Synchrotron Radiation News

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gsrn20

Advancing Discovery with Artificial Intelligence and Machine Learning at NSLS-II

Andi Barbour, Stuart Campbell, Thomas Caswell, Masafumi Fukuto, Marcus Hanwell, Andrew Kiss, Tatiana Konstantinova, Ricarda Laasch, Phillip Maffettone, Bruce Ravel & Daniel Olds

To cite this article: Andi Barbour, Stuart Campbell, Thomas Caswell, Masafumi Fukuto, Marcus Hanwell, Andrew Kiss, Tatiana Konstantinova, Ricarda Laasch, Phillip Maffettone, Bruce Ravel & Daniel Olds (2022): Advancing Discovery with Artificial Intelligence and Machine Learning at NSLS-II, Synchrotron Radiation News, DOI: 10.1080/08940886.2022.2114716

To link to this article: https://doi.org/10.1080/08940886.2022.2114716

Published online: 16 Sep 2022.

TECHNICAL REPORT

Advancing Discovery with Artificial Intelligence and Machine Learning at NSLS-II

Andi Barbour[1], Stuart Campbell[1], Thomas Caswell[1], Masafumi Fukuto[1], Marcus Hanwell[1], Andrew Kiss[1], Tatiana Konstantinova[1], Ricarda Laasch[1], Phillip Maffettone[1], Bruce Ravel[2] and Daniel Olds[1]

[1]*Brookhaven National Laboratory, Upton, New York, USA* *dolds@bnl.gov*

[2]*National Institute of Standards and Technology, Gaithersburg, Maryland, USA*

With the National Synchrotron Light Source II (NSLS-II) coming online in 2015 as the brightest source in the world, the imminent upgrades at the Advanced Photon Source, Advanced Light Source, and Linear Coherent Light Source, and advances in detector technology, the data generation rates at the U.S. Department of Energy (DOE) Basic Energy Sciences' X-ray light sources are skyrocketing. At NSLS-II, over 1 petabyte of raw data was produced last year, and that rate is expected to increase as the facility matures [1]. Despite such huge data generation rates, approaches to both experimental control and data analysis have not kept pace. Consequently, data collected in seconds to minutes may take weeks to months of analysis to understand. Due to such limitations, knowledge extraction is often divorced from the measurement process. The lack of real-time feedback forces users into flying blind at the beamline, leading to missed opportunities, mistakes, and inefficient use of beamtime as a resource—as all beamlines are oversubscribed. This is a challenge facing nearly all users of light sources. One promising path forward to solve this challenge—both during data collection and post-experiment analysis—is the use of artificial intelligence (AI) and machine learning (ML) methods [1, 2].

In this contribution, we review recent developments employing AI/ML methods at the NSLS-II, tackling the outlined challenges. These innovations have covered the whole life cycle of an experiment, from improving daily beamline operation, accelerating the pace of data analysis, and even automating experimentation with advanced AI-driven feedback loops. Here, we describe several developments in each of these areas that have been pioneered by scientists at the NSLS-II.

Improved operations

Foundational infrastructure for data access and artificial intelligence

One foundational cornerstone to these developments is the common underlying Bluesky control system [3] that spans NSLS-II beamlines, allowing development and deployment of AI/ML methods for both beamline control and data analysis in a streamlined and extensible fashion. Of note are recent developments of the Bluesky Queueserver and Tiled data access system. These developments are charting a new course for the Bluesky and Data Broker projects, respectively, where client-server architectures are now being developed to enable greater flexibility, extensibility, and reliability at beamlines. The Queueserver project offers a queue-based API for acquisition planning that closely maps to how many scientists plan their experiments, with the ability for both human and autonomous agents to edit the queue as the experiment is in progress. Tiled enables distributed remote access to the data, both raw and processed, promptly as it is acquired and reduced. Both projects expose programming language agnostic interfaces that fully encapsulate these services. This offers far greater freedom for researchers to try new things, such as AI/ML agent-directed data acquisition and streaming data analysis.

Automated data guidance for XAFS

One of the places AI/ML is finding use at the NSLS-II is streamlined operations, where the methods act like digital assistants to the human researchers by automating tedious or repetitive tasks. At NSLS-II's Beamline for Materials Measurement (BMM), a supervised learning model is used to make an initial evaluation of every X-ray Absorption Fine Structure (XAFS) spectrum as it is measured. The purpose of this evaluation is to distinguish measurements that look like XAFS spectra from failed measurements. A failed measurement might be due to network error, failure of a detector, or even something as prosaic as a poorly mounted sample falling from the beam. To build this tool, a dataset of over 800 hundred spectra has been tagged by beamline staff as either successful or failed measurements. Using this corpus, a classifier was trained to evaluate newly measured data. Multi-layer Perceptron and Random Forest classifiers have proven highly successful, with no false positive results and less than 2% rates of false negatives [4]. This evaluation tool is now incorporated into all XAFS measurements at the beamline. Measurements recognized as successful are subjected to further data reduction before being presented to the user. A negative evaluation triggers an alert to staff for potential operational problems at the beamline. Failed measurements are also logged for later evaluation by beamline staff and may be used to improve upon the training corpus.

Enhanced XRF imaging and tomography

At the Submicron Resolution X-ray Spectroscopy (SRX) beamline, pre-trained 3D convolutional neural networks (CNN) are being used to improve reconstruction methods and provide autonomous data acquisition for X-ray fluorescence (XRF) imaging and mapping. XRF is a

powerful way to spatially resolve the elemental distributions in materials. The fluoresced X-rays are collected to generate a hyper-spectral image, where individual element fluorescent lines can be identified simultaneously. Unlike full-field imaging techniques, in order to build up an image, the sample is rastered through a focused X-ray beam. This technique is sufficient for 2D imaging; however, it is difficult to compete with the speed of full-field X-ray detectors for 3D imaging and tomography. For tomographic reconstructions, where 2D images are collected at many different angles, this is a lengthy process that can consume an entire beamtime. In order to improve this imaging technique, advanced algorithms use AI/ML methods to improve the reconstruction quality and data collection efficiency. In this project, called HyperCT, NSLS-II is employing super voxel model-based tomographic reconstruction (svMBIR) algorithms that require fewer projections to resolve the sample reconstruction quality. Combining this with artificial intelligence and adaptive scans that can identify future projections that will have the greatest impact on improving reconstruction quality will decrease time necessary to collect hyperspectral, multi-element volumes of new and exciting materials and samples. The integration requirements for this project are met by Bluesky and Queueserver projects described earlier.

Streaming analysis

Phase transition detection in PDF data

ML methods can automate many of the specialized analysis methods, often at speeds on par with the data generation rates. Unsupervised learning approaches are of particular interest, as they can grant model-free

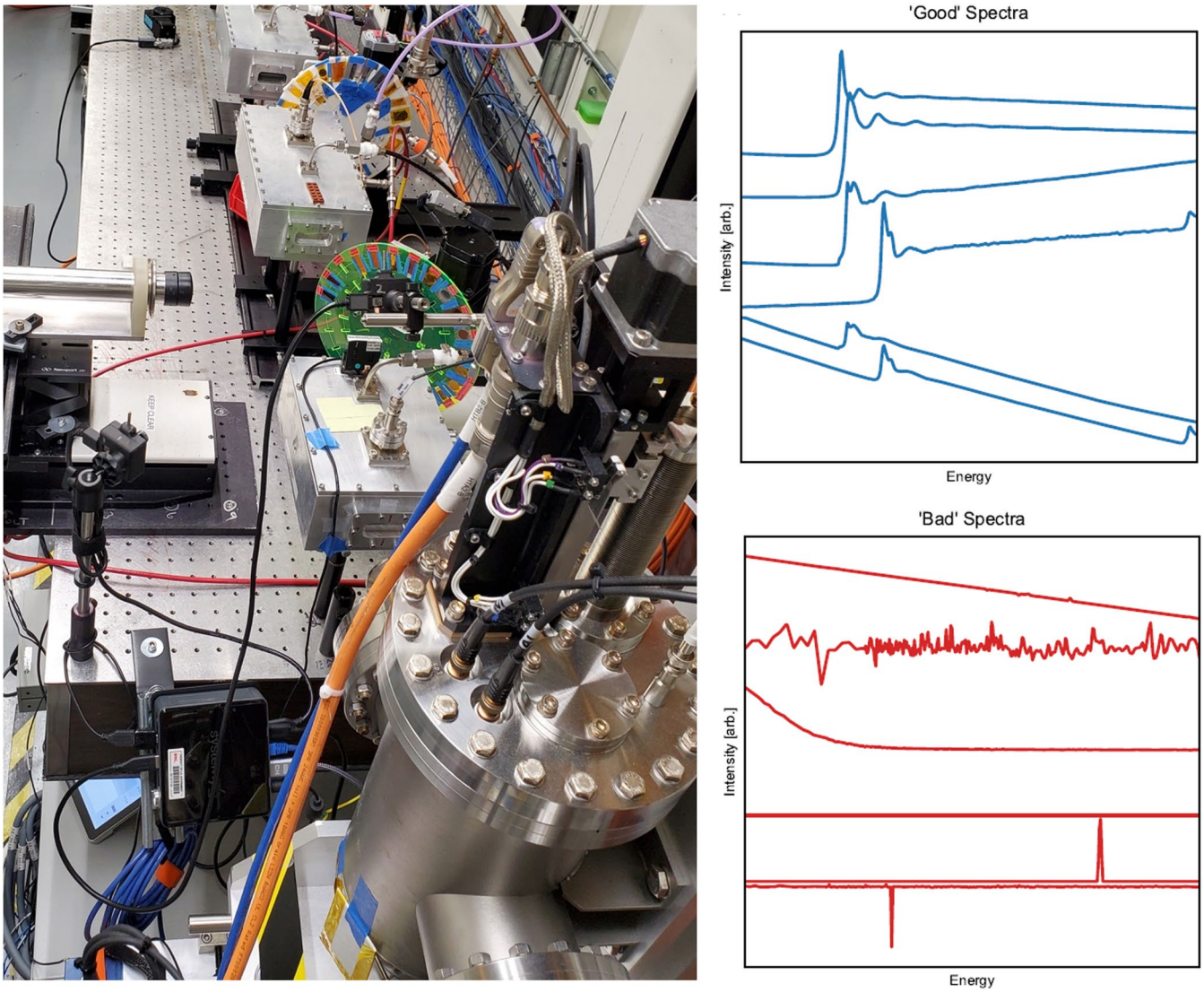

Figure 1: Examples of "good" and "bad" spectra used to train the classifier at the BMM beamline (shown in photograph). Note that data measured on any element are transformed onto a common, unit-less abscissa as required for classification by the machine learning model. The examples show the transformed data. [4]

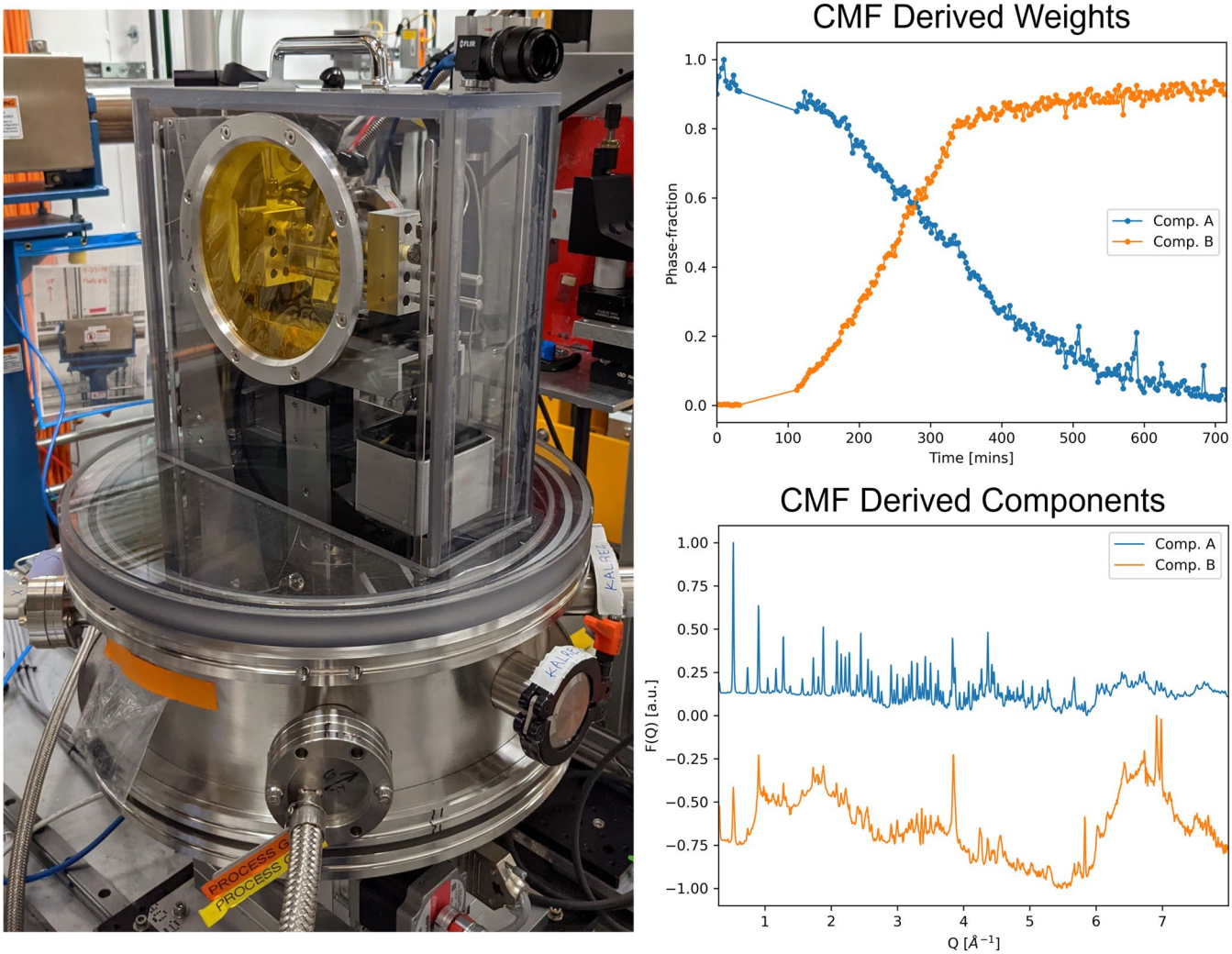

Figure 2: Metal organic frameworks (MOFs) are an important class of materials in gas storage and purification, heterogeneous catalysis, and as negative thermal expansion materials. However, characterizing their atomic-level structure can be difficult due to the co-existence of crystalline features, amorphized framework, weakly correlated pore guest molecules, and other defects. Researchers performed in-situ PDF studies of ZIF-8 to understand the origins of lost pore accessibility in the MOF during acid gas streams (e.g., flu gas, NOx) and the associated structural degradation mechanisms in a specialized sample enviornment on the PDF beamline (shown in photograph). However, the characteristic scattering signal of the degradation products was unknown a priori, making real-time analysis difficult. Constrained Matrix Factorization was used to monitor the state of the reaction, extracting components and weights in real time, ensuring all critical information obtained prior to completion of the study.

insights to researchers regardless of *a priori* knowledge. Non-negative matrix factorization (NMF) is an appealing class of methods for performing unsupervised learning on streaming spectral data. However, canonical NMF methods have no underlying requirement that the reconstruction uses components or weights that are representative of the true physical processes. By constraining a subset of the NMF weights or components as rigid priors, provided as known or assumed values, we can provide significant improvement in revealing true underlying phenomena in a method we refer to as Constrained Matrix Factorization (CMF) [5]. An example of CMF being applied to degradation of MOF materials is shown in Figure 2.

Classifying powder diffraction spectra

Another example of AI for analysis developed at NSLS-II is the crystallography companion agent (XCA). This is an open-source package initially developed for the feed forward classification of diffraction experiments [6]. The classification approach is pseudo-unsupervised, in that it does not require labeled data to be trained. Instead, XCA simulates a realistic dataset that encompasses the perturbations and physics of the measurement. Starting only from proposed crystalline phases, XCA overcomes the challenge of degenerate solutions by training an ensemble of Convulutional Neural Network (CNN) agents that can accurately predict phase existence

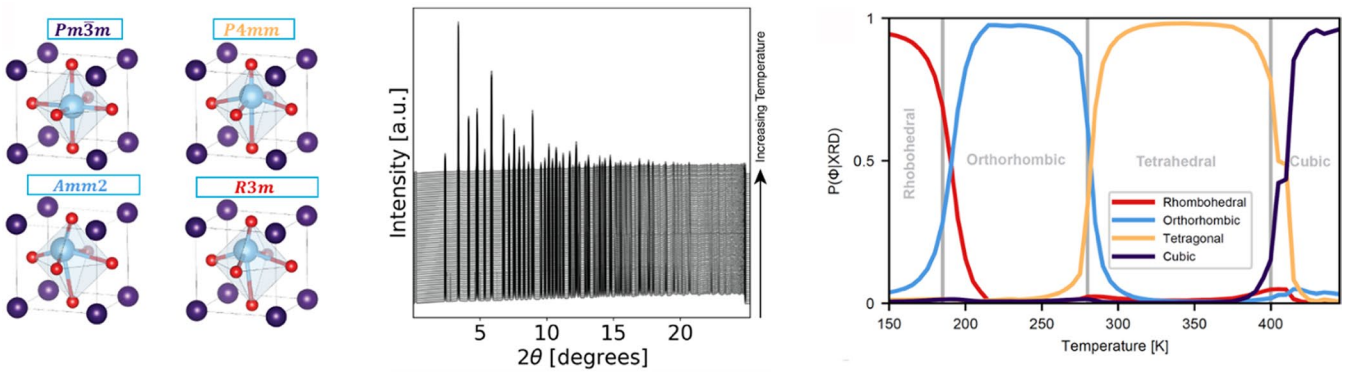

Figure 3: Phase transitions in $BaTiO_3$ involve symmetry breaking from a Ti translation (left). Plotting the diffraction patterns of the material over variable temperature measurement (center) shows the subtlety of the differences between the four phases. XCA rapidly produces a probabilistic temperature-dependent phase mapping of $BaTiO_3$ that is more accurate than current refinement techniques. Dotted lines show the expected transition temperatures, and each colored line corresponds to the probability of a given phase existing [6].

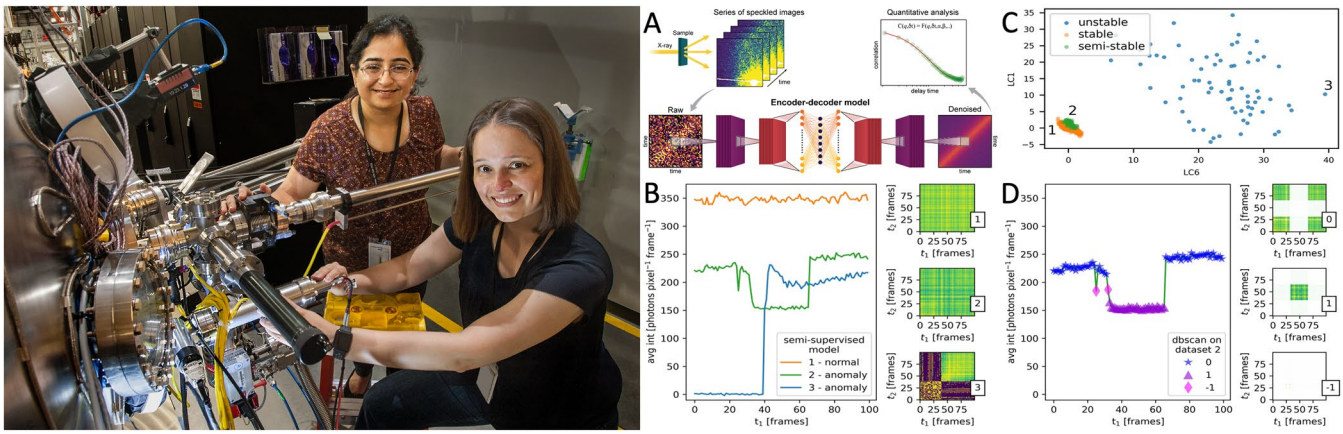

Figure 4: The CSX beamline with examples of AI/ML for XPCS analysis. (A) XPCS workflow with deep learning. (B) Application of semi-supervised outlier models (IFT, LOD, EE) from derived 1D timeseries (average intensity) to streamline review of automated C2 results [4]. (C) Identifying truly anomalous data with unsupervised clustering on latent space coordinates from CNN-ED for C2 denoising [8]. (D) Unsupervised clustering of derived 1D timeseries data (dataset 2) to dynamically mask/unmask pre-computed C2 results.

and uncertainty. The approach has been applied in searching organic porous materials for predicted phases, mapping a ternary alloy's phase diagram, and identifying the subtle phase changes in a ferroelectric material. To accompany the predictive model, XCA can also train a variational autoencoder (VAE), which provides insightful visualization of the experimental measurement [7]. The latent space of a VAE and the reconstruction error can be monitored as proxies for distributional drift, alerting the user when an experiment is beyond the scope of the simulated data used to train an agent. The combination of the Shannon entropy of the output prediction from the CNN ensemble and the reconstruction error of a VAE trained with the same data serves as an explainable AI for identifying outliers and distributional drift.

Automating XCPS analysis

X-ray photon correlation spectroscopy (XPCS) performed at the Coherent Soft X-ray Scattering (CSX) beamline is used to quantitatively characterize sample dynamics. This method is based on X-ray speckle intensity correlations within a timeseries of coherent X-ray scattering images, typically using an ensemble approach (Figure 4a) [9]. Robust, reliable real-time analysis is required to ensure maximal data and beamtime usage in both high-throughput and long duration experiments. Typical XPCS algorithms [1, 10] cannot distinguish sample dynamics and intensity changes induced by instrumental effects, adversely impacting result accuracy. AI/ML methods are ideally suited to aid users in real-time analysis of XPCS data because artifact-induced dynamics have recognizable patterns or signatures [1]. At NSLS-II, we

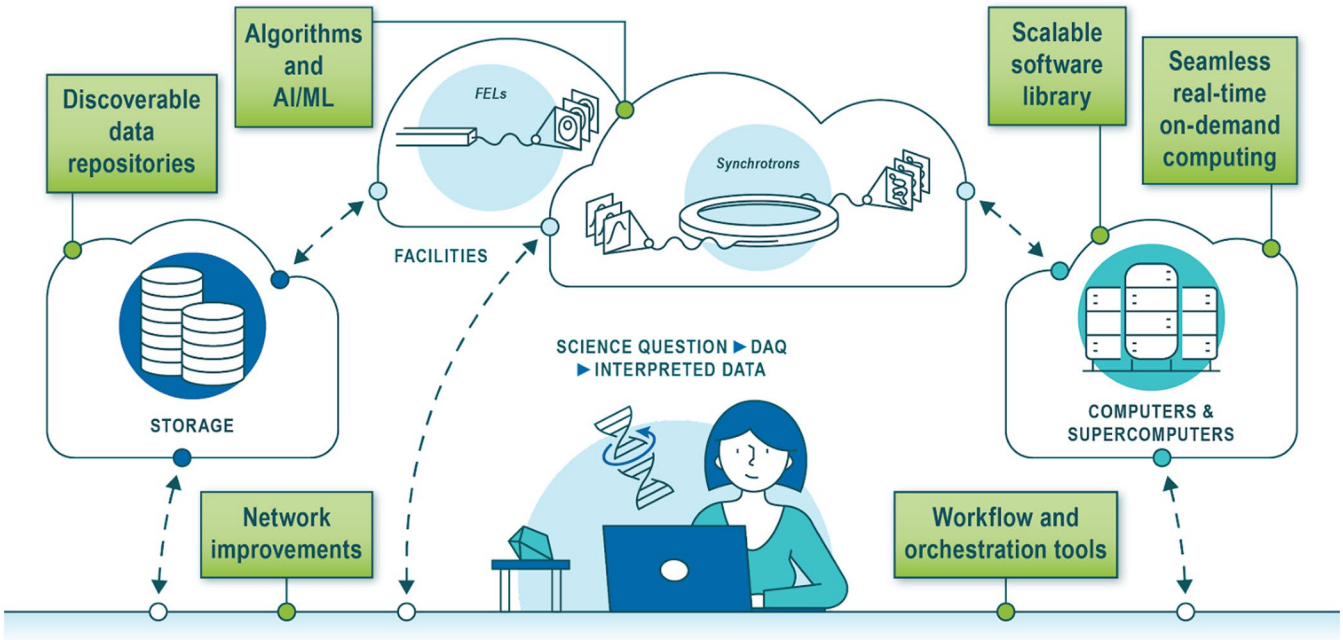

Figure 5: A complete suite of tools and capabilities in data infrastructure and computing needs to be created and developed in order to achieve our vision. In addition to experimental control and data acquisition, these tools and capabilities include algorithms and AI/ML, scalable software libraries, workflow and orchestration tools, seamless real-time on-demand computing, network improvements, and discoverable data repositories [14].

have developed models that are time, sample, and instrument agnostic using derived data and novel normalization schemes to detect artifacts, remove noise, and extract quantitative results with improved resolution for non-equilibrium dynamics [8].

The C2 calculation, one typical XPCS algorithm, is the most informative, but it is computationally expensive. AI/ML real-time monitoring of data quality (e.g., "good" or "bad") can be supported with limited computational power by operating on reduced 1D times series representations of the data. Recently developed semi-supervised outlier detection for normal XPCS data sets are being tested at NSLS-II, with the binary characterization tuned to potentially label good data as an "anomaly" [4]. Figure 4b illustrates usage of these outlier models on 3 XPCS datasets. These light and easy-to-use models are integrable into a variety of data collection and results reporting schemes, like immediate notification of potentially suspect data. However, human evaluation of data quality is more nuanced. Encoding rich information from C2s in a low-dimensional space using CNN-based denoising models opens new opportunities (Figure 4a). We plan to exploit the encoded representation of C2s for improved anomaly detection, classification, and other applications. Figure 4c corroborates that low-dimensional space representations of C2s align with an expert interpretation of each C2, which are a fully salvaged dataset 2 and far less reliable dataset 3 [10].

Finally, we show how one may employ unsupervised clustering of frames using the 1D time series. Grouping in this way leads to more efficient C2 computation, but it increases the number of datasets to investigate and makes integration with other AI/ML models more challenging since training data is often derived from "good" data with no missing values. However, unsupervised grouping methods remain part of the essential toolbox since data masking may be applied to the full C2 result (Figure 4d).

Adaptive and autonomous measurements

Axis of control requirements

A promising application of AI/ML methods at large user research facilities is in enabling autonomous, or "self-driving," experiments. Discovery and optimization of modern materials require an ability to efficiently probe vast, multi-dimensional material and processing parameter spaces, going beyond intuitive or exhaustive high-throughput approaches. Efficient design of adaptive feedback systems can be considered along three axes: coordination motif, source of signal, and required timescale of the loop. The coordination motif describes how the experiment and the algorithm interact. In some applications, the algorithm needs to be synchronous "in-the-loop" of the experiment, while in other applications we may need to asynchronously couple the algorithm to the experiment. An example of synchronous control is optimizing the plan of a 2D mapping experiment. Examples of asynchronous coupling include watching data quality, tuning a temperature ramp rate, suspending and resuming data acquisition based on beam conditions, or using multiple agents to collect and analyze data.

Source of signal describes the nature of the data being used in the feedback system. Often, the information that we want to feedback on can be classified as either engineering values or scientific values. Engineering values are raw machine measurements, such as motor positions or

shutter states, and do not require any knowledge of the sample under study to interpret and act upon. Science signals are very tightly coupled to the current experiment and their interpretation is often context dependent.

The final dimension of adaptive experiments is the required timescale for the feedback. For example, a control loop to maintain hardware alignment may need to run at sub-ms timescales whereas an algorithm to select the next sample to synthesize in a day-long reaction has a much larger time budget.

Each combination of communication motif, signal class, and timescale places considerations and restrictions and constraints on the sort of approaches that can be used for adaptive feedback. Selecting the right tools for a given task can be a critical consideration for successful deployment of such methods. Medium-speed, synchronous "in the loop" feedback on scientific values is the combination of these axes that is of the most interests to beamline operations, and we will now discuss two examples of this type.

Autonomous X-ray scattering experiments

At the Complex Materials Scattering (CMS) and Soft Matter Interfaces (SMI) beamlines at NSLS-II, a collaborative effort to enable intelligent material exploration during autonomous X-ray scattering experiments has been developed [11]. The developed system consists of three key elements: automated sample handling and data acquisition, based on Bluesky; real-time data processing; and decision-making algorithms, which take processed data as input and select the next experimental point to be measured. The algorithms, based on gpCAM, utilize Gaussian process regression in a Bayesian optimization framework, generating surrogate models and uncertainly distributions, using them to define an acquisition function for the given experimental objective, and locating its maximum to make the decision. This approach has been successfully applied to imaging heterogeneous materials and to materials parameter space exploration based on combinatorial sample libraries, combined with a robotic sample exchanger where appropriate to further expand the size of parameter space that can be explored in a given experiment.

Reinforcement learning driven high-throughput PDF measurements

An additional area of AI/ML being explored at the NSLS-II is the use of reinforcement learning (RL) being used to automate and optimize high-throughput data collection. Unlike supervised learning methods, where a very large, labeled set of data is employed to train, RL is trained via a dynamic interaction loop whereby the AI can learn "on-the-fly" as it is interacting with an environment. AI are trained with simulated experiments that recreate raw and processed data streams. It is important that the simulated environment encountered by the AI is sufficiently similar to the real-world data streams encountered at the beamline. Thus, in the initial deployment of the so-called "BadSeed" agent [12], trained to run high-throughput diffraction experiments on the PDF beamline, a data quality metric is employed, which is calculated independently from the AI. Thus, the decision-making logic learned by the RL agent is more broadly generalized to any scenario, which can output a similar formatted data quality metric [13]. By using this RL agent, the beamline was shown to be 100% more efficient at measuring critical structural information than standard iterative data collection plans.

Outlook and conclusions

Today, scientists at NSLS-II are using AI/ML methods to improve operations, expedite analysis, and facilitate discovery. Whether acting as digital beamline assistants or automating data processing, these tools are now in use on several beamlines around the facility. However, it is clear that we are only scratching the surface of what is possible, and additional investment is need.

To further these developments, NSLS-II has been working closely with the other four DOE light source facilities, coordinated through the 5-way Light Source Data and Computing Steering Committee. Our vision in this area is to establish a transformative computational fabric that covers the full lifecycle of the data generated at the DOE light sources, including theory, simulation, experiment design, data generation, data analysis, and publication. This will also connect 200+ instruments at light sources with a multi-tiered computing landscape, including edge computing, high-performance computing centers, data repositories, and cloud computing. This will best serve the 10,000+ DOE light source users per year.

Acknowledgments

The authors gratefully acknowledge the efforts of fellow staff and collaborators at Brookhaven National Laboratory, Oak Ridge National Laboratory, Lawrence Berkeley National Laboratory, and Purdue University.

Funding

This article highlights research that used the BMM, CHX, CMS, CSX, PDF, SMI, and SRX beamlines of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory (BNL) under Contract No. DE-SC0012704.

Some projects were supported by a BNL Laboratory Directed Research and Development (LDRD) projects 20-032 "Accelerating materials discovery with total scattering *via* machine learning" and 20-038 "Machine Learning for Real-Time Data Fidelity, Healing, and Analysis for Coherent X-ray Synchrotron Data." This research also used resources from the DOE project "Intelligent Acquisition and Reconstruction for Hyper-Spectral Tomography Systems (HyperCT)". ■

References

1. S. I. Campbell et al., *Mach. Learn: Sci. Technol.* **2** (1), 013001 (2021). doi:10.1088/2632-2153/abbd4e
2. BES User Facilities Data Management and Analysis Resource Needs, BES/ASCR Data Call, 2019.
3. D. Allan et al., *Synchrotron. Radiat. News* **32** (3), 19 (2019). doi:10.1080/08940886.2019.1608121
4. P. M. Maffettone et al., *Digit. Discov.* **1** (4), 413 (2022). doi:10.1039/D2DD00014H

5. P. M. Maffettone et al., *Appl. Phys. Rev.* **8** (4), 041410 (2021). doi:10.1063/5.0052859
6. P. M. Maffettone et al., *Nat. Comput. Sci.* **1** (4), 290 (2021). doi:10.1038/s43588-021-00059-2
7. L. Banko et al., *NPJ Comput. Mater.* **7** (1), 1 (2021). doi:10.1038/s41524-021-00575-9
8. T. Konstantinova, L. Wiegart, M. Rakitin, A. M. DeGennaro, A. M. Barbour, *Phys. Rev. Res.* (in press).
9. O. G. Shpyrko, *J. Synchrotron Rad.* **21** (5), 1057 (2014). doi:10.1107/S1600577514018232
10. A. Fluerasu et. al., *Phys. Rev. E.* **76** (1), 1 (2007). doi:10.1103/PhysRevE.76.010401
11. M. M. Noack et al., *Nat. Rev. Phys.* **3** (10), 685 (2021). doi:10.1038/s42254-021-00345-y
12. P. M. Maffettone et al., *Mach. Learn: Sci. Technol.* **2** (2), 025025 (2021). doi:10.1088/2632-2153/abc9fc
13. D. Olds et al., *3rd Annual Workshop on Extreme-Scale Experiment-in-the-Loop Computing (XLOOP)* (IEEE, New York, NY, USA 2021), pp. 36–42.
14. NSLS-II Strategic Plan, 2022.