Machine Learning: Science and Technology

PERSPECTIVE • OPEN ACCESS

Outlook for artificial intelligence and machine learning at the NSLS-II

To cite this article: Stuart I Campbell *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 013001

View the article online for updates and enhancements.

You may also like

- The Latest Status of NSLS-II Insertion Devices
Toshi Tanabe, Charles Kitegi, Ping He et al.

- A NIST facility for resonant soft x-ray scattering measuring nano-scale soft matter structure at NSLS-II
Eliot Gann, Thomas Crofts, Glenn Holland et al.

- Lossless crossing of a resonance stopband during tune modulation by synchrotron oscillations
G M Wang, T Shaftan, V Smaluk et al.

This content was downloaded from IP address 68.196.193.203 on 30/01/2022 at 22:28

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
30 June 2020

REVISED
3 September 2020

ACCEPTED FOR PUBLICATION
1 October 2020

PUBLISHED
31 March 2021

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PERSPECTIVE

Outlook for artificial intelligence and machine learning at the NSLS-II

Stuart I Campbell, Daniel B Allan, Andi M Barbour, Daniel Olds, Maksim S Rakitin, Reid Smith and Stuart B Wilkins

National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY 11973, United States of America

E-mail: scampbell@bnl.gov

Keywords: synchrotron x-ray, scientific Python, large scale facilities, strategy

Abstract

We describe the current and future plans for using artificial intelligence and machine learning (AI/ML) methods at the National Synchrotron Light Source II (NSLS-II), a scientific user facility at the Brookhaven National Laboratory. We discuss the opportunity for using the AI/ML tools and techniques developed in the data and computational science areas to greatly improve the scientific output of large scale experimental user facilities. We describe our current and future plans in areas including from detecting and recovering from faults, optimizing the source and instrument configurations, streamlining the pipeline from measurement to insight, through data acquisition, processing, analysis. The overall strategy and direction of the NSLS-II facility in relation to AI/ML is presented.

1. Introduction

The National Synchrotron Light Source II (NSLS-II) [1] (figure 1), is the newest light source in the US Department of Energy (DOE) complex delivering unprecedented brightness to advanced beamlines, employing the latest detectors and beamline instrumentation. A schematic plan view of the facility can be seen in figure 2. The source brightness coupled with advanced multidimensional detector technology leads to experiments being performed (a) much faster and with a higher throughput than ever before, (b) with higher resolution for both imaging and spectroscopy, (c) with an unprecedented signal to noise ratio thereby enabling studies of previously unobservable signals. In addition to this, NSLS-II has demonstrated that this all can be done under incredibly stable and reliable operating conditions.

Artificial intelligence and machine learning (AI/ML) is a key emerging technology that will allow us to harness the brightness of the source. It will enable us to perform experiments:

- More efficiently and intelligently with science-knowledge informed decision making.
- With optimal setup using automatic alignment.
- With clever monitoring and fault detection.
- With direct feedback from incredibly fast physics-based simulations.
- More efficiently and with higher sensitivity with intelligent systematic error reduction.
- Under incredibly stable operating conditions with real-time feedback to the experimental control and end-user.

Modern synchrotron lightsource x-ray beamlines are highly complex and sensitive, machines. They utilize some of the most advanced and cutting edge technology available to focus concentrated beams of light with a brightness greater than the Sun onto a point a million times smaller than the head of a pin. Synchrotron beamtime is an incredibly valuable resource. For a user to obtain beamtime, they must request time through a highly competitive, peer-reviewed proposal. Each synchrotron beamline at the NSLS-II is highly oversubscribed, that is many more experiments are requested than performed. It is therefore in both the users and facilities best interest to maximize the throughput and minimize the downtime of every beamline in order to achieve the highest scientific output. Like all complex machines—sometimes—things break. When

© 2021 The Author(s). Published by IOP Publishing Ltd

Figure 1. Aerial photograph of the National Synchrotron Light Source II situated at Brookhaven National Laboratory. (Courtesy of Brookhaven National Laboratory).

they do, it can be a non-trivial task requiring expert scientists and engineers with years of experience to diagnose, fix, and return the experiment to optimal performance, often under considerable time constraints.

Due to careful design and engineering, most breakdowns are only short-lived. The actual recovery of the operating conditions may not take more than a few minutes once the problem is diagnosed, such as an component realignment, a motor which has lost it is position, a reset of the control system, or just a sample that has fallen out of alignment. However, that lost time can be extremely detrimental to the experiment at hand, particularly if it occurs during some critical stage of an in-situ or time-dependent measurement. In addition, when breakdowns occur out of office hours there is the human cost of the 'call-in', often requiring an expert to physically come to the facility. If this diagnostic knowledge can be encoded into an AI/ML process, commonly encountered failure modes can be autonomously diagnosed and perhaps even remedied with only very minor interruption to the experiment at hand.

Critical to developing such approaches is the availability of large labeled training-sets of data, which capture the observable of any labeled failure condition. These observable should include everything measurable, from the obvious (telemetry data on motor positions, detector images, beam position monitors) to the seemingly unimportant (network traffic levels, local weather) to make a truly extensive data set spanning both normal and well-understood failure modes of operation. It is the generation of these labeled datasets that presents the greatest challenge to seeing this capability realized.

To this end, there are ongoing projects at the NSLS-II to develop and standardize far-reaching observable-capture tools that are readily deployed across NSLS-II. These tools are meant to greatly automate the tasks required, so that a wide view of observable states can be catalogued in a unified database. Thus, even normal operations will continuously contribute to the total training dataset, with minimal additional staff effort required to catalog system failures. Through this task, we hope to greatly lower the barrier to entry towards development of these tools across the whole facility.

2. Enabling technologies

Applying AI/ML effectively at light sources requires managing the large volumes of scientific data generated at the facility. In the past, sychrotron data has been captured and stored using a wide variety of disparate systems that evolved independent at their respective instruments. This variety makes large-scale systematic

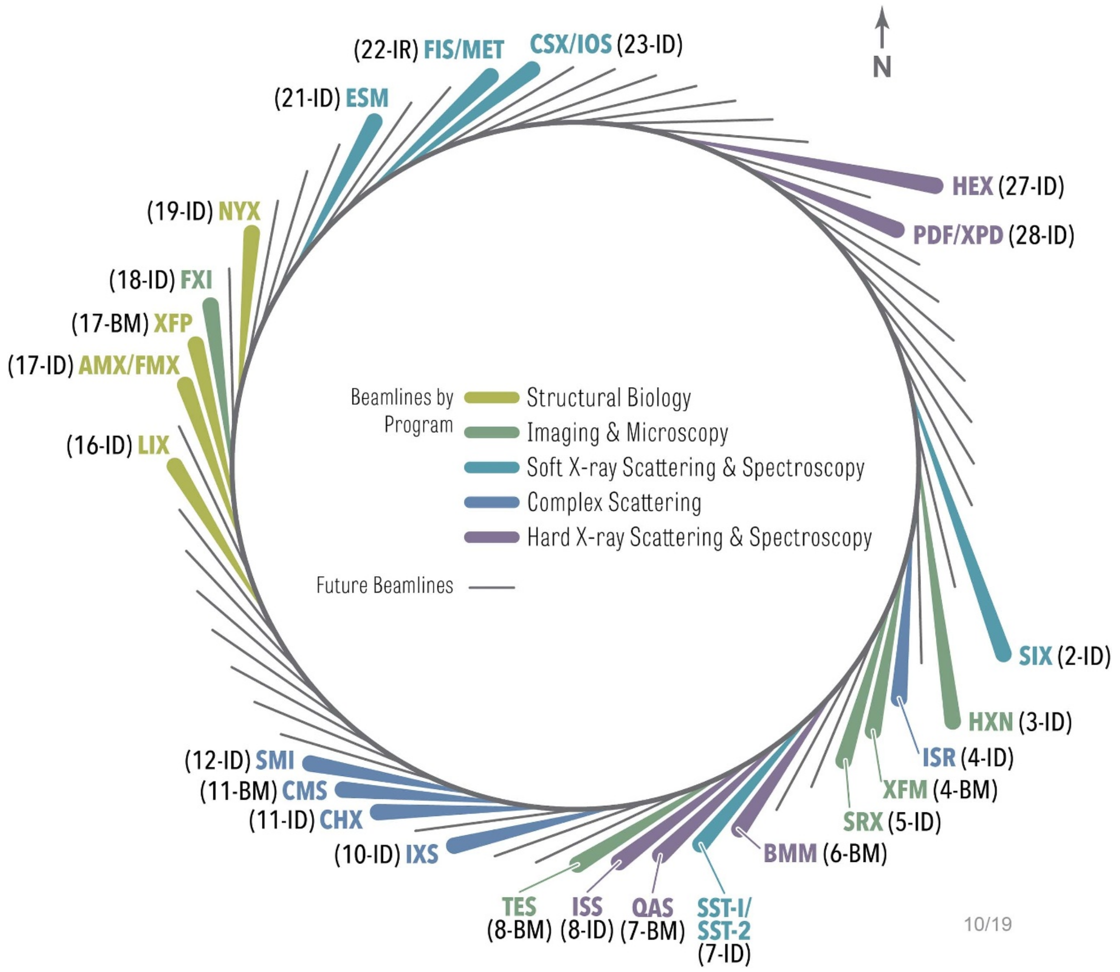

Figure 2. Layout of the individual beamlines at the NSLS-II, indicating the different scientific areas covered. (Courtesy of Brookhaven National Laboratory).

studies prohibitively labor-intensive at best. Through the Bluesky Project [2, 3], working across instruments and facilities, we have built a system that provides a shared, unified core of scientific data acquisition, management, and access software, while leaving room for the necessary variety between instruments, facilities and techniques.

Beginning with data acquisition, the raw measurements must be captured along with sufficient physical context and what the intent of the experimenter was. In the past, this context has often stored in an ad hoc fashion—in paper notes, cryptic file names, or even mental recollections. To operate at the scale where AI/ML models perform well, everything must be captured in a predictable, machine-readable fashion. This includes primary measurements (e.g. images or spectra), secondary measurements (e.g. motor positions, beam current, sample temperature), hardware configuration (e.g. exposure time), and physical details of the hardware (e.g. pixel size). Metadata about sample composition and preparation is also necessary to contextual measurements. There is no single metadata schema or ontology that can suit the breadth of synchrotron research, but we can enable scientific domains, instruments, and groups can define and enforce well-suited schemas, in order to make this problem manageable. In addition, we can perform prompt data quality checking, first-pass processing, and tagging, to ensure that data is well-labeled at acquisition time, such that all the relevant context is captured in machine-readable form before the experiment has concluded.

Next, the data must be made accessible in a uniform way. We contend that it is not possible to standardize on one data *format* because of the range of access patterns and performance requirements across techniques and because of the constraints imposed by hardware vendors. However, it is possible to place data behind a common programmatic interface (API) upon which shared user-facing tools can be built for searching and accessing scientific data for visualization, analysis, or egress to other systems.

Finally, it is clear that a reliable and flexible IT infrastructure is essential in order to actually deploy and use any of these tools. The scale and the type of compute (e.g. CPU, GPU, FPGA) or data storage (e.g. high performance parallel file-systems, cloud base object stores) should be chosen with the particular use case in mind. For a certain number of AI/ML problems, such as supervised learning, the high computational resources are only needed during training before the experiment and then only relatively modest or optimized compute is required during the actual experiment. At the NSLS-II the aim is to ensure a consistent yet flexible tiered infrastructure that ranges from local dedicated hardware for particular beamlines or experiments (e.g. edge computing resources), through facility and laboratory central capabilities (e.g. institutional clusters), onto larger resources such as the commercial cloud or high performance computing facilities. The use cases for the light sources are being included when defining the requirements for the next generation exascale computing facilities.

These shared tools for search, visualization, and analysis will provide a unified user experience across facilities. Through this effort, we will better fulfill our mission by expanding the service we provide to our users further into data analysis. Data management and computation will become a problem owned by the facilities, not their users, and in this way the facilities will expand their reach into user communities less fluent in the minutia of synchrotron data acquisition or analysis.

3. Dealing with faults

In recent years NSLS-II has been able to run with greater than 96% reliability. In a year with 5000 operating hours, this translates to 150–200 h of downtime during scheduled operations. While it meets our performance goal and is competitive among other synchrotron light sources globally, these downtime hours still can have large impacts on experimenters—especially those with complex experimental setups or using the beam in a narrow time-window. Keeping reliability in that range also leaves a very small margin for downtime—especially for long failures. An AI/ML implementation in accelerator operations [4–6] would reduce unexpected interruptions for users by focusing on two specific goals: forecasting and reducing the number of beam-dumps and reducing downtime from diagnosis and recovery.

Forecasting and reducing the number of beam-dumps: The first goal is to identify any subsystem degradation during operations that would likely lead to a beam dump. In our experience, these include gradual drifts in temperature, water flow, vacuum pressure, power-supply output or ground-current, beam position, and other changes that result in an Equipment Protection System trip and are slow enough to act on prior to reaching the trip limit. These systems contain thousands of signals that rely heavily on experts to monitor subsystem performance, observe trends, and dictate maintenance activities. We display and monitor drift-values for some of these slow-moving parameters, and it does give us more lead-time to plan preventive maintenance and avoid operation interruption. Moving further in that direction with machine-learning (ML) would let us identify preventable trips (much better than with threshold alarms). Moreover, when the subsystems enter states previously associated with potential impending faults, we may broadcast cautionary messages to users such that experimental plans could be adjusted accordingly. Looking further out, an ML algorithm could give time estimates of systems or components whose parameters are slowly-changing, and be used to schedule preventative maintenance tasks.

Reduce downtime from diagnosis and recovery: Quick failures such as RF cavity arcs, power supply trips, or controls-network failures often occur instantly and cannot necessarily be predicted or prevented. The second purpose of an operations neural network deals with un-prevented faults by identifying the source of the fault. Reducing time and effort spent diagnosing a problem leads to faster repairs and recoveries, less cool-down of beamline equipment, and less time lost. This is significant because, in cases of non-standard faults, identifying the system or piece of equipment at-fault can take longer and contribute more Downtime than the repair itself. In FY18, the NSLS-II Mean Time To Recovery was 1.5 h (which includes diagnosing failure sources, devices' repair, setting the machine back to operation state), yet 57% of Downtime came from only nine dumps (of 326 faults). In the most extreme cases we lost more than 24 h to diagnosing and repairing power-supplies, which is 24 h that users are stuck waiting. An algorithm using previous trip-conditions to pinpoint the potential cause(s) of a beam dump can drastically help operator and expert to focus on limited devices and reduce downtime spent diagnosing problems.

Data methods: The NSLS-II machine archive is comprehensive, and tracks 106 000 individual variables at their own update speeds (up to 10 kHz). We store nearly 47 TB/year of archive data. Our fault-report archive is 1400 entries in a database, occupies about 1 MB. At present, we are developing two types of ML algorithm that will work in tandem. The first method uses grouping techniques of individual device histories, to identify & flag 'abnormality' of device-values compared with 'good' Operations (defined by sustained nominal beam current). The second method uses Fault history to pinpoint moments at/before known faults and manually identify to the algorithm what machine parameters do during specific failure types. Combining the techniques should enable associations and pattern-identification beyond what an Operator or system expert would notice by looking at graphs or setting single-variable alarms.

Stability forecast: Finally, we see the ability to extend these 'Fault prediction' and 'abnormality detection' tools beyond Storage Ring trips. By using AI/ML to identify when injector conditions are drifting, Operations staff could alert beamline staff when conditions may be unstable, even beam being present. Years of traceable parameters like electron-gun temperature, injection/extraction kicker magnet stability, and power spectrum distributions will all be trained into a 'stability prediction'. For more sensitive beamlines, knowing when fluctuations in electron beam current, x-ray flux, or x-ray beam position could greatly assist in avoiding collecting unusable data.

4. Alignment

All of the Light Source user facilities have a common problem, the lack of automated means to quickly and accurately align and optimize optical components of x-ray beamlines to achieve a photon beam with the desired characteristics. Manual alignment/optimization of a beamline can only be reliably performed with a small number of (preferably independent) degrees of freedom, which rarely results in the globally optimal configuration. In addition, such manual optimization is hard to perform when the system varies as a function of time, for example when the electron beam changes, requiring constant readjustment of the alignment. With the increased degree of sophistication of modern beamlines, incorporating dozens of equipment units (mirrors, transfocators, monochromators, slits, stages, etc), it is crucial to establish procedures to perform on-line beamline optimization in a semi- or fully-automated fashion, to expedite beamline preparation and increase scientific output by dramatically reducing beamline tuning-time and downtime, and potentially achieve better results of global optimization.

The Bluesky data collection software is used extensively both at the NSLS-II and has seen adoption at other User Facilities worldwide, such as the Linac Coherent Light Source located at SLAC National Accelerator Laboratory, the advanced photon source (APS) located at Argonne National Laboratory, and the Australian Synchrotron. It has also some adoption and testing at both the advanced light source (ALS) located at Lawrence Berkeley National Laboratory and the Stanford Synchrotron Radiation Lightsource located at SLAC National Accelerator Laboratory. The bluesky library handles experiment control and scientific data collection, allowing to control an instrument irrespective of its software/hardware configuration. It uses a concept of plans, which define the experimental procedures to run. The plans allow the implementation of complex procedures in a way convenient for the scientist. Users can either use predefined plans, or custom plans may be designed if the pre-existing plans do not satisfy the experimental requirements. The RunEngine is a core part of the bluesky library, which passes instructions from the plans to the lower hardware levels (e.g. to move motors and/or read detectors) using the ophyd hardware abstraction library, and collects experiment information during a run into documents. The documents can be configured for storing in a database, and/or can be used for live visualization/analysis of the data being collected. The databroker library, a part of the Bluesky project, provides convenient access methods to the documents containing data and metadata from experimental measurements and operate on them in the form of standard scientific Python data structures (Pandas dataframes, NumPy arrays, etc). The relationships between the libraries are depicted in the figure 3.

A long-term goal is to extend this software infrastructure to the simulation tools available at DOE National Laboratories, like Sirepo [7], OASYS [8], and LUME [9], since it proved to be flexible and extendable. A flexible optimizer for beam intensity based on a differential evolutionary (DE) algorithm has been implemented. This optimizer operated on three backends: the Tender Energy X-Ray Absorption Spectroscopy (TES) beamline at the NSLS-II [10, 11], a set of simulated EPICS [12] motors provided via a Docker container [13], and the synchrotron radiation workshop (SRW) simulations via the Sirepo framework [7] (see figure 4 for the TES virtual beamline representation in Sirepo).

The system could be used to optimize an entire beamline or just a specific part, like a nanofocusing KB mirror or a zoom optics system. The optimization plan requires information about the motors to be optimized and their bounds, the detector to use, and the type of a scanning procedure to use. The TES beamline has already been used as a model to create a prototype code that used simulated EPICS motors and Sirepo simulations.

Simulated motors can be used to test code without using a real beamline. DE optimization using the Docker-run simulated motors proved to work in a very similar fashion to the optimization of a real beamline. Both the real motors and the simulated motors are operated using the EPICS control system protocol, so the motors act very similarly and can be easily swapped. Since the simulated motors did not have a detector to read, it is necessary to create a computed signal to simulate detector intensity. This methodology has already been successfully tested on the TES beamline layout, by modeling three motors and a detector signal satisfying the expected distribution based on the motor positions.

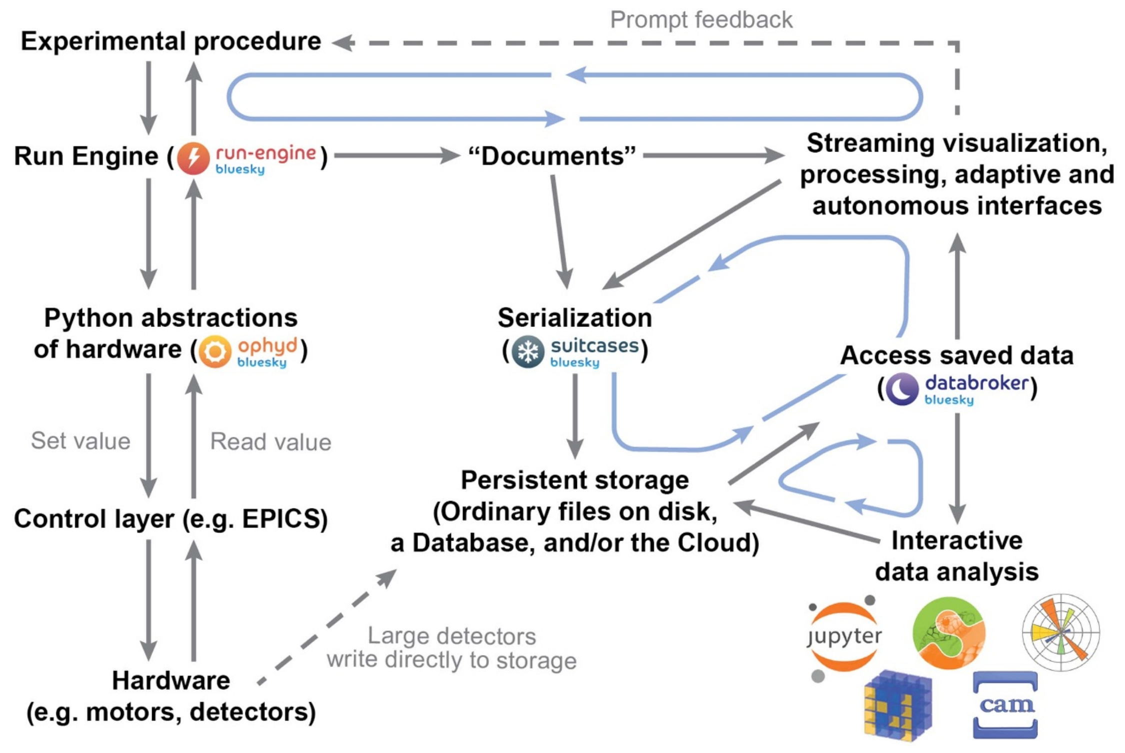
Prompt feedback
Experimental procedure
Run Engine (run-engine bluesky)
"Documents"
Streaming visualization, processing, adaptive and autonomous interfaces
Python abstractions of hardware (ophyd bluesky)
Serialization (suitcases bluesky)
Access saved data (databroker bluesky)
Set value
Read value
Control layer (e.g. EPICS)
Persistent storage (Ordinary files on disk, a Database, and/or the Cloud)
Interactive data analysis
Large detectors write directly to storage
Hardware (e.g. motors, detectors)
jupyter
cam

Figure 3. The relationships between the software libraries in Bluesky. Each component is individually useful, and they operate cohesively to leverage scientific Python for data acquisition and analysis. (Courtesy of Brookhaven National Laboratory).

The methodology has been proved to be effective by creating the Sirepo-Bluesky interface and tested. Sirepo, combined with the SRW code, can simulate beamline configurations and is the third backend that can be optimized. The Sirepo-Bluesky interface [14, 15] was implemented to programmatically change multiple parameters of simulations and submit them to a local or remote Sirepo server. The integration was done in such a way that the results from simulations were recorded and made available using the databroker interface. A Sirepo model of the TES beamline was created and optimized using the enhanced DE algorithm. That approach provided an avenue for the ML and deep learning (DL) based optimization.

The project was based on previous developments of an enhanced genetic algorithm and a DE algorithm to accomplish the automatic beamline optimization [16, 17]. Despite the advantages of evolutionary algorithms, the optimization speed is still insufficient (at the level of approximately 10 min), and optimization is difficult because of multiple sources of noise (mechanical backlash, building temperature, etc). The ML and DL models can make the optimization faster, linearize the behavior of adaptive optics, and increase reliability for automated alignment. The benefits of that approach were demonstrated in a recent paper [18], where a deep artificial neural network was used to produce solutions for the planar chaotic three-body problem orders of magnitude faster based on the existing training and validation data set. This type of approach to beamline optimization allows the production of training data sets for ML/DL models automatically by using the DE method. Alternatively, realistic simulated models of x-ray beamlines can provide enough training data for ML/DL methods. Such models and the corresponding results were demonstrated in recent publications (e.g. [19, 20]), which employed a modern beamline simulation framework Sirepo with the SRW backend [7].

We believe the most appropriate ML method for beamline optimization is deep reinforcement learning (DRL). This is the method used to train models that learn to choose actions to reach a goal, such as playing video games [21], playing board games [22], and developing robot locomotion [23]. New DRL algorithms are under intense research by the ML community, but we expect to use well-known algorithms such as *PPO* [24] and *A3C* [25]. These have been demonstrated to work on non-trivial problems and a body of practical knowledge for training with these algorithms has accumulated in journal articles, technical reports, and books, and with computational scientists at NSLS-II, ALS, and APS. Code for these methods is available from open-source projects such as *Tensorforce* [26] and *rlpyt* [27], based on *TensorFlow* and *PyTorch* respectively. These choices will restrict the project to developing the game-like training process, specifically defining a reward scheme to train models to efficiently optimize a beamline.

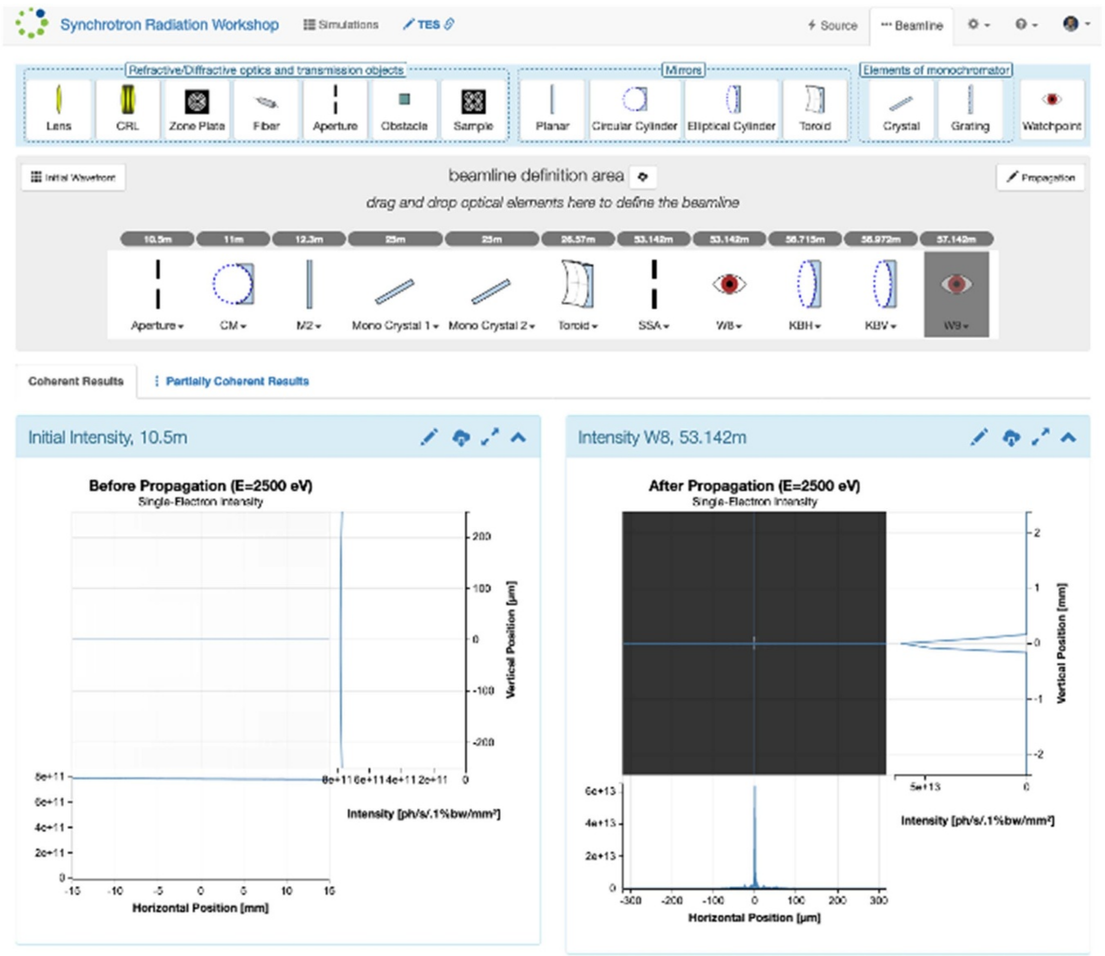

Figure 4. NSLS-II TES virtual beamline representation in Sirepo.

5. Data fidelity

X-ray facilities are striving to produce increasingly coherent light because coherent x-ray scattering techniques produce data with unique scientific insight. X-ray photon correlation spectroscopy (XPCS)[28, 29] is one such technique, which is primarily used to understand dynamics in scientific samples by analyzing 2D images that were recorded at a regular time interval using signal processing methods. Beamlines performing XPCS experiments have a very high requirement for optical stability because it is impossible from XPCS alone to ascertain if the derived dynamics are inherent to the sample or are induced by an artifact. Currently, expert researchers invest time and effort to manually inspect data to assess its quality, or more aptly, ensure that data are of high fidelity. Once the data's fidelity is known to be acceptable, then the researchers become sufficiently confident to extract quantitative dynamics and continue the experiment. Given that x-ray facilities are delivering brighter sources and advances in detector technology permit 11.8 kHz data production of multi-speckled x-ray patterns [30], the data fidelity can no longer be reliably or practically assessed by a human-in-the-loop decision-making process. Clearly, this represents one of the most serious bottlenecks to scientific progress. Fortunately, advances in data science and AI/ML make it possible to mine large quantities of data and discover underlying patterns that enable automated evaluation of data quality and predictive analysis without requiring humans eyes to evaluate individual pieces of data. Perhaps more interesting from an AI/ML perspective is evaluating if known AI/ML methods can be applied to time-series data like XPCS data.

We justify the applicability of AI/ML to XPCS data reduction with the analysis of a model dataset collected at the CSX beamline of NSLS-II [31]. The dataset's original purpose was to provide a better understanding of experimental results from a $La_{1.875}Ba_{0.125}CuO_4$ single crystal and increase confidence in the XPCS analysis of the associated standard time-series measurement. The researchers published the subsequent results in October 2016, approximately 1.5 years after the first dataset was collected and

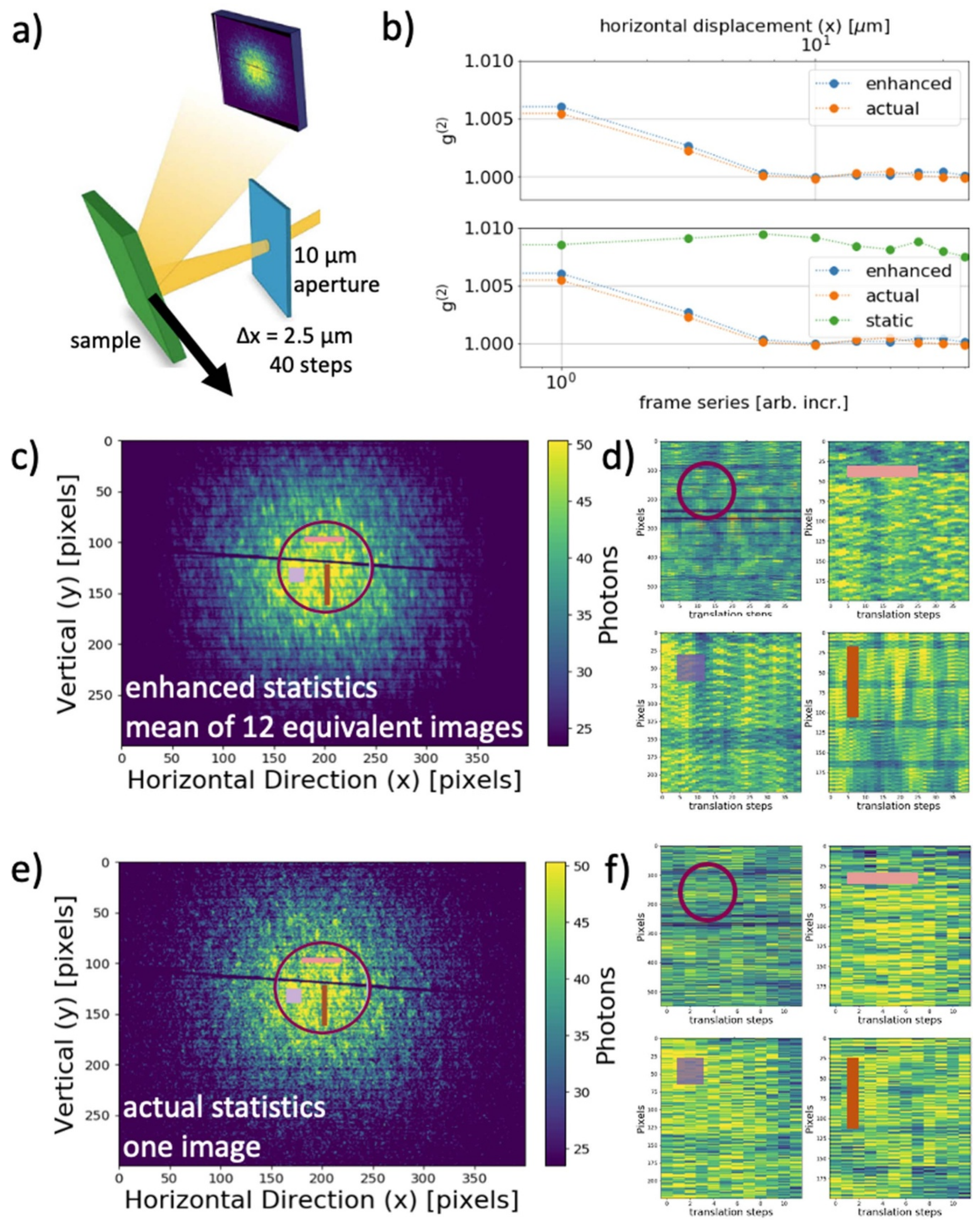

Figure 5. Model experiment to illustrate XPCS's lack of dynamics specificity. (a) Illustration of model experiment with forty 2.5µm horizontal sample displacement displacements. Twelve 4 s exposures were collected at each step. (b) Standard one-time auto-correlation for data represented in c-f accompanied by the artifact-free analysis on the first 12 images (fixed position). (c) Average speckle pattern of the first 12 images (enhanced statistics) with 4 ROIs (overlays). (d) Speckleograms on the data represented in c; inset denotes the ROI. (e) Speckle pattern of first image (actual statistics) with 4 ROIs (overlays). (f) Speckleograms on the data represented in e; inset denotes the ROI.

demonstrated the presence of static domains [32–34][1]. In regards to the model experiment described by figure 5(a), it represents a single dataset consisting of 41 equivalent points. The most basic analysis in XPCS is applying the one-time auto-correlation function

$$g^{(2)}(q,t) = \frac{\langle I(q,t)I(q,t+\delta t)\rangle}{\langle I(q)\rangle^2}, \tag{1}$$

[1] The findings of this work support the temporal stability of the charge density wave found in $La_{1.875}Ba_{0.125}CuO_4$ and that the interplay between the charge and lattice degrees of freedom maybe be important to understanding superconductivity in cuprates.

where q represents a collection of pixels (i.e. multiple signals), each with intensity, I. However, in this model example, we perform the measurement and analysis as a function of sample displacement (x) instead of time (t) at a fixed interval of δx (δt). As figure 5(a) depicts, the sample was translated beneath the x-ray beam spot (10µm diameter) in order to observe 'dynamics' during the measurement, albeit purposely induced dynamics. The standard auto-correlation analysis (equation (1)) for XPCS is shown in figure 5(b). However, figures 5(c)–(d) represent 3.5 times better data compared the observed statistics in the published findings, which is illustrated by figure 5(e)–(f). [32] As expected in the presented model, the auto-correlation analysis shows that the speckles' intensity fully de-correlate after ≈10µm total displacement, which is when an entirely different area of the sample is illuminated. Any naive, automatic analysis without user-in-the-loop evaluation does not reveal the source of the dynamics, which is in fact induced by a moving sample.

There are methods to establish the data's level of fidelity with respect to artifacts (or correlated noise). Given that a user has no *a priori* knowledge of the sample's dynamics, it is clear that the choice of selecting an ensemble of pixels or a region of interest (ROI) is important. First, the ROI size and shape can affect the analysis tools used to further evaluate the data. Speckleograms (often referred to as waterfall plots or kymographs) are one such tool that shows how a given pixel intensity evolves, and four speckleograms of equal pixel population are shown in figure 5(c). The square and horizontal rectangular ROIs most clearly demonstrate an isotropic, repetitive speckle evolution with distance, or generically with time. Such an observation is indicative of sample motion, which can happen accidentally or as a consequence of temperature control or energy injection. We can further illustrate how the changes are subtle and not immediately recognized by expert researchers. Consider that the results of Chen et al are based on a time series of a fixed sample position with the same photon counting statistics in figure 5(e) and b. Additionally, the measured time window may be similar to the extracted dynamics (figure 5(f)), which is especially true in the case of slow dynamics in solid materials [35, 36]. Not to mention that datasets consist of more frames by a factor of 100-500. To this end, real data are messy, and they often require an experienced eye or much effort to understand if the data should be analyzed as is, especially when real and artificial dynamics are present. Therefore, we plan to implement AI/ML methods to automate data fidelity decisions and guide users in understanding the fidelity of the data and its subsequent analysis. Furthermore, with two beamlines like CSX [31] and CHX [37] and tools like SRW for *in silico* beamlines (see section 4), it is possible to develop a data-driven approach with AI/ML to identify the signature of specific artifacts that affect the XPCS results [19, 20]. Armed with the knowledge of an artifact's root cause, it may be possible to account for the artifact using AI/ML advancements in neural networks and extract a meaningful result, thus realizing a fully AI/ML guided XPCS analysis.

We recognize the that our data-driven approach with labeled training data will require time and effort to develop, but that does not address two problems: (1) automated analysis pipelines are needed now (2) we require a robust method to firstly broadly classify collected data as 'good' or 'bad'. Depending on the nature of the convoluted artifact and inherent sample dynamics, 'bad' XPCS data may be salvaged with user-defined parameters. Figure 6 illustrates this issue in the field at the CSX beamline. Figures 6(a) and (b) shows the data and standard auto-correlation analysis associated with anti-ferromagnetic ordering in a complex nickel oxide [38]. The first pass analysis includes all data (every frame), and there are clear oscillations accompanied by what may be 'very slow' dynamics of some origin. It is possible to observe oscillations in the one-time auto-correlation for real scientific samples and one may derive more descriptive equations to extract meaningful physical parameters. [39–42] However, it is also possible that the first pass analysis is not appropriate. Upon further inspection, the diffraction peak's center of mass is shifting in time (figures 6(c) and (d)). While this could be inherent to the sample's magnetic ordering, further investigation of data archived [43] in parallel with the bluesky [2, 3] data acquisition, it is clear there is a correlation between the peak's center of mass and the pressure of the deionized water circuit that serves one-fifth of the accelerator's ring and possible beamlines. We can attempt to salvage the data by removing outliers, and in this case, we only considered data near the statistical mode (within 0.6–1.0 pixels). The results of eliminating the outlier (or bad) frames from the analysis are presented in figures 6(a) and (c)–(f).

Since culling data can sometimes salvage XPCS measurements, automation of outlier detection is an attractive and time saving proposition; however, simple algorithmic implementations often require tuning for the observed signal to noise, potentially eliminate useful data, and do not often consider the different measurement geometries and scattering patterns. In obvious cases of sparse outliers, standard AI/ML techniques anomaly detection can salvage data. We want to employ these and related techniques in unsupervised and semi-supervised ML to supplement user data reduction and analysis. If successful, we will use these tools to support verification and validation of a user labeled data repository to support our training data-driven approach to XPCS artifact identification and artifact removal without eliminating data. However, it is important to stress that by eliminating many points in time, we loose the ability to observe fast dynamics. Therefore both AI/ML approaches we have outlined here are necessary.

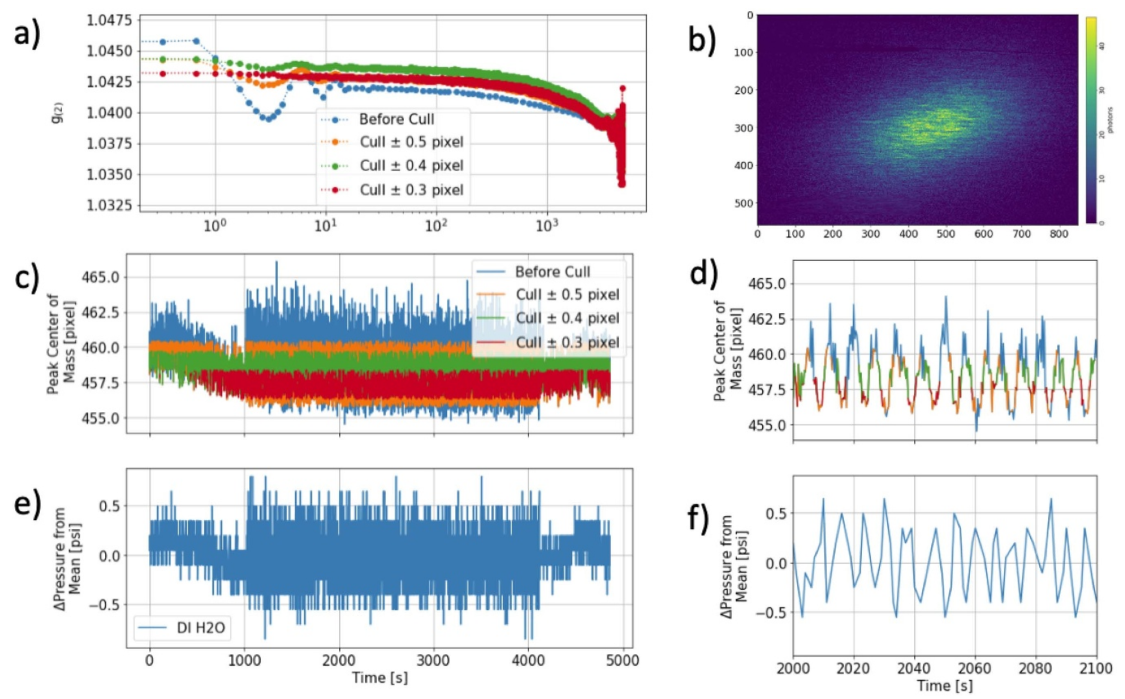

Figure 6. Data collected during user operations at the CSX beamline. (a) One-time auto-correlation of raw and culled based on a manual outlier identification process. (b) Diffraction signal from anti-ferromagnetic ordering in a complex nickel oxide. (c,d) Diffraction peak's horizontal center of mass as a function of time (or image). (e,f) Likely source of beamline instability (optics cooling water pressure), which was archived during the collection time of the data discussed in a–d.

To summarize, applying AI/ML to data collected during XPCS experiments cannot only fast track the discovery and correction of measurement instabilities, but it will also reduce user-in-the-loop decision making. As a result, users will be able to concentrate on the experiment at hand and have time to think about how to best utilize the allocated time. From a facility perspective, we plan to develop these tools to work within the NSLS-II's Data Acquisition, Management and Analysis infrastructure at the CSX and CHX beamlines. These beamlines utilize different detector technologies and specialize in different scientific areas. We will, therefore, have a diverse breadth of scientific data. Additionally, our customized AI/ML tools for XCPS will be compatible with other NSLS-II beamlines that occasionally perform these challenging experiments.

6. Efficient data collection

It is feasible to generate a populated knowledge database before the user begins their experiment that could be used to train an initial model. The information on the experiment and sample that the user submits as part of their proposal could be used to perform simulations ahead of time and to extract information from other information sources such as materials databases [44] and natural language processing of literature [45–47].

A classic scenario facing users of relatively high-throughput instruments at NSLS-II, such as the powder diffraction beamline, PDF [48] and the absorption spectroscopy beamline, BMM [49], is dynamic scheduling of samples brought for study. Users often bring a library of different samples related to their initial proposal, and based on some preliminary screening measurements, try to plan out how to most effectively use their allocated beamtime. However, these plans frequently change as the data accumulates and something unexpected or unanticipated is revealed. For example, a critical sample may not be scattering as strongly as expected or an *in situ* setup may take longer than anticipated to setup. This is to say nothing of the classic 'when have I measured enough' problem in these statistic-dependent techniques, where increased measurement time will improve data quality only up to some condition-limited level (e.g. a detector noise-floor is met). The answer to how long one should measure ends up far too often based on remaining shift-allocation and yet unmeasured hopes, instead of statistically sound decisions of data quality.

With the advent of ML methods, the expert-knowledge and signal-processing methods required to reliably make determinations on measurement quality can be trained into a decision-making agent through either conventional supervised learning or reinforcement learning methods. The former is somewhat straightforward conceptually but requires curating a large database of measured patterns, appropriately labeled with an associated data quality metric. Once trained, such a tool could monitor the summation of cumulative data on a single data condition (e.g. temperature point, sample coordinate), and signal when appropriate data quality condition has been met. This signal would trigger the acquisition plan to advance to the next sample or condition point. The disadvantage of this approach is the supervised training would likely result in an agent only effective on the specific beamline configuration used in training, and almost certainly not applicable across different techniques. In order to successfully train a model, we require approximately 10 000 to 100 000 labeled datasets.

Reinforcement learning methods have been recently grown in popularity due to their ability to train supervised-like learning agents in a dynamic experience loop, essentially creating their own training data as they interact with a system. This can be of particular use in cases where either when large curated datasets are not readily available or the conditions of the system may change over time. In the case of efficient data collection, the beamline acts as the environment with the agent directing the measurement scheduling via Bluesky. Although the employed loss metric between techniques may vary, the approach of training the agent across the beamlines will be the same. In this way, well scoped reinforcement learning methods have broad appeal as a potential 'universal data collection advisor' at NSLS-II.

In realizing either of these approaches, we will be able to offer users real-time feedback on their individual data quality or simply automate the process entirely. This is of particular interest for high-throughput or mail-in program applications, as it can free up staff time that might otherwise be spent babysitting data quality metrics at the beamline. Both supervised learning and reinforcement learning methods for these tasks are under development at the facility, with prototypes that have been successfully deployed on the 28-ID-1 (PDF) beamline.

7. Data analysis

After collection, data from the NSLS-II must still be processed and analyzed to extract and utilize the inherent information. Although the details of these measurement-to-information pipelines vary between both techniques and beamlines, all can in principle be aided by the development of automated processing and analysis tools. This becomes especially true in those measurements that generate very large datasets, where manual individual-driven analysis approaches become tedious or impossible.

One common challenge encountered during measurements is to rapidly identify underlying sample changes as a function of experimental coordinate, such as phase-transitions occurring during a temperature-series. Often, researchers are most interested in first mapping where these transitions occur and then studying in detail the material behavior about this transition. Unsupervised learning approaches, such as hierarchical-clustering, can be particularly helpful in these situations.

To demonstrate, we here present an example of barium titanate ($BaTiO_3$) measured on the pair-distribution function (PDF) beamline in an oscillatory temperature series spanning 450 to 150 K resulting in 119 individual datasets. The crystalline structure of $BaTiO_3$ is known to transition between four different phases across that temperature range (cubic, tetragonal, orthorhombic, and rhombohedral). However, beside the phase-transitions, material thermal expansion will cause the diffraction peaks to shift positions as a function of temperature (moving to lower-Q at higher temperatures). Combined with the overall subtle nature of these phase-transitions, it can be a challenge to quickly and accurately identify these transitions as they occur from the raw data, even when we are expecting them [50]. For reference, the raw data from the detector is provided on the figure 7(a).

By employing hierarchical clustering methods [51], the data was correctly categorized into regions corresponding to the four phases, despite having no prior information about the material. The processed data, colored according to theses clustering results, is shown in figure 7(b), along with the temperature profile of the experiment. This example employed unsupervised learning methods on a well known system, the strength of this approach is the ability to just as accurately detect transitions or suggested groupings on data from materials with unknown transitions. The inherent limitations of unsupervised learning approaches is that on their own they apply no physically derived labels (such as space-group) to the clustered data. Employing such methods to process and analyze big data or real-time streaming results is one of the overarching goals in utilizing AI/ML methods at NSLS-II.

8. Conclusion

Herein, we have highlighted our general strategy to utilize AI/ML methods at NSLS-II for improved operations, automation, and analysis. The examples listed were meant to demonstrate some of the different areas of development by scientists at NSLS-II, but in no way represent an all-encompassing list of work. We

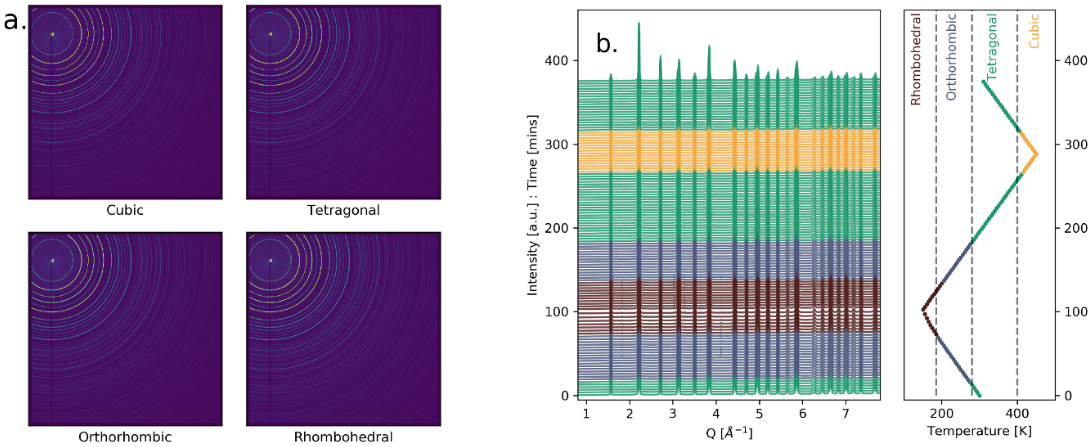

Figure 7. Unprocessed data from the detector (a.) at the PDF beamline (NSLS-II) of $BaTiO_3$ in four different phases: cubic, tetragonal, orthorhombic, and rhombohedral. The processed data (b.-left) from the measurements taken across a temperature series (b.-right). Note that the Hierarchical clustering methods employed sorted the data into the four colored groupings, which were then labeled according to their crystal structure. No prior input about the expected phases was included in the approach.

hope this contribution serves to help foster the growth of such tools by the greater computational science community, and helps to open a dialogue with external researchers, fostering collaborations.

While we are not looking to develop a specific API to develop and train AI/ML tools for synchrotrons, we are looking to expose the hooks to enable those tools to interact with Bluesky. In this way, individual AI/ML components that are developed either internally by scientists at BNL or externally by the greater scientific community can readily interface with the beamlines via the Python-enabled control *à la* Bluesky.

Whilst the development of each project and tool will follow a more agile approach, the overall effort will be coordinated across the NSLS-II to ensure it is in line with both the strategic direction and shared community efforts.

The NSLS-II strategy can broadly be defined as utilizing AI/ML methods in the areas of the following.

Autonomous experiments. and control with physics-based decision making to make and modify materials in for example additive manufacturing and to provide automatic alignment of complex optical systems such as high resolution spectrometers.

Prompt and automatic data analysis. The speed of modern synchrotron data collection has outpaced typical data analysis methods for years. As the quantity of data produced from these experiments is much greater than can be handled with conventional analysis approaches, ML and artificial intelligence (AI) methods will be employed to provide rapid analysis at the beamline.

Improving experiment performance. from source to sample. This includes solutions such as providing predictive state monitoring of the accelerator to minimize beam dumps and maximize user time, adaptively adjusting experiments to optimize on beamline and accelerator performance and providing intelligent tools for automatic beamline alignment.

Data quality optimization. such as removing systematic errors for example identifying and removing artifacts in coherent scattering data and imaging data to maximize both efficiency and quality of data.

For each of these areas, the NSLS-II is ensuring that as we develop these capabilities we are working together and coordinating with the other DOE light sources. With the added challenges originating from the COVID-19 pandemic and the associated transition from a predominantly onsite user model to one focused on remote-access for the foreseeable future, the development of these tools is made even more critical.

It is clear that AI and ML techniques have the potential to make a great impact on the diverse science being performed at synchrotrons.

Acknowledgments

The authors would like to collectively thank their colleagues at NSLS-II and BNL for the work underlying the highlights in this paper and deeply regret the space requirements that prevented their individual acknowledgement.

This research used PDF, CSX, and TES beamlines and resources 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 and under 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'. We would like to acknowledge Anthony DeGennaro who is a co-PI for LDRD 20-38 from BNL Computer Science Initiative (CSI).

ORCID iDs

Stuart I Campbell https://orcid.org/0000-0001-7079-0878
Daniel B Allan https://orcid.org/0000-0002-5947-6017
Andi M Barbour https://orcid.org/0000-0003-2631-7500
Daniel Olds https://orcid.org/0000-0002-4611-4113
Maksim S Rakitin https://orcid.org/0000-0003-3685-852X
Reid Smith https://orcid.org/0000-0002-2538-8924
Stuart B Wilkins https://orcid.org/0000-0003-1191-3350

References

[1] National Synchrotron Light Source II (NSLS-II) website (https://www.bnl.gov/ps/)
[2] Allan D, Caswell T, Campbell S and Rakitin M 2019 *Synchrotron Radiation News* **32** 19–22
[3] Bluesky website (https://blueskyproject.io)
[4] Nielsen J 2018 Fault detection and alarm systems for the CERN Technical Infrastructure *Machine Learning Applications for Particle Accelerators* (SLAC National Accelerator Laboratory) (https://indico.fnal.gov/event/16327/)
[5] Huang X 2018 Facility needs: Synchrotrons *Machine Learning Applications for Particle Accelerators* (SLAC National Accelerator Laboratory) (https://indico.fnal.gov/event/16327/)
[6] Colocho X 2018 Machine Learning Algorithms in use at SLAC *Workshop for Accelerator Operations* (New York: Stonybrook University) (https://indico.bnl.gov/event/3878/)
[7] Rakitin M S, Moeller P, Nagler R, Nash B, Bruhwiler D L, Smalyuk D, Zhernenkov M and Chubar O 2018 *J. Synchrotron Radiat.* **25** 1877–92
[8] Rebuffi L and Sánchez del Río M 2016 *J. Synchrotron Radiat.* **23** 1357–67
[9] LUME: Lightsource Unified Modeling Environment (https://github.com/slaclab/lume)
[10] Northrup P 2019 *J. Synchrotron Radiat.* **26** 2064–74
[11] Tender Energy X-ray Absorption Spectroscopy (TES) beamline at National Synchrotron Light Source II (NSLS-II) website (https://www.bnl.gov/ps/beamlines/beamline.php?r=8-BM)
[12] EPICS collaboration website (https://epics-controls.org)
[13] `MCAG-setupMotionDemo` as a Docker image (https://hub.docker.com/r/mikehart/motorsim/)
[14] Sirepo-Bluesky repository (https://github.com/NSLS-II/sirepo-bluesky)
[15] Rakitin M S *et al* 2020 Introduction of the Sirepo-Bluesky interface and its application to the optimization problems *Proc. SPIE* 209–26
[16] Xi S, Borgna L S and Du Y 2015 *J. Synchrotron Radiat.* **22** 661–5
[17] Xi S, Borgna L S, Zheng L, Du Y and Hu T 2017 *J. Synchrotron Radiat.* **24** 367–73
[18] Breen P G, Foley C N, Boekholt T and Zwart S P 2020 *Mon. Not. R. Astron. Soc.* **494** 2465–70
[19] Chubar O, Rakitin M, Chen-Wiegart Y C, Fluerasu A and Wiegart L 2017 Simulation of experiments with partially coherent x-rays using Synchrotron Radiation Workshop *Proc. SPIE* **10388** 1038811
[20] Wiegart L, Rakitin M, Zhang Y, Fluerasu A and Chubar O 2019 *AIP Conf. Proc.* **2054** 060079
[21] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D and Riedmiller M 2013 *Playing Atari With Deep Reinforcement Learning* (Preprint 1312.5602)
[22] Silver D *et al* 2017 *Nature* **550** 354–9
[23] Schulman J, Levine S, Moritz P, Jordan M I and Abbeel P 2015 *Trust Region Policy Optimization* (Preprint 1502.05477)
[24] Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O 2017 *Proximal Policy Optimization Algorithms* (*Preprint* 1707.06347)
[25] Mnih V, Badia A P, Mirza M, Graves A, Lillicrap T P, Harley T, Silver D and Kavukcuoglu K 2016 *Asynchronous Methods for Deep Reinforcement Learning* (Preprint 1602.01783)
[26] Tensorforce: a TensorFlow library for applied reinforcement learning (https://github.com/tensorforce/tensorforce)
[27] Stooke A and Abbeel P 2019 *Rlpyt: A Research Code Base for Deep Reinforcement Learning in Pytorch* (Preprint 1909.01500)
[28] Shpyrko O G 2014 *J. Synchrotron Radiat.* **21** 1057–64
[29] Sinha S K, Jiang Z and Lurio L B 2014 *Adv. Mater.* **26** 7764–85
[30] Zhang Q, Dufresne E M, Grybos P, Kmon P, Maj P, Narayanan S, Deptuch G W, Szczygiel R and Sandy A 2016 *J. Synchrotron Radiat.* **23** 679–84
[31] Coherent Soft X-ray Scattering (CSX) beamline at national synchrotron light source II (NSLS-II) website (https://www.bnl.gov/ps/beamlines/beamline.php?r=23-ID-1)
[32] Chen X *et al* 2016 *Phys. Rev. Lett.* **117** 167001
[33] Thampy V *et al* 2017 *Phys. Rev.* B **95** 241111
[34] Chen X *et al* 2019 *Nat. Commun.* **10** 1–6
[35] Kukreja R, Hua N, Ruby J, Barbour A, Hu W, Mazzoli C, Wilkins S, Fullerton E E and Shpyrko O G 2018 *Phys. Rev. Lett.* **121** 177601
[36] Yue L *et al* 2020 *Nat. Commun.* **11** 1–8
[37] Coherent Hard X-ray Scattering (CHX) beamline at national synchrotron light source II (NSLS-II) website (https://www.bnl.gov/ps/beamlines/beamline.php?r=11-ID)
[38] Lee S, Jiang J, Fabbris G, Mazzoli C, Disa A, Dean M, Walker F and Ahn C 2020 Antiferromagnetic Domain Dynamics in Nickelate Heteorstructures APS March Meeting March 2020 G47.00006 Denver, Colorado

[39] Gutt C, Ghaderi T, Chamard V, Madsen A, Seydel T, Tolan M, Sprung M, Grübel G and Sinha S 2003 *Phys. Rev. Lett.* **91** 076104

[40] Fluerasu A, Kwasniewski P, Caronna C, Destremaut F, Salmon J B and Madsen A 2010 *New J. Phys.* **12** 035023

[41] Rogers M C, Chen K, Andrzejewski L, Narayanan S, Ramakrishnan S, Leheny R L and Harden J L 2014 *Phys. Rev.* E **90** 062310

[42] Urbani R, Westermeier F, Banusch B, Sprung M and Pfohl T 2016 *J. Synchrotron Radiat.* **23** 1401–8

[43] EPICS archiver appliance (https://slacmshankar.github.io/epicsarchiver_docs)

[44] Jain A *et al* 2013 *APL Mater.* **1** 011002

[45] Pouchard L, Juhas P, Billinge S, Wright C, Campbell S, Park G, Stavitski E and Van Dam H 2019 *Handbook on Big Data and Machine Learning in the Physical Sciences* vol 2

[46] Park G, Rayz J T and Pouchard L 2020 Figure descriptive text extraction using ontological representation *Florida Artificial Intelligence Research Conf.* (https://aaai.org/ocs/index.php/FLAIRS/FLAIRS20/paper/view/18418)

[47] Park G and Pouchard L 2019 Scientific Literature Mining for Experiment Information in Materials Design *2019 New York Scientific Data Summit (NYSDS)* pp 1–4 (New York, NY, USA: IEEE) (https://ieeexplore.ieee.org/document/8909726/)

[48] Pair Distribution Function (PDF) beamline at National Synchrotron Light Source II (NSLS-II) website (https://www.bnl.gov/ps/beamlines/beamline.php?r=28-ID-1)

[49] Beamline for Materials Measurement (BMM) beamline at National Synchrotron Light source II (NSLS-II) website (https://www.bnl.gov/ps/beamlines/beamline.php?r=6-BM)

[50] Olds D, Peterson P F, Crawford M K, Neilson J R, Wang H W, Whitfield P S and Page K 2017 *J. Appl. Crystallogr.* **50** 1744–53

[51] Peterson P F, Olds D, Savici A T and Zhou W 2018 *Rev. Sci. Instrum.* **89** 093001