate, to expand the mmCIF dictionary to support such extended deposition and archiving.

Crucially, unmerged data collected by the rotation method can preserve instrumental metadata about the image number and the detector position at which each diffraction spot was located and integrated, providing a broader decision-making scope over the way it is incorporated into the scaling and merging process. This opens a wide range of possibilities for improving any initially performed scaling/merging steps and for extraction of further data. We will present examples touching upon the following areas:

1. production of full validated data quality metrics that are often incomplete or inconsistent in deposited merged data;
2. detection of problematic images and image ranges, and remediation by their selective exclusion from scaling/merging;
3. anisotropic diffraction limit analysis (or re-analysis) with STARANISO, if not already performed;
4. extraction of previously unexploited anomalous signal and computation of anomalous difference Fourier maps;
5. 'reflection auditing' by tracing outliers detected at the refinement stage back to their unmerged contributors in terms of specific image numbers and detector positions, thus diagnosing ice rings, poor beamstop masks, angular overlaps, *etc.*;
6. detection of radiation damage *via* $F_{\rm early}-F_{\rm late}$ maps; adapting parametrisation to patterns of structural radiation damage.

We are grateful to the PDBx/mmCIF Subgroup on Data Collection and Processing, especially Aaron Brewster, Ezra Peisach, Stephen Burley and David Waterman, for a stimulating collaboration that provided a context for presenting these investigations.

[1] Kleywegt, G. J., Harris, M. R., Zou, J.-Y., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
[2] Joosten, R. P., Womack, T., Vriend, G. & Bricogne, G. (2009). *Acta Cryst.* D**65**, 176–185.

Keywords: raw diffraction data, unmerged diffraction data, scaling and merging methods

imgCIF as a solution for automated processing of raw crystallographic data

James R. Hester

ANSTO, Locked Bag 2001, NSW 2232, Australia
Email: jxh@ansto.gov.au

Laboratories and large-scale facilities currently produce a torrent of raw crystallographic data from a variety of bespoke and off-the-shelf instruments. Ideally, arbitrary raw data sets from these instruments could be automatically processed without time-consuming user intervention to determine the appropriate format and instrument setup. Unfortunately, the plethora of experimental geometries and data formats, the lack of a predictable link between

DOI and the raw data URL, and the unwieldiness of working with non-local data sets all lead to serious challenges in creating such a hands-off machine-interoperable data ecosystem.

The imgCIF format [1], which was originally envisioned as an archival container for raw data, has been expanded and repurposed to improve automatic interoperability. An imgCIF file may now refer to raw data frames located externally to the imgCIF file, while precisely specifying the geometry of the instrument for each data frame, thus allowing automated raw data processing *via* an imgCIF file with no manual intervention. This approach relies for its effectiveness on previous work on crystallographic raw data standards: in particular, widespread storage of raw data in either the Crystallographic Binary Format (CBF) [2] or HDF5 [3,4] standards alleviates most issues associated with correctly converting streams of bytes into images, with the imgCIF file then relating the image pixels to laboratory coordinates. The 'CheckCIF for raw data' work developed under the auspices of IUCr Journals [5] is a simple example of the automated computation on raw data enabled by such imgCIF files. The current approach is not inherently limited to single-crystal X-ray data from flat detectors, and so, for example, can be used to describe raw powder diffraction data from curved or flat area-detector instruments from both X-ray and neutron sources with no further changes to the imgCIF standard.

A number of limitations are evident for this imgCIF-based approach to raw data. The need to use pointers to data frames that, unlike DOIs, are not guaranteed to be stable over time makes imgCIF files containing raw data pointers somewhat fragile for archival purposes. In addition, the range of raw data frame formats that are defined as accessible *via* an imgCIF file is currently restricted to that large subset of current files for which support is easy to implement due to simplicity, the availability of cross-platform libraries, or well-specified standards: CBF, HDF5, and SMV. Other raw formats should be converted to one of these formats (typically CBF) when depositing. The utility of this approach is also limited by the recognition of imgCIF files by common data analysis software, and availability of tools to produce imgCIF files; work to create such tools is ongoing [6].

The author acknowledges the imgCIF mailing list and the IUCr Journals raw data working group for feedback on this work.

[1] Bernstein, H. J (2006). *International Tables for Crystallography, Volume G* edited by S. R. Hall and B. McMahon, pp 199–205
[2] Bernstein, H. and Hammersley, A. P. (2006). *International Tables for Crystallography, Volume G*, edited by S. R. Hall and B. McMahon, pp 37–43
[3] The HDF Group (1997–2022). Hierarchical Data Format, version 5. https://www.hdfgroup.org/HDF5
[4] Könnecke, M., et al (2015). *J. Appl. Cryst.* **48**, 301–305.
[5] Kroon-Batenburg, L. M. J., Helliwell, J. R. & Hester, J. R. (2022). *IUCrData* **7**, x220821
[6] https://github.com/COMCIFS/instrument-geometry-info/tree/main/Tools

Keywords: raw data, FAIR, interoperability

Raw data, metadata and the experimental narrative: reuse of time-of-flight neutron diffraction data

M. Guthrie

Oak Ridge National Laboratory, Oak Ridge, TN 37830
Email: Guthriem@ornl.gov

Modern Time-Of-Flight (TOF) neutron diffractometers consist of wide angular banks of highly pixelated detectors. The TOF of arriving neutrons, which are emitted in precisely controlled pulses, is proportional to energy and, thus, each pixel resolves energy by precisely recording the arrival time of detected neutrons. At the Spallation Neutron Source at Oak Ridge National Laboratory (ORNL), the resulting raw data sets are recorded in "event mode", consisting of lists containing pixel id, TOF and the absolute time of the generating pulse for every neutron detection event. In parallel to the neutron events, raw data sets also contain a complete set of metadata including both experimentally logged process values and a full mathematical description of the instrument. The data are stored according to the nexus standard [1]. The ORNL Neutron facilities follow a principle of hosting and persists all raw data, which are owned by the experimental team and can be made openly accessible at their request.

However, in order that the raw data be compatible with FAIR principles, they must be linked with calibration data and the calibration parameters derived from these. Of equal importance is standardized data-reduction procedures

and persistence of the parameters these use. The latter is particularly pertinent for TOF diffraction for two main reasons. Firstly, the data set contains simultaneous measurement of volumes of reciprocal space that can be integrated and reintegrated *via* various schemes (often exchanging counting statistics for diffraction resolution and range). Secondly, wavelength-dependent attenuation corrections can be complicated and use algorithms that continue to evolve with time. Thus, details of the entire reduction workflow must also be captured both for provenance and to allow future re-processing with algorithms that may improve over time.

A last, critical component of the data generation is something that may be called the "experimental narrative". This is the flow of reactive decision making and responsive, real-time adjustment or optimization of the instrumentation that forms the context of otherwise adjacent datasets in the catalogue. Often, by nature of experimental science, where the answers are not known *a priori*, the extraction of analytic conclusions – by human or machine actors – is impossible without this accompanying information. In this presentation, I will highlight some approaches that are employed in the ORNL Neutron Facilities and discuss some opportunities for future improvements of these.

[1] Könnecke, M., et al (2015). *J. Appl. Cryst.* **48**, 301–305.

80 Years of the Powder Diffraction File™ (PDF®): A Database Perspective

S. Kabekkodu and T. Blanton

International Centre for Diffraction Data, 12 Campus Blvd, Newtown Square, PA19073, USA
Email: Kabekkodu@icdd.com

Crystallographic databases play a vital role in materials research, influencing materials development and providing a reference for materials characterization. Design, data curation, and data management are all critical factors in developing a successful and useful database.

The International Centre for Diffraction Data (ICDD®) Powder Diffraction File (PDF) is a powerful database for materials characterization that has been used extensively by the scientific community. Starting with 1000 entries on printed cards in 1941, the database has grown to contain over 1 million unique material data sets. The Powder Diffraction File has a wealth of information that a materials scientist can take advantage of for materials identification, characterization, computation and design. The Powder Diffraction File in Relational Database (RDB) format contains extensive chemical, physical, bibliographic and crystallographic data including atomic coordinates enabling characterization and computational analysis.

Proper database structure, data validations and phase-type classifications are crucial in making any database useful and reliable. While using a database, it is important to know the quality of the crystal structure, diffraction pattern data and any data field of interest found in the database. With the varying quality of published data in the literature, the PDF database editorial review processes require rigorous data evaluation methods to define data based on its quality.

This presentation will focus on various aspects of data archival, curation and classifications. The current progress and challenges in archiving raw powder patterns for the future reusability will be covered in detail.

Keywords: crystallographic database, powder diffraction, raw data archival, phase identification

Posters

Fast and efficient compression algorithms for macromolecular crystallography experiments

D. F. Kreitler[1], H. J. Bernstein[2] and J. Jakoncic[1]

[1]*National Synchrotron Light Source II, Brookhaven National Laboratory, Bldg 745, Upton, NY 11973-5000, USA*
[2]*Ronin Institute for Independent Scholarship, c/o NSLS-II, Brookhaven National Laboratory, Bldg 745, Upton, NY 11973-5000, USA*
Email: hbernstein@bnl.gov

Structural biology experiments benefit significantly from state-of-the-art synchrotron data collection. One can acquire macromolecular crystallography (MX) diffraction data on large-area photon-counting pixel-array detectors at framing rates exceeding 1000 frames per second, using 200 Gbps network connectivity or higher when available. In extreme cases this represents a raw data throughput of about 25 GB/s, which is nearly impossible to transmit, process and store uncompressed at reasonable cost.

Our field has used lossless compression for decades to make such data collection manageable. In the near future increasing data rate and volumes will make it necessary to seriously consider increasing use of lossy compression in which less effort is put into ensuring retention of all background information than is put into ensuring retention of all Bragg reflections crossing the "entropy limit" to save storage space.

Most MX beamlines are now fitted with DECTRIS Eiger detectors, which are all delivered with optimized compression algorithms by default, and they perform well with moderate framing rates and typical diffraction data. Recently, better lossless compression algorithms have been developed and are available to the research community. Here we discuss one of the latest and most promising lossless algorithms on a variety of diffraction data like those routinely acquired at state-of-the-art MX beamlines and compare the results to those obtained when backgrounds between Bragg reflections are handled with lossy compression.

Keywords: data compression, bslz4, zstd, lossless, lossy

CODATA, IUCr, PDBj collaboration for medical-protein crystal structure definitive versions of data files

J. R. Helliwell[1]*, G. Kurisu[2] and L. Kroon-Batenburg[3]

[1]*University of Manchester, UK;* [2]*University of Osaka and PDBj, Japan;* [3]*University of Utrecht, The Netherlands*
Email: john.helliwell@manchester.ac.uk

Envisaging a process akin to a journal's peer review we have set up an IUCr and PDBj collaboration for medical-protein crystal structure definitive versions of data files within the CODATA GOSC Case Studies.† Quoting from our Case Study webpage 'The overall reproducibility of the diffraction data and their linked molecular model is the overarching guide. The scope of this challenge, in general terms, can be judged by the fact that the FAIR movement (FAIR=Findable, Accessible, Interoperable, Reusable) did not include data quality in its criteria. In the spirit of scientific reproducibility, we introduce a term somewhere between reusability and reproducibility, namely definitive reusability.' That PDBj had launched a raw diffraction images data archive XRDa https://xrda.pdbj.org/ was pivotal as it would allow a combined evaluation of raw data, processed structure factors and derived protein molecular model. This also would lead to general community benefit beyond medical pandemic challenges, although of course very important, to the whole of macromolecular crystallography. Feedback on a PDBj deposition is made by JRH and LKB to GK and who then can decide, like a journal editor exactly what feedback is made to a depositor to PDBj for a possible reversioning of a PDBj deposition. Progress of this initiative will be described and spans Covid-19 and other medically important proteins (*e.g.* see [1,2,3]. As an example, we choose to scrutinise the various metrics used to judge the resolution limit for refinement of a protein model. The availability of the raw diffraction data in PDBj's XRDa allows a direct comparison of these metrics to be made by a raw diffraction data reuser *versus* that chosen by a depositor into PDBj.

[1] Brink, A. and Helliwell, J. R. (2022). *IUCrJ*, **9**, 180–193.
[2] Hanau, S. and Helliwell, J. R. (2022). *Acta Cryst.* F**78**, 96—112.
[3] Helliwell, J. R. (2021). *Acta Cryst.* F**77**, 388–398.

†See https://codata.org/initiatives/decadal-programme2/global-open-science-cloud/case-studies/diffraction-data/ .

Keywords: raw data reuse, PDBj XRDa archive, macromolecular crystallography

CODATA, IUCr, PDBj collaboration for medical-protein crystal structure definitive versions of data files

Helliwell, J R*, Kurisu, G, and Kroon-Batenburg, L

University of Manchester, UK; University of Osaka and PDBj, Japan; University of Utrecht, The Netherllands

Introduction

At the Research Data Alliance Plenary 17 (https://www.rd-alliance.org/plenaries/rda-17th-plenary-meeting-edinburgh-virtual) concerns were voiced that "multiple versions of covid-19-proteins crystal structure data were useless". I (JRH) replied that it was a form of mayhem but not useless. But how to improve? The multiple versions had arisen from well-meaning multiple task forces offering improved versions on their own personal websites but in the end largely ignored by the depositors at the PDB with only 30 revised versions out of more than 1000 depositions as of IUCr Prague Congress. Clearly this suggested to JRH that a direct collaboration with the PDB would be an improved approach.

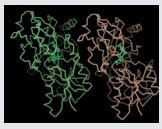
Aim

So, envisaging a process akin to a journal's peer review we set up a collaboration within the CODATA GOSC Case Studies, formally endorsed by the IUCr (see https://codata.org/initiatives/decadal-programme2/global-open-science-cloud/case-studies/diffraction-data/). Progress of this initiative has been made and spans covid-19 and other medically important proteins (*e.g.* see Helliwell 2021).

Method

That PDBj had launched a raw diffraction images data archive XRDa https://xrda.pdbj.org/ was pivotal as it would allow a combined evaluation of raw data, processed structure factors and derived protein molecular model. This also would lead to general community benefit beyond medical pandemic challenges, although of course very important, to the whole of macromolecular crystallography. Feedback on a PDBj deposition is made by JRH and LKB to GK and who then can decide, like a journal editor exactly what feedback is made to a depositor to PDBj for a possible reversioning of a PDBj deposition.

Reprocessing results on a test case, 7ccy



This is an exemplary case:-

No difference map peaks at 5.00 sigma

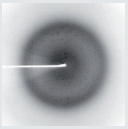


Table 1 Let's check the resolution limit various indicators

Dials reprocessing in ccp4i2 (Beilsten-Edmands et al 2020) using the PDBj 7ccy diffraction images as an example (Sato et al 2021). It gives a useful range of diagnostics of the possible resolution limits (and what a large range those possible resolution limits are!).
Resolution cut off estimates:-

resolution of all data	: 1.913
based on CC(1/2) >= 0.33	: 1.946
based on mean(I/sigma) >=2.0	: 3.037
based on R-merge < 0.5	: 2.411
based on R-meas < 0.5	: 2.497
based on completenes >=90%	: 2.335
based on completeness >=50%	: 2.155

Using iMosflm reprocessed raw data to 2.0Å (Battye et al 2020) PDB Redo estimate based on ***Diederichs and Karplus***:-

******** Testing resolution cut-offs: 2.40Å 2.29Å 2.16Å 2.00Å.**

o Testing resolution 2.40

o Testing resolution 2.29

*** LL-free deteriorated**

o Testing resolution 2.16

*** R-free deteriorated**

*** Weighted R-free deteriorated**

-High resolution cut-off:2.29Å

Conclusions

Dials makes several recommendations about the resolution limit for protein model refinement. Sato et al used a diffraction resolution limit of 2.40Å, which is a good choice across the several parameters listed above.

For clarity we suggest that it would be better to apply the best model as the criterion which Diederichs and Karplus have shown should be via the paired refinement method (Maly et al 2020 and references therein). This is available for instance at https://pdb-redo.eu/ (Joosten et al 2014).

Also, the (Fo-Fc) peaks' list should be acted upon before deposition in the PDB. To try to ensure this the PDB Validation Report could include a list of (Fo-Fc) peaks that have not been dealt with in the model to openly advise the depositor. That would ensure that the model likely does not need post publication peer review.

Acknowledgement

We are very grateful to Sato et al for depositing their raw, processed and derived data files at PDBj and XRDa.

References

Battye, T.G.G., Kontogiannis, L., Johnson, O., Powell, H.R. & Leslie, A.G.W. (2011) Acta Cryst. **D67**, 271-281.

Beilsten-Edmands, J., Winter, G., Gildea, R., Parkhurst, J., Waterman, D. & Evans, G. (2020). Acta Cryst. D76, 385-399.

Helliwell, J R (2021) Acta Cryst F77, 388-398.

Joosten RP, Long F, Murshudov GN, Perrakis A. IUCrJ. 2014 May 30;1(Pt 4):213-20.

Maly, M., Diederichs, K., Dohnalek, J. & Kolenko, P. (2020). IUCrJ 7, 681-692.

Sato, H., Sugishima, M., Tsukaguchi, M., Masuko, T., Iijima, M., Takano, M., Omata, Y., Hirabayashi, K., Wada, K., Hisaeda, Y. & Yamamoto, K. (2021). Biochem. J. 478, 1023–1042.

Microsymposium A119: Interoperability of Crystallographic Data and Databases

Monday 28 August 2023, Room 212/213, 4:00 pm – 6:30 pm

Applications of metadata collection and analysis to parallel crystallisation and the ENaCt technique

T. Smith, M. Probert and M. Hall

School of Natural and Environmental Sciences, Bedson Building, Newcastle University, NE1 7RU, Newcastle Upon Tyne, UK
Email: t.smith7@newcastle.ac.uk

The analysis of large datasets forms the backbone of many aspects of today's scientific literature. Modern modes of data interrogation, such as machine learning techniques, have become ever more popular alongside the explosion of experimental data size. These methods can enable the unbiased extraction of information that may be overlooked or simply ignored by the human eye. This trend towards large data use and its analysis is a trend that is expected to grow and flourish to become a dominant aspect of modern research. Behind the glamour of the headline results and algorithms lies an ugly truth: large datasets are hard to manage, and the associated difficulties only increase with the size and scope of the intended interrogation. Proper data management techniques and tools are often an afterthought, and only considered in earnest when datasets begin to become unruly – the FlexTape™ patch for an already sinking vessel. We aim to show that by considering data management early, and investing time in quality tooling, it is possible to achieve a more harmonious existence that alleviates the need for patches and disentanglement problems during analysis.

High-throughput crystallisation experiments using the ENaCt technique[1] have the potential to generate significant amounts of direct data (images, success/failure, *etc.*). However, collecting these data can be tedious and error-prone when left in the hands of humans. Herein is described an automated optical microscope coupled to a desktop application which significantly increases the reliability of collected data. In combination with an industrial machine vision camera containing a built-in grid of polarising filters, a real-time image processing system has been developed to automatically extract information about crystal quality from captured images and save that alongside relevant experimental metadata in a consistent and structured way.

Using a combination of a central relational database; python-powered JSON API; and custom-designed and implemented Query Language; we were able to provide team members with varying levels of programming ability a mechanism to efficiently access this dataset. As a result, they can make more informed decisions about which experiments to perform to enhance the likelihood of overall success.

Thanks to Newcastle University for funding this project.

[1] Tyler et al., (2020). *Cell Chem.* **6**, 1755.

Keywords: crystallisation, automation, software

DIMAS: A web-based service crystallography submission and data management system

Toby Blundell[1]* and Oleg Dolomanov[2]

[1]*Department of Chemistry, Durham University, South Road, Durham, DH1 3LE, UK*
[2]*OlexSys Ltd, Durham University, South Road, Durham, DH1 3LE*
**Email: Toby.j.blundell@durham.ac.uk*

The Department of Chemistry at the University of Durham, UK has a long and prestigious reputation in the field of single-crystal X-ray crystallography. As the premier technique allowing full structural identification of crystalline materials, crystallographers at Durham University work with researchers in numerous fields including materials science, chemistry, physics, biology and engineering.[1]

In collaboration with OlexSys Ltd [2] and Labsafe [3] we have developed a web-based platform to allow researchers to submit crystalline samples for X-ray diffraction analysis to the in-house crystallography service. DIMAS guides

users through the process of submitting samples allowing them to include details about the synthesis and crystallisation conditions. There is an embedded chemical drawing module to input the expected compound and users can keep track of all their submissions in one convenient location.

Following collection, a full set of crystallographic files, including data reduction outputs, structure solution and refinement files, as well as rendered images of the structures, can be uploaded. DIMAS automatically extracts the relevant unit-cell parameters, symmetry and data statistics as well as allowing the user to download the archived files and giving research group leaders oversight over their research group data.

Fig. 1. DIMAS, a collaboration between Durham University, OlexSys Ltd and Labsafe.

We acknowledge Dr Dmitry Yufit and OlexSys Ltd, particularly Prof. Horst Puschmann and Prof. Judith Howard.

[1] https://crystallographygroup.webspace.durham.ac.uk/
[2] https://www.olexsys.org/
[3] https://www.olexsys.org/labsafe/

Keywords: service crystallography, database, user management

Protein–Ligand Binding Database (PLBD) of Crystal Structures and Intrinsic Thermodynamic Parameters

D. Lingė[1], M. Gedgaudas[1], A. Merkys[2], V. Petrauskas[1], A. Vaitkus[2], A. Grybauskas[2], V. Paketurytė[1], A. Zubrienė[1], A. Zakšauskas[1], A. Mickevičiūtė[1], J. Smirnovienė[1], L. Baranauskienė[1], E. Čapkauskaitė[1], V. Dudutienė[1], E. Urniežius[1], A. Konovalovas[3], E. Kazlauskas[1], K. Shubin[4], H. B. Schiöth[5], W.-Y. Chen[6], J. E. Ladbury[7], S. Gražulis[2] and D. Matulis[1]*

[1] *Department of Biothermodynamics and Drug Design, Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio 7, Vilnius, LT-10257, Lithuania*
[2] *Sector of Crystallography and Cheminformatics, Institute of Biotechnology, Life Sciences Center, Vilnius University, Saulėtekio 7, Vilnius, LT-10257, Lithuania*
[3] *Department of Biochemistry and Molecular Biology,, Institute of Biosciences, Life Sciences Center, Vilnius University, Saulėtekio 7, Vilnius, LT-10257, Lithuania*
[4] *Latvian Institute of Organic Synthesis, Riga LV-1006, Latvia*
[5] *Functional Pharmacology and Neuroscience, Department of Surgical Sciences, Upp6ala University, Uppsala, Sweden*
[6] *Department of Chemical and Materials Engineering, National Central University Jhong-Li 320, Taiwan*
[7] *School of Molecular and Cellular Biology, University of Leeds LC Miall Building, Leeds, LS2 9JT, United Kingdom*
* *Email: daumantas.matulis@bti.vu.lt*

Here we introduce a Protein-Ligand Binding Database (PLBD) of thermodynamic and kinetic data of protein interaction with small molecule compounds, available at https://plbd.org. The binding data are linked to the same protein-ligand co-crystal structures, enabling the structure-thermodynamics correlations both in terms of protein structure and compound chemical formula. Currently, the database contains over 5500 binding datasets of 556 sulfonamide compound interactions with the 12 catalytically active human carbonic anhydrase (CA) isozymes determined by the fluorescent thermal shift assay, isothermal titration calorimetry, inhibition of enzymatic activity, and surface plasmon resonance [1,2]. In the PLBD, we emphasize the intrinsic thermodynamic parameters that account for the binding-linked protonation reactions. In addition to the protein-ligand binding affinities, the database provides calorimetrically measured binding enthalpies, enhancing the understanding of reaction mechanisms. The database also contains 127 X-ray crystal structures of six CA isozyme complexes with ligands [1,2]. The database has been built using the FAIR data principles and the model was originally developed for exchanging

crystallographic data in the CIF framework. The database schema and deposited data have revision and versioning systems providing historical traces of database evolution. The PLBD is useful for the investigations of protein-ligand recognition principles and could be applied for small molecule drug design.

This research was funded by a grant S-LLT-20-2 from the Research Council of Lithuania, Lithuanian-Latvian-Taiwan Cooperation Programme.

[1] Matulis, D. (Editor) (2019). *Carbonic Anhydrase as Drug Target: Thermodynamics and Structure of Inhibitor Binding*. Springer Nature.
[2] Linkuvienė, V., Zubrienė, A., Manakova, E., Petrauskas, V., Baranauskienė, L., Zakšauskas, A., Smirnov, A., Gražulis, S., Ladbury, J. E. & Matulis, D. (2018). Thermodynamic, Kinetic, and Structural Parameterization of Human Carbonic Anhydrase Interactions toward Enhanced Inhibitor Design. *Q. Rev. Biophys.* **51**, 1–48.

Keywords: protein–ligand binding, X-ray crystallography, thermal shift assay, differential scanning fluorimetry, isothermal titration calorimetry, carbonic anhydrase, sulfonamide

Database interoperability: a powder diffraction perspective

S. Kabekkodu and T. Blanton

International Centre for Diffraction Data, 12 Campus Blvd, Newtown Square, PA19073, USA
Email: Kabekkodu@icdd.com

Crystallographic database interoperability is essential to support growing data driven innovation in materials research. ICDD's Powder Diffraction File™ (PDF®) is the database that has been used by all of the major powder diffractometer manufactures for phase identification for several decades. Maintaining database interoperability to work with all of the public and proprietary phase identification software is crucial. The ICDD PDF database was first published in 1941, and since 1967 (first digital crystallographic database) the PDF has been configured keeping in mind software developers in terms of data interoperability. ICDD adopted a RDBMS (Relational Database Management System) in the year 2000. RDBMS provides flexible access to the database for programming and a relational database construct allows ICDD to adhere to FAIR (Findable, Accessible, Interoperable, Reusable) principles. Software developers are able to maintain database interoperability even as there are continuous and dynamic changes to the PDF database. Metadata also plays a very important role in adhering to FAIR principles. In powder diffraction there is a challenge as there is no common raw data format among the diffractometer manufacturers. The PDF is accessible independent of these varying formats.

The ICDD PDF is a curated database with every entry evaluated using a combination of computer and human editorial review and presented using a quality mark nomenclature that provides the user with a systematic process for understanding data entry quality and a consistent approach in assessing the goodness-of-match for phase identification. These quality marks are also important from the interoperability perspective, given the existence of multiple automated software routines used for phase identification. An additional feature of the Powder Diffraction File it the inclusion of 18,000+ entries with raw data powder diffraction patterns that can be used for direct comparison to user data.

ICDD PDF database interoperability in terms of structure and semantics will be discussed in this presentation.

Keywords: crystallographic database, interoperability, powder diffraction, raw data archival, phase identification

Interoperability of Databases as viewed by the Publisher

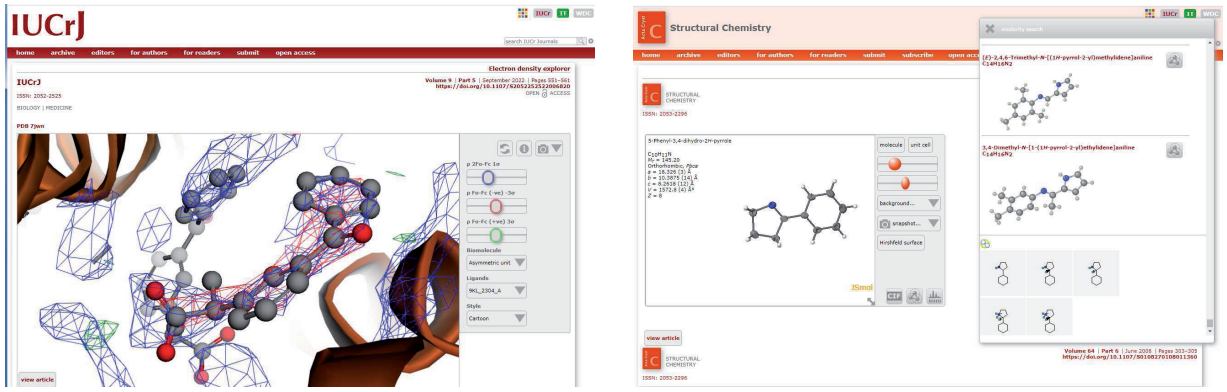
Brian McMahon

International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK

The emergence of interoperable protocols that allow curated databases and publishers to exchange research data sets and associated metadata brings great benefits for the way that databases and journals can complement each other. The International Union of Crystallography (IUCr) is a publisher of several data-rich journals, and has taken full advantage of these protocols. Indeed, it has sponsored the Crystallographic Information Framework (CIF) project, which provides crystallography with greater interoperability than most other scientific domains.

Structural databases began by manually harvesting tabulated data from printed journals, in arbitrary formats and needing much effort to re-keyboard the data, an error-prone exercise. The introduction of the CIF format in 1991 [1] provided a standard machine-readable presentation that could be easily harvested over the network and ingested error-free into the databases' own internal formats. While practice has evolved so that many data sets are now deposited directly with databases before publication, the availability of these data sets, still mandated by IUCr journals, from the publisher's own website allow future harvesting by new and diverse databases.

IUCr journals routinely provide links to the structure representation views of published structures in the relevant databases. They now also import data sets from external databases, allowing readers of published structural articles to manipulate value-added visual representations of the structures within the online article environment itself (Fig. 1).



(*a*) (*b*)

Fig. 1. Value-added structure exploration within IUCr journals. (a) Interactive electron density maps. (b) Three-dimensional ellipsoid visualization and similarity searches against both other articles published in IUCr journals and the PubChem database.

As IUCr journals become rich repositories of deposited structural data sets, so they have the potential to host novel data discovery systems that do not use the formal relational schemas of the conventional databases [2]. New approaches using artificial intelligence techniques open the possibility of even greater synergy between traditional publishing and database platforms. These may, further, help to address the challenge of making scientific connections across different disciplines, where there are not as yet widely available tools to make semantic links between different domain-specific metadata standards such as those catalogued by the Research Data Alliance [3].

[1] Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Cryst.* A**47**, 655—685.
[2] Özer, B., Karlsen, M. A., Thatcher, Z. et al. (2022). Towards a machine-readable literature: finding relevant papers based on an uploaded powder diffraction pattern. *Acta Cryst.* A**78**, 386—394.
[3] Research Data Alliance (2023). Metadata Standards Catalog. https://rdamsc.bath.ac.uk/

Keywords: Databases, Interoperability, Scientific publishing

The Cambridge Structural Database: a multidisciplinary resource

S. C. Ward and M. P. Lightfoot

The Cambridge Crystallographic Data Centre, Cambridge, UK
Email: ward@ccdc.cam.ac.uk

The Cambridge Structural Database (CSD) was founded on a vision that collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments. The existence of a common crystallographic language, an ability to understand crystallographic information, and an awareness of the importance of data interoperability have enabled that vision to come true. The database is now a valued multidisciplinary resource used extensively in academia as well as in industry particularly as part of drug development and materials design.

In crystallography we are extremely fortunate that a standard file format has been adopted by researchers, software creators and publishers alike providing a strong foundation for data sharing. As a proponent of the FAIR data principles, the Cambridge Crystallographic Data Centre (CCDC) has supported the community by developing a workflow from structure deposition to data sharing. This workflow seeks to promote these guiding principles by enabling depositors to provide information which renders the datasets more Findable, Accessible, Interoperable and Reusable. As the curators of a domain specific repository, we also work to ensure this crystallographic data can be utilised across disciplines to enable the advancement of science as well as crystallography.

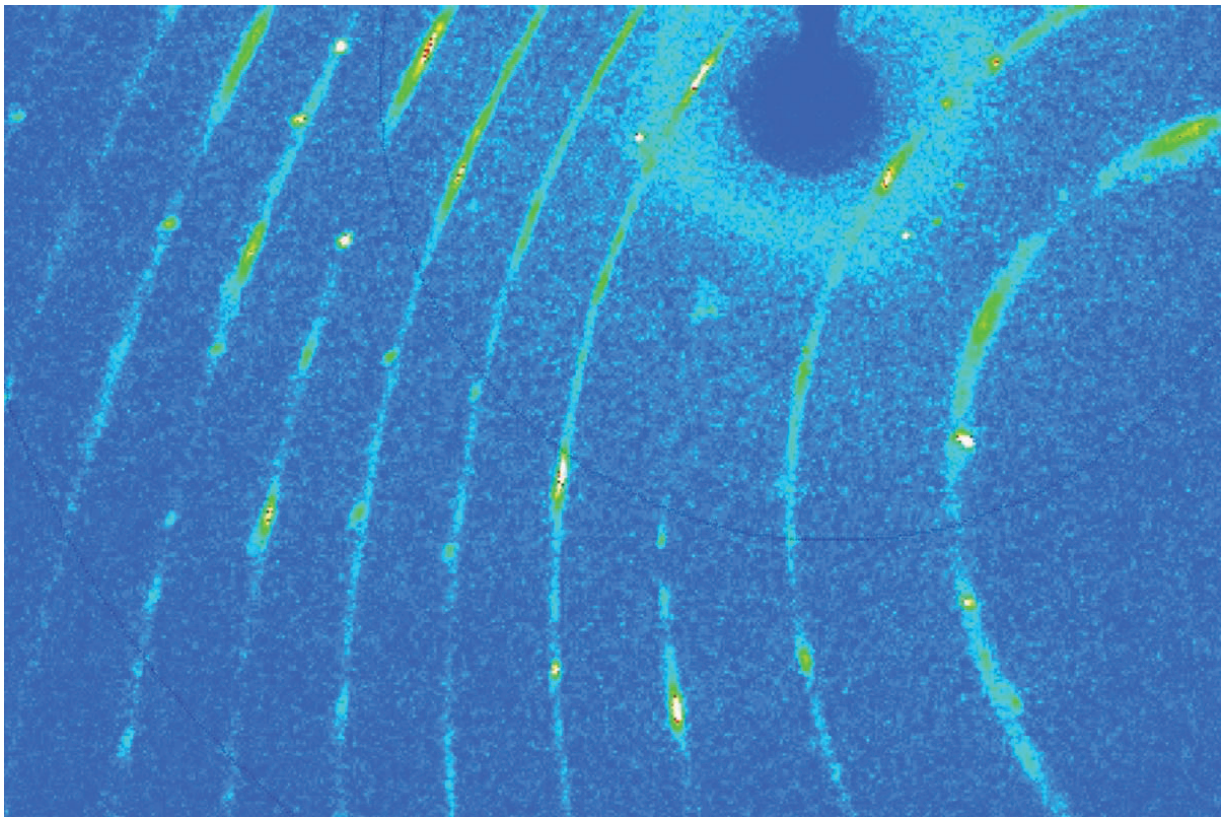
This talk will explore the part we can play in helping the reliable sharing of data between disciplines and how we are supporting efforts to adhere to new best practices for data management that will enable researchers to get the most from crystal structure data. It will focus on the steps the CCDC has made to increase interoperability, as well as some of the opportunities and challenges we face as we work to ensure that the wealth of information contained in 1.2 million structures in the CSD can exploited by machines as well as people from across disciplines.

Keywords: CSD, FAIR, interoperability

Feature article *IUCr Newsletter* (2020), Vol. **28**, no. 1

Raw data availability: the small-molecule crystallography perspective

Simon Coles and Amy Sarjeant



A diffraction image from a challenging sample that surprisingly yielded a structure sufficient to inform the chemists of their reaction outcome. However, the refinement was unstable and produced unreliable derived geometric parameters.

In a research climate that encourages the application of 'FAIR' data principles (that scientific data be Findable, Accessible, Interoperable and Reusable), crystallography has been able to hold its head high. Development of the Crystallographic Information Framework has led to the standardisation of datafile formats (and, more importantly, the precise codification of machine-readable terms for describing all useful attributes of data and associated metadata). Journals have required deposition of derived data in the form of atomic positional coordinates and displacement parameters, and, in many cases, the structure factors or Rietveld profiles from which models have been derived. The value of collecting these results as aggregated collections of searchable structural models in databases and journals has been well demonstrated over the last half-century.

However, in recent years, in line with concerns about research reproducibility (a 2018 Special Collection of *Nature* articles illustrates concerns in a variety of disciplines), focus in crystallography has shifted towards the desirability – or need – to retain and make available the primary experimental data ('raw' data) coming from the instruments. The IUCr commissioned a Diffraction Data Deposition Working Group (DDDWG) to consider the motivation and value of routinely storing and making available raw data sets.

The DDDWG's final report (2017) spanned all the IUCr Commissions and the opportunities for these communities to harness the massive increases in archiving capabilities, even for raw data. The first of 14 Recommendations was:

> 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.

Several case studies across biological and chemical crystallography and powder diffraction were published

to demonstrate the value of preserving raw data (Helliwell *et al.*, 2017).

At the inaugural meeting of the IUCr Committee on Data (CommDat) during the August 2017 IUCr Congress in Hyderabad, India, a start was made on discussing the Commissions' reactions to the DDDWG's final report. Subsequent detailed discussion in the Commission on Biological Macromolecules led to an article published in IUCr Journals with a specific implementation plan encouraging the systematic archival of experimental data with trusted repositories, such as the PDB, and linking to raw data from publications where possible (such linking is available from IUCr Journals, for example) (Helliwell *et al.*, 2019). The powder diffraction community has also begun to react to the DDDWG's final report, see for example, Aranda (2018).

However, for structural chemistry, a general hypothesis was aired during the CommDat meeting that it is not necessary for small-molecule crystallographers to make all raw data available, although there are many clear cases where this is desirable. In most cases, small-molecule data are clean and simple. If, in an ideal situation, it is possible to demonstrate that all Bragg diffraction has been accounted for, that there is nothing of interest remaining in the images and that they have been processed into structure factors appropriately, then why would it be necessary to retain the raw data?

Surveying the field

Here follows a summary of our exploration of this hypothesis, the first step of which was a survey (announced in the *IUCr Newsletter* and sent to a number of individuals) questioning current raw data management practices, as these underpin the ability to make the data available in the first place. Following our summary of the survey results presented here, we will hold a chemical crystallographers' workshop at the 2020 IUCr Congress and General Assembly from which we intend to develop initial guidance on best practices for archiving small-molecule crystallographic raw data and making it publicly available.

While our personal experiences led us to believe that the small-molecule community lagged behind other disciplines in their readiness to store and share raw data, we were curious to know if this was truly the case and, more importantly, why. The methodology used was the design of a short online survey in consultation with CommDat colleagues, which was then disseminated through various crystallographic and social networks. A total of 193 responses were received from around the world, representing a good cross-section of academia, industry, government laboratories, researchers, professors and staff crystallographers. The questions, responses and raw survey data have been deposited in the Chemical Crystallography Community grouping of the Zenodo repository (https://doi.org/10.5281/zenodo.3673958).

Current raw data archiving practice

While a majority of survey respondents, nearly 80%, claim they do archive their raw data in response to a binary yes/no question, more probing questions reveal that our initial suspicions were largely correct. Few people archive their raw diffraction data in a systematic, searchable, robust and secure way. In the main, data archiving is predominately performed 'in-house' – that is, on laboratory or office computers and with little backup, either off-site or to a secondary hard drive. Only a very few respondents store their raw data on facility/institutional archives, and almost nobody uses an independent or commercial cloud-based approach.

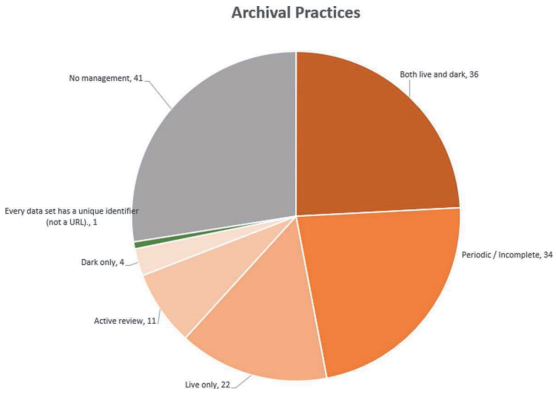


Fig. 1. Data management strategies amongst chemical crystallographers.

For those who do archive their raw data, more than a quarter have no management of their archives. The rest claim some management practices, though these are often 'dark', *i.e.* not accessible by others and to be used only for disaster recovery, or they are intermittent and incomplete (Fig. 1). The majority of archives are not only inaccessible to all but the facility manager but also unsearchable. By unsearchable, we mean that there is no structuring of the information or indexing based on a number of key identifiers/fields. Any search capability that does exist relies on operating-system-based tools and pivots on a single identifier, such as folder name (which invariably relates to a sample identifier). Sometimes, there is additional grouping, but this is in arbitrary categories such as year of data collection or which diffractometer was used (although such categorisation may suggest suitable metadata for structuring an efficient distributed search/indexing system). This behaviour leads to a situation where only those 'in the know' are able to find data for a particular experiment. Though it is admirable that the majority of respondents to the survey do currently take some steps

to back up their raw data, clearly more can be done and there are likely to be a number of 'quick wins' that can be achieved at little or no cost (a considerable contributing factor) simply by raising awareness.

The availability of raw data

The predominant culture in small-molecule crystallography appears to be somewhat protective: 152 out of 186 respondents declare that data are not made readily available to collaborators. This is likely to reflect the fact that many of these facilities are relatively small-scale enterprises run by, at most, a handful of people, many of whom have to charge service fees in order to continue operating. This is generally in sharp contrast to macromolecular crystallography facilities. Macromolecular crystallography has taken the lead in making experimental data freely available, but while many great initiatives from this community have been followed by other crystallography subdisciplines in the past, this may be an approach that is less attractive in small-molecule crystallography. Another aspect of raw data availability is compliance with mandates imposed by funding institutions to make all experimental data openly available. Of our survey respondents, more than half claimed to be unaware of funding mandates, yet suspected that they must exist. Most respondents would also be willing to follow policies to make raw data sets for funded research accessible at some time after measurement, roughly evenly split between those who would do so for all their raw data sets and those who would do so only where mandatory. There were 17 respondents who claimed that they would not comply with such a policy, even if such compliance were mandatory. Generally, the aspirations of funders in adopting these approaches are well founded, in that they want data (generally funded by taxpayers) to be more widely exploited. However, the state of policies varies significantly around the world, particularly in terms of implementation and policing. These mandates can, therefore, be very polarising in the research community and often grudgingly followed at minimum compliance levels only – they are rarely embraced for their aspirations or intentions.

To understand why some may be resistant to or unable to comply with data archiving, it is revealing to look at the reasons why facilities might not currently archive or make their data available. Forty percent of those who do not archive their raw data (14 out of 35) believe this is an unnecessary measure. However, the most common reason people don't back up their data appears to be a lack of appropriate infrastructure. Some respondents declare a lack of ability as a reason, and we suspect, therefore, that 'infrastructure' refers not only to supplying space for storage but also to the tools required to manipulate, transform, annotate, search and validate raw data. A lack of finances is also a clearly stated factor, with more than half of respondents stating that they are unwilling to absorb the costs associated with archiving and managing raw data. The current funding and operating models of small-molecule crystallography facilities are quite simply unable to cater to these aspects – this is unlikely to change and points to the requirement for community-level and centralised solutions and approaches.

The value of making raw data available properly

A key factor often overlooked in data management activities is incentivisation. Initially, there is often resistance to changes in routine practice, and when these changes are imposed by external institutions or funding agencies, they are all too often viewed as punitive. It is, therefore, important to understand what our community sees as the value of good data management and to articulate this clearly. Responses to the survey in this regard were generally positive, with many of the opinion that making raw data available would enable new scientific insights (Fig. 2). Validation of results is a clear driver for making data available, while training and methods development were also considered worthy incentives.

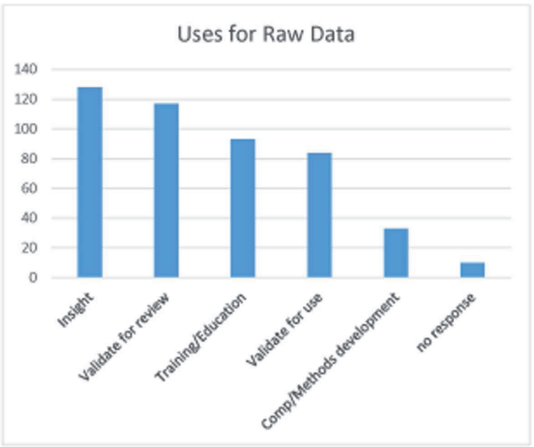


Fig. 2. The benefits of having access to raw data, as identified by the survey respondents.

It is also worth noting that this discussion has centred around service crystallography – the predominant mode of operation for most facilities. However, there is a large and growing element of crystallography that we will group here as 'advanced techniques', including, but by no means limited to, quantum crystallography, dynamic crystallography (covering numerous methods), neutron scattering, electron diffraction, nuclear magnetic resonance crystallography and diffuse-scattering analysis. These advanced techniques are becoming foundational methods for many areas of research – and generally produce diffraction data of a lower standard

(or requiring a deeper level of analysis) than that commonly accepted in service crystallography. Arguably, it is these advanced techniques that will have the strongest need for robust raw data management, validation and sharing – although there will of course be commonalities with some of the tougher service crystallography examples such as disordered and incommensurate structures. In all of these cases, the drivers for sharing will be to generate further insight – or that future methods might be able to extract more information without the need to repeat the experiment.

Centralised neutron facilities have had an established approach to raw data preservation for some time and this has more recently led to a greater degree of availability as the ability to 'publish' has developed, *i.e.* repository capability of assigning DOIs and opening up to the internet – see, for example, ILL[1] and ISIS[2] data management and DOI policies. In the macromolecular community, raw data archiving is becoming just another part of the process of publication – largely in the face of fraud and out of a sense of being able to do a better job of modeling the structure oneself. Most small-molecule crystallographers would agree that routine well-behaved structures don't need re-refining just to measure bond distances or angles, let alone re-integrating the entire dataset. However, the respondents of this survey do seem to agree that in cases of difficult refinements, disorder, twinning and modulation, having access to the raw diffraction images would benefit the community. Additionally, more than half of all respondents felt access to raw data was essential in examining pathological samples or for validating scientific claims and quality (Fig. 3).

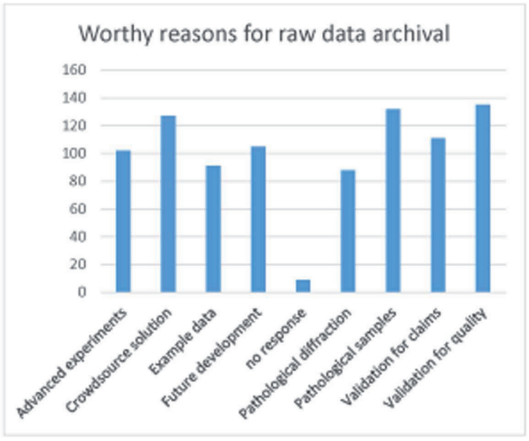


Fig. 3. Motivation for archiving raw data sets.

Clearly, while the community does seem to agree that raw data archiving is a worthwhile practice, most lack the funds, capability, infrastructure and motivation to do so in a structured way (particularly one that would then readily enable wider sharing). Yet, as already proven by the macromolecular community, there is a real benefit to having raw data available. When we look back over the evolution of crystallographic data sharing, more crystallographers made more of their experimental data available once a standard format for data became available – the CIF (Crystallographic Information File). Though standard formats for sharing experimental data, *e.g.* FCF, existed around the same time as the CIF, it did not become routine for crystallographers to share their structure factors until relatively recently when this became an automated output of the refinement and embedded in the CIF. In the same vein, we should now consider how sharing of raw data can be made an automatic part of publishing crystallographic data. What steps can we, as a community, take to ensure this valuable practice becomes part of everyday operations? Who should bear the costs involved and the responsibility for maintaining such an archive?

In conclusion, we focus on two factors in this article – data management practice and the sharing of data to support scientific assertions or findings. While we use the term archiving to refer to data management practices in general, we understand that this includes the aspect of locking data away to keep it safe, as opposed to sharing it for the greater good. With a modest amount of culture change and relatively little extra money or effort, it should be possible for small-molecule crystallography facilities to manage their raw data so that it is easy to make it more widely available. However, without clear guidance on when to make data available and without centralized tools and infrastructure to support the process, widespread data sharing will continue to be elusive in small-molecule crystallography. Where is that guidance to come from, and who will create and maintain the necessary tools and infrastructure?

The authors would like to acknowledge John Helliwell and Brian McMahon for their significant input during the preparation of this article.

References

Aranda, M. A. G. (2018). *J. Appl. Cryst.* **51**, 1739–1744.

Helliwell, J. R., McMahon, B., Guss, M. & Kroon-Batenburg, L. M. J. (2017). *IUCrJ*, **4**, 714–722.

Helliwell, J. R., Minor, W., Weiss, M. S., Garman, E. F., Read, R. J., Newman, J., van Raaij, M. J., Hajdu, J. & Baker, E. N. (2019). *Acta Cryst.* D**75**, 455–457.

[1] https://www.ill.eu/users/user-guide/after-your-experiment/data-management/
[2] https://www.isis.stfc.ac.uk/Pages/Digital-Object-Identifiers-(DOIs)-for-ISIS-Data.aspx

IUCrData

ISSN 2414-3146

IUCrData launches Raw Data Letters

L. M. J. Kroon-Batenburg,[a]* J. R. Helliwell[b] and J. R. Hester[c]

[a]Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands, [b]Department of Chemistry, The University of Manchester, Manchester M13 9PL, United Kingdom, and [c]Australian Nuclear Science and Technology Organisation, Locked Bag 2001, Kirrawee DC, NSW 2232, Australia. *Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Keywords: Raw Data Letters; imgCIF.

IUCrData, the peer-reviewed open-access data publication from the International Union of Crystallography (IUCr), is launching a new section – Raw Data Letters. This is a collaborative innovation of IUCr Journals with the IUCr Committee on Data. Future raw data sets will become increasingly large and no one group will be able to analyze all of the scientific content in a timely manner. As already occurs in other scientific disciplines (*e.g.* astronomy, particle physics) others will need to have access to raw crystallographic data sets so that open-science-based research can proceed at a rapid pace. However, proper credit needs to be attributed fairly among those who design experiments and collect data, and those who subsequently use the data to establish new results.

With these points in mind, the new section will publish short descriptions of crystallographic raw data sets from X-ray, neutron or electron diffraction experiments, in the biological, chemical, materials science or physics fields, and provide a persistent link to the location of the raw data. The letters in this section will describe interesting features in raw data sets, allowing researchers to attract attention to particular aspects of the data that could be of interest to methods and software developers for purposes such as reanalysis by newer methods or may be relevant to the structural interpretation. We envisage different types of Raw Data Letters. The structure could have been solved and published elsewhere, but the letter describes interesting features that were observed but ignored in the original structure determination. The letter could describe remarkable features but no attempt is made as to their interpretation; in this way the data attract attention and the original authors get credit for their work. The raw data described in a letter show Bragg reflections to a reasonable resolution but the structure could not be solved. Again the authors would get credit for their work. Also letters describing the reuse of publicly available data by improved methods are welcome. In general, publication in Raw Data Letters promotes data retrieval by other scientists and will enhance visibility of the data. Raw Data Letters support Open Science policies: no research data should be lost but should be made available to the research community according to the FAIR principles, for which the correctness and completeness of the metadata are crucial, and these will be central to the reviewing process.

Science funders and policy makers are working increasingly towards the Open Science model to make science useful for society. Good scholarship demands openness and transparency of protocols and scientific results as well as proper data management and validation of scientific knowledge. Many open science platforms have seen the light, *e.g.* the OpenAIRE project (https://www.openaire.eu) and the European Open Science Cloud (EOSC, Jones, 2015) promoting the sharing of data. Guidelines for proper data management are described in *The FAIR principles for scientific data management and stewardship* by Wilkinson *et al.* (2016), which requires research data to be Findable, Accessible, Interoperable and Reusable.

What does this mean for crystallographic data?

Findable: the data are assigned a persistent identifier, they come with metadata by which they are indexed in a searchable resource.

Accessible: the data should be retrievable through well established communication protocols.

Interoperable: the data use a shared (documented) broadly applicable language for knowledge representation.

Reusable: the metadata should accurately describe experimental attributes of the data, and are released with a clear and accessible data usage license. Data formats should be described.

Table 1
Core metadata list and their data names in imgCIF.

Minimal metadata	imgCIF data name
Data binary format	_array_structure.byte_order, _array_structure_compression_type, _array_structure.encoding_type
Number of pixels, pixel size (binning mode)	_array_structure_list.index _array_structure_list.dimension _array_structure_list_axis.displacement_increment
Beam center (mm)	_axis.offset[1..3] (needs _axis category, see below) _diffrn_scan_frame_axis.displacement
Origin of data frame	_array_structure_list.precedence (1 or 2) _array_structure_list.direction (increasing or decreasing)
Wavelength	_diffraction_radiation_wavelength.value(/wavelength) or _diffraction_radiation.type
Rotation axis	_diffrn_scan_axis.axis_id _diffrn_scan_axis.angle_start _diffrn_scan_axis.angle_range
Rotation range per frame/number of frames	_diffrn_scan_axis.angle_increment, _diffrn_scan.frames _diffrn_scan_frame.frame_number
Axes and offsets	_axis.id, _axis.type, _axis.depends_on, _axis.vector[1] _axis.vector[2] _axis.vector[3] _axis.offset[1] _axis.offset[2] _axis.offset[3]
Detector-to-sample distance	_diffrn_scan_axis.displacement_start _diffrn_scan_axis.displacement_range
Conditions	_diffrn.ambient_temperature

The IUCr has always taken a leading position in data sharing by linking publications to coordinates and structure factors as well as validation reports. In chemical crystallography the checkCIF tool, as part of the submission system, ensures consistency and integrity of the data. Likewise, in macromolecular crystallography a paper describing a crystal structure has to link to the Protein Data Bank (PDB) entry, while the wwPDB deposition system generates a validation report. The IUCr established a working group in 2011 (DDDWG) to address 'the growing calls within the crystallographic community for the deposition of primary diffraction images, with some mechanism that allows their retrieval by other scientists for such purposes as reanalysis, software and methods development, validation and review' (see https://forums.iucr.org/). The final report of the DDDWG made a series of recommendations, the first two of which are as follows.

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.

The coordinating and advisory role of the DDDWG has been continued by the Committee on Data (CommDat), which was established by the IUCr in 2016. This committee plays a significant role in the current initiative.

A series of papers appeared in *Acta Cryst. D* in 2014 (Guss & McMahon, 2014; Kroon-Batenburg & Helliwell, 2014; Terwilliger & Bricogne, 2014; Meyer *et al.*, 2014) discussing the possibilities of raw-data depositions either in centralized facilities or distributed repositories. Whereas at that time the possibilities of data transfer and estimated costs of storage and curation were seen as a barrier, in recent years several freely accessible repositories have become available that make routine raw-data deposition feasible, and the bandwidth of internet connections has also increased substantially. A recent editorial from IUCr journals (*FAIR diffraction data are coming to protein crystallography*, Helliwell *et al.*, 2019) encourages authors to provide a DOI for their original raw data when submitting their article.

In a topical review in *IUCrJ* (Kroon-Batenburg & Helliwell, 2014), the requirements for metadata are discussed. Without correct and complete metadata we would certainly not adhere to the FAIR principles as the reusability would be compromised.

We also note and have warmly welcomed the PDBj initiative in 2021 to launch its own raw diffraction data archive

(https://xrda.pdbj.org/), which is integrated with PDBj, and which opens up the FAIR principles and data record right back to the time of measurement in a depositor's research studies.

With Raw Data Letters we want to elicit re-use of the original data. Diffraction data can be from various disciplines: macromolecular crystallography, chemical crystallography, XFELs, synchrotron serial crystallography, materials science powder diffraction *etc*. These come with many different data formats and varying metadata quality. To ensure reusability, metadata should be accurate and complete and at least sufficient. For single-crystal data we have made a list of core metadata; their presence is a key requirement for correct (automated) reprocessing of the data. Our list is a superset of the NeXus/HDF5 NxMx Gold Standard that was developed by Bernstein *et al.* (2020). We decided to capture metadata in imgCIF (Bernstein & Hammersley, 2005; Hammersley *et al.*, 2005), which is well known to the community via its CBF variant, and already includes the appropriate data names for our core metadata list (see Table 1). imgCIF also has the advantage of being a plain-text format, which allows editing with familiar tools and provides excellent guarantees of readability over the long term.

The current core metadata list has been specifically developed for single-crystal data; it will have to be extended for powder diffraction and high-pressure data; XFEL data will also need additional information.

There are two aspects to metadata that need attention after having established the core metadata list: (1) we need to generate the metadata from the original raw data format, possibly completed by user-supplied metadata, and (2) we need a checking procedure for checking consistency and correctness that can be used on the IUCr webserver.

A separate working group has been developing tools for extracting metadata information from raw images. A key problem is that raw data, unlike structural data, may be deposited in one of a multitude of formats currently in use. The situation for large-scale facilities is somewhat different to that for home diffractometers. Large-scale facilities often have PILATUS detectors that use CBF binary data format and miniCBF (ASCII) headers or EIGER detectors that use the Nexus/HDF5 data structure; at the same time CCD detectors may still be in use. Home diffractometers often have manufacturer-developed detectors and binary image formats. Most home diffractometers also have multi-circle diffractometers that make the description of metadata more complicated.

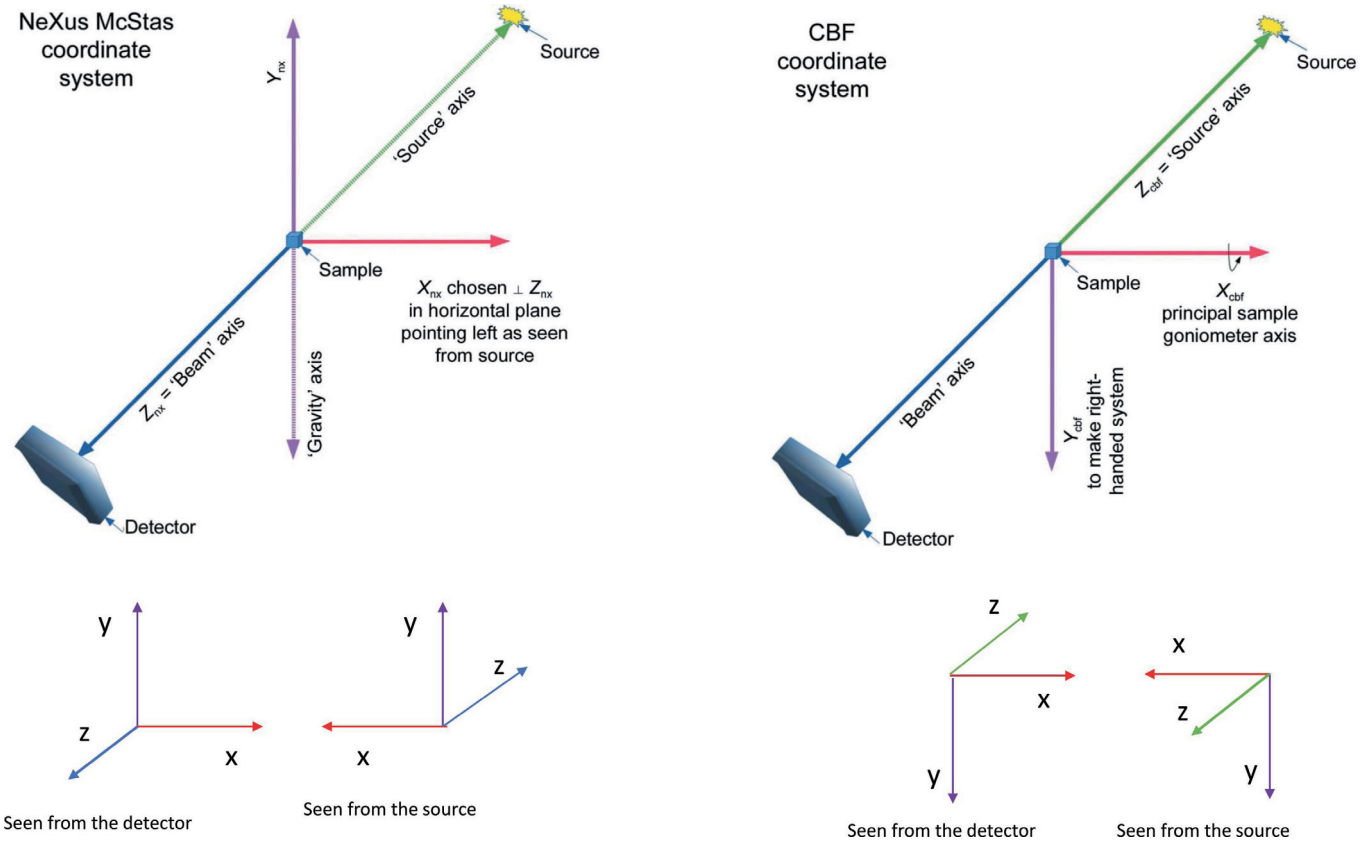


Figure 1
The Nexus McStas axis system is different from the CBF laboratory system. Seen from the source in McStas, *Z* points away from the source and the *X* axis points to the left, while in CBF *Z* points back into the source and *X* is along the principal goniometer axis, *i.e.* points into the goniometer block. A vector defined in McStas is brought by a 180° rotation around Nexus_X to the CBF system. Note that if the goniometer is at the opposite side, the transformation is around Nexus_Y (image courtesy of the Gold Standard paper, Bernstein *et al.*, 2020).

Practical issues have forced us to compromise on the DDDWG recommendations above. We have adopted an approach where the metadata, in the form of an imgCIF file, is separated from the raw data frames, to avoid the need for unfeasibly large data downloads during the validation process. The imgCIF file required during submission instead contains internal pointers to the raw data as URLs. The use of URLs, rather than more robust DOIs, is an additional compromise resulting from the requirement that checks be able to access the actual data files containing the data frames; there is no standardized way of finding the data files from data DOIs as the DOI usually resolves to an informational 'landing page'. To mitigate against the URL fragility that DOIs were designed to avoid, the accessibility of all repositories referenced by archived imgCIF files held by the journal will be regularly checked, and in the rare case that URLs become inaccessible they will be manually updated using the data DOI provided by the authors during submission. Once the data DOI specifications mature to the point that data files can be directly referenced, our approach envisages that the URL form of a data DOI would be used in the submitted imgCIF file. We understand that our solution is necessarily a compromise, and will continue to refine the way in which raw data are handled as we gain experience and receive feedback from the community.

For the first letters published, we generated complete and consistent imgCIF files using the CBF laboratory axis system. For the macromolecular letter (Neviani *et al.*, 2022), we used a Python script written by Fabio Dall'Antonia and Julian Hörsch (European XFEL) to convert the metadata information from a Nexus/HDF5 master file into CBF/imgCIF. This particular Diamond Light Source Nexus/HDF5 (meta)data follows the 'Gold Standard' defined by the HDRMX working group (Bernstein *et al.*, 2020). The CBF laboratory axis system is different from that of Nexus McStas. Fig. 1 explains the details.

Most diffractometer manufacturers provide software for conversion to CBF, mostly miniCBF headers and CBF binary data. Bruker AXS was very helpful in providing a tool for image conversion to full CBF files. We used these files as the basis for our imgCIF file in the letter describing the twinned form of *o*-nitroaniline (Lutz & Kroon-Batenburg, 2022).

Procedures and tools for the generation of imgCIF files from the available image data files are being incorporated into the IUCr Journals submission system. This is ongoing work, as we have to deal with many different formats, and contributions from the community are welcome. Further information on tools for authors will available shortly.

Submissions are now being welcomed – updated Notes for authors and submission instructions are available from the *IUCrData* website. This is early days for the Raw Data Letters section. Our Co-editors (see https://iucrdata.iucr.org/x/services/editors.html) are very keen to work with authors to facilitate publication of their data. We look forward to receiving your Raw Data Letters.

Acknowledgements

We thank the members of the working group for discussion and constructive contributions to the development of the format of the Raw Data Letters and checkCIF tools: Fabio Dall'Antonia, Julian Hörsch, Andy Götz, Brian McMahon, and the IUCr Journals Editorial Office. We thank Joerg Kaercher (Bruker AXS) for providing a tool for conversion from Bruker sfrm format to full CBF, and for fruitful discussions.

References

Bernstein, H. J., Förster, A., Bhowmick, A., Brewster, A. S., Brockhauser, S., Gelisio, L., Hall, D. R., Leonarski, F., Mariani, V., Santoni, G., Vonrhein, C. & Winter, G. (2020). *IUCrJ*, **7**, 784–792.

Bernstein, H. J. & Hammersley, A. P. (2005). *International Tables For Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 37–43. Chester: International Union of Crystallography.

Guss, J. M. & McMahon, B. (2014). *Acta Cryst.* D**70**, 2520–2532.

Hammersley, A. P., Bernstein, H. J. & Westbrook, J. D. (2005). *International Tables For Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 444–458. Chester: International Union of Crystallography.

Helliwell, J. R., Minor, W., Weiss, M. S., Garman, E. F., Read, R. J., Newman, J., van Raaij, M. J., Hajdu, J. & Baker, E. N. (2019). *Acta Cryst.* D**75**, 455–457.

Jones, B. (2015). *Towards the European Open Science Cloud.* http://doi.org/10.5281/zenodo.16001.

Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). *Acta Cryst.* D**70**, 2502–2509.

Lutz, M. & Kroon-Batenburg, L. M. J. (2022). *IUCrData*, **7**.

Meyer, G. R., Aragão, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). *Acta Cryst.* D**70**, 2510–2519.

Neviani. V., Lutz, M., Oosterheert, W., Gros, P. & Kroon-Batenburg, L. M. J. (2022). *IUCrData*, **7**, x220852.

Terwilliger, T. C. & Bricogne, G. (2014). *Acta Cryst.* D**70**, 2533–2543.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). *Sci. Data*, **3**, 160018.

IUCrData

ISSN 2414-3146

Crystal structure of the second extracellular domain of human tetraspanin CD9: twinning and diffuse scattering

Viviana Neviani, Martin Lutz, Wout Oosterheert, Piet Gros and Loes Kroon-Batenburg*

Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands. *Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Received 5 July 2022
Accepted 24 August 2022

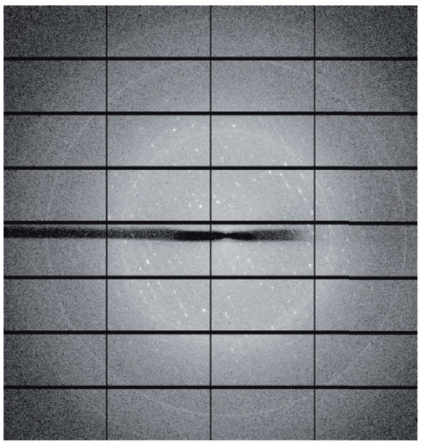
Edited by J. R. Helliwell, University of Manchester, United Kingdom

Keywords: twinning; diffuse scattering; tetraspanin $CD9_{EC2}$; raw data.

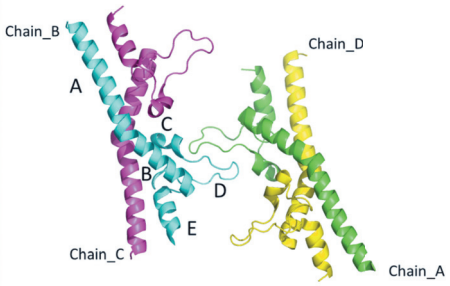
Structural data: full structural data are available from iucrdata.iucr.org

Remarkable features are reported in the diffraction pattern produced by a crystal of the second extracellular domain of tetraspanin CD9 (deemed $CD9_{EC2}$), the structure of which has been described previously [Oosterheert *et al.* (2020), *Life Sci. Alliance*, **3**, e202000883]. $CD9_{EC2}$ crystallized in space group *P*1 and was twinned. Two types of diffuse streaks are observed. The stronger diffuse streaks are related to the twinning and occur in the direction perpendicular to the twinning interface. It is concluded that the twin domains scatter coherently as both Bragg reflections and diffuse streaks are seen. The weaker streaks along **c*** are unrelated to the twinning but are caused by intermittent layers of non-crystallographic symmetry related molecules. It is envisaged that the raw diffraction images could be very useful for methods developers trying to remove the diffuse scattering to extract accurate Bragg intensities or using it to model the effect of packing disorder on the molecular structure.

Raw data



Structure



HDF5 and CBF data files: https://doi.org/10.5281/zenodo.5886687

Metadata imgCIF file: https://doi.org/10.1107/S2414314622008525/he4557img.cif

Data processing and refinement

This letter gives a detailed description of the raw diffraction data that were used for analysis and structure determination of the second extracellular domain of tetraspanin CD9 ($CD9_{EC2}$) as previously reported (Oosterheert *et al.*, 2020). The raw diffraction images show streaked diffuse scattering, and this feature is detailed here to serve as an example for archiving and re-analysis of raw diffraction data. $CD9_{EC2}$ crystallized in

space group *P*1, has four molecules in the asymmetric unit, arranged as a dimer of domain-swapped dimers (Fig. 1). The crystal was non-merohedrally twinned with a twofold rotation about **a*** + **b*** as the twinning operation.

With the *EVAL* software suite (Schreurs *et al.*, 2010) two lattices could be indexed and reflections of the largest domain were integrated while overlapping reflections of the second domain could be largely deconvoluted, leaving only 21.5% of reflections overlapping.

Initial de-twinning was performed with the *TWINABS* software (Sheldrick, 2009) based on observed structure factors. These data were used for structure solution by molecular replacement as described previously (Oosterheert *et al.*, 2020). Final detwinning was based on calculated structure factors and final refinement rounds in *Refmac*5 yielded R_{work}/R_{free} = 23.9/27.9%.

Structural details can be found in (Oosterheert *et al.*, 2020) and in the Protein Data Bank under the accession code 6rlr. Data collection details and statistics are listed in Table 1.

Data description

Data were collected at Diamond Light Source (DLS) beamline I-04, in total 3600 images in 0.1° fine sliced mode using a single rotation axis. The diffraction data were written in HDF5 format. Since, at the time of data processing, *EVAL* could not read the HDF5 files, they were converted by a tool at DLS to CBF format, using mini-cbf headers, and these were processed by *EVAL*. Both HDF5 (.h5) and CBF (.cbf) data files have been deposited in Zenodo. Indexing of peaks in the diffraction data with *DIRAX* (Duisenberg, 1992) indicated non-merohedral twinning of the crystal with a twofold rotation around the **a*** + **b*** diagonal as the twinning operation (Fig. 2). Concurrent with twinning, diffuse streaks are seen in the diffraction. Diffraction images were mapped to reciprocal space, to a resolution of 4 Å, using *IMG2HKL* in *EVAL* (Schreurs *et al.*, 2010), first by merging images in groups of 5, and then carefully redistributing intensities while correcting for Lorentz and polarization factors. The reciprocal space map was merged using Laue symmetry −1. The reciprocal space

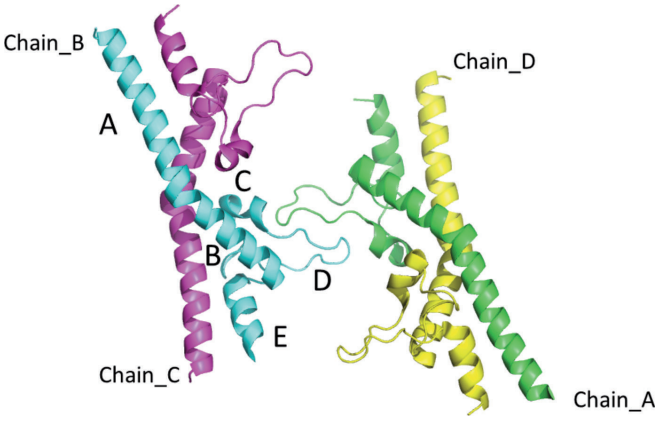


Figure 1
Asymmetric unit of the twinned $CD9_{EC2}$ crystal, coloured by protein chain. Domains are labelled *A*–*E* in the cyan-coloured chain. The *D* loop is flexible as follows from comparison of different structures involving $CD9_{EC2}$ and structures of EC2 domains from other tetraspanins (Oosterheert *et al.*, 2020). The direction of view is approximately along the non-crystallographic (NCS) twofold axis that coincides with the twofold twin axis **a*** + **b*** (see text). The NCS operation transforms chain *A* into chain *B*, and chain *C* into chain *D*.

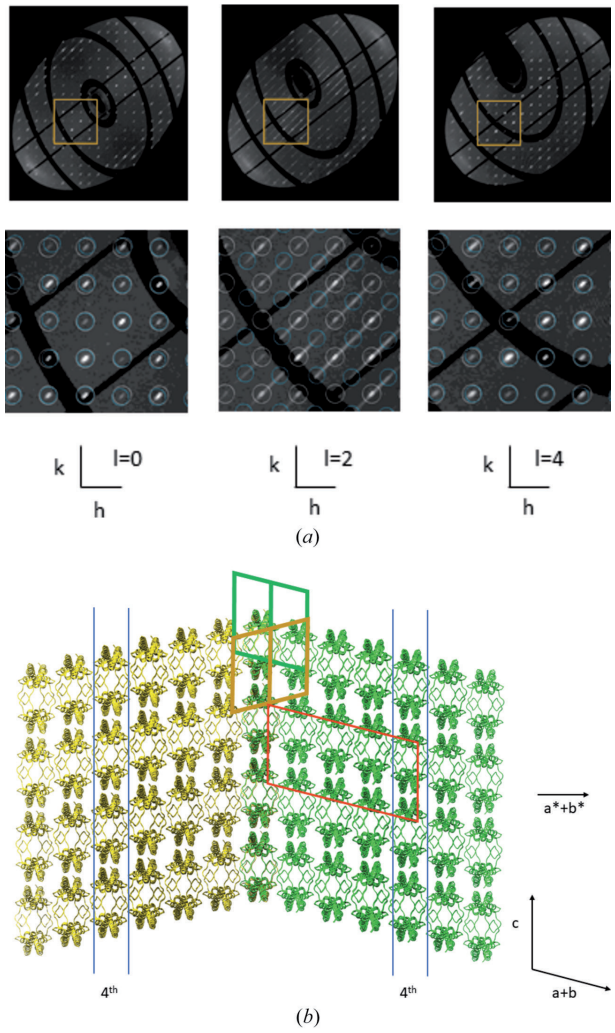


Figure 2
Twinned $CD9_{EC2}$ crystal. (*a*) Reciprocal space reconstructions, *hk*0, *hk*2 and *hk*4. The lower panels are zoomed in at the area of the yellow box, where Bragg reflections originating from lattice 1 and 2 are coloured white and cyan, respectively. In slice *hk*2, streaks along **a*** + **b*** are evident. All slices $hk(l = 4n)$ are ordered and the spots indexed by the two matrices nearly overlap, while in slices $hk(l = 4n + 2)$ they have maximum separation. (*b*) Green and gold structures represent the two twin domains in the crystal. The second domain is rotated by 180° around **a*** + **b*** with respect to the first. The twin interface is the plane in the middle with base vectors **c** and **a** − **b** for either lattice. In the figure, molecules of the two domains are overlayed on this layer. The 180° twin rotation applied to domain 1 causes chains *A* and *C* of the molecules in domain 1 (green) to superimpose on chains *B* and *D* of domain 2 (gold), respectively. Starting from the interface the 4th layers (between the blue lines) in the two domains are the same and form an ordered array. A super cell (red) can be constructed with transformation matrix (1,−1,0/0,0,1/−2,−2,1) on which the two twin lattices overlap. The consequence is that in reciprocal space reflections of the twin lattices overlap for every $l = 4n$.

map is available as a *CCP*4 map (https://doi.org/10.5281/zenodo.6961763), that can *e.g.* be viewed with *Chimera* (Pettersen *et al.*, 2004). Fig. 2(*a*) shows three sections through reciprocal space on an *hkl* grid, where evident streaks are running in the **a*** + **b*** direction, particularly in the *hk*2 section.

In the following we draw conclusions on the origins of the diffuse scattering features we observed. A variety of diffuse scattering features in macromolecular crystals of different origin is discussed by Glover *et al.* (1991). For an extensive treatment of diffuse scattering in proteins, we refer to the paper on the Gag protein by Welberry *et al.* (2011). Geometrical frustration of the packing of two molecular configurations of the Gag protein led to circular diffuse scattering features. The origin of our diffuse scattering of $CD9_{EC2}$, though, is different; we only have streaked diffuse scattering that is caused by stacking disorder of layers. We summarize here what we concluded in the Oosterheert paper on the origin of the diffuse scattering (see Fig. 2 for details).

The twinning interface is a layer with base vector **c** and **a** − **b**. The two twin domains each grow from this interface along their respective **a*** + **b*** directions. For every fourth layer the two structures exactly overlap.

Reflections can be indexed on a so-called stacking lattice (Dornberger-Schiff, 1956; Lutz & Kroon-Batenburg, 2018) with dimension 1/4*c*. On this lattice the twin structure is completely ordered and as a result the reciprocal space slices at l = 4*n* have only ordered Bragg spots.

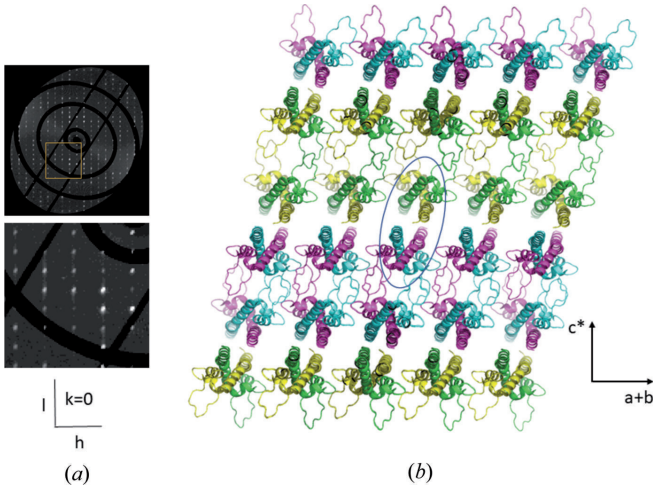


Figure 3
Diffuse scattering along **c***. (*a*) Reciprocal space reconstruction of *h*0*l* layer and zoomed in at the area of the yellow box. Intensities are on the same scale as in Fig. 2(*a*). Streaks are observed along **c***, but much weaker than those along **a*** + **b***. (*b*) Model of a single domain in which one layer is replaced by a rotation copy generated by the NCS rotation operation, which coincides with the twinning, *i.e.* 180° rotation around **a*** + **b***. The colouring of the chains is the same as in Fig. 1, but the asymmetric unit used here (indicated by an ellipse) is such that the loops are on the outside. Inclusion of a layer of rotated copies will disrupt the periodicity perpendicular to the *ab* layer which is the **c*** direction, leading to diffuse streaks in the diffraction pattern.

Table 1
Experimental details.

Data for the highest resolution shell are given in parentheses.

Raw data	
DOI	https://doi.org/10.5281/zenodo.5886687
Data archive	Zenodo
Data format	HDF5 and CBF
Data collection	
Beamline	Diamond I04
Detector type	EIGER 16M
Radiation type	Synchrotron X-ray source
Wavelength (Å)	0.979491
Beam centre (mm)	−166.87, 172.50
Detector axis	$-Z$
Detector distance (mm)	287.22
Pixel size (mm)	0.075 × 0.075
No. of pixels	4148 × 4362
No. of scans	1
Scan axis	ω, X
Start angle, increment per frame (°)	0.0, 0.1
Scan range (°)	360
No. of frames	3600
Exposure time per frame (s)	0.01
Crystal and refinement data	
Resolution range (Å)	29.02–2.0 (2.07–2.0)
Space group	*P*1
Cell dimensions *a*, *b*, *c* (Å)	39.986, 39.998, 63.643
Cell angles α, β, γ (°)	80.39, 76.29, 68.15
Total no. of reflections	76175 (6985)
No. of unique reflections	22863 (2180)
Completeness (%)	95.6 (93.0)
Multiplicity	3.4 (3.4)
$I/\sigma(I)$	4.8 (0.8)
R_{merge}	0.10 (1.15)
$R_{p.i.m.}$	0.07 (0.73)
$CC_{1/2}$	0.998 (0.322)

All the other slices have streaks in **a*** + **b*** which is the direction of the packing disorder.

A twofold NCS (non-crystallographic symmetry) operation transforms the independent molecules into one of the others; the corresponding axis co-aligns with twin axis **a*** + **b***. Chain *A* superposes with chain *B* (r.m.s.d. 0.482 Å) and chain *C* with chain *D* (r.m.s.d. 0.460 Å). The $CD9_{EC2}$ molecule has a flexible *D* loop (see Fig. 1, where the *D* loop is marked for chain *B*). In the structure this loop is located at the twin interface. A crystal consisting of small twin domains is coherently scattering over a length scale determined by the coherence length of the X-rays (see Thompson, 2017 for a discussion on order–disorder and twinning). This is the case for our crystal, which gives rise to both Bragg peaks and diffuse streaks [Fig. 2(*a*)].

It was noticed by a reviewer that streaks are also seen in the **c*** direction. This is indeed the case, and they occur for every *hkl* layer containing **c***, but are significantly weaker than the **a*** + **b*** streaks [Fig. 3(*a*)]. The origin of the diffuse streaks lies within a single domain and is unrelated to the twinning. Our reasoning is that due to NCS, local rotation of molecules can occur, without serious clashes. It may be that within a single domain, a rotated copy of an entire *ab* layer is included in the lattice [see Fig. 3(*b*)], which is conceivable because the molecules are packed through the *D* loops. This leads to

disruption of periodicity in stacking of *ab* layers and to diffuse streaks in the **c*** direction.

We present here the raw diffraction data and the most likely explanation for the diffuse features. Any interested researcher can generate detailed models of disorder, calculate the diffuse scattering and compare them with our data.

References

Dornberger-Schiff, K. (1956). *Acta Cryst.* **9**, 593–601.

Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92–96.

Glover, I. D., Harris, G. W., Helliwell, J. R. & Moss, D. S. (1991). *Acta Cryst.* B**47**, 960–968.

Lutz, M. & Kroon-Batenburg, L. M. J. (2018). *Croat. Chem. Acta*, **91**, 289–298.

Oosterheert, W., Xenaki, K. T., Neviani, V., Pos, W., Doulkeridou, S., Manshande, J., Pearce, N. M., Kroon-Batenburg, L. M. J., Lutz, M., van Bergen en Henegouwen, P. M. P. & Gros, P. (2020). *Life Sci. Alliance*, **3**, e202000883.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *J. Comput. Chem.* **25**, 1605–1612.

Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2010). *J. Appl. Cryst.* **43**, 70–82.

Sheldrick, G. (2009). *TWINABS*. University of Göttingen, Germany.

Thompson, M. C. (2017). *Identifying and Overcoming Crystal Pathologies: Disorder and Twinning*, ch. 8, *Protein Crystallography: Methods and Protocols* in *Methods in Molecular Biology*, Vol. 1607. Clifton: Springer.

Welberry, T. R., Heerdegen, A. P., Goldstone, D. C. & Taylor, I. A. (2011). *Acta Cryst.* B**67**, 516–524.

IUCrData

ISSN 2414-3146

Accurate intensity integration in the twinned γ-form of *o*-nitroaniline

Martin Lutz and Loes Kroon-Batenburg*

Department of Chemistry, Structural Biochemistry, Bijvoet Centre for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands. *Correspondence e-mail: l.m.j.kroon-batenburg@uu.nl

Received 5 July 2022
Accepted 3 November 2022

Edited by S. J. Coles, University of Southampton, United Kingdom

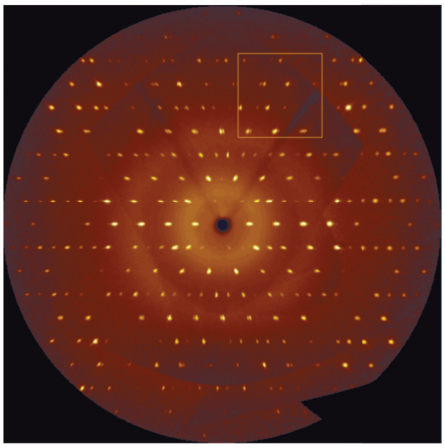
Keywords: twinning; scattering; twin interface; raw data.

CCDC reference: 2217206

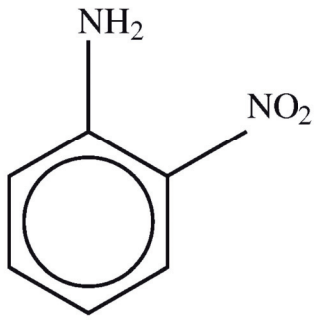
Structural data: full structural data are available from iucrdata.iucr.org

o-Nitroaniline, $C_6H_6N_2O_3$, is known to be polymorphic. The α-form is probably amorphous, while the β- and γ-forms are crystalline. Difficulties with the unit-cell determination of the γ-form were reported as a consequence of twinning. In this paper, newly recorded diffraction data of the γ-form of o-nitroaniline are described that were processed taking into account the two twin lattices. Data were partly deconvoluted and much better agreement was obtained in terms of R_1 values and C—C bond precision. The availability of raw data and proper reprocessing using twin lattices is by far superior to efforts to de-twin processed structure factors.

Raw data



Chemical



Bruker SMART data files and CBF files: https://doi.org/10.5281/zenodo.7193538

Metadata imgCIF file: https://doi.org/10.1107/S2414314622010598/ii4001img.cif

Introduction

o-Nitroaniline is known to be polymorphic (Aakeröy *et al.*, 1998*a*,*b*). The α-form is probably amorphous, while the β- and γ-forms are crystalline. Difficulties with the unit-cell determination of the γ-form were reported as a consequence of twinning. The unit cell appears to be *C*-centered orthorhombic, but was determined to be a pseudo-merohedral monoclinic twin by Herbstein (1965), who also observed diffuse streaks along a^*. Pseudo-orthorhombic twinning together with unusual extinctions was discussed by Dunitz (1964). While he assumed a twin obliquity of 0°, we here show an example where this is not exactly the case, *i.e.* the twin obliquity is 0.743°. The structure was determined before, supposedly from data of untwinned crystals, but the R_1 values of 10.9, 7.01 and 7.98% remain large (Dhaneshwar *et al.*, 1978; Nieger, 2007; Zych *et al.*, 2007). In this paper we describe newly recorded diffraction data of the γ-form of *o*-nitroaniline (I) and

Table 1
Experimental details.

Raw data			
DOI	https://doi.org/10.5281/zenodo.7193538		
Data archive	Zenodo		
Data format	CBF		
Data collection			
Diffractometer	Bruker Kappa APEXII		
Temperature (K)	150		
Detector type	APEXII CCD		
Radiation type	Mo $K\alpha$		
Wavelength (Å)	0.71073		
Beam center (mm)	−30.401, −30.637		
Detector axis	$-Z$		
Detector distance (mm)	41		
Swing angle (°)	−21.52		
Pixel size (µm)	0.12 × 0.12		
No. of pixels	512 × 512		
No. of scans	7		
Exposure time per frame (s)	10		
Scan axis	Start angle, increment per frame (°)	Scan range (°)	No. of frames
ϕ, $-X$ (ω = 164.659°, κ = 46.226°)	74.659, −0.300	−360	1200
ω, $-X$ (κ = −73.760°, ϕ = 10.746°)	−169.393, −0.300	−118.2	394
ω, $-X$ (κ = −73.760°, ϕ = −91.253°)	−169.393, −0.300	−118.2	394
ω, $-X$ (κ = 88.307°, ϕ = 160.033°)	−157.189, −0.300	−82.2	274
ω, $-X$ (κ = 88.307°, ϕ = 58.033°)	−157.189, −0.300	−82.2	274
ω, $-X$ (κ = −73.760°, ϕ = −40.253°)	−169.393, −0.300	−118.2	394
ω, $-X$ (κ = −73.760°, ϕ = 166.747°)	−169.393, −0.300	−118.2	394
Crystal data			
Chemical formula	$C_6H_6N_2O_2$		
M_r	138.13		
Crystal system, space group	Monoclinic, $P21/a$		
a, b, c (Å)	15.2066 (5), 10.0938 (4), 8.3580 (2)		
β (°)	106.693 (3)		
V (Å^3)	1228.82 (7)		
Z	8		
μ (mm^{-1})	0.12		
Crystal size (mm)	0.37 × 0.30 × 0.16		
Data processing			
	Twin	Single lattice	
Absorption correction	Multi-scan (*TWINABS2012/1*; Sevvana *et al.*, 2019)	Multi-scan (*SADABS*; Krause *et al.*, 2015)	
T_{min}, T_{max}	0.683, 0.746	0.628, 0.746	
No. of measured, independent and observed [$I > 2\sigma(I)$] reflections	26077, 2864, 2629 2145 overlapping and 842 single reflections and 123 systematic absences	24760, 2816, 2547	
R_{int}	0.028	0.034	
(sin $\theta/\lambda)_{max}$ (Å^{-1})	0.655	0.655	
Refinement			
No. of reflections	2864	2816	
No. of parameters	198	197	
H-atom treatment	N—H refined freely; C—H refined with a riding model	N—H refined freely; C—H refined with a riding model	
$R[F^2 > 2\sigma(F^2)]$, $wR(F^2)$, S	0.0314, 0.0860, 1.085	0.0787, 0.2545, 1.154	
Twin fraction BASF	0.2003 (10)		
Weighting scheme	a = 0.0449, b = 0.2210	a = 0.0702, b = 6.2812	
$\Delta\rho_{max}$, $\Delta\rho_{min}$ (e Å^{-3})	0.23, −0.22	0.35, −0.36	
Bond precision C—C (Å)	0.0017	0.0062	

process these by taking into account the two twin lattices. We show that the availability of raw data and proper reprocessing using twin lattices is by far superior to efforts to de-twin processed structure factors.

Data processing and refinement

Data were processed in two ways: by using a single lattice, ignoring the second lattice completely, and by using two twin lattices. *EVAL* software (Schreurs *et al.*, 2010) was used for both, followed by *SADABS/TWINABS* (Krause *et al.*, 2015; Sevvana *et al.*, 2019) for scaling. Splitting of the radiation in $K\alpha_1$ and $K\alpha_2$ in a 2:1 ratio is taken into account in the *EVAL* model profiles for either lattice. The statistics for the two approaches are given in Table 1. Several indicators in the single-crystal data processing show that the crystal is not a single crystal. The first real sign of an alarm occurs when it comes to the space-group determination: the most likely space

group is $P2_1/a$ but systematic absences for the *a*-glide plane are clearly violated (reflection condition $h0l$; $h = 2n$). Structure refinement with *SHELXL* (Sheldrick, 2015) converges at high residuals $R_1[I > 2\sigma(I)] = 0.0787$ and wR_2(all refl.) = 0.2545 and the proposed weighting scheme is rather unusual. The two twin lattices are related by a twofold rotation about *c*, resulting in the twin matrix (−1 0 −1 / 0 −1 0 / 0 0 1) (see below). With only single-crystal structure factors it is still possible to use the knowledge of the twin matrix. Inclusion of this matrix in the *SHELXL* refinement assumes that the lattices overlap exactly and the obliquity would be 0°. In reality, not all reflections overlap and thus the refinement results improve only slightly {$R_1[I > 2\sigma(I)] = 0.0678$ and wR_2(all refl.) = 0.2390}. As a last resort, one can de-twin the merged data with *TWINROTMAT* in *PLATON* (Spek, 2020). This produces an HKLF5-type file for refinement in *SHELXL* in which each reflection is either overlapped or single (935 reflections are overlapping). The structure refinement improved to $R_1[I > 2\sigma(I)] = 0.0465$ and wR_2(all refl.) = 0.1332. As we will see below, it is a poor approach for resolving the twinning issue with *processed* data, clearly *raw* diffraction data are needed to reprocess with two lattices.

To show the advantage of proper processing we used two matrices. One twin component was clearly the largest and we processed the data with this lattice while including the second lattice as interfering in *EVAL*. Reflections are deconvoluted when the covariance of the overlapping intensities is below a given threshold. This led to 2145 overlapping and 842 single reflections and 123 systematic absences (Table 1). The agreement factors of the *SHELXL* refinement are much improved to $R_1[I > 2\sigma(I)] = 0.0314$ and $wR2$(all refl.) = 0.0860 (Table 1) and the displacement ellipsoids of the two independent molecules are perfectly reasonable (Fig. 1).

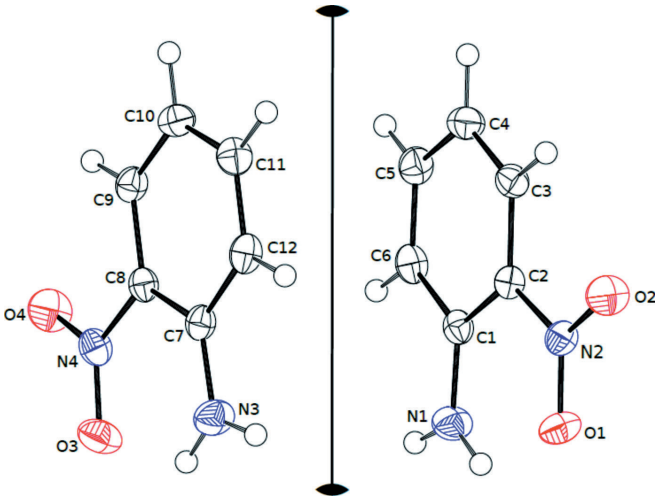


Figure 1
Molecular structure of the two independent molecules of (I) in the crystal. Displacement ellipsoids are drawn at the 50% probability level. Hydrogen atoms are drawn as small spheres of arbitrary radii. The molecules are related by a non-crystallographic twofold axis approximately along *a**.

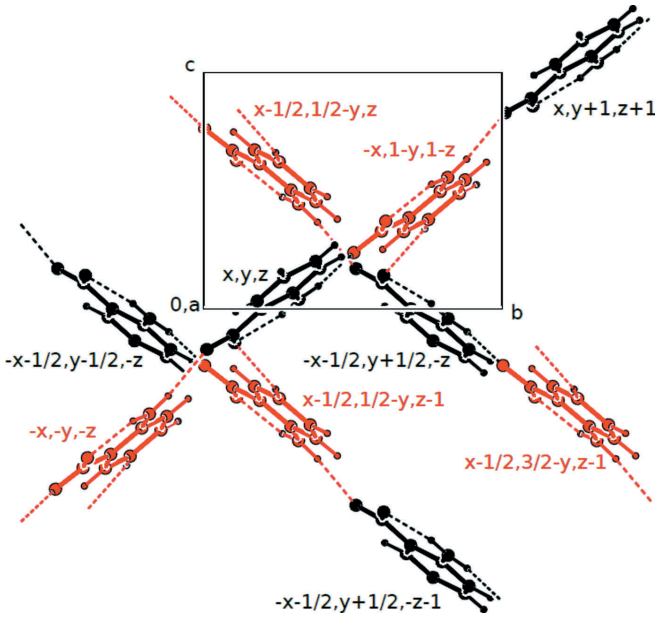


Figure 2
Hydrogen-bonded layers in the monoclinic structure.

The crystal structure has $P2_1/a$ symmetry with two independent molecules, which are shown in Fig. 1. The molecules are connected by hydrogen bonds, forming two-dimensional layers in the *bc* plane (Fig. 2).

Data description

Data were collected on our in-house *APEXII* diffractometer with Mo *Kα* radiation, with multiple scans (Table 1). In total 3324 images were recorded. The unit cell was determined with *DIRAX* (Duisenberg, 1992) and two lattices were found that could be transformed into each other with a nearly twofold rotation. Pseudo-orthorhombic twinning is characterized by a base-centered orthorhombic twin lattice derived from a monoclinic-*P* crystal lattice (Dunitz, 1964). The monoclinic *P* cell of (I) can be transformed to a near orthorhombic *B* lattice [the non-standard setting of the space group $P2_1/a$ was chosen for compatibility with earlier literature (Herbstein, 1965); base-centered *B*-orthorhombic was chosen so as to leave *b* and *c* unchanged] with the following operation:

$$\begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \\ \mathbf{c}' \end{pmatrix} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{pmatrix}$$

giving cell parameters $a' = 29.1340\ (10)$, $b' = 10.0938\ (4)$, $c' = 8.3580\ (2)$, Å, $\alpha' = 90$, $\beta' = 90.743\ (3)$, $\gamma' = 90°$. The *c* axis was chosen as the twofold twin rotation axis. Clearly we find a twin obliquity of 0.743 (3)° and a non-merohedral twin. As a consequence, the orthorhombic *B* lattices of the individual twin components do not exactly overlap. If one overlooks the twinning and indexes the spots as *B*-centered orthorhombic in space group $B22_12$, the reflection conditions appear to be: hkl; $h + l = 2n$, $0k0$; $k = 2n$, which is usual for the space group,

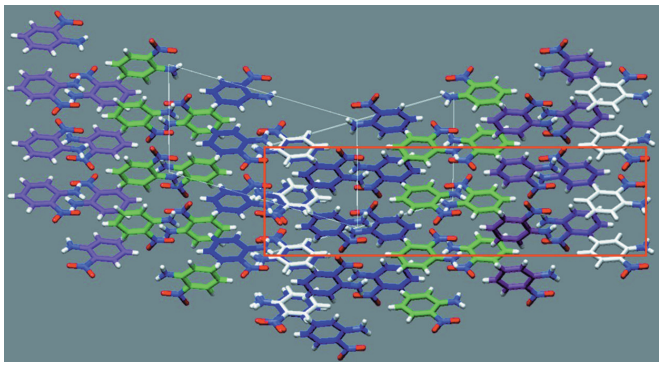


Figure 3
Twin domains viewed down the monoclinic *b*-axis, with alternate layers colored in white, blue, green and magenta. The layers have a width of two hydrogen-bonded molecules (see Fig. 2) and have hydrophobic faces. The second lattice (left) has a row of molecules in the blue layer in common with that of the white layer in the first lattice (right). The second domain is generated by a twofold rotation around *c* and a shift over *1*/2*a*. The pseudo-orthorhombic unit cell is shown in red. The hydrophobic interactions between the layers are almost completely conserved across the twin interface.

and *h*00; *h* = 4*n*, 00 *l*; *l* = 4*n*. The latter two are non-space-group extinctions and exactly such observations are considered as a signal for pseudo-orthorhombic twinning of an underlying monoclinic lattice (Dunitz, 1964). Processing the data as single-component orthorhombic does not result in a structure solution, notably because the *a* glide plane is absent in $B22_12$.

The twin rotation about the *c*-axis results in stacking faults of the hydrogen-bonded layers. In Fig. 3, the two domains and the twinning interface is shown. The second lattice was generated by 180° rotation around *c* followed by a translation over *1*/2*a*, by which the two independent molecules are interchanged in the position. In fact, the two molecules can almost be transformed into each other by a twofold rotation along *a*′, the long orthorhombic axis, showing that this axis is a near orthorhombic twofold axis. The twinning and stacking faults follow the OD theory as proposed by Dornberger-Schiff (1966) for similar systems.

As a consequence of the twin obliquity ω = 0.743 (3)°, reflections are split in reciprocal space. This can be seen in simulated precession photographs that were generated with

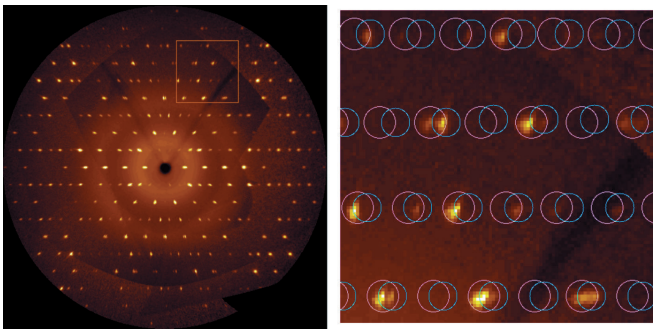


Figure 4
Left: simulated precession photograph in the *h*0*l* plane of (I) up to a resolution of 0.9 (Å). The reconstruction is based on seven scans with a total of 3324 raw images. Right: zoomed image, is from the yellow square in the left image. White circles are the predicted impacts for the first twin component, blue circles for the second.

the program *PRECESSION* in the *EVAL* package. In the 0th layers, this mainly affects layer *h*0*l* (Fig. 4). In the layers, *hk*0 and 0*kl* reflections remain nearly unaffected.

References

Aakeröy, C. B., Nieuwenhuyzen, M. & Price, S. L. (1998*a*). *J. Am. Chem. Soc.* **120**, 8986–8993.
Aakeröy, C. B., Beatty, A. M., Nieuwenhuyzen, M. & Zou, M. (1998*b*). *J. Mater. Chem.* **8**, 1385–1389.
Dhaneshwar, N. N., Tavale, S. S. & Pant, L. M. (1978). *Acta Cryst.* B**34**, 2507–2509.
Dornberger-Schiff, K. (1966). *Acta Cryst.* **21**, 311–322.
Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92–96.
Dunitz, J. D. (1964). *Acta Cryst.* **17**, 1299–1304.
Herbstein, F. H. (1965). *Acta Cryst.* **19**, 590–595.
Krause, L., Herbst-Irmer, R., Sheldrick, G. M. & Stalke, D. (2015). *J. Appl. Cryst.* **48**, 3–10.
Nieger, M. (2007). *CSD Communication* (CCDC No. 655336). CCDC, Cambridge, England.
Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2010). *J. Appl. Cryst.* **43**, 70–82.
Sevvana, M., Ruf, M., Usón, I., Sheldrick, G. M. & Herbst-Irmer, R. (2019). *Acta Cryst.* D**75**, 1040–1050.
Sheldrick, G. M. (2015). *Acta Cryst.* C**71**, 3–8.
Spek, A. L. (2020). *Acta Cryst.* E**76**, 1–11.
Zych, T., Misiaszek, T. & Szostak, M. M. (2007). *Chem. Phys.* **340**, 260–272.

IUCrData

ISSN 2414-3146

Cyclohexane plastic phase I: single-crystal diffraction images and new structural model

Sylvain Bernès* and Sebastian Camargo

Instituto de Física Luis Rivera Terrazas, Benemérita Universidad Autónoma de Puebla, 18 Sur y San Claudio S/N, Puebla, Pue. 72570, Mexico. *Correspondence e-mail: sylvain_bernes@hotmail.com

Received 11 November 2022
Accepted 7 February 2023

Edited by L. M. J. Kroon-Batenburg, Utrecht University

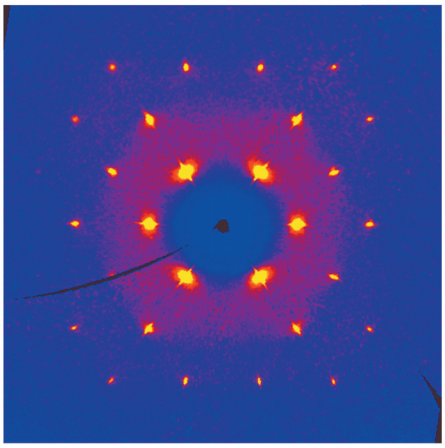
Keywords: cyclohexane; plastic crystals; X-ray diffraction; diffraction image; disorder.

CCDC reference: 2240539

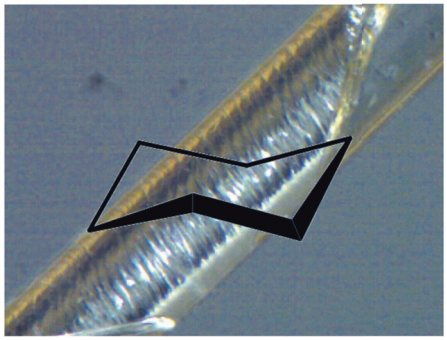
Structural data: full structural data are available from iucrdata.iucr.org

The plastic phase of cyclohexane (polymorph I) was studied by Kahn and co-workers, without achieving a satisfactory determination of the atomic coordinates [Kahn *et al.* (1973). *Acta Cryst*. B**29**, 131–138]. The positions of the C atoms cannot be determined directly as a consequence of the disorder in a high-symmetry space group, an inherent feature of plastic materials. Given this situation, the building of a polyhedron describing the disorder was the main tool for determining the molecular structure in the present work. Based on the shape of reflections {111}, {200} and {113} in space group $Fm\bar{3}m$, we assumed that cyclohexane is disordered through the action of rotation group 432. The polyhedral cluster of disordered molecules is then a rhombic dodecahedron centred on the nodes of an fcc Bravais lattice. The vertices of this polyhedron are the positions of C atoms for the cyclohexane molecule, which is disordered over 24 positions. With such a model, the asymmetric unit is reduced to two C atoms placed on special positions, and an acceptable fit between the observed and calculated structure factors is obtained.

Raw data



Chemical scheme



Stoe Stadivari data files and CBF files: https://doi.org/10.5281/zenodo.7154725

Metadata imgCIF file: https://doi.org/10.1107/S2414314623001141/iq4001img1.cif

Introduction

The concept of an organic 'plastic crystal' was first stated by J. Timmermans in 1938, although the name was coined ten years later by A. Michils, due to the mechanical softness of these materials. They share many physicochemical features with liquid crystals, and were indeed first described as a new mesomorphic state of matter. Thermodynamically, they are characterized by a very low entropy of fusion, $\Delta S_m < 5$ eu (1 eu = 4.185 J mol^{-1} K^{-1}), which was interpreted as the signature that a quasi-isotropic state,

similar to that of liquids, is set just below their melting-point temperature (Timmermans, 1961). Mechanically, these materials behave as plastic metals and can be extruded at quite low pressures (Michils effect; Michils, 1948). Most often, the molecules concerned belong to high-symmetry point groups, and present a more or less globular shape while being orientationally disordered around their rotation axis; they have a marked propensity for polymorphism, with the form close to the melting point crystallizing in a high-symmetry space group, usually in the cubic crystal system, with a highly disordered crystal structure.

Cyclohexane, C_6H_{12}, is an emblematic example of such crystals. The ground state of the molecule is the rigid-chair conformer, belonging to $\overline{3}m$ (D_{3d}) point group. The high-temperature phase I, in space group $Fm\overline{3}m$, undergoes an isothermal transition at 186 K to the low-temperature ordered phase II, in space group $C2/c$ (Kahn *et al.*, 1973). The entropy of fusion, $\Delta S_m = 2.29$ eu, is much lower than that measured for the II→I transition, $\Delta S_{\mathrm{II}\rightarrow\mathrm{I}} = 8.66$ eu (Ruehrwein & Huffman, 1943).

Early literature regarding the crystallographic characterization of cyclohexane phase I (Hassel & Sommerfeldt, 1938; Oda, 1948; Renaud & Fourme, 1966; Kahn *et al.*, 1973), systematically complains about technical hurdles related to the very nature of plastic crystals: (i) a very rapid fall-off of diffraction intensity with increasing Bragg angle; (ii) a scattering background blackening the photographic plates and masking weak reflections; (iii) the extreme difficulty of obtaining a reliable set of atomic coordinates, as a direct consequence of the previously mentioned issues. Indeed, only one article explicitly suggests a structural model based on atomic coordinates (Kahn *et al.*, 1973), which is discussed further below.

During our work on the structure prediction and crystallographic characterization of cycloalkanes that are liquids at room temperature (Camargo, 2018), we were able to obtain diffraction frames for the plastic phase I of cyclohexane. A careful examination of the reciprocal space rebuilt from these raw data offers greater insight into how molecules behave in the plastic phase, and allowed us to propose a new simple model explaining how molecules are disordered about the nodes of the fcc Bravais lattice.

Crystallization and data collection

Anhydrous cyclohexane (reference 227048, Sigma-Aldrich, 99.5%) has a melting point close to 279 K. The end of a 0.4 mm diameter glass Lindemann capillary tube was filled with liquid cyclohexane and the head of the capillary sealed with wax, while avoiding any contamination of cyclohexane. The capillary was mounted on a standard goniometer head, and cyclohexane was crystallized *in situ*, on a Stoe Stadivari diffractometer equipped with an Oxford Cryosystems Cobra cooling device. No head spinning was applied during crystallization, and a key condition was to keep the capillary horizontally ($\chi = -90^\circ$), in order to have the N_2 flow approximately normal to the capillary. In a first step, successive cycles of cooling/heating ramps with different rates were applied to obtain a powdered sample: from room temperature to 260 K at 360 K h^{-1}, and then to 270 K at 200 K h^{-1}. These microcrystals were then carefully merged by heating the sample to 273 K (60 K h^{-1}) and then to 274 K (2 K h^{-1}). Once a single crystal is stabilized in the capillary, the sample can be cooled to 260 K at 10 K h^{-1} and then to 250 K at 20 K h^{-1}. We found that this methodology affords large and good-quality single crystals in a reproducible manner.

Diffraction intensities for one crystal were collected at 255 K with Ag $K\alpha$ radiation (AXO microfocus source equipped with multilayer ASTIX-f optics) and a PILATUS 100 K detector (487 × 195 pixels), accumulating 1139 frames over 19 h, each one being collected over 60 s, with a scan range of 1° in ω. Another crystal was collected at 245 K over 91 h. For this experiment, a very long exposure time of 1800 s per frame was used, with a scan range of 2° in ω. A set of 183 frames was collected for this crystal. Both data sets afford virtually the same structure refinement. The refinement reported in this paper is based on the first data set. The second data set is used herein for Fig. 3 only.

Data processing

Reciprocal space for each crystal was built using all collected frames, with the dedicated *X-AREA* tool (Stoe & Cie, 2019). A cubic 3D array centred on the origin of reciprocal space, with boundaries at −0.75 and +0.75 Å^{-1} ($2\theta_{\max} = 42.7^\circ$) and a pixel resolution of 0.003 Å^{-1} was computed. Each detector pixel was divided into 10 subpixels in the plane of the detector, and into 20 subpixels in the direction normal to that plane. The resulting 3D arrays contain approximately 125×10^6 voxels. Images in Figs. 1–3 are plotted using a conventional blue/yellow heat map.

Structure factors were obtained by integrating the 1139 frames collected on the first crystal. Elliptical integration masks are used, with the smallest diameter given by $W = A + B\tan\theta$ and the largest diameter calculated as $W/\cos 2\theta + (\Delta\lambda/\lambda)\tan\theta$, with $A = 5$ and $B = -8$. A rather large mosaic spread parameter was applied (ems = 0.048 rad), to take into account the plastic nature of the crystal. Finally, the background area was systematically limited to one pixel around the peak area. Intensities were scaled in the $m\overline{3}m$ Laue class in a standard way.

Data description

The ($hk0$) layer built with 1139 frames (Fig. 1) clearly shows that a single crystal was grown. Bragg peaks are well defined, although the resolution is, as expected, very low: the last observed reflections in the full pattern are (333) and (511), corresponding to a resolution of 1.67 Å. That resolution is not improved if frames are collected over 1800 s instead of 60 s. Moreover, this is exactly the same resolution as that obtained by Kahn *et al.* in 1973, and should thus be regarded as an intrinsic limit imposed by the plastic nature of the material. On the other hand, a homogeneous background is visible for

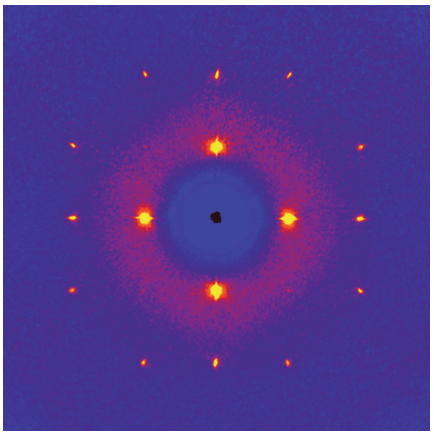


Figure 1
The (*hk*0) layer of reciprocal space for the plastic phase of cyclohexane, at 255 K. Direction [100] runs on the right, direction [010] runs upward.

$2\theta < 12.5°$, indicative of a degree of disorder, or, at the very least, indicative of large atomic motions in the crystal. Another projection of the reciprocal space (Fig. 2) shows an interesting feature: the reflections with highest intensities, {111}, display a rod-like diffuse scattering streak along ⟨111⟩, which should be related to the main direction for disorder (Welberry & Butler, 1995; Welberry & Goossens, 2014). These diffuse streaks are better visualized using data collected with long-exposure frames (Fig. 3). Assuming that disorder occurs exclusively through the crystallographic rotation axis, involved symmetry elements that are not parallel to ⟨111⟩ in space group $Fm\bar{3}m$ are two- and fourfold axes. However, threefold axes are also used for disordering the molecule, as showed by the diffuse halo wrapping the {111} peaks. Although, to a lesser extent, {200} and {113} reflections also show diffuse streaks along ⟨200⟩ and ⟨113⟩, respectively, which we assign to rotations around the two-, three- and fourfold axes (Figs. 1 and 2). Given that only low-angle reflections are involved, we assume that observed streaks do not originate from α_1/α_2 radiation splitting or other experimental artefacts.

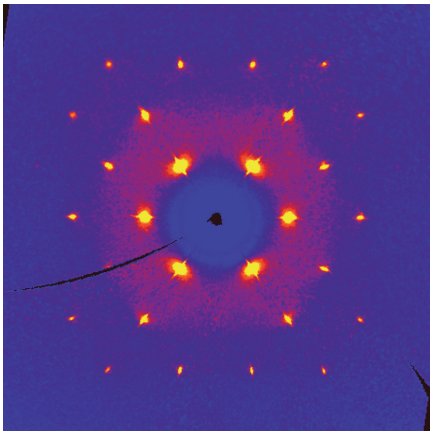


Figure 2
Reciprocal space of the plastic phase of cyclohexane, at 255 K, viewed in a projection normal to [011]. Direction [100] runs on the right, direction [01$\bar{1}$] runs upward.

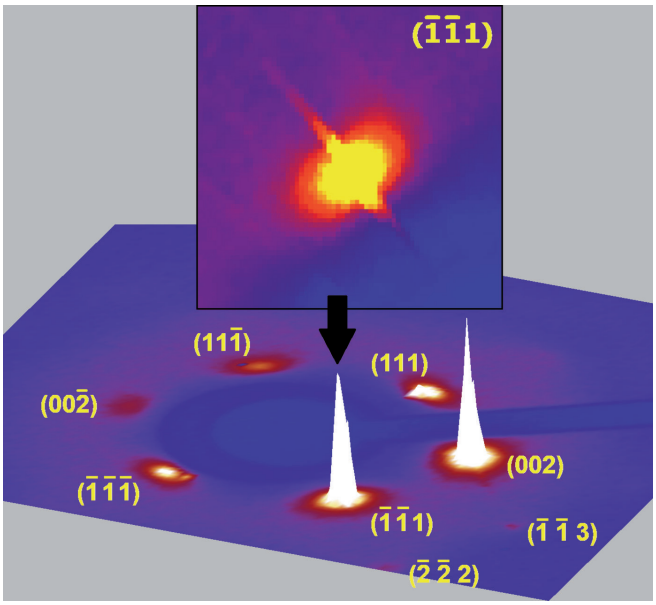


Figure 3
A single frame collected over 2° in ω over 1800 s for the plastic phase of cyclohexane at 245 K (bottom). Only the small part of the frame containing Bragg peaks is shown. The 3D plot of the frame (CBF format) was obtained using *CAP Frame View* v. 1.3 (Rigaku OD, 2015). Top inset: Reflection ($\bar{1}1\bar{1}$) rebuilt using 183 collected frames. The projection is normal to [0$\bar{1}$1], with [100] running on the right and [0$\bar{1}$1] running upward. This image was generated using *X-AREA* (Stoe & Cie, 2019).

The simplest model based on the hypothesis of a rigid chair cyclohexane disordered through all crystallographic rotations includes two sites for the C atoms. Atom C1 is placed on the fourfold axis, with coordinates (x, 0, 0), corresponding to the Wyckoff position 24*e* (4*m.m*) in space group $Fm\bar{3}m$, while atom C2 is placed on the threefold axis, with coordinates (x', x', x'), corresponding to the Wyckoff position 32*f* (.3*m*). By placing this asymmetric unit close to the origin, a set of 14 C atoms are connected, forming a rhombic dodecahedron, a well-studied convex polyhedron with Euler characteristic $\chi = 2$ (Fig. 4). This polyhedron is centrosymmetric, and its centre coincides with the crystallographic inversion centre. As this polyhedron belongs to the family of edge-transitive polyhedra, all C—C bonds are equivalent and have the same

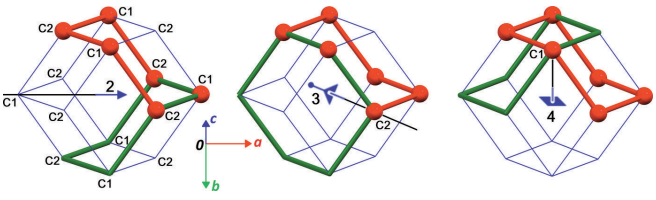


Figure 4
The rhombic dodecahedron (blue edges) describing the 24 disordered positions of cyclohexane at 255 K. On the left, atom labels are given, omitting their symmetry codes for clarity. Red and green molecules are related by twofold (left), threefold (middle) and fourfold (right) rotations. The twofold axis is viewed in the plane of projection, while the three- and fourfold axis are inclined with respect to that plane. The projection is viewed along [012], and the unit-cell origin is coincident with the centre of the polyhedron.

Table 1
Experimental details.

Raw data				
DOI	https://doi.org/10.5281/zenodo.7154725			
Data archive	Zenodo			
Data format	CBF			
Data collection				
Diffractometer	Stoe Stadivari			
Temperature (K)	255			
Radiation type	Ag $K\alpha$			
Detector type	Dectris Pilatus 100 K R			
Wavelength (Å)	0.56083			
Beam centre (mm)	41.882, 16.7			
Detector axis	$-Z$			
Detector distance (mm)	40.0			
Pixel size (mm)	0.172 × 0.172			
No. of pixels	195 × 487			
No. of scans	17			
Exposure time per frame (s)	60.0			
Swing angle (°)	Scan axis	Start angle, increment per frame (°)	Scan range (°)	No. of frames
−23.4	ω, X (χ = −55.183°, φ = −55.0°)	142.261, 1.0	47.0	47
−23.4	ω, X (χ = −30.183°, φ = −5.0°)	142.261, 1.0	47.0	47
23.4	ω, X (χ = −35.183°, φ = 150.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −20.183°, φ = −85.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −30.183°, φ = −105.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −35.183°, φ = −35.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −35.183°, φ = 75.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −50.183°, φ = −160.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −20.183°, φ = 25.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −45.183°, φ = −55.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −50.183°, φ = 15.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −50.183°, φ = 125.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −25.183°, φ = −110.0°)	−40.679, 1.0	77.0	77
23.4	ω, X (χ = −55.183°, φ = −170.0°)	−40.679, 1.0	67.0	67
23.4	ω, X (χ = −40.183°, φ = 140.0°)	−40.679, 1.0	67.0	67
23.4	ω, X (χ = −45.183°, φ = −150.0°)	−25.679, 1.0	32.0	32
23.4	ω, X (χ = −20.183°, φ = −100.0°)	−25.679, 1.0	32.0	32
Crystal data				
Chemical formula	C_6H_{12}			
M_r	84.16			
Crystal system, space group	Cubic, $Fm\bar{3}m$			
a (Å)	8.712 (4)			
V (Å^3)	661.1 (9)			
Z	4			
μ (mm^{-1})	0.03			
Crystal size (mm)	0.40 × 0.30 × 0.30			
Data processing				
Absorption correction	Multi-scan (*X-AREA;* Stoe & Cie, 2019)			
T_{min}, T_{max}	0.558, 1.000			
No. of measured, independent and observed [$I > 2\sigma(I)$] reflections	1690, 31, 10			
R_{int}	0.018			
(sin $\theta/\lambda)_{max}$ (Å^{-1})	0.487			
Refinement				
$R[F^2 > 2\sigma(F^2)]$, $wR(F^2)$, S	0.080, 0.190, 1.10			
No. of reflections	31			
No. of parameters	5			
No. of restraints	3			
H-atom treatment	H-atom parameters constrained			
$\Delta\rho_{max}$, $\Delta\rho_{min}$ (e Å^{-3})	0.07, −0.16			

Computer programs: *X-AREA* (Stoe & Cie, 2019), *SHELXL2018/3* (Sheldrick, 2015) and *Mercury* (Macrae *et al.*, 2020).

bond length, as expected for cyclohexane. The rhombic faces, with configuration v3.4.3.4, display obtuse angles of arccos(−1/3) = ±109.47°, which accommodate sp^3-hybridized C atoms. The dihedral angle between edge-sharing rhombus is 120°, affording the expected C—C—C—C torsion angles of ±60° in cyclohexane.

Most importantly, the rhombic dodecahedron has full octahedral symmetry ($m\bar{3}m$ or *432), and its rotation group is the chiral octahedral group 432. The chair conformation of cyclohexane, with symmetry $\bar{3}m$, is thus compatible with the rhombic dodecahedron, and the full polyhedron is indeed generated by rotation of one chair about the elements of the

Table 2
Refined structure of plastic cyclohexane at 255 K. Refined parameters are x, x', U(C1) and U(C2).

Refined parameters	
Position of C1: (x, 0, 0), sof, U_{iso}	x = 0.2030(18), 3/56, U(C1) = 0.41 (3) Å^2
Position of C2: (x', x', x'), sof, U_{iso}	x' = 0.1019 (10), 1/14, U(C2) = 0.35 (3) Å^2
Cyclohexane geometry	
C—C bond length	1.534 (8) Å
C—C—C bond angles	109.9 (17), 109.2 (8)°
C—C—C—C torsion angles	−59.7 (10), 59.7 (10)°

rotation group 432, as reflected in the shape of the Bragg reflections, as discussed above. The molecule is then disordered over 24 positions (order of the rotation group). Symmetry-related molecules in this polyhedral cluster are depicted in Fig. 4.

Once the polyhedron describing the disorder in the plastic phase has been laid down, the structure refinement is straightforward. A single geometric parameter should actually be refined, that is the bond length C1—C2 = d. Since both atoms lie on special positions, only two positional parameters are used, x and x'. Using the structure factors extracted as described in the previous section, we refined an isotropic model with *SHELXL* (Sheldrick, 2015; refinement against F^2, no extinction parameter refined), including three restraints for the geometry of the polyhedron: d = 1.54 (1) Å, and a couple of restraints for 1,3-distances: C1···C1′ = $d(2\sqrt{2/3})$ and C2···C2′ = $d(2\sqrt{2/3})$, with standard deviations of 0.03 Å, and with primed atoms generated by suitable symmetry operations. Site occupancy factors (sof) are calculated considering the Wyckoff positions and assuming that each of the 14 vertices in the polyhedron has the same probability to be occupied: sof(C1) = (24/192) × (6/14) = 3/56 and sof(C2) = (32/192) × (6/14) = 1/14. Finally, all H atoms were added in idealized positions, corresponding to special positions 96*k* (H1 bonded to C1), and 96*k* and 32*f* (H2*A* and H2*B* bonded to C2), with C—H = 0.95 Å, and with calculated displacement parameters U_{iso}(H) = 2.8U_{iso}(carrier C).

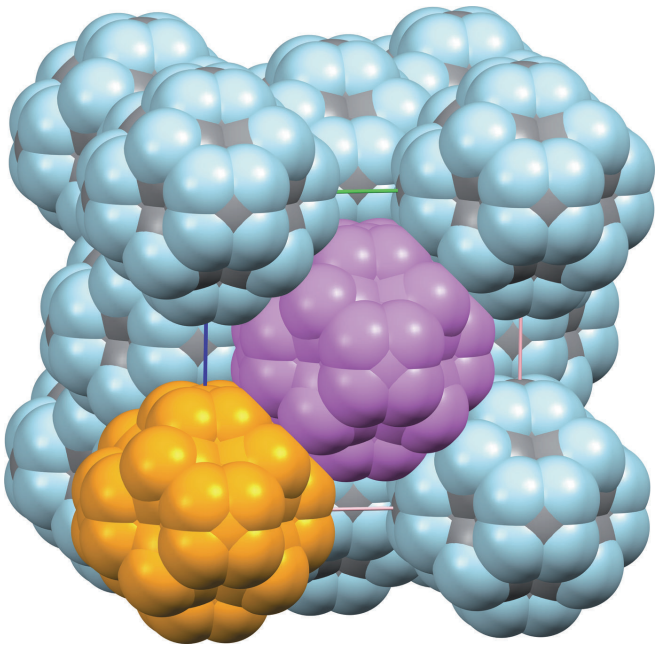


Figure 5
Packing structure of cyclohexane at 255 K, in a space-filling representation. All disordered sites for C (grey) and H (blue) atoms in one unit cell are represented with their van der Waals radii (Macrae *et al.*, 2020). Two neighbouring clusters of disordered molecules in the fcc lattice are shown in orange and magenta, with the purpose of emphasizing the contact between the two clusters. The interface separating the clusters is a 'diagonal' mirror plane m in space group $Fm\bar{3}m$.

The structure is then refined (Table 1) using five parameters and 31 independent reflections, of which ten have $F_o > 4\sigma(F_o)$, converging towards the expected geometry (Table 2). Notably, the refined C1—C2 bond length of 1.534 (8) Å is identical to that determined by electron diffraction, 1.535 (2) Å (Ewbank *et al.*, 1976). Displacement parameters are very high, reflecting the motions of C atoms bouncing from vertex to vertex in the polyhedral cluster. Actually, Figs. 1–3 reflect accurately the idea of Timmermans about plastic crystals: they are solids behaving like liquids over short distances (one polyhedron). From the crystallographic point of view, plastic cyclohexane can be seen as a liquid with long-range order, affording a diffraction pattern. The dynamic disorder being identical for every node in the lattice, the crystal structure emulates a close-packed arrangement (Fig. 5), in which the atomic sites have very low occupancies (see Table 2). As a consequence, the density is also very low, 0.85 g cm^{-3}. The non-plastic phase II of cyclohexane has a more regular density of 1 g cm^{-3}.

Discussion and conclusions

Strangely enough, Kahn *et al.* were unable to move towards the model we propose in Table 2, probably because they did not realize that C atoms could lie on special positions. Instead, they used an asymmetric unit including three C atoms close to the origin, all in *general* positions. With such a model, the 144-vertex polyhedron describing the disorder is hugely complex, and individual cyclohexane molecules are hardly discernible. Actually, their polyhedron has a shape close to that of a sphere, which has Euler characteristic χ= 2, like any (convex) polyhedron whose boundary is topologically equivalent to a sphere. It is thus not surprising that they could obtain a satisfactory agreement between observed and calculated structure factors, although their structural model is far from satisfactory.

It is worth noting that the notion of 'refinement' for such plastic structures is of little sense, especially if least-squares methods are involved, since the data-to-parameter ratio rapidly drops to too low values. Even the identification of a suitable asymmetric unit cannot rely on mainstream approaches like direct methods, since atomic resolution is not achievable. Instead, a careful examination of data in reciprocal space, in particular the shape of the Bragg peaks, can be helpful. In 1973, this perspective was not considered by Kahn *et al.* In contrast, the 1948 article of Tutomu Oda, of limited impact because written in Japanese, is noteworthy. The abstract mentions: '*Besides the Bragg reflections, we observed remarkable diffuse scattering of considerable intensity, similar*

to that shown by cyclohexanol. Namely, there appear on the Laue and oscillation photographs a number of so-called diffuse spots and apparently circular diffuse haloes, which resemble to the liquid diffraction haloes'. Nowadays, computer simulations allow the interpretation of the diffuse scattering observed in many materials. This may be achieved either in reciprocal space by considering the material as a modulated phase, or with a correlation method in direct space, using short-range chemical and atomic displacement pair-correlation parameters (Rosenkranz & Osborn, 2004; Welberry, 2022). In the case of molecular crystals, Monte Carlo and reverse Monte Carlo simulations are also a very promising approach, since they are applicable to disorder of any complexity (Welberry, 2022). However, only a few such simulations have been carried out for plastic crystals to date (for example, for α-CBr_4; Folmer *et al.*, 2008), and the molecular dynamics associated with the disorder in these materials is not fully understood.

We also extended this study to cycloheptane phase I and cyclooctane phase I (both in space group $Pm\bar{3}n$). Preliminary results can be found in the Master's thesis of the last author (Camargo, 2018; available online). We also plan to collect data at temperatures as close as possible to the melting points of these materials, and to use Cu $K\alpha$ radiation for collecting frames.

Acknowledgements

We thank Dr Paolo Celani (Stoe & Cie GmbH, Darmstadt), for providing scripts allowing the exportation of original frames to CBF (crystallographic binary file) format, and the Editor of *Raw Data Letters*, for the development of suitable tools for preparing imgCIF files.

Funding information

The following funding is acknowledged: Consejo Nacional de Ciencia y Tecnología (studentship No. CVU-784426; grant No. 268178).

References

Camargo, S. (2018). Master's thesis, Benemérita Universidad Autónoma de Puebla, Puebla, Mexico, https://repositorioinstitucional.buap.mx/handle/20.500.12371/8287.
Ewbank, J. D., Kirsch, G. & Schäfer, L. (1976). *J. Mol. Struct.* **31**, 39–45.
Folmer, J. C. W., Withers, R. L., Welberry, T. R. & Martin, J. D. (2008). *Phys. Rev. B*, **77**, 144205.
Hassel, O. & Sommerfeldt, A. M. (1938). *Z. Phys. Chem.* **40B**, 391–395.
Kahn, R., Fourme, R., André, D. & Renaud, M. (1973). *Acta Cryst.* B**29**, 131–138.
Macrae, C. F., Sovago, I., Cottrell, S. J., Galek, P. T. A., McCabe, P., Pidcock, E., Platings, M., Shields, G. P., Stevens, J. S., Towler, M. & Wood, P. A. (2020). *J. Appl. Cryst.* **53**, 226–235.
Michils, A. (1948). *Bull. Soc. Chim. Belg.* **57**, 575–617.
Oda, T. (1948). *X-RAYS*, **5**, 26–30.
Renaud, M. & Fourme, R. (1966). *J. Chim. Phys.* **63**, 27–32.
Rigaku OD (2015). *CAP Frame View*. Rigaku Oxford Diffraction, Yarnton, England.
Rosenkranz, S. & Osborn, R. (2004). *Neutron News*, **15**, 21–24.
Ruehrwein, R. A. & Huffman, H. M. (1943). *J. Am. Chem. Soc.* **65**, 1620–1625.
Sheldrick, G. M. (2015). *Acta Cryst.* C**71**, 3–8.
Stoe & Cie (2019). *X-AREA* and *X-RED32*. Stoe & Cie, Darmstadt, Germany.
Timmermans, J. (1961). *J. Phys. Chem. Solids*, **18**, 1–8.
Welberry, T. R. (2022). *Acta Cryst.* B**78**, 344–355.
Welberry, T. R. & Butler, B. D. (1995). *Chem. Rev.* **95**, 2369–2403.
Welberry, T. R. & Goossens, D. J. (2014). *IUCrJ*, **1**, 550–562.

full crystallographic data

IUCrData (2023). **8**, x230114 [https://doi.org/10.1107/S2414314623001141]

Cyclohexane plastic phase I: single-crystal diffraction images and new structural model

Sylvain Bernès and Sebastian Camargo

cyclohexane

Crystal data

C_6H_{12}
M_r = 84.16
Cubic, $Fm\bar{3}m$
a = 8.712 (4) Å
V = 661.1 (9) Å^3
Z = 4
F(000) = 192
D_x = 0.845 Mg m^{-3}
Melting point: 279 K
Ag $K\alpha$ radiation, λ = 0.56083 Å
Cell parameters from 664 reflections
θ = 3.2–9.6°
μ = 0.03 mm^{-1}
T = 255 K
Rod, colourless
0.40 × 0.30 × 0.30 mm

Data collection

Stoe Stadivari diffractometer
Radiation source: Sealed X-ray tube, Axo Astix-f Microfocus source
Graded multilayer mirror monochromator
Detector resolution: 5.81 pixels mm^{-1}
ω scans
Absorption correction: multi-scan X-AREA 1.88 (Stoe & Cie, 2019)
T_{min} = 0.558, T_{max} = 1.000
1690 measured reflections
31 independent reflections
10 reflections with $I > 2\sigma(I)$
R_{int} = 0.018
θ_{max} = 15.9°, θ_{min} = 3.2°
h = −8→8
k = −8→8
l = −8→7

Refinement

Refinement on F^2
Least-squares matrix: full
$R[F^2 > 2\sigma(F^2)]$ = 0.080
$wR(F^2)$ = 0.190
S = 1.10
31 reflections
5 parameters
3 restraints
0 constraints
Hydrogen site location: inferred from neighbouring sites
H-atom parameters constrained
$w = 1/[\sigma^2(F_o^2) + (0.0326P)^2 + 1.826P]$ where $P = (F_o^2 + 2F_c^2)/3$
$(\Delta/\sigma)_{max} < 0.001$
$\Delta\rho_{max}$ = 0.07 e Å^{-3}
$\Delta\rho_{min}$ = −0.16 e Å^{-3}

Fractional atomic coordinates and isotropic or equivalent isotropic displacement parameters (Å^2)

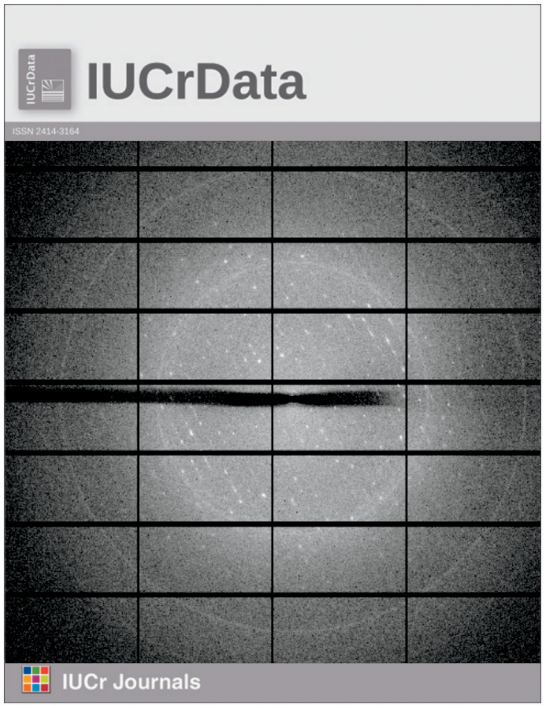
	x	*y*	*z*	U_{iso}*/U_{eq}	Occ. (<1)
C1	0.2030 (18)	0.000000	0.000000	0.41 (3)*	0.4286
H1	0.265179	−0.062430	0.062430	1.162*	0.2143
C2	0.1019 (10)	0.1019 (10)	0.1019 (10)	0.35 (3)*	0.4286

H2A	0.038557	0.164370	0.038557	0.992*	0.2143
H2B	0.163367	0.163367	0.163367	0.992*	0.2143

Geometric parameters (Å, º)

C1—C2^{i}	1.534 (8)	C1—H1^{iii}	0.9407
C1—C2^{ii}	1.534 (8)	C1—H1^{i}	0.9407
C1—C2^{iii}	1.534 (8)	C2—H2A	0.9516
C1—C2	1.534 (8)	C2—H2B	0.9269
C1—H1	0.9406	C2—H2A^{iv}	0.9516
C1—H1^{ii}	0.9407	C2—H2A^{v}	0.9516
C2^{i}—C1—C2^{ii}	109.9 (17)	C1^{v}—C2—H2A	109.2
C2^{iii}—C1—C2	109.9 (17)	C1—C2—H2A	109.2
C2^{iii}—C1—H1	109.3	C1^{v}—C2—H2B	109.7
C2—C1—H1	109.3	C1^{iv}—C2—H2B	109.7
C2^{i}—C1—H1^{ii}	109.3 (4)	C1—C2—H2B	109.7
C2^{ii}—C1—H1^{ii}	109.3 (4)	H2A—C2—H2B	109.9
C2^{iii}—C1—H1^{iii}	109.3 (4)	C1^{iv}—C2—H2A^{iv}	109.2 (4)
C2—C1—H1^{iii}	109.3 (4)	C1—C2—H2A^{iv}	109.2 (4)
H1—C1—H1^{iii}	109.7	H2A—C2—H2A^{iv}	109.1
C2^{i}—C1—H1^{i}	109.3 (4)	H2B—C2—H2A^{iv}	109.9
C2^{ii}—C1—H1^{i}	109.3 (4)	C1^{v}—C2—H2A^{v}	109.2 (4)
H1^{ii}—C1—H1^{i}	109.7	C1^{iv}—C2—H2A^{v}	109.2 (4)
C1^{v}—C2—C1^{iv}	109.2 (8)	H2A—C2—H2A^{v}	109.1
C1^{v}—C2—C1	109.2 (8)	H2B—C2—H2A^{v}	109.9
C1^{iv}—C2—C1	109.2 (8)	H2A^{iv}—C2—H2A^{v}	109.1
C2^{i}—C1—C2—C1^{v}	−0.4 (15)	C2^{i}—C1—C2—C1^{iv}	−119.9 (5)
C2^{ii}—C1—C2—C1^{v}	119.9 (5)	C2^{ii}—C1—C2—C1^{iv}	0.4 (15)
C2^{iii}—C1—C2—C1^{v}	59.7 (10)	C2^{iii}—C1—C2—C1^{iv}	−59.7 (10)

Symmetry codes: (i) $x, -y, z$; (ii) $x, y, -z$; (iii) $x, -y, -z$; (iv) z, x, y; (v) y, z, x.



IUCrData

iucrdata.iucr.org

IUCrData is a peer-reviewed open-access data journal from the IUCr. The first phase of this innovative venture, launched in January 2016, enabled authors to publish brief Data Reports on crystal structures of inorganic, metal–organic or organic compounds. Data Reports include the crystallographic data (CIF and structure factors), a data validation report, figures and a text representation of the data. The journal has now entered its second phase with a second primary article category, Raw Data Letters.

Raw Data Letters

In 2022, *IUCrData* launched a new section – Raw Data Letters, a collaborative innovation of IUCr Journals with the IUCr Committee on Data. This section publishes short descriptions of crystallographic raw data sets from X-ray, neutron or electron diffraction experiments, in the biological, chemical, materials science or physics fields, and provides a persistent link to the location of the raw data. Information for each dataset includes an imgCIF containing core metadata, a diffraction image, figures and a description of the data and their processing. See the infographic overleaf.

Editor

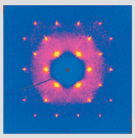
Loes Kroon-Batenburg

Co-editors

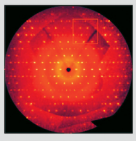
Miguel Aranda
Elena Boldyreva
Aaron Brewster
Simon Coles
John Helliwell

See the **Editorial** and useful links at
iucrdata.iucr.org/x/services/rawdataletters.html

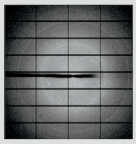
Recent Raw Data Letters



IUCrData (2023). **8**, x230114
Cyclohexane plastic phase I: single-crystal diffraction images and new structural model
S. Bernès and S. Camargo



IUCrData (2022). **7**, x221059
Accurate intensity integration in the twinned γ-form of *o*-nitroaniline
M. Lutz and L. Kroon-Batenburg



IUCrData (2022). **7**, x220852
Crystal structure of the second extracellular domain of human tetraspanin CD9: twinning and diffuse scattering
V. Neviani, M. Lutz, W. Oosterheert, P. Gros and L. Kroon-Batenburg