BROOKHAVEN NATIONAL LABORATORY

DECTRIS
detecting the future

Gold Standard for Macromolecular Crystallography Diffraction Data

Herbert J. Bernstein
Ronin Institute for Independent Scholarship,
c/o NSLS II, Brookhaven National Laboratory, Upton, NY USA

based on

Herbert J. Bernstein, Andreas Förster, Asmit Bhowmick, Aaron S. Brewster, Sandor Brockhauser, Luca Gelisio, David R. Hall, Filip Leonarski, Valerio Mariani, Gianluca Santoni, Clemens Vonrhein, Graeme Winter
(2020)
“Gold Standard for macromolecular crystallography diffraction data”, IUCrJ 7:5,
https://doi.org/10.1107/S2052252520008672

Work Supported in part by Dectris Ltd, US Department of Energy Offices of Biological and Environmental Research and of Basic Energy Sciences (grant Nos. DE-AC02-98CH10886 and E-SC0012704), National Institutes of Health (grant Nos. P41RR012408, P41GM103473, P41GM111244, R01GM117126, P30GM133893 and R21GM129570) and the Hungarian government (grant No. GINOP 2.2.1-15-2016-00012)

The World has Changed

After two decades of effort, agreement has been reached on an updated specification of data and metadata for diffraction images to be produced at light sources:

- to facilitate the processing of data sets by tools available to users at a wide range of institutions, including at their home institutions as well as at light sources other than those at which they were collected, and

- to ensure that software and algorithms developed in the future can be used to extract additional and new information from the raw archived data with a complete experimental description (Kroon-Batenburg & Helliwell, 2017).

What has Changed?

- **Macromolecular crystallography (MX)** is the dominant means of determining the three-dimensional structures of biological macromolecules. Over the last few decades, most MX data have been collected at synchrotron beamlines using a large number of different detectors produced by various manufacturers and taking advantage of various protocols and goniometries.
- These **data came in their own formats**: sometimes proprietary, sometimes open. The **associated metadata rarely reached** the degree of completeness required for FAIR (**Findability, Accessibility, Interoperability and Reusability)** data management.
- Efforts to **reuse old data** some time later were often frustrated. In the culmination of an effort dating back more than two decades, a large portion of the research community concerned with high data-rate macromolecular crystallography (**HDRMX**) has now agreed to an updated specification of data and metadata for diffraction images produced at **synchrotron light sources and X-ray free-electron lasers (XFELs).**

What is the Gold Standard?

- **This Gold Standard will facilitate the processing of data sets independent of the facility at which they were collected and enable data archiving according to FAIR principles, with a particular focus on interoperability and reusability.**
- It builds on the **NeXus/HDF5 NXmx application definition** and the International Union of Crystallography (IUCr) **imgCIF/CBF dictionary**, and it is compatible with major data processing programs and pipelines and should be applied to all detectors used for crystallography.
- Whether we are dealing with CBF files or NeXus/HDF5 files, the information in a Gold Standard data set is the same: **one or more diffraction-image data arrays of pixels along with sufficient metadata to allow software to determine exactly where in the laboratory coordinate system each pixel was located** and when the intensity recorded in that pixel was recorded, so that the software can locate spots, index them and integrate them.

The Gold Standard (cont.)

- **In a Gold Standard data set, the necessary data and metadata for processing a reasonable range of use cases is recorded in the data set.** Although the data set will normally consist of multiple files, these files should be packaged together in an appropriate container, for example a single folder in the file system at the collecting facility or under a single DOI in a data-set repository.
- **The specification of which metadata need be retained with the data depends on the experiment being performed and the software that will be used for processing, i.e. the 'use case'. It is intended to be adequate for single-axis rotation experiments at synchrotrons and stills collected at XFELs and synchrotrons**, and to work properly with the data-reduction programs DIALS (Waterman *et al*., 2013; Winter *et al.*, 2018), XDS (Kabsch, 2010a,b), MOSFLM (Battye *et al*., 2011), HKL-2000 (Otwinowski & Minor, 1997), the data processing pipelines xia2 (Winter, 2010) and autoPROC (Vonrhein *et al*., 2011), as well as future versions of OnDA (Mariani *et al.*, 2016).

imgCIF/CBF vs. NeXus/HDF5

- In 1995, Andrew Hammersley proposed a ‘Crystallographic Binary Format’ which, after considerable discussion and revision, was adopted by the IUCr in 2005 (Bernstein, 2005; Bernstein & Hammersley, 2005; Ellis & Bernstein, 2005). The resulting ‘imgCIF/CBF’ format, metadata and supporting software was adopted by Dectris for the then-new PILATUS detector in 2007 (Powell *et al.*, 2007). In subsequent years it became clear that changes would be needed to this format to support higher data rates and institutional policies (Bernstein, 2010). For the Dectris EIGER detectors, CBF was integrated with the Hierarchical Data Format (HDF5) and became the new NeXus/HDF5 NXmx format (Donath *et al.*, 2013; Könnecke *et al.*, 2015; Hester, 2016; Bernstein, 2017). **Everything in the NeXus version of the Gold Standard has an equivalent in imgCIF/CBF.**

What about structure factors?

- This standard is focused on raw diffraction images rather than the structure factors, since in modern MX data collection, diffraction images are the primary raw data and structure factors are derived data.

- Structure factors are very important, and even if they are derived data they should of course be recorded, not least because since 2008 they have been mandatory for PDB depositions using the appropriate mmCIF definitions (Jiang *et al.*, 1999).

- If structure factors are available, they should be added to Gold Standard files for storage, archiving and deposition. In mmCIF the REFLN category is used. In NeXus/HDF5 the NXreflections category is used

Where and When

- While each data set should contain all of the data and metadata necessary for processing, it also should clearly identify where and when it was collected by specifying the scientific instrument or beamline and the facility at which it was collected and the times of collection.

- In the NXmx Gold Standard, the full name of the scientific instrument or beamline is carried in the /(entry):NXentry/(instrument):NXinstrument/name field and the name of the facility is carried in the /(entry):NXentry/(source):NXsource/name field.

- The full and precise UTC ISO 8601 (Wolf & Wicksteed, 1998) time/date of the first data point collected is carried in the /(entry):NXentry/ start_time field and an estimate of the likely time of collection of the last data point is carried in the /(entry):NXentry/ end_time_estimated field.

Experimental Geometry

- One of the most important sets of metadata used in processing is information on where the components of the experimental setup are positioned and oriented relative to one another.

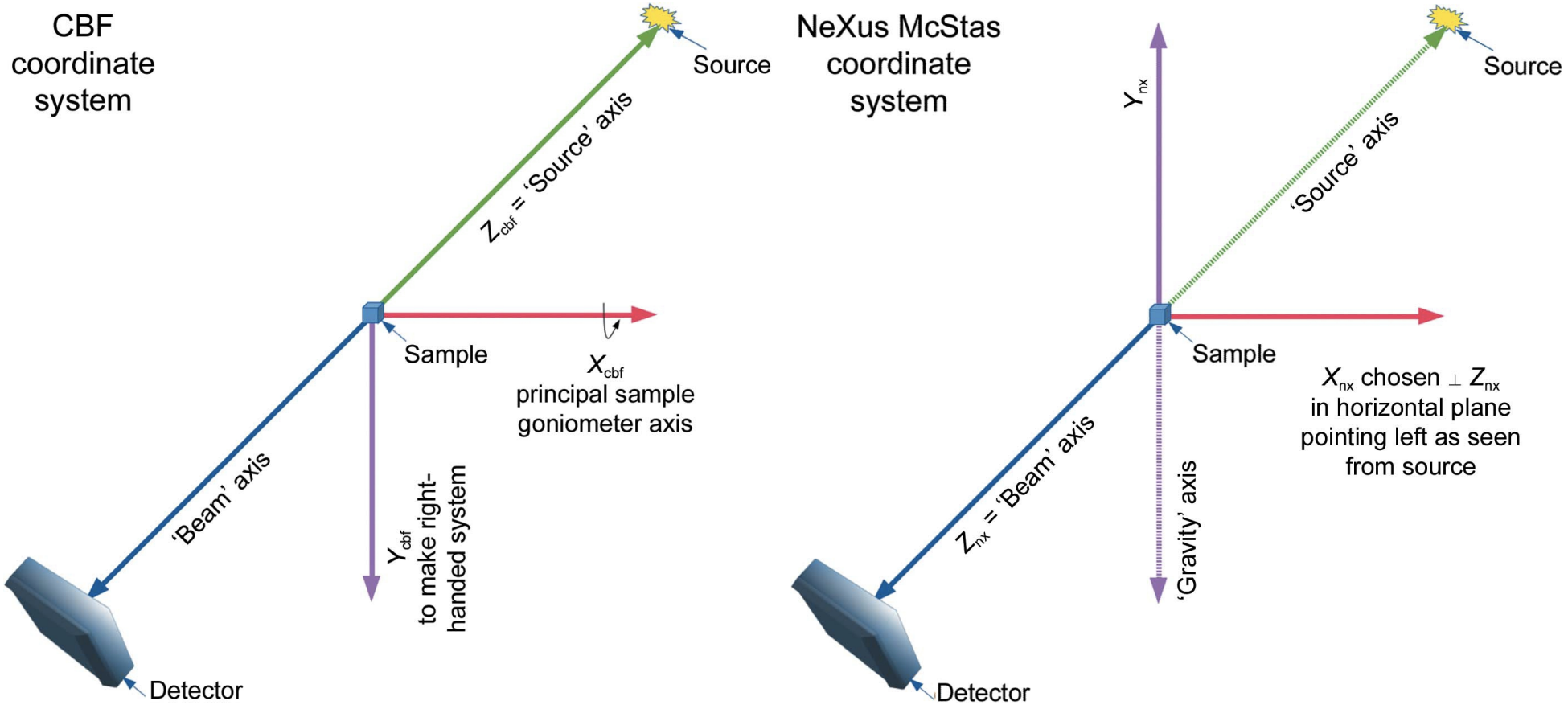
- We need to precisely map the events recorded in a pixel to reciprocal space, which implies a need to know or infer the sample orientation, detector position and characteristics, beam wavelength and direction at the very least.

- We need to know how the sample is positioned and oriented relative to the incident beam, where the detector is positioned and oriented relative to the sample, where in the plane of the detector the incident beam would have hit and where the various sensor modules of the detector are positioned relative to one another.

- Essentially, we need a blueprint of the experimental setup.

Coordinate Frames

- The NeXus/HDF5 files specify axes in the NeXus McStas coordinate system. It is important to note that imgCIF/CBF uses a different coordinate system. In most cases the two coordinate frames are related by a 180° rotation around the vertical axis.



Handling Axes

- The set of metadata used for handling axes both in CBF and in NeXus/HDF5 describes fixed or variable positioning axes in terms of directional vectors in nested lists with optional offset vectors between pairs of axes.

- For an experiment with both a detector and a sample goniometer, we need to provide the nested chains of axes that determine the position and orientation of the detector and of the sample.

- The more complex the design of the experiment and the more varied the non-default choices permitted by the software, the more different metadata may be required to ensure correct processing at a wide range of facilities.

- The Gold Standard is only the minimum set of metadata upon which we have agreed so far.

Where it all Fits

- General organization: A NeXus/HDF5 Gold Standard file consists of a nested tree of groups. The outermost is NXentry, which contains the groups NXdata, NXsample, NXinstrument and NXsource

- NXsample contains the group NXtransformations. NXinstrument contains the groups NXattenuator, NXdetector_group, NXdetector, and NXbeam. NXdetector contains the groups NXtransformations, NXcollection,and NXdetector_module

- For details on the standard, see the NXmx application definition. Earlier versions of the NeXus NXmx application definition have been available since 2014. The latest version prior to formal adoption is available from

  http://github.com/HDRMX/definitions

Where find Examples

- The HDRMX version will be updated as needed to reflect changes during and after adoption.
- To date, the applicability of the Gold Standard has been demonstrated both for single-axis rotation data at a synchrotron
  https://doi.org/10.5281/zenodo.3484187
  and for serial crystallography data at an XFEL
  https://doi.org/10.5281/zenodo.3352357

Overall Structure

BROOKHAVEN NATIONAL LABORATORY
DECTRIS detecting the future

- The overall structure of a NeXus/HDF5 file represents a tree of nested groups. Each group may contain fields and/or groups. Each field or groups may have attributes.
- The top level of the tree is either an NXentry or an Nxsubentry.
- Looking at just the top two levels of each group:
- Group:NXentry
  - ...

Overall Structure

BROOKHAVEN NATIONAL LABORATORY
DECTRIS detecting the future

- The overall structure of a NeXus/HDF5 file represents a tree of nested groups. Each group may contain fields and/or groups. Each field or groups may have attributes.
- The top level of the tree is either an NXentry or an Nxsubentry.
- Looking at just the top two levels of each group:
- Group:NXentry
  - Group:NXdata
  - Group:NXsample
    - ...
  - Group:NXinstrument
    - ...
  - Group:NXsource

Overall Structure

BROOKHAVEN NATIONAL LABORATORY
DECTRIS detecting the future

- The overall structure of a NeXus/HDF5 file represents a tree of nested groups. Each group may contain fields and/or groups. Each field or groups may have attributes.
- The top level of the tree is either an NXentry or an Nxsubentry.
- Looking at just the top two levels of each group:
- Group:NXentry
  - Group:NXdata
  - Group:NXsample
    - NXtransformations
  - Group:NXinstrument
    - ...
  - Group:NXsource

Overall Structure

- The overall structure of a NeXus/HDF5 file represents a tree of nested groups. Each group may contain fields and/or groups. Each field or groups may have attributes.
- The top level of the tree is either an NXentry or an Nxsubentry.
- Looking at just the top two levels of each group:
- Group:NXentry
  - Group:NXdata
  - Group:NXsample
    - ...
  - Group:NXinstrument
    - Group:NXattenuator
    - Group:NXdetector_group
    - Group:NXdetector
    - Group:NXbeam
  - Group:NXsource

NXdetector with Fields

BROOKHAVEN NATIONAL LABORATORY

DECTRIS detecting the future

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional

- field:data recommended
- field:description recommended
- field:time_per_channel optional
- Group:NXdetector_module required
- field:distance
- field:distance_derived recommended
- field:dead_time optional
- field:count_time recommended
- field:beam_center_derived optional
- field:beam_center_x recommended
- field:beam_center_y recommended
- field:angular_calibration_applied optional
- field:angular_calibration optional
- field:flatfield_applied optional
- field:flatfield optional
- field:flatfield_error optional
- field:pixel_mask_applied optional
- field:pixel_mask recommended
- field:countrate_correction_applied optional
- field:bit_depth_readout recommended
- field:detector_readout_time optional
- field:frame_time optional
- field:gain_setting optional
- field:saturation_value optional
- field:underload_value optional
- field:sensor_material required
- field:sensor_thickness type required
- field:threshold_energy optional
- field:type optional

NXdetector with Fields

BROOKHAVEN NATIONAL LABORATORY
DECTRIS detecting the future

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
      - field:sensor_material required
      - field:sensor_thickness type required
      - field:threshold_energy optional
      - field:type optional

NXdetector with Fields

BROOKHAVEN NATIONAL LABORATORY
DECTRIS detecting the future

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
      - field:sensor_material required
      - field:sensor_thickness type required
      - field:threshold_energy optional
      - field:type optional

NXdetector with Fields

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
      - field:sensor_material required
      - field:sensor_thickness type required
      - field:threshold_energy optional
      - field:type optional

NXdetector with Fields

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
      - field:sensor_material required
      - field:sensor_thickness type required
      - field:threshold_energy optional
      - field:type optional

NXdetector with Fields

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
      - field:sensor_material required
      - field:sensor_thickness type required
      - field:threshold_energy optional
      - field:type optional

NXdetector with Fields

BROOKHAVEN NATIONAL LABORATORY
DECTRIS detecting the future

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
      - field:sensor_material required
      - field:sensor_thickness type required
      - field:threshold_energy optional
      - field:type optional

NXdetector with Fields

- Group:NXentry
  - Group:NXinstrument
    - Group:NXdetector
      - field:depends_on optional
      - Group:NXtransformations required
      - Group:NXcollection optional
      - field:data recommended
      - field:description recommended
      - field:time_per_channel optional
      - Group:NXdetector_module required
      - field:distance
      - field:distance_derived recommended
      - field:dead_time optional
      - field:count_time recommended
      - field:beam_center_derived optional
      - field:beam_center_x recommended
      - field:beam_center_y recommended
      - field:angular_calibration_applied optional
      - field:angular_calibration optional
      - field:flatfield_applied optional
      - field:flatfield optional
      - field:flatfield_error optional
      - field:pixel_mask_applied optional
      - field:pixel_mask recommended
      - field:countrate_correction_applied optional
      - field:bit_depth_readout recommended
      - field:detector_readout_time optional
      - field:frame_time optional
      - field:gain_setting optional
      - field:saturation_value optional
      - field:underload_value optional
    - field:sensor_material required
    - field:sensor_thickness type required
    - field:threshold_energy optional
    - field:type optional

References I

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

(Bernstein, 2005) Bernstein, H. J. (2005). International Tables For Crystallography, Vol. G, edited by S. R. Hall & B. McMahon, pp. 199 – 205. Chester: International Union of Crystallography.

(Bernstein & Hammersley, 2005) 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.

(Bernstein, 2010) Bernstein, H. J. (2010). HDF5 as Hyperspectral Data Analysis Format Workshop, 11 – 13 January 2010, ESRF, Grenoble, France.

(Bernstein, 2017) Bernstein, H. J. (2017). Acta Cryst. A73, a189.

(Donath *et al.*, 2013) Donath, T., Rissi, M. & Billich, H. (2013). Synchrotron Radiat. News, 26, 34 – 35.

(Ellis & Bernstein, 2005) Ellis, P. J. & Bernstein, H. J. (2005). International Tables For Crystallography, Vol. G, edited by S. R. Hall & B. McMahon, pp. 544 – 556. Chester: International Union of Crystallography.

References II

(Hester, 2016) Hester, J. (2016). Data Sci. J. 15, 12.

(Jiang *et al.*, 1999) Jiang, J., Abola, E. & Sussman, J. L. (1999). Acta Cryst. D55, 4.

(Kabsch, 2010a) Kabsch, W. (2010a). Acta Cryst. D66, 125 – 132.

(Kabsch, 2010b) Kabsch, W. (2010b). Acta Cryst. D66, 133 – 144.

(Könnecke *et al.*, 2015) Könnecke , M., Akeroyd, F. A., Bernstein, H. J., Brewster, A. S., Campbell, S. I., Clausen, B., Cottrell, S., Hoffmann, J. U., Jemian, P. R., Männicke, D., Osborn, R., Peterson, P. F., Richter, T., Suzuki, J., Watts, B., Wintersberger, E. & Wuttke, J. (2015). J. Appl. Cryst. 48, 301 – 305.

(Kroon-Batenburg & Helliwell, 2017) Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). Acta Cryst. D70, 2502 – 2509.

(Mariani *et al.*, 2016) Mariani, V., Morgan, A., Yoon, C. H., Lane, T. J., White, T. A., O'Grady, C., Kuhn, M., Aplin, S., Koglin, J., Barty, A. & Chapman, H. N. (2016). J. Appl. Cryst. 49, 1073 – 1080.

(Otwinowski & Minor, 1997) Otwinowski, Z. & Minor, W. (1997). Methods Enzymol. 276, 307 – 326.

(Powell *et al.*, 2007) Powell, H., Leslie, A. & Battye, G. (2007). CCP4 Newsl. Protein Crystallogr. 46, contribution 1.

References III

(Vonrhein *et al.*, 2011) Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). Acta Cryst. D67, 293 – 302.

(Waterman *et al.*, 2013) Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K., Evans, G. & Rosenstrom, P. (2013). CCP4 Newsl. Protein Crystallogr. 49, 13–15.

(Winter, 2010) Winter, G. (2010). J. Appl. Cryst. 43, 186–190.

(Winter *et al.*, 2018) Winter, G.,Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). Acta Cryst. D74, 85 – 97.

(Wolf & Wicksteed, 1998) Wolf, M. & Wicksteed, C. (1998). Status for Date and Time Formats. https://www.w3.org/1998/.status/NOTE-datetime-19980827/status.