Best practices for high data-rate macromolecular crystallography (HDRMX)

Herbert J. Bernstein[a,b], Lawrence C. Andrews[c], Jorge Diaz[d], Jean Jakoncic[e], Nicholas K. Sauter[f], Alexei Soares[g] and Maciej R. Wlodek[h]

[a]Rochester Institute of Technology, c/o NSLS-II Bldg 745, Brookhaven National Laboratory, Upton, NY 11973, USA, yayahjb@gmail.com
[b]Ronin Institute for Independent Scholarship, 5 Brewster Lane, Bellport, NY 11713, USA, yayahjb@gmail.com
[c]Ronin Institute for Independent Scholarship, 9515 NE 137th St, Kirkland, WA, 98034 USA, lawrence.andrews@ronininstitute.org
[d]Saint Joseph's College, Patchogue, NY, 11772, USA, jorgediazjr7@gmail.com
[e]NSLS-II Bldg 745, Brookhaven National Laboratory, Upton, NY 11973, USA, jjakoncic@bnl.gov
[f] Lawrence Berkeley National Laboratory, Biosciences, Berkeley, CA 94720, USA, nksauter@lbl.gov
[g]NSLS-II Bldg 745, Brookhaven National Laboratory, Upton, NY 11973, USA, soares@bnl.gov
[h]Stony Brook University, Stony Brook, NY 11794, USA maciej.r.wlodek@gmail.com

Enabled by changes in technology, macromolecular crystallography increasingly is able to extend its focus from the averaged state observed in a single crystal or in a few merged crystals to studies of families of distinct structural states observed by single-shot or small wedge probes of a large ensemble of tiny crystals or micro-focus probes of one or more larger crystals. This transition is driving a series of disruptive changes in the way diffraction data is collected, processed, and archived. Hardware improvements, such as fast high resolution detectors, high brilliance x-ray micro-beams, and automated sample handling, are generating high data-rate and high data-volume data streams that conventional software packages and pipelines designed for simple single crystal experiments and one-node serial processing are not able to support or even keep up with. Past practice must be re-examined to ensure the quality of results and timely delivery. Networks and computational resources have had to be upgraded. Bottlenecks in pipelines must be removed, often by converting serial execution to parallel execution on multiple nodes, but those changes themselves can generate yet more network and computational load. Higher flux, smaller beams and faster detectors open the door to experiments with very large numbers of very small samples that can reveal polymorphs and dynamics, but require re-engineering of approaches to clustering of images both at synchrotrons and XFELs. The need for management of orders of magnitude more images and limitations of filesystems favor a transition from simple one-file-per-image systems such as CBF to image container systems such as HDF5. This further increases the load on computers and networks and requires a re-examination of the presentation of metadata. In this talk we discuss three important components of this problem – improved approaches to clustering of images to better support experiments on polymorphs and dynamics, recent and upcoming changes in metadata for Eiger images, and software to rapidly validate images in the revised Eiger format.

This work has been supported in part by funding from Dectris, Ltd., NIH and DOE.