A Novel GUI for Serial Data Classification Using Machine Learning Approaches

Gihan K Ketawala[1], Professor Petra Fromme[1], Ast Professor Sabbine[1]
[1]Botha Arizona State University
gihan.ketawala@asu.edu

The field of biological structure determination is currently experiencing a groundbreaking transformation driven by advances in serial X-ray diffraction using X-ray free-electron laser (XFEL) sources. The development of serial femtosecond crystallography (SFX) has been a pivotal milestone in this revolution. By enabling the measurement of micrometer-sized crystals at room temperature, SFX allows time-resolved studies to map the progression of biochemical reactions with unprecedented temporal resolution. This remarkable feat is made possible by ultrafast X-ray pulses that effectively circumvent the detrimental effects of radiation damage.

Such pulses can deliver X-ray doses more than a thousand times higher than those achievable with conventional X-ray sources[1]. The high-intensity X-ray pulses have spurred the development of state-of-the-art X-ray detectors, which are continually being refined and improved. However, experimental parameters can sometimes push these detectors beyond their reliable operating range, resulting in individual frames within a single run of data collection that may be unreliable. Incorporating intensities from these problematic frames into the merged structure factors can introduce inaccuracies in the final reported intensities. This can be especially detrimental for applications such as anomalous phasing or time-resolved difference density calculations, where highly accurate recordings are essential.

In addition to the challenges posed by detector limitations, researchers must contend with the vast amounts of data generated by XFEL experiments. A single experiment can produce multiple terabytes of data, which must be rapidly processed and analyzed. While software tools for online data monitoring and reduction have been developed over the past decade (e.g., OM[2], Cheetah[3], CrystFEL[4]), these programs primarily focus on identifying crystal hits rather than classifying data based on spurious, often unquantifiable artifacts.

To address these challenges, we report the development of a new data sorting tool that incorporates a diverse range of Machine Learning algorithms. The data sorting tool is designed to be flexible and adaptable, catering to the specific requirements of individual experiments and addressing the challenges posed by data artifacts and detector limitations. This tool can be trained to sort data either through manual sorting by the user or by profile-fitting the expected intensity distribution on the detector based on the experiment.

Our data sorting tool is integrated into an intuitive graphical user interface (GUI) that is specifically tailored to support the detectors, file formats, and software utilized at prominent XFEL facilities such as the Linac Coherent Light Source (US) and the European XFEL (Germany). This user-friendly interface streamlines the data analysis process, enabling researchers to process and classify their data more accurately and efficiently rapidly. By improving the quality and reliability of the data produced by XFEL experiments, our data sorting tool promises to advance the field of biological structure determination further andcontribute to a deeper understanding of the fundamental processes that govern life at the molecular level.

In summary, the ongoing revolution in the field of biological structure determination is being driven by the advent of serial femtosecond crystallography (SFX) and X-ray free- electron laser (XFEL) sources. These technologies have enabled researchers to probe biochemical reactions with unparalleled temporal resolution, but they also generate vast amounts of data and pose unique challenges for data processing and analysis. To address these issues, we have developed a new data sorting tool that leverages machine learning algorithms and is integrated into an easy-to-use graphical user interface (GUI). This tool offers improved accuracy and efficiency in processing and classifying data from XFEL experiments, further enhancing our understanding of the complex molecular processes that underpin life.

References

{1} Chapman, Henry N, et al. Nature 470, no. 7332 (2011): 73-77.
{2} Barty, Anton, et al. Journal of applied crystallography 47, no. 3 (2014): 1118-1131.
{3} Mariani, Valerio, et al. Journal of applied crystallography 49, no. 3 (2016): 1073-108
{4} White, Thomas A., et al. Journal of applied crystallography 45, no. 2 (2012): 335-341.

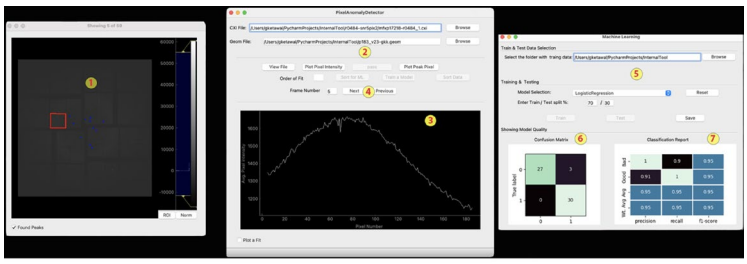

Figure 1. The user interface of the PixelAnomalyDetector. *The* loaded *image file* is displayed in the image viewer (1), and a vertically averaged pixel intensity plot (3) is readily available in the *main window* (2). Seamlessly navigate through HDF5 files with the next and previous buttons (4). The *tool* offers a wide range of options to train your model, including model selection and customizable train-test splits (5), *with* the ability to display model performance via a confusion matrix (6) and classification report (7).

Figure 1