Building Latent Spaces to Sort Massive X-ray Diffraction Datasets

SLAC NATIONAL ACCELERATOR LABORATORY

Deeban Ramalingam[1], Chun Hong Yoon[2], Frédéric Poitevin[2]

[1]University of California Los Angeles, Los Angeles, CA 90095 USA.

[2]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA.

[+]Contact: fpoitevi@slac.stanford.edu

Introduction

Single Particle Imaging (SPI) by X-ray Free-electron Lasers (XFELs) is used to reconstruct 3D electron density models from single-particle diffraction patterns. It is important that the patterns used correspond to exactly one particle diffracting while in practice the recorded images might correspond to any integer number of particles diffracting at the same time. In this work, we apply dimensionality reduction techniques to build latent spaces that enables efficient sorting of single-particle hits from other types of hits. Since large-scale experimental datasets are not available yet and lack ground-truth information about the image content and the orientations of the particle being imaged, to validate and calibrate our method, we first simulate massive synthetic datasets with pysingfel[1], a simulator developed in the Advanced Methods group at LCLS. We have also developed user-friendly visualizers that allow to explore the latent spaces generated.

Keywords: SPI, XFEL, latent spaces

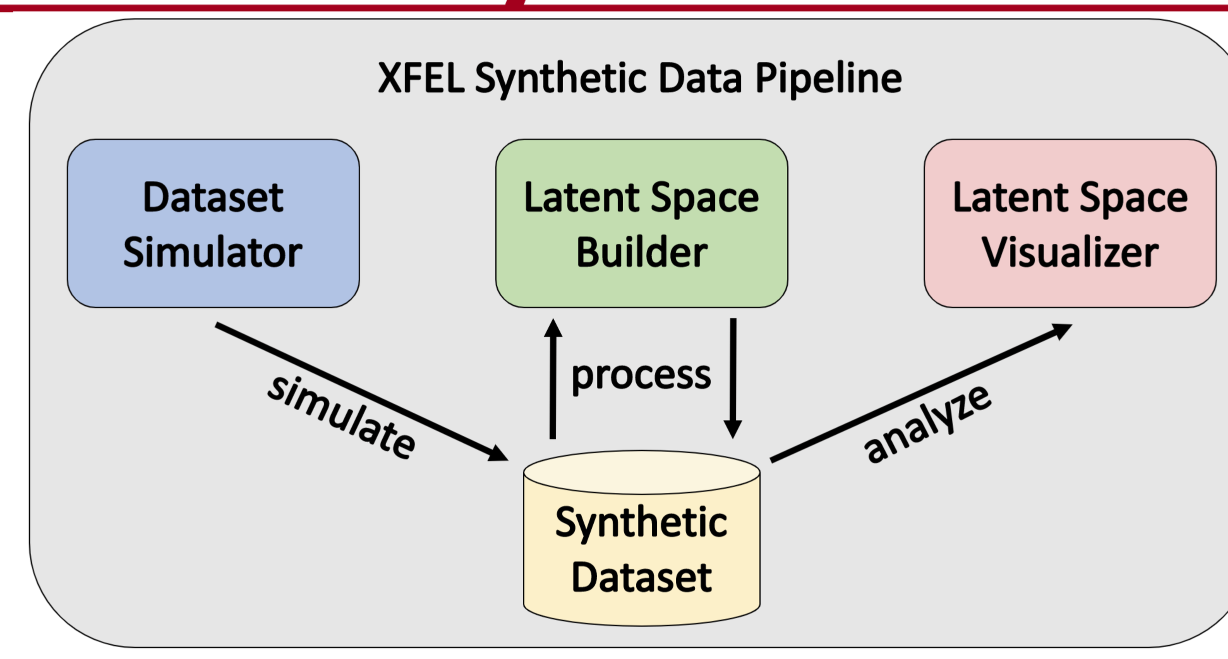


Fig 1. The XFEL Synthetic Data Pipeline simulates, processes, and analyzes X-ray diffraction data to validate and calibrate dimensionality reduction techniques for sorting X-ray diffraction patterns.

Dataset Simulator

The Dataset Simulator generates several diffraction patterns and ground-truth orientations which are stored in the Synthetic Dataset. We used pysingfel on 3iyf[2] to create a series of mixed-hit datasets of various mixture ratios of single-particle hits to outliers (80:20, 90:10, 95:5, 99:1) and sizes (10K and 100K). When simulating the diffraction patterns, three objects need to be defined. Their parameters are specified below (For more details, see [3].):

Beam parameters	
Energy (eV)	4600.0
Fluence (photons per second)	1.0e12
Radius (m)	0.1e-6
Detector	
Type	pnCCD
Particle	
System	3iyf
Hydrated	No

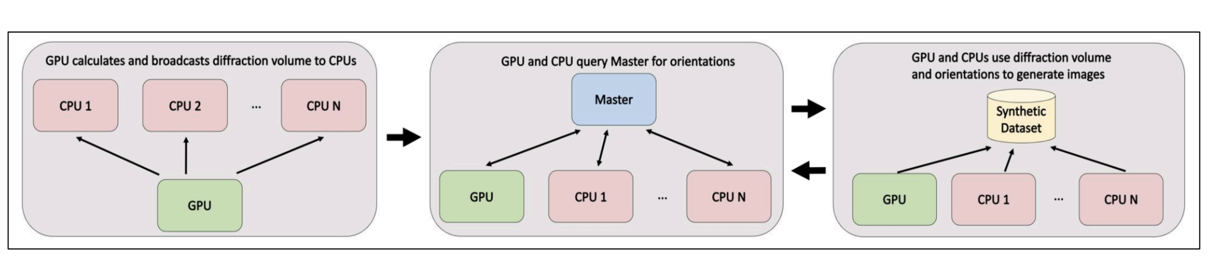


Fig 2. MPI[4] model for efficiently simulating large-scale synthetic diffraction datasets.

Latent Space Builder

The Latent Space Builder constructs the latent spaces from the Synthetic Dataset of diffraction images using dimensionality reduction methods. The latent spaces are then stored in the Synthetic Dataset. We construct the latent space using Incremental PCA. We flag outliers using Elliptic Envelope. For 3iyf-99-100K, 90K single-hits and 1K outliers were correctly identified. There were no false positives.

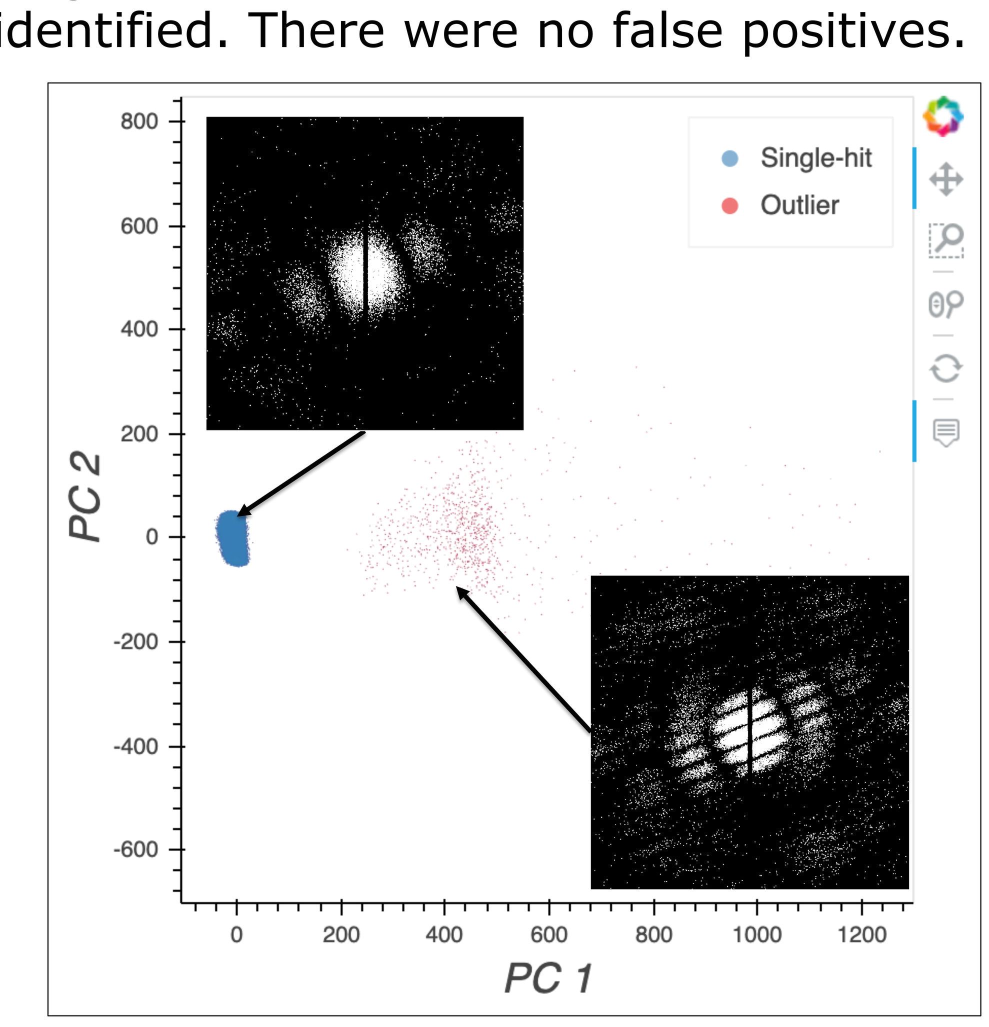


Fig 3. Built latent space separates single-particle hits from other types of hits.

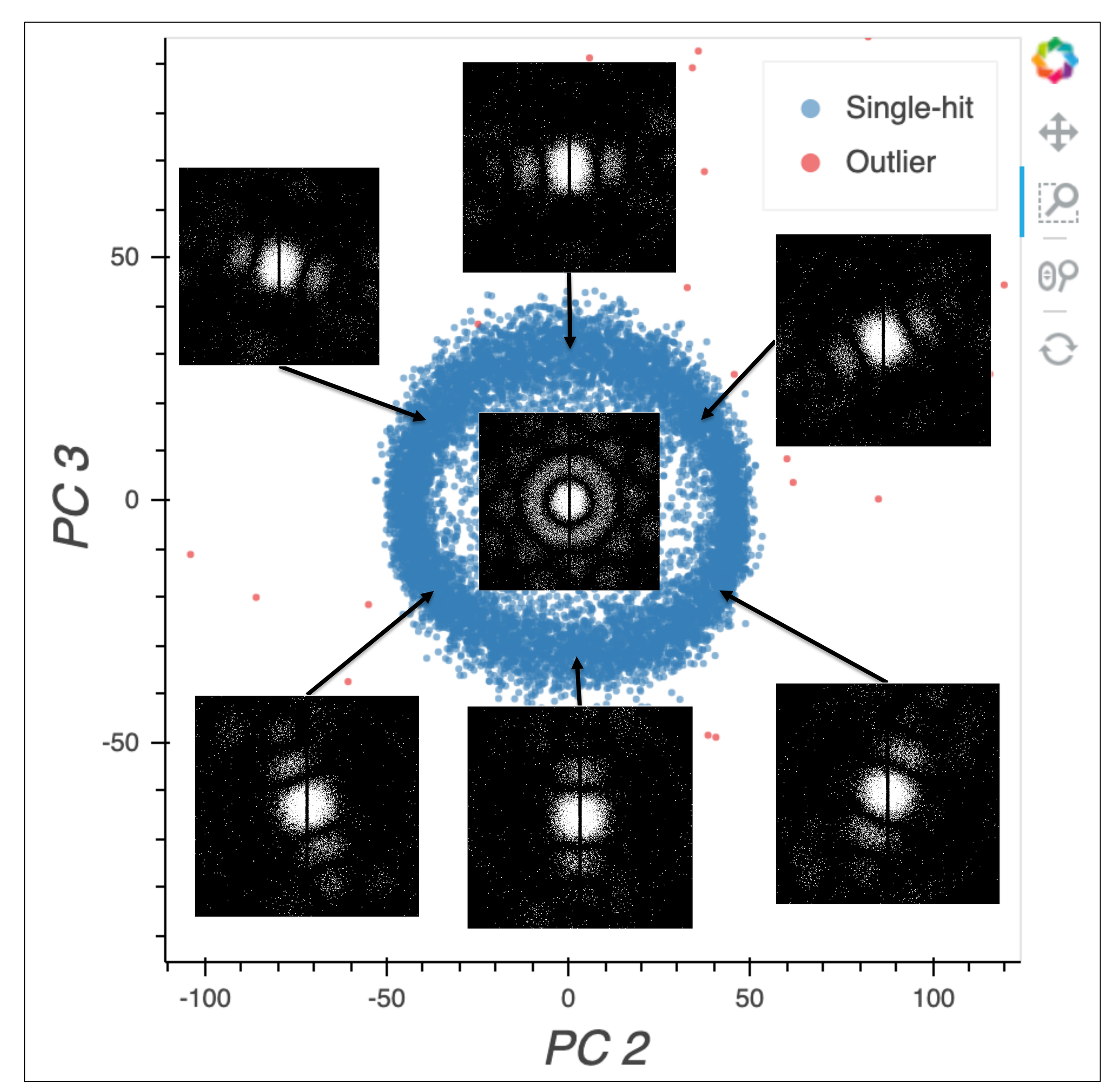


Fig 4. Built latent space captures orientation information associated with each diffraction pattern.

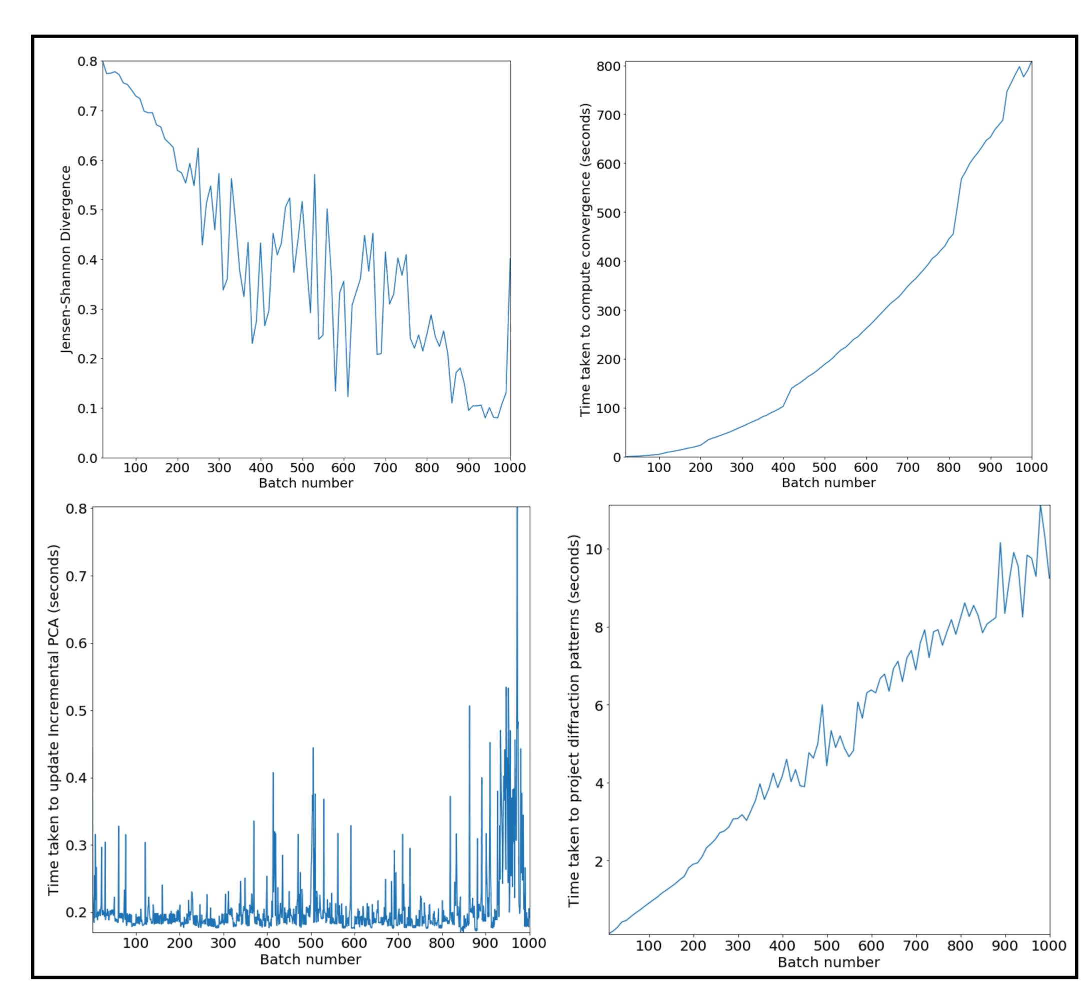
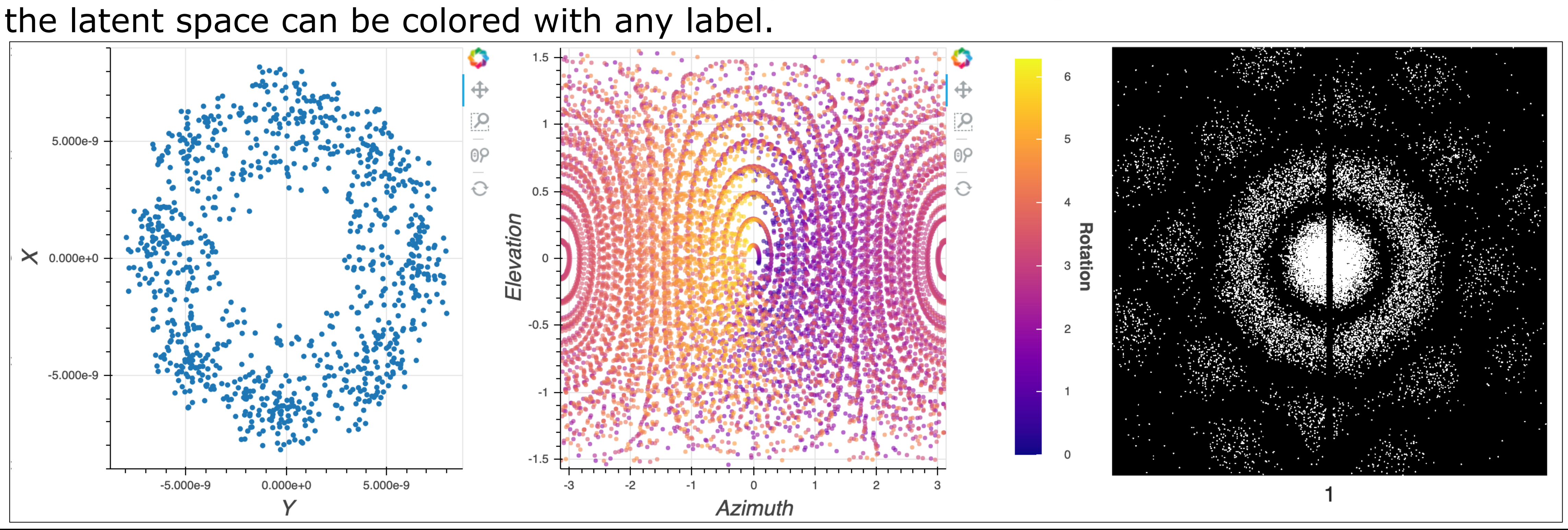


Fig 5. Convergence and timing analysis for building the latent space.

Latent Space Visualizer

Written with Bokeh. Displays both real- and reciprocal space images and latent spaces. The points in the latent space can be colored with any label.



Conclusions

We built a latent space for a large number of simulated diffraction patterns. This latent space not only allows us to sort single-particle hits from other types of hits, but also reveals the orientation information associated with each diffraction pattern. This method would potentially enable efficient processing of large-scale diffraction data in SPI by XFELs for LCLS.

Acknowledgments

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

Date: 09/10/2020