Contribution ID: **25** Type: **Submitted talk**

Data reduction in serial crystallography

Wednesday 20 September 2023 11:30 (30 minutes)

Serial crystallography (SX) has become an established technique for protein structure determination, especially when dealing with small or radiation-sensitive crystals and investigating fast or irreversible protein dynamics. The advent of newly developed multi-megapixel X-ray area detectors, capable of capturing over 1000 images per second, has brought about substantial benefits. However, this advancement also entails a notable increase in the volume of collected data. Today, up to 2 PB of raw data per experiment could be easily obtained under efficient operating conditions. The combined costs associated with storing data from multiple experiments provide a compelling incentive to develop strategies that effectively reduce the amount of data stored on disk while maintaining the quality of scientific outcomes. Lossless data compression methods are designed to preserve the information content of the data but often struggle to achieve a high compression ratio when applied to experimental data that contains noise. Conversely, lossy compression methods offer the potential to greatly reduce the data volume. Nonetheless, it is vital to thoroughly assess the impact of data quality and scientific outcomes when employing lossy compression, as it inherently involves discarding information. The evaluation of lossy compression effects on data requires proper data quality metrics.

Our focus here is to evaluate different lossless and lossy data compression methods and determine the appropriate metrics for evaluating the impact of lossy compression on the final SX data quality. Our research found that effective strategies for lossy data reduction in SX are: non-hits rejection (in the case of strongly diffracting crystals), binning (in the case of crystals with small unit cell) and reduction in the precision of the measured diffraction pattern intensities, especially in a non-uniform way (saving only several most significant bits). At the same time, we demonstrate the potential risks associated with particular lossy data reduction schemes, such as: reduction in the number of stored diffraction patterns, saving only the intermediate results (.mtz files), or saving only the regions around the detectable Bragg peaks.

Some of the lossy compression schemes, that we have developed and/or tested, can be implemented either in hardware or as HDF5 plugins for application in crystallography as well as for data generated using other techniques.

Website

Author: GALCHENKOVA, Marina (FS-CFEL-1 (Forschung mit Photonen Experimente 1))

Co-authors: CHAPMAN, Henry (FS-CFEL-1 (Forschung mit Photonen Experimente 1)); YEFANOV, Oleksandr (FS-CFEL-1 (Forschung mit Photonen Experimente 1))

Presenter: GALCHENKOVA, Marina (FS-CFEL-1 (Forschung mit Photonen Experimente 1))

Session Classification: Day 2