Joint APS/CNM Workshop 1: Advancing Data Analysis for XRD/XCT in Post-APS-U: FAIR, Automated Analysis Pipelines and Graph Deep Learning

Tuesday, May 7, Afternoon

1:30 – 1:40 Donald Brown (Lawrence Berkeley National Laboratory)
Users Perspective on Synchrotron HEXRD Data Analysis and Challenges

1:40 – 2:00 Daniela Ushizima (Lawrence Berkeley National Laboratory)
Analyzing Lithium Structures in Batteries Using Iterative Residual U-Net and Operando *XCT*

2:00 – 2:20 Roger French (Case Western Reserve University)
Data-centric AI for Synchrotron Science: Enabled by FAIR, Pipelines, Graph Learning

2:20 – 2:40 Pawan Tripathi (Case Western Reserve University)
CRADLE Data Explorer: Advanced Visualization Platform for HEXRD and XCT Image Sequences and Deep Learning Analysis of HEXRD and XCT Image Sequences

2:40 – 3:00 Hemant Sharma (Argonne National Laboratory)
High-energy X-ray Diffraction and High-energy Diffraction Microscopy (HEDM) Data Analysis Tools

3:00 – 3:10 Break

3:10 – 3:30 Reeju Pokharel (Los Alamos National Laboratory)
Machine Learning for Faster Microstructure Analysis

3:30 – 3:50 Rafael Vescovi (Argonne National Laboratory)
Streamlining Data Analysis for Synchrotron Datastreams Using Globus Platform Services

3:50 – 4:10 Brian Toby (Argonne National Laboratory)
Early Work on Powder Diffraction Data Collection/Reduction Pipelines and Recent Work on Data Analysis Tools

4:10 – 4:30 Dan Savage (Los Alamos National Laboratory)
Streamlining Engineering Diffraction Analysis Using the MAUD Interface Language Kit (MILK)

4:30 – 4:50 Jana B. Thayer (SLAC National Accelerator Laboratory)
Edge to Exascale Workflows to Support Massive-scale Data Analytics at LCLS

4:50 – 5:00 Conclusion and Closing Remarks

5:00 Adjourn

Analyzing Lithium Structures in Batteries Using Iterative Residual U-Net and *Operando* XCT

Daniela Ushizima[1,2,3], Iryna Zenyuk[1,4], Dula Parkinson[1], and Pavel Shevchenko[5]

[1]Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[2]Berkeley Institute of Data Sciences, University of California, Berkeley, CA 94720
[3]Bakar Institute, University of California, San Francisco, CA 94143
[4]National Fuel Cell Research Center, University of California, Irvine, CA 92697
[5]Argonne National Laboratory, Lemont, IL 60439

The study introduces batteryNET, a computational framework [1] based on an Iterative Residual U-Net architecture, designed for detecting lithium (Li) structures in *operando* datasets of lithium-metal batteries imaged at the synchrotron facilities such as the APS. A key challenge addressed is the low x-ray attenuation of Li metal combined with the large size of the data. This approach enables the monitoring of Li structure evolution within pouch cells using XCT by capturing changes in Li-related components with semantic segmentation and providing detailed visualizations of inactive Li. Additionally, the study quantifies the volume and effective thickness of electrodes, as well as the deposited and redeposited Li, uncovering new insights into the spatial relationships among these components. This deep learning framework offers valuable data for assessing battery performance, and a new way to compare current samples with future battery designs. Furthermore, the developed semantic segmentation technique demonstrates versatility, showing potential applicability to other datasets featuring filamentous structures [2,3].

This work was supported by AMLXD and CAMERA projects led by Lawrence Berkeley National Laboratory under contract number DE-AC02-05CH11231 with the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility, located at Lawrence Berkeley National Laboratory.

[1] Huang, Perlmutter, Su, Quenum, Shevchenko, Parkinson, Zenyuk, Ushizima, “Detecting lithium plating dynamics in a solid-state battery with *operando* x-ray computed tomography using machine learning”. Nature Computational Materials (2023).
[2] Quenum, Zenyuk, Ushizima, “Lithium Metal Battery Quality Control via Transformer–CNN Segmentation.” Journal of Imaging (2023).
[3] Ushizima, Sordo, Andeer, Sethian, Northen, “Rhizonet: Image Segmentation for Plant Root in Hydroponic Ecosystem,” bioRxiv 2023.11. 20.565580 (2023).

Data-centric AI for Synchrotron Science: Enabled by FAIR, Pipelines, Graph Learning

Roger H. French[1], Alexander H. Bradley[2], Balashanmuga Priyan Rajamohan[2], Arafath Nihar[2], Thomas G. Ciardi[2], Weiqi Yu[2], Redad Medhi[1], Erika I. Barcelos[1], Pawan K. Tripathi[1], Frank Ernst[1], and Matthew A. Willard[1]

[1]Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH 44106
[2]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106

APS-U presents new opportunities for deep learning from massive, multimodal synchrotron datasets. By leveraging existing datasets and datastreams, technologically informed “AI for Synchrotron Science” can be developed using a “data-centric AI" approach, as opposed to the “model centric AI” of things such as Large Language Models. Data-centric AI, trained on Terabyte datasets, creates models which can naturally address complex materials science questions on performance, structure, properties, and reliability. But the large datasets and datastreams required for training present substantial challenges for scientists working on their notebook computers or even in traditional HPC clusters. These data-centric AI challenges can be addressed by integration of distributed computing with high performance computing as we do in CRADLE, combined with data FAIRification, and development of automated analysis pipelines that self-document the analysis by FAIRifiying the intermediate and final results. CRADLE Data Explorer, for example, enables real-time exploratory data analysis across petabytes of data, to inform scientific investigations. For image or movie datasets, a natural learning framework is convolutional neural networks. More complex problems often naturally fall into a spatiotemporal analysis framework, with deep learning by spatiotemporal graph neural network (st-GNN) models with metadata feature vectors. A st-GNN model trained on spatial and temporal data, will utilize the intrinsic spatial and temporal coherence of the problem. These deep learning neural network models therefore can encapsulate all the information from the experiments in an st-graph and a knowledge graph (k-graph). This trained k/st-graph model serves as a data-driven Digital Twin of the problem it trained on. Data-centric k/st-graph deep learning AI enables new approaches to complex problems and massive datasets, while also providing concise insights into scientific findings.

CRADLE Data Explorer: Advanced Visualization Platform for HEXRD and XCT Image Sequences and Deep Learning Analysis of HEXRD and XCT Image Sequences

Pawan Tripathi[1], Tommy Ciardi[2], Arafath Nihar[2], Weiqi Yu[2], Redad Mehdi[1], Finley Holt[1], Gabriel Ponon[1], Matthew Willard[1], Frank Ernst[1], and Roger French[1]

[1]Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH 44106
[2]Department of Computer and Data Sciences, Case Western Reserve University, Cleveland, OH 44106

The CRADLE Data Explorer is a cutting-edge visualization platform built on an integrated approach utilizing distributed and high-performance computing, for visualizing High-Energy X-ray Diffraction (HEXRD) and X-ray Computed Tomography (XCT) image sequences. It provides users with immersive 3D views of HEXRD and XCT data, enabling exploration down to the pixel level. Leveraging advanced visualization techniques, researchers can navigate through intricate image sequences with unparalleled precision, gaining insights into the structural characteristics of materials under investigation. The integrated approach of the CRADLE Data Explorer ensures seamless interaction with distributed and high-performance computing resources. This enables researchers to handle large and complex datasets with ease, while maintaining optimal performance and scalability. Another aspect of CRADLE Data Explorer lies in its ability to detect and analyze subtle features within the data, including noise and other anomalies. By employing sophisticated algorithms and data processing techniques, researchers can identify and quantify these features, facilitating a deeper understanding of the underlying materials properties.

Additionally, we demonstrate automated analysis pipelines based on deep learning techniques learning from 2D HEXRD image sequences, enabling the detection and characterization of phase transformations with remarkable accuracy and pitting corrosion from XCT image sequences paving the way towards *in-situ* analysis during data collection at synchrotron facilities.

High-energy X-ray Diffraction and High-energy Diffraction Microscopy (HEDM) Data Analysis Tools

Hemant Sharma[1]

[1]Computational Sciences and Artificial Intelligence Group, X-ray Science Division, Argonne National Laboratory, Lemont, IL 60439

I will cover a range of tools developed at the APS for analysis of high-energy x-ray diffraction data. These include automated setup calibration, fast azimuthal integration for powder diffraction, and algorithms for the reconstruction of high-energy diffraction microscopy (HEDM) data. The talk will also include a brief demo of the various capabilities.

Machine Learning for Faster Microstructure Analysis

Reeju Pokharel

[1]Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87544

Multimodal scattering data holds valuable insights into material microstructure and properties, yet analyzing it requires extensive expert user interventions. We apply tailored machine learning (ML) methods to accelerate microstructure analysis from diffraction data. Here we present proof-of-concept results to rapidly determine orientations, enhance information extraction from noisy measurements, and enable real-time simulations and reconstructions. Overall, the current ML methods demonstrate orders of magnitude speed-ups in analysis. The overarching goal is to combine these methods to guide experiments in real-time by providing instant microstructural feedback. This will maximize scientific insights extractable from experiments within limited beamtime.

Streamlining Data Analysis for Synchrotron Datastreams Using Globus Platform Services

Rafael Vescovi[1]

[1]Argonne National Laboratory, Lemont, IL 60439

The intersection of advanced computing technologies and the increasing data volumes from scientific instruments poses both challenges and opportunities. Advanced detectors employed by modern experimental facilities routinely collect data at multiple GB/s, necessitating online analysis methods to enable the collection of interesting subsets of such massive data streams. New methods are required for configuring and running distributed computing pipelines—what we call flows—that link instruments, computers (e.g., for analysis, simulation, AI model training), edge computing (e.g., for analysis), data stores, metadata catalogs, and high-speed networks. This talk will describe how Globus platform services, particularly Globus Flows and Compute, can be used to streamline the analysis of data generated by scientific instruments through the integration with High-Performance Computing (HPC) resources. I will demonstrate how Flows can be used to construct a secure analysis pipeline to automate and outsource the management, analysis, and publication of APS data. I will explain how such flows can be deployed, monitored, and managed at both the APS and ALCF and demonstrate their application.

Early Work on Powder Diffraction Data Collection/Reduction Pipelines and Recent Work on Data Analysis Tools

Brian Toby[1]

[1]Advanced Photon Source, Argonne National Laboratory, Lemont, IL 60439

In the mid-2000s, beamline 11-BM was initially deployed for mail-in high-resolution powder diffraction, at a time when mail-in access to user facilities was nearly unknown. To manage this, I developed a proposal-to-disposal sample management system that included a data collection pipeline. The pipeline provided data post-processing and also automated instrument calibration. Most instrument operation, including alignment, was also automated. This work will be reviewed. Since that time, much of my effort, combined with that of Robert Von Dreele, has been to establish GSAS-II, a general-purpose diffraction data analysis package, for all types of diffraction measurements, including single-crystal, powder diffraction, pink-beam, small-angle, and reflectometry from high-resolution, imaging, and lab x-ray sources, as well as TOF and CW neutron sources. This includes Rietveld analysis, as well as image detector calibration, masking, and integration. While GSAS-II was initially designed as a GUI-based tool, it also has Python API. The API, along with the platform independence offered by Python, has allowed GSAS-II to become the tool of choice for projects wishing to embed diffraction data reduction, simulations, or fitting into a pipeline.

Streamlining Engineering Diffraction Analysis Using the MAUD Interface Language Kit (MILK)

Dan Savage[1]

[1]Los Alamos National Laboratory, Los Alamos, NM 87545

Diffraction experiments often result in tens to thousands of patterns from which information can be extracted ranging from microstructure to equation of state. The extraction happens by fitting models of the diffraction process which describe the instrument (e.g., sample-to-detector distances and angles), the sample's crystal structure (e.g., crystallographic space group and lattice parameters), and the microstructure (e.g., phase fractions and texture). Finding the set of parameters that best describes a measurement is hard to do in a robust way without expert intervention and in a way that leverages parallel resources (critical for real-time analysis). Benchmark examples of the open-source Python scripting framework for MAUD Rietveld software (MILK) will be presented with the Python Rietveld optimization and uncertainty quantification package (Spotlight) which together enable the scalable Rietveld optimization strategy and advanced analysis workflows (e.g., sampling uncertainty in key parameters).

Edge to Exascale Workflows to Support Massive-scale Data Analytics at LCLS

Jana B. Thayer[1]

[1]SLAC National Accelerator Laboratory, Menlo Park, CA 94025

The LCLS-II Data System architecture addresses the challenges in data acquisition, data processing, data management, and workflow orchestration posed by the increase in data rate, volume, and complexity generated by the Linac Coherent Light Source upgrade. However, the exponential increase in the scale and speed of the data is prohibitive to traditional data analysis workflows, which rely on scientists painstakingly tuning parameters during live experiments to guide data collection and analysis. Instead, the automated delivery of actionable information about the experiment in real-time and near-real-time is needed to enable experiment steering and experiment design. Data processing and feature extraction at the edge are a strategic solution to producing actionable information that can feed sophisticated machine-assisted experiment steering feedback loops. Edge to exascale workflows spawn new requirements on metadata collection and provenance tracking.