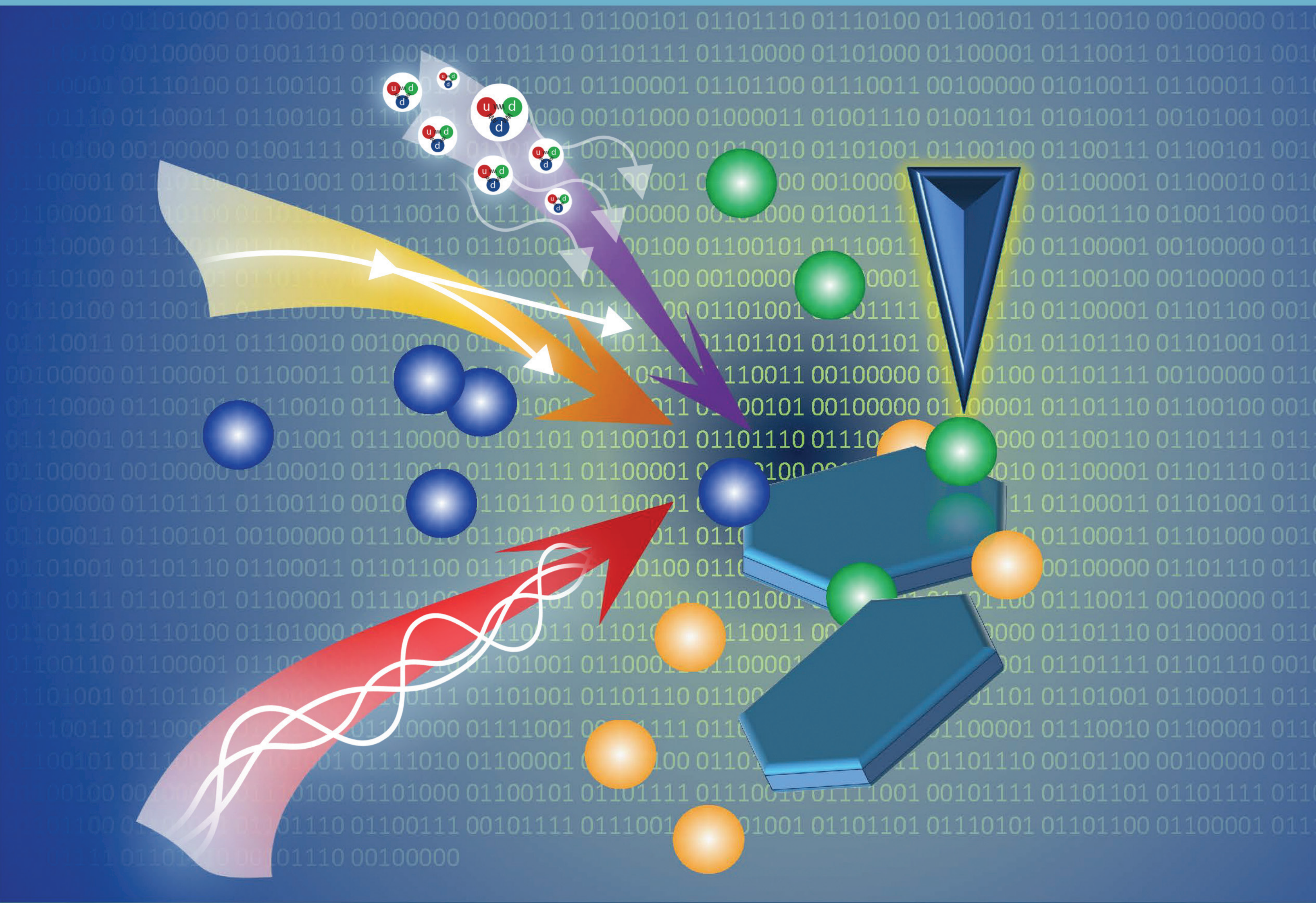
Facilities' Current Status and Projections for Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning



Companion Document for the Basic Energy Sciences Roundtable on
Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning
October 22–23, 2019

The artwork on the cover is a conceptual image of a molecular system being interrogated by multiple US Department of Energy Basic Energy Sciences scientific user facility probe modalities. The data are then fused and interpreted by new capabilities enabled by artificial intelligence/machine learning (AI/ML). The arrows represent the available probe modalities: light (red is photons, yellow is x-rays [both soft and hard]); neutrons (the purple arrow shows the particle representation of the up(u) down(d) down(d) quarks of a neutron); and imaging and nanoscale (i.e., local) modalities (blue triangle). The backdrop of binary numbers connotes the underpinning of high-performance computing and AI/ML-aided information inference and data analytics.

Image courtesy of Oak Ridge National Laboratory

Facilities' Current Status and Projections for Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning

Chairs – Daniel Ratner and Bobby G. Sumpter

SLAC National Accelerator Laboratory, Oak Ridge National Laboratory

with contributions from
Argonne National Laboratory
Brookhaven National Laboratory
Lawrence Berkeley National Laboratory
Sandia National Laboratories
Oak Ridge National Laboratory
SLAC National Accelerator Laboratory
…

April 2019

Table of Contents

List of Figures

Abbreviations, Acronyms, and Initialisms

2D	two-dimensional
3D	three-dimensional
4D	four-dimensional
AD	automatic differentiation
AI	artificial intelligence
ALS	Advanced Light Source
AMO	atomic, molecular, and optical
APS	Advanced Photon Source
APS-U	Advanced Photon Source–Upgrade
ASCR	Office of Advanced Scientific Computing Research (DOE)
BCDI	Bragg coherent diffraction imaging
BES	Office of Basic Energy Sciences (DOE)
CAMERA	Center for Advanced Mathematics for Energy Research Applications
CBIR	content-based image retrieval
CDI	coherent diffraction imaging
CERN	European Organization for Nuclear Research
CFN	Center for Functional Nanomaterials
CMC	ceramic matrix composite
CMS	Complex Materials Scattering
CNN	convolutional neural network
DFT	density functional theory
DOE	US Department of Energy
DWP	dispersion wave parameter
EM	electron microscopy
EPICS	Experimental Physics and Industrial Control System
ESnet	Energy Sciences Network
FACET	Facility for Advanced Accelerator Experimental Test
FEL	free electron laser
FPGA	field programmable gate array
FTS	Spallation Neutron Source First Target Station
GISAXS	grazing incidence small angle x-ray scattering
GPU	graphics processing unit
HFIR	High Flux Isotope Reactor
HPC	high-performance computing
ICEMAN	Integrated Computational Environment for Modeling and Analysis
ID	insertion device
k-NN	k-nearest neighbors
LCLS-II	Linac Coherent Light Source-II
LFA	learned function accelerator
MD	molecular dynamics
Micro-CT	microtomography
ML	machine learning
MS-D	mixed-scale dense
NERSC	National Energy Research Scientific Computing Center

NN	neural network
NSLS-II	National Synchrotron Light Source–II
NSRC	Nanoscale Science Research Center
ORNL	Oak Ridge National Laboratory
PCA	principal component analysis
PRD	priority research direction
RL	reinforcement learning
RMS	root mean square
SANS	small-angle neutron scattering
SASE	self-amplified spontaneous emission
SLAC	SLAC National Accelerator Laboratory
SLIC	simple linear iterative clustering
SNS	Spallation Neutron Source
SPEAR3	Stanford Positron Electron Asymmetric Ring
SSRL	Stanford Synchrotron Radiation Lightsource
STS	Spallation Neutron Source Second Target Station
SUF	scientific user facility
TMF	The Molecular Foundry
TOF	time-of-flight
UED	Ultrafast Electron Diffraction
XANES	x-ray absorption near edge structure
XAS	x-ray absorption spectroscopy
XFEL	x-ray free electron laser

1. Introduction

The US Department of Energy (DOE) operates some of the world's largest scientific user facilities (SUFs), producing unprecedented quantities of data (see Figure 1). Reaching the full potential of these facilities will require innovations to solve a variety of technical challenges associated with data acquisition, simulations, control, analysis, and curation for artificial intelligence (AI) and machine learning (ML) applications. This document highlights challenges and AI/ML opportunities at the light sources, Nanoscale Science Research Centers (NSRCs), and neutron facilities that make up the DOE Office of Basic Energy Sciences (BES) SUFs.

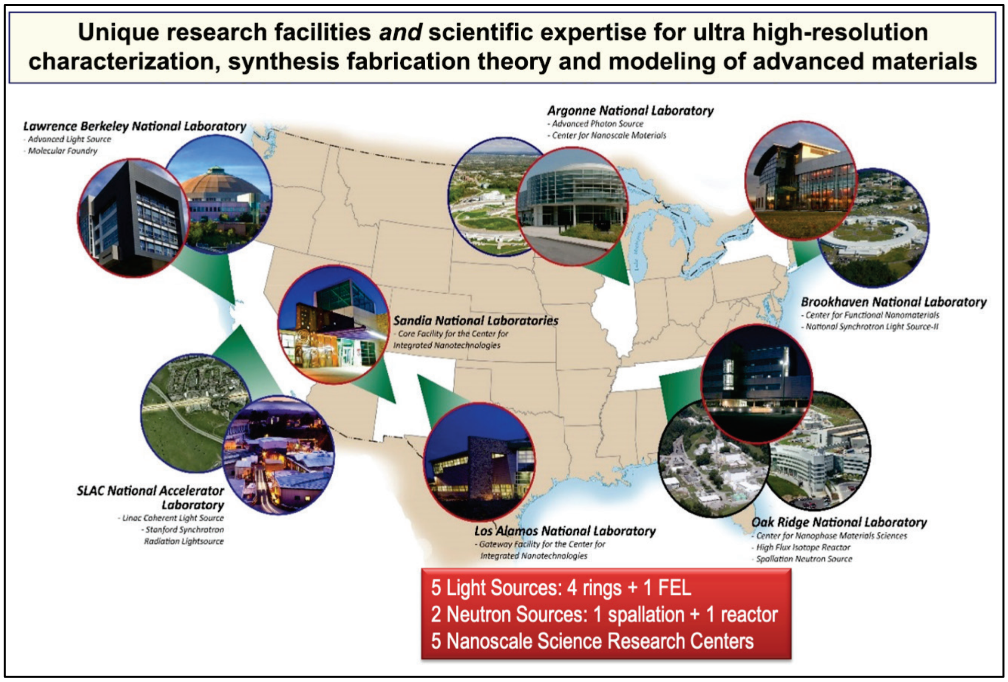


Figure 1. US Department of Energy scientific user facilities. *Source: US Department of Energy.*

To give an example of the challenges facing BES SUFs, the Linac Coherent Light Source–II (LCLS-II) will be the largest-ever EPICS (Experimental Physics and Industrial Control System) project, with 2 million process variables. Beamlines at the LCLS-II-HE (High Energy) are expected to generate megapixel images at megahertz repetition rates. Challenges span speeding up high-fidelity simulations for online models, fast tuning in high-dimensional space, anomaly/breakout detection to protect sensitive superconducting modules, virtual diagnostics that can operate at megahertz repetition rates, smart data acquisition schemes, and sophisticated compression/rejection data pipelines operating at the edge (i.e., next to the detector) to capture the highest-value data from user experiments. The newer synchrotron facilities—the National Synchrotron Light Source–II (NSLS-II), Advanced Light Source (ALS), and Advanced Photon Source (APS) upgrades—will face similar challenges.

Another example is the Spallation Neutron Source (SNS), the world's highest-flux pulsed neutron source, located at Oak Ridge National Laboratory (ORNL). The 19 instruments in its First Target Station (FTS) in time-of-flight (TOF) mode using a broad spectrum of neutrons, coupled with a large solid-angle coverage of position-sensitive detectors, produce large datasets in minutes. The ongoing SNS Proton Power Upgrade project will further increase the neutron flux at the FTS and will eventually serve the SNS Second Target Station (STS) with 22 additional instruments. TOF instruments such as the VENUS imaging instrument, under construction at SNS, will produce 2 TB and 20 TB datasets per instrument configuration; the need to measure several such datasets for tomographic reconstruction poses challenges for data acquisition, storage, curation, and analysis.

Maximizing the efficiency and scientific impact of the sources will require the integration of data collection, data reduction, and online real-time data analysis and curation. Large-scale computation applications such as molecular dynamics (MD) simulations for comparison with neutron scattering data, density functional theory (DFT) for comparison with neutron spectroscopy data, Monte Carlo ray tracing for simulating instrument and complex sample effects, diffuse scattering modeling for investigating the defects in solids, and large-scale tomographic reconstruction will require the development of automated scientific workflows, surrogate models, and novel data science approaches.

Similarly, the five NSRCs—the Center for Functional Nanomaterials (CFN), Center for Nanoscale Materials, Center for Integrated Nanotechnologies, Center for Nanophase Materials Sciences, and the Molecular Foundry (TMF)—have a large and expanding scope of capabilities driven by in situ and in operando characterization that require integrated data management and deep data analysis. These include data acquisition modes in scanning probe and transmission electron microscopy (EM) and "on-the-fly" data analysis and feedback for control as data volumes approach 400 GB/s per instrument. In addition, workflows for diverse modalities of synthesis that allow informed decision-making for streamlining the overall process are needed, alongside approaches for materials characterization using multiple modalities (e.g., optical, electrical, scattering, gravimetric, microscopy), both in situ and ex situ.

How Can Artificial Intelligence and Machine Learning Help?

AI and ML hold promise to address the critical challenges described in this document. The following are some examples of how these technologies could meet scientific research needs going forward.

- **Edge-ML for triggering/data reduction.** At present, there is no solution to enable large-array detector acquisition at full beam rate for future light sources. The most promising path is implementing convolutional neural networks (CNNs) on field programmable gate arrays (FPGAs), for example, for pulse timing and Bragg peak extraction. The situation is similar for EM at the NSRCs, especially considering the newly developed ultrafast detector at ALS and TMF.
- **Detectors.** High-fidelity diagnostics that cannot run at full beam rate can be supplemented by "virtual diagnostics," using ML to infer measurements from low-fidelity diagnostics. ML methods can be used to better characterize detectors and improve processing of raw data (e.g., photon counting from patterns).
- **Smart data acquisition.** New acquisition methods can use ML for autonomous, real-time experimental control. Adaptive techniques can focus on high-value regions and identify rare events. Multiplexing and ML-enabled reconstruction (e.g., compressive sensing) can reduce acquisition time, increase the signal-to-noise ratio, lower radiation doses, and improve resolution.
- **Online optimization/prognostics.** AI/ML optimization is now common at light sources, increasing beam delivery and freeing operators for more complex tasks. Anomaly/breakout detection can predict failures and diagnose faults to improve facility uptime or alert users when performance drifts.
- **Model optimization.** ML will help determine the most appropriate models to describe the measured data at BES SUFs, enabling automated model refinement to greatly enhance the scientific impact and throughput.
- **Optimized operations.** ML will have significant impact on accelerator operation, target design and operation, and design of future neutron scattering instruments.
- **Multimodal learning.** ML can help enable online analysis during an experiment when more than one type of characterization probe is being used.

- **Surrogate modeling/simulation.** ML surrogate models of expensive simulations or calculations can give orders-of-magnitude speedup, for example, by upgrading physics-based models such as tight-binding to the level of DFT or quantum Monte Carlo to gain efficiency and retain accuracy, or by developing inverse models.
- **On-the-fly analysis and feedback for control.** ML can enable very efficient analysis of streaming data, reducing computational needs and allowing real-time control. For example, in EM, ML analysis is fast enough to provide feedback for scanning or manipulation of atoms by design.

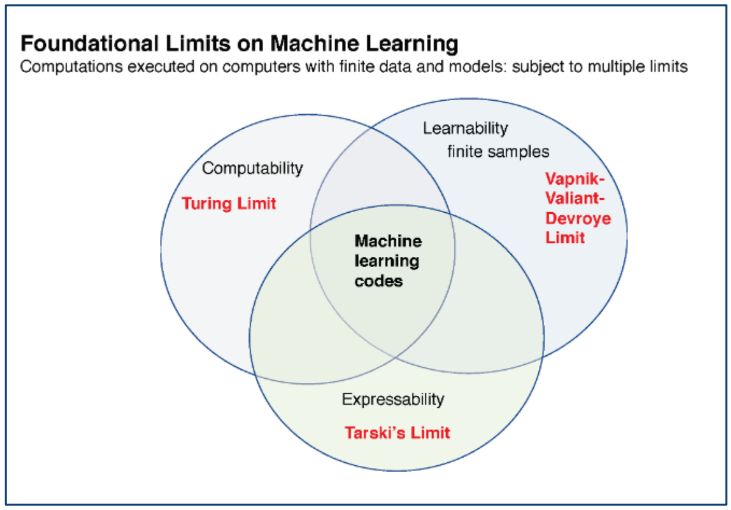


Figure 2. Venn diagram highlighting limits on machine learning. *Source: Nagi Rao, Oak Ridge National Laboratory.*

As the research community moves forward, carefully applying ML will help ensure success because ML techniques are not suited to all problems (see Figure 2). Optimal application of ML occurs at the intersection of situations that satisfy three computational limits: the Turing limit (computability), the Vapnik-Valiant-Devroye limit (learnability of finite sample sizes), and the Tarski's limit (expressability). Finite and sparse data can often pose complications in reliably incorporating ML, and bounding of confidence in an ML-based model (i.e., uncertainty quantification) will always be important, especially if the problem is outside the overlap of the three limits.

Why Now?

In January 2018, the DOE Office of Advanced Scientific Computing Research (ASCR) hosted a Basic Research Needs workshop focused on ML for science, resulting in priority research directions (PRDs) for interpretability, inference, robustness, and scientific computing tools (*Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*, https://www.osti.gov/servlets/purl/1478744). Although the workshop highlighted significant investment in ML for the analysis of big data, there has been relatively little activity regarding the generation of such datasets, a critical need as the DOE SUFs begin commissioning LCLS-II and the APS Upgrade. PRD 6 from that workshop—intelligent automation and decision support—is highly relevant to the SUFs, especially with new facilities pushing the limits of current technology. Timely advances in AI and ML are critical to enable the SUFs' full scientific potential. To successfully use ML at the SUFs, critical challenges need to be addressed including handling new data types, archiving metadata and preserving provenance, creating workflows to manage data transfer to and integration with high-performance computing (HPC) facilities, developing new software stacks, and undertaking uncertainty quantification to identify regions of model validity.

Because DOE facilities have mature control systems that can support large data production rates, they are ideal test beds for developing and demonstrating AI and ML tools. As a result, in addition to aiming to improve SUF operations, research on ML for scientific facilities can also lead to the development of new tools that will benefit complex facilities even beyond the DOE. There may be broad opportunities for collaboration with ASCR and industry, including fabrication, power grids, and design of high-reliability organizations.

2. Current Status, Challenges, and Opportunities at the Scientific User Facilities

AI/ML Needs and Opportunities for Data-Intensive Detector Development

Detectors at ALS are producing ever-increasing amounts of data as x-ray storage ring brightness increases. The same is true for EM at the NSRCs, which are transitioning from "snapshots" to "movies" (see section on AI/ML for Data Production and Analysis at the NSRCs). Detector data will also increase considerably at the future high-repetition-rate x-ray free electron lasers (XFELs).

Two decades ago, the results of an experiment could generally be carried away from the laboratory on a portable archival storage device. One decade ago, in many cases, data volumes had increased so dramatically that this was no longer true. "Write all data to disk" was a plausible, if not always effective, strategy because large volumes of data might be difficult to transport or analyze. Over time, networking speed and density have increased, as have detector data volumes. The speed of archival storage media, however, has not. We are now at the point where writing all data to disk is no longer even possible. Something else must be done.

Several recent studies chart the growth in light source detector data, but interpreting the data can be complicated given differences in factors such as source capabilities and diversities of beamlines. As a simpler illustration, Figure 3 shows the raw EM data rate at the National Center for Electron Microscopy at TMF. For both x-rays and electrons, detectors have gone from tools such as film and fiber-coupled phosphor charge-coupled device detectors to pixilated semiconductor detectors.

AI/ML techniques are promising candidates to reduce raw data volumes (i.e., coming out of the detector) to those that can and should be stored. Schematically, a detector consists of the elements shown in Figure 4. Sensors, together with signal capture electronics, transform an incident probe particle into a measurable quantity. The measurement is performed by front-end processing electronics, and the corresponding (digital) data are aggregated and then transported.

AI/ML techniques are already in use in computation (after the network) and could readily reduce data volumes for primitives at the aggregation stage. For example, Lawrence Berkeley National Laboratory staff are studying ML methods to take detector hits that span multiple pixels and reduce them to single (x,y) hit coordinates. Aggregation is generally performed by FPGAs, so effort is required to customize any AI/ML technique to be operable in an FPGA. Furthermore, this is the earliest possible stage for "image-level" analysis and feature extraction.

Pushing AI/ML techniques into front-end processing would be quite challenging but may be along the lines of future research considered for next-generation microelectronics. While edge computing encompasses activities taking place before the network is engaged, the techniques to deploy computational methods are dramatically different depending on where they occur.

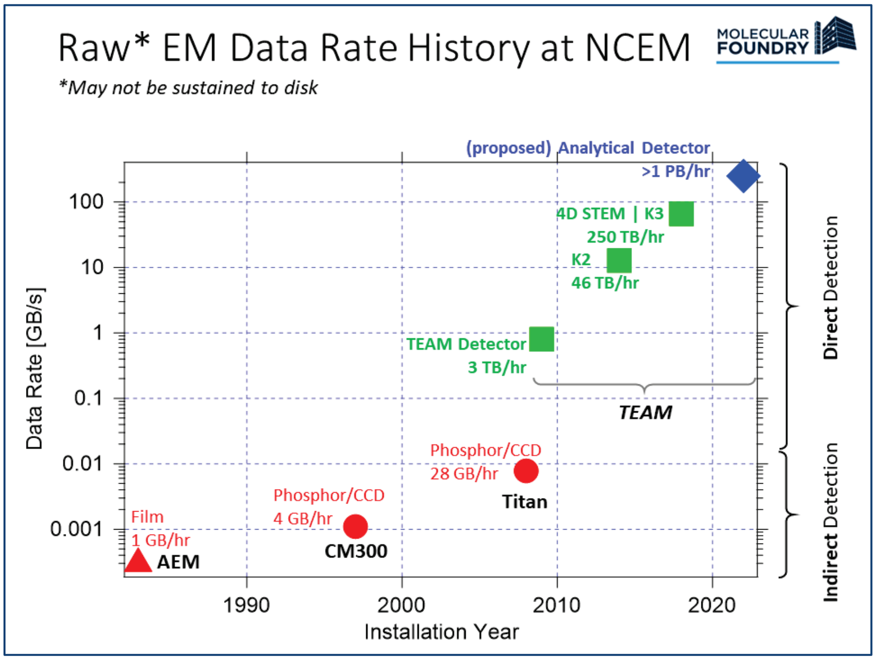


Figure 3. Electron microscopy data rate at the National Center for Electron Microscopy/ The Molecular Foundry. *Source: Lawrence Berkeley National Laboratory.*

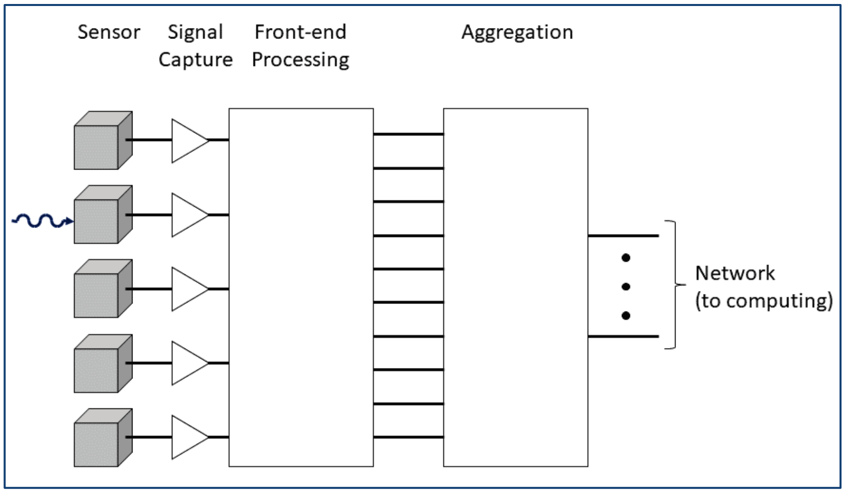


Figure 4. Elements of a detector. *Source: Peter Denes, Lawrence Berkeley National Laboratory.*

An example use case at ALS is tomography: the detector limitations/challenges are as previously described. Adaptive acquisition methods are appealing for the following reasons:

- For time-resolved experiments, samples are often bigger than the field of view at the desired resolution. Therefore, one would like to start at low resolution and, as events of interest begin to occur, predict/detect the area of interest, zoom in on that region, and keep the interesting area in the middle of the field of view.
- Samples/experiments vary significantly so the general framework would need improved tunability.

ML methods designed for adaptive acquisition on variable samples are in the early stages, but they already provide value, and users are excited.

Finally, even at reduced volumes, quasi-real-time data transfer to and analysis at HPC facilities are areas ripe for expansion. The ALS, working with TMF, recently deployed a hundred thousand–frame EM detector, with a direct 400 GB/s connection to the National Energy Research Scientific Computing Center (NERSC). Independent of actual analysis software, the mechanics of data transport and storage have been more challenging than initially foreseen. LCLS-II will present similar needs on a potentially larger scale. As high-volume, quasi-real-time data transfer to HPC facilities becomes more common, there will be opportunities to improve the interface.

AI/ML for Data Production and Analysis at the NSRCs

Numerous opportunities exist for AI methods and ML tools to address the production, mining, analysis, and control of large and diverse datasets generated at the NSRCs. Nanoscience already leverages the powerful work being done in AI and ML, but the NSRCs also bring new challenges stemming from a rich set of highly diverse data, including data from material and chemical syntheses. Work at the NSRCs to date has used the intrinsic scientific rigor of the BES SUFs to address the challenge of diverse data, and the centers are providing physics-aware ML models that are fast, accurate, and verifiably correct. Specific areas of NSRC work and identified challenges and opportunities are briefly outlined below.

1. **AI/ML methods for accelerating nanoscience data analysis**
   - **Unsupervised deep learning for microscopy data.** As shown in Figure 3, the current data acquisition rates of EM (up to 400 GB/s) surpass the storage and processing capabilities needed to enable continuous operation. AI/ML is ideally suited to enable in-line feature extraction and classification of data. In particular, unsupervised learning is promising for a major speed-up of the analysis, as well as providing the potential for detecting features that can go unnoticed by an instrument user.
   - **Incorporating user-generated data for on-the-fly surrogate model training.** Users performing first-principles calculations could use high-quality, dynamical data generated on the NSRC computational infrastructure to train accurate, physics-based surrogate models for deployment to exascale facilities. Examples include automatically retraining a classical force field from the dynamical matrices computed along an ab initio MD trajectory, and training tight-binding models and orbital-free density functionals from each electronic iteration of DFT calculations. Another example is developing methods for replacing expensive function calls with fast AI/ML-generated surrogates. Building learned function accelerators (LFAs) on the fly can minimize computational cost by targeting data collection only to regions of interest and can improve accuracy by reverting to the original function if LFA reliability checks fail. The overall goal is to facilitate user access to exascale HPC facilities through the development of surrogate models compatible with scalable methods at leadership facilities (e.g., MD). This goal can be accomplished by (1) existing mid-range computing facilities at NSRCs enhanced with on-the-fly learning (e.g., GPU-based cores) and (2) the competency of the user community and its focus on BES-relevant systems of interest.
   - **Inverse design, decision trees, active learning, and reinforcement learning (RL) for optimization of materials.** Bayesian, RL, and Monte Carlo tree searches could be used to optimize the design of molecules and materials in dynamic settings and to provide rapid feedback for decision-making. RL could be used to automate decision-making in high-throughput synthesis and multimodal in situ characterization techniques. This will be particularly relevant for fast x-ray and electron detectors with unique geometries that are under development.
   - **Inverse problems in imaging.** AI/ML is being used as a functional approximator to learn inverse operators where forward modeling exists (e.g., dynamical scattering). Predictions of the AI/ML approximator could be refined through automatic differentiation (AD) when the forward computation is inexpensive (e.g., coherent imaging). An AI/ML approximator could be used for

direct data interpretation or to assist iterative solutions when forward computation is expensive (e.g., core-level spectroscopy).

2. **Development of AI/ML for nanoscience physics and chemistry learning**
   - **Physics-based learning for AI/ML.** The connection between AI/ML and the underlying statistical physics or imaging transform needs to be developed. In particular, causal inference must be allowed, perhaps by imposing symmetry and other physical constraints, including those based on in-line theory and modeling, during the learning process.
   - **Multifidelity-scale bridging, model formulation, and surrogate modeling/simulation.** ML is being used for surrogate models to gain efficiency, for fast approximants, and for upgrading physics-based models (e.g., tight-binding to the level of DFT or quantum Monte Carlo to gain efficiency and retain accuracy). Other high-priority areas include developing inverse models, force fields, and exchange correlation functionals and leap-frogging timescales in MD simulations. AI/ML can bridge the electronic, atomistic, and mesoscopic scales for materials modeling and design (e.g., to train coarse-grained or atomistic MD models against first principles–derived or experimental datasets). Additionally, clear potential exists to better explore active learning and evolutionary strategies for sampling training and test datasets.
   - **Quantum ML.** AI/ML offer the opportunity to use the large Hilbert space provided by quantum systems to go beyond classical ML for pattern recognition and data analysis.

3. **AI/ML for nanoscience experiments analysis, monitoring, control, and design**
   - **AI and ML in edge computing and integrated experimental instruments (e.g., AI/ML edge solutions).** This would potentially allow on-the-fly analysis with feedback during an experiment for maximizing information gain. ML is already helping to interpret data from multimodal probes and analysis across platforms, including registration and scaling (e.g., pan-sharpening) for structure–property mapping. Clear potential exists to move this forward for future capabilities.
   - **Current work pushing toward automating aspects of experiments (e.g., tuning environment, importance sampling, next-experiment recommendation).** Autonomous smart instrumentation and synthesis are areas that future work could significantly impact (see item 4 below). Challenges include that automated experimentation quickly generates datasets for inclusion in databases (e.g., EM images). New workflows are also needed because experimenters will typically augment experimental data with data derived from simulations, including atomistic classical or quantum simulations, and forward modeling of experimental data (i.e., materials projects and/or NSRC/SUF user-generated datasets). More details on data challenges and workflows are highlighted in the Common Data Challenges sidebar.

     In situ multimodal analysis. AI/ML is being used to implement online analysis during an experiment in which more than one type of probe is being used. Efficient materials and device characterization are critical elements in the materials discovery workflow. Therefore, NSRCs' characterization capabilities are constantly used to determine chemical composition, materials structure, physical properties, and overall functionality. In general, this determination involves (1) an analytical step to confirm that the target chemicals and/or materials are produced; (2) characterization of the physical properties, morphologies, defects, and interfaces of the functional materials by multiple probes/techniques; and (3) characterization of the functional properties (i.e., in situ and in operando) in devices. New analysis will be required across all these platforms, including registration of data from different instruments and scaling (e.g., pan-sharpening) for structure–property mapping.

4. **Enabling autonomous smart synthesis**

The NSRCs differ from light and neutron sources in that they provide materials synthesis capabilities. Modern synthesis incorporates a wide range of design rules and theories, alongside advanced characterization tools capable of observing synthesis processes on size and timescales at which they occur. However, these traditional modes of synthetic exploration often have difficulty in scaling to handle the complexity envisioned for next-generation materials, which is multicomponent and hierarchical—incorporating molecular, nanoscale, and mesoscale components—to exhibit collective, emergent, and responsive properties. Therefore, new modes of synthesis are needed. A promising direction is autonomous smart synthesis, which would integrate all aspects of the materials discovery loop, from material preparation through characterization to data interpretation and feedback to minimize the experimental trials needed to achieve the desired property. This preparation could allow vastly more challenging materials problems to be tackled. AI/ML is core to this vision to automate the model-building and decision-making aspects of the experimental loops to enable machine-guided synthesis, processing, and materials discovery. We envision being able to address synthesis and control of the following:

- **Design of pathways to metastable phases and materials that persist out of equilibrium.** These materials enable access to a diversity of properties beyond the limits drawn by equilibrium thermodynamics. For example, photon-driven chemistry or optically driven materials processes could provide more control and lead to new materials, such as metastable phases or new low-dimensional materials with dynamics controlled by in-plane heterogeneity rather than layer stacking order. Another example is self-assembly, in which transient (nonequilibrium) intermediate states frequently appear, and control of assembly pathways can enable improved structural control.
- **Interfacial processes and properties.** Controlling interfaces in materials often relies on precise control of atomic bonding at the joint between two dissimilar materials. The ideal strategy to avoid performance-limiting defects is to minimize perturbation of the atomic order at the interface by preserving a high degree of crystallographic order (e.g., epitaxy).

Common Data Challenges

The optimal data workflow architecture—compute, storage, network, algorithms, middleware, and software—to properly handle diverse data and to properly leverage AI/ML requires improved data curation.

- **Data formats:** standardized data and metadata formats
- **Data governance:** experiment links, processing stages, as well as storage, ownership, and access policy
- **Data interfaces:** between experiments and supercomputing/exascale computing
- **Data integration:** across instruments, modalities, experiments (including for large-volume streaming data), and simulations
- **Data coordination:** tools to easily capture and manage metadata, track data provenance, record data processing history, and establish links between databases
- **Data reduction:** rapid extraction of physically meaningful parameters from new, fast detectors
- **Better data analytics tools:** both real-time and post-acquisition, including real-time feedback loops based on analytics and modeling
- **Data quality:** indicators, prescreening, edge computing
- **Improved data and analytics access:** through cloud technologies
- **Data systems:** allow for volume, searchability, and discoverability, and backdrops to HPC resources

- **Design of materials for quantum information sciences.** Such materials include potential solid-state qubits, photon sources, and quantum sensing systems. One example is the Quantum Materials Press for robotic generation of layered heterostructures of 2D materials. This system will generate rich structural, heterointerface, and functional property datasets that will require AI/ML analysis.
- **Heterogeneity in complex systems.** This can be done using high-throughput nanomaterial synthesis and automated atomic-scale, multimodal characterization. The aim is to broadly understand how population diversity influences growth and behavior, with the ultimate goal of creating a closed-loop materials property prediction, synthesis, and characterization. Understanding and controlling heterogeneity may enable the design of multifunctional and self-regenerating catalytic systems.

NSRC Strategy

The ongoing work and proposed activities, opportunities, and challenges will require new pilot projects in some cases and scale-up of successful pilot projects in other cases. They will critically depend on continued development of AI/ML methods for science and will need dedicated staff members. Future work in AI/ML will require access to increasingly large computing resources. Each facility will need to continue investing in advanced local hardware for development and to serve the needs of the user community, especially in conjunction with experiments. Simultaneously, access to powerful computing in burst (i.e., on-demand) modes will become necessary. Therefore, we anticipate the need for a new paradigm for elastic/on-demand access to DOE supercomputing and exascale computing resources.

AI/ML Developments at the Advanced Light Source

Pattern recognition for images and volumes

There are three main challenges for light source data: variety, veracity, and volume. The impact of these challenges results in users taking years to extract scientific insight through a manual process. Therefore, enabling ALS users to extract quantitative information from acquired data plays a critical role in determining the value of the data acquired. For beamlines acquiring image and volumetric data, ML-based feature extraction, detection, segmentation, and classification techniques fulfill the need to extract quantitative information by providing near-real-time solutions using computer vision. At the ALS, ML already plays an instrumental role in aiding users to transition to more autonomous modes of operations, leveraging both expert-developed heuristics and pretrained data. ALS scientists, together with the Center for Advanced Mathematics for Energy Research Applications (CAMERA), have developed and deployed software infrastructure at beamlines to ensure users can leverage trained ML models with minimal effort. Figure 5 illustrates a user-guided ML-driven process by highlighting the pipeline experienced by users analyzing ceramic matrix composites (CMCs). Starting with the acquisition of the tomography data at the microtomography (micro-CT) beamline at the ALS, the software stack exercises a CNN using TensorFlow to extract frequency distribution of microstructures and visualization of the material composition. Additionally, the current data analysis using the ALS framework retains knowledge (e.g., metadata, data-driven models) that will benefit future users at the beamline by improving the interpretability of the ML models. Finally, adding the ability to scan large datasets to find inconsistencies among real structures known a priori, and results derived from automated detection algorithms, provides immediate scientific feedback to users and software developers. The framework can be generalized to a wide variety of science domains, covering everything from materials discovery (e.g., exploration of morphologies in extreme environments) to manufacturing.

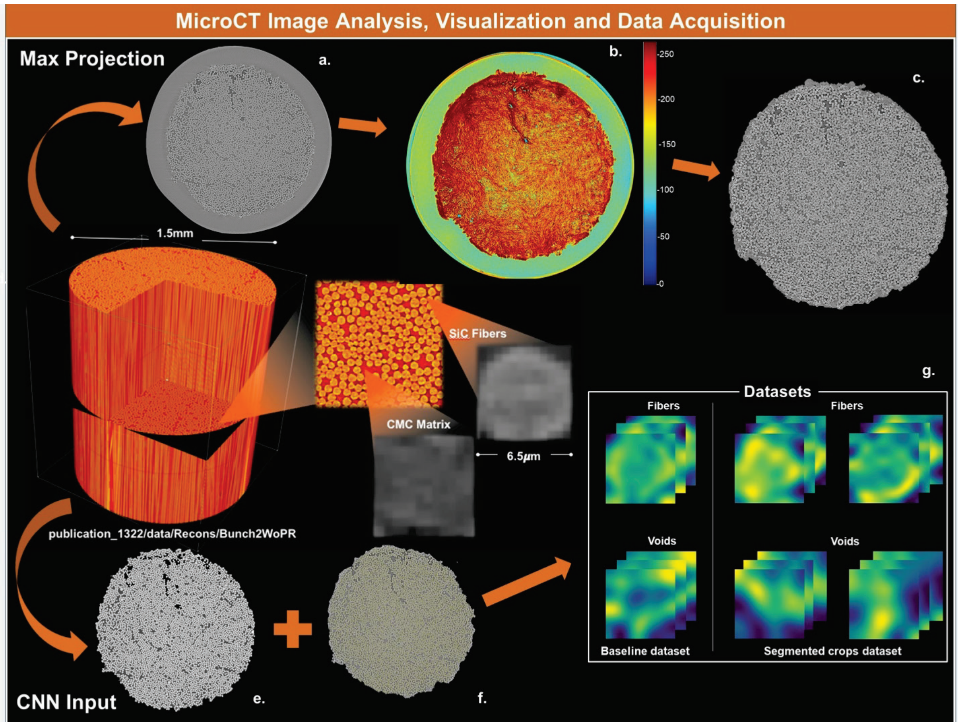


Figure 5. Multiscale image analysis (a–g) from global analysis of ceramic matrix composites to local detection of fiber profiles with convolutional neural networks leveraging knowledge from previous experiments at the ALS. *Source: Image courtesy of Daniela Ushizima, Lawrence Berkeley National Laboratory.*

Extension to pattern recognition for scattering patterns

As the rate of data acquisition increases, analyzing large amounts of data in a timely manner becomes challenging. A collaborative effort between the ALS and CAMERA scientists uses AI/ML-based algorithms and approaches to assist in data analysis in a high-throughput and automated fashion for scattering data. Similar to pattern recognition for real-space images, CNNs are being applied to aid data analysis for x-ray scattering.

Nanostructured thin films have a variety of applications in modern materials science; for example, they may be used in waveguides, gaseous sensors, organic photovoltaics, and piezoelectric devices. Thin-film structure characterization is commonly done by grazing incidence small-angle x-ray scattering (GISAXS). Analysis of scattering patterns (not limited to GISAXS experiments) has proven to be a challenging and time-consuming task because of the nature of reciprocal space data. To highlight the potential of neural networks (NNs) in this domain, CNNs were trained to quickly analyze scattering patterns and identify underlying crystal structures, crystal orientation, and film thickness. CNNs based on the AlexNet architecture were trained using over 7 million simulated x-ray scattering patterns. The resulting models recognized GISAXS scattering pattern with a success rate of 98% from simulated data. Figure 6 shows how the CNN classified the various thin-film motifs. Principal component analysis (PCA) was done on the final connected layer of the CNN and highlights the separations the CNN identified among the various scattering patterns. Various noise sources (e.g., shot noise, Gaussian blur) were also introduced, and the success rate dropped dramatically.

The use of these trained models on such datasets will greatly impact the x-ray and neutron science communities by speeding up GISAXS data analysis of new materials. However, the current predictive models for real data still require improvements to enable use in real-world environments. The presented developments will provide guidance for GISAXS users to increase curated public datasets and push toward real-time data analysis at high-brightness facilities. The efforts undertaken by ALS scientists highlight the opportunities that AI/ML provides in impacting these real-world use cases and addressing the greater big-data challenges for user facilities.

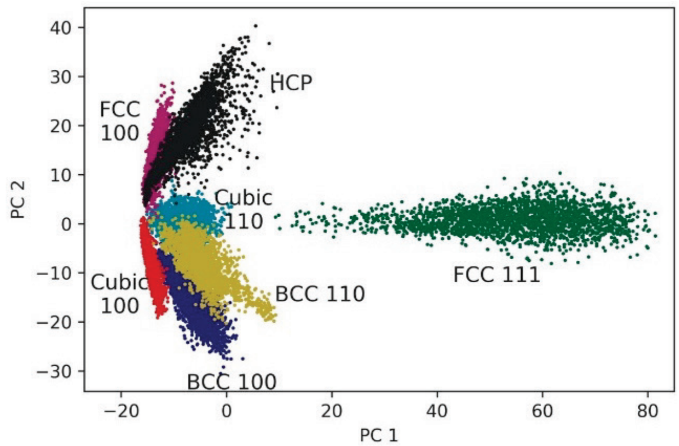


Figure 6. Principal component analysis of the final fully connected layer of the convolutional neural network. PCA shows a clear distinction in the classified FCC 111 structure and orientation compared with the other crystal structures and orientations. *Source: Image reproduced with permission from Liu et al. MRS Communications, 9(2), 586–592. doi:10.1557/mrc.2019.26.*

Machine learning for beam size correction inside the ALS storage ring

The success of storage ring light sources lies in their stability, resulting in constant position, angle, and intensity of radiation delivered at tunable wavelengths and narrow bandwidths. Transverse beam stability currently sits at 10% of the root mean square (RMS) emittance. This deviation is starting to be detected in experiments being conducted at ALS. Stability requirements will be even more stringent with the planned upgrade of the ALS to a diffraction-limited light source.

Traditionally, the ALS relies on a forward loop to correct for changes in the transverse beam size that result from the undulator gap change using the dispersion wave parameter (DWP). The effect of the DWP is periodically measured and stored in a look-up table to be used over a period of weeks or months. The ALS measures the 2D beam size at a high refresh rate at a dedicated target endstation. The availability of high-quality data for the beam size, together with a complete set of metadata from the storage ring, allows the construction of an ML model to correct for beam size fluctuations. To aid in the transverse beam stabilization, a correction loop has been implemented based on NNs. The vertical beam size has extreme nonlinear relationships among all the insertion device (ID) gap and phase settings. The NN for the accelerator was trained by varying the gaps of all undulators inside the ALS storage ring and recording the resulting beam size changes while also sweeping through all possible DWP values. The results of a trained NN are shown in Figure 7. The area with a white background represents the NN being disengaged and results in high fluctuations in the vertical beam size due to undulator and ID gap changes. When the NN was activated, the DWP value was chosen to counteract all the undulator motions and stabilize the vertical beam size. The RMS of the vertical beam was decreased by an order of magnitude, as shown in the portion between the arrows in Figure 7.

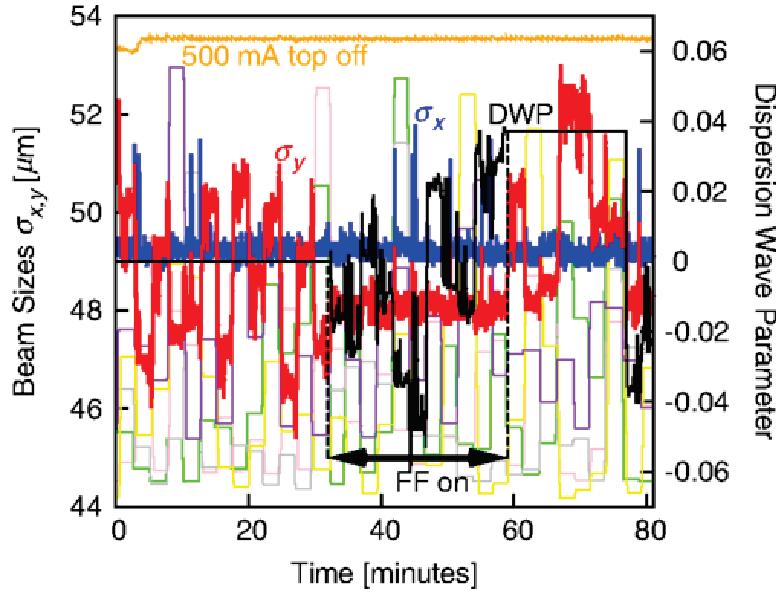


Figure 7. Vertical beam size (red) is shown to stabilize under application of a feed-forward neural network. The area between the arrows highlights the influence of the neural network on the beam size and the dispersion wave parameter. *Source: Permission to use image granted by Alexander Hexemer. Licensed under Creative Commons Attribution 4.0 International License.*

APS Needs and Progress on AI/ML Applications

The APS is one of the largest data producers among the DOE BES facilities. Every beamline is optimized for a different type or set of measurement(s), and at present, nearly every beamline has a different stack of control, workflow, and analysis protocols. Even so, the instruments can be grouped into a much smaller set of categories (around a dozen) that can share future development thrusts because they share measurement types and scientific goals. Most of these areas are shared with other BES light and neutron facilities. The APS is currently embarking on an upgrade (called APS-U) that will greatly enhance the performance of most beamlines and result in data rates increasing in many cases by two or more orders of magnitude. In particular, techniques that use coherent x-rays will make full use of the upgrade's vast improvement in source properties. These coherent scattering techniques are expected to generate far larger volumes of data than what can be processed through conventional methods on today's computing resources. For example, in the APS-U era, the PtychoProbe feature beamline alone is expected to generate ~130 PB of raw ptychographic data per year. Over 30 petaflops of continuous computing power will be needed to keep up with this anticipated data generation rate using current ptychographic reconstruction algorithms. (In comparison, it is anticipated that the APS will produce ~1 PB of raw ptychographic data per year by 2022, requiring <1 petaflop of continuous computing power.) Likewise, the upgraded instrumentation of the accelerator and storage ring will result in the recording of over two orders of magnitude more accelerator control, diagnostics, and orbit information.

This section discusses APS needs, surveys areas in which APS projects employ ML, and outlines some barriers to more widespread use of AI/ML. AI/ML approaches can greatly enhance APS operations. In fact, it will be impossible for the APS-U to manage beamline operations and process the raw data generated at the ~70 beamlines without adopting machine-enhanced beamline control and supervision—and advanced supervised and unsupervised ML techniques—to accelerate data analysis. Likewise, operation of the APS will benefit tremendously from a fast-feedback control system functioning at a level beyond the current state of the art to more reliably meet photon delivery specifications.

Data production rates are not the only motivation for AI/ML. Imaging techniques need to employ an advanced computational methodology to reduce the number of required sample doses by combining domain-dependent knowledge with experimental measurements, because the required number of photons is determined by resolution. Even under the most favorable conditions, the absorbed radiation for resolving a 10 nm feature is estimated to be ~10^{10} rad, which is at the borderline of inducing irreversible damage to biological specimens. Preservation methods such as chemical fixation, dehydration, and cryogenic cooling are often applied to improve the tolerance of the sample to radiation damage; however, they also can alter the underlying morphology and chemical composition of the sample. AI/ML-aided reconstruction provides an effective path to overcome this dose limit on spatial resolution, a crucial milestone for reaching the promised diffraction-limited resolution. Denoising and deblurring while preserving spatial resolution has been an active AI/ML research area in the image processing community, and these techniques need to be adopted for x-ray imaging applications. Many of these techniques need to be extended from two dimensions to three dimensions and higher (i.e., time plus external stimuli).

Machine operations

The APS accelerator complex is a complicated, integrated system that includes vacuum, magnets, radio frequency cavities, mechanical systems, water cooling, power supplies, and other subsystems. During operation, values for tens of thousands of settings and measurements are logged each second. Applying AI/ML to these data can allow optimization and increased reliability of the APS accelerator complex. Diagnosis of the occasional electron beam loss and beam dumps that occur require extended analysis efforts from experts to correlate failure events to abnormal conditions of one or more technical subsystems. Application of AI/ML can offer advance warning of failure and perhaps mitigation of these accelerator operation problems.

The APS-U accelerator is a challenging next-generation machine on which the beam size is greatly reduced via stronger focusing and more damped excitations. To optimize the performance of the APS-U accelerator, an ongoing effort is applying AI/ML techniques to the data generated by accelerator physics simulations. It is worth mentioning that AI/ML optimization methods (e.g., genetic algorithms and particle swarm optimization) have already been successfully applied for several years to improve beam lifetime, transverse coupling, and injection efficiency in APS operations.

Data review/preprocessing

Many APS operations require human intervention to review the quality of datasets. As an example, when the high-resolution/high-throughput powder diffractometer instrument was commissioned for its very popular mail-in access process approximately a decade ago, a small fraction of datasets were found to exhibit problems that would not be obvious to users once the data were preprocessed. These include issues such as ice crystals building up on the sample and materials degrading or transforming in the beam. The issues could not be detected easily in software at the time but could easily be spotted by an operator through a quick visual inspection, and a manual review of all data was implemented. This sort of operation is an obvious target for ML, but there are many more.

Another logical target is the human intervention needed to visually separate signals. ML models such as CNNs, which have demonstrated superhuman performance on image classification and processing tasks, are being developed to parse the raw data to identify and quantify features of interest and retain only relevant portions of the raw data. Current work in this area includes detection of spurious scattering signals in powder diffraction or pair distribution function experiments employing area detectors. The unwanted signals can be single-crystal diffraction spots from the sample container (e.g., diamond anvil cell) or from polycrystalline components in the sample or container. The intended and unintended parts of images can be very similar in appearance, but contextual information obvious to a trained observer allows for their separation. The need for automation of this process is illustrated by a common type of experiment, in situ battery cycling. In these experiments, unwanted lithium or sodium crystal spots commonly appear, but their positions on the images constantly change over the thousands of images collected with time. Manually masking each image is almost impossible. An ML-based automasking process is being developed. Another task for which human effort is currently needed to separate signals is masking diffraction from crystals other than the intended subject in Bragg coherent diffraction imaging (BCDI). Post APS-U, data volumes will make this impossible.

Data reconstruction/analysis

Traditionally, synchrotron sources have been more concerned with progressing beyond state-of-the-art instrumentation than with improving computational data analysis techniques. Given order-of-magnitude improvements in source brightness and coherent flux, this approach is no longer sustainable; data analysis methods must be revolutionized. Additionally, light and neutron facilities are hosting an ever-increasing number of users from an extremely diverse range of scientific specializations who are not expert users. Thus, mechanisms for data analysis that are accessible to nonspecialists must be an increasing focus. Beamline staff bring domain-specific knowledge for data analysis but must depend on finding expertise in modern methods in computation and applied mathematics; that is why multimission labs such as Argonne National Laboratory are ideal environments for improving data analysis methodologies. This situation is illustrated by several of the example cases below in which initial collaborative efforts in data reconstruction demonstrate the potential value of ML in APS data analysis.

Neural networks for low-dose x-ray tomography

Low-dose imaging is a crucial computational development goal for soft matter imaging. CNN is one of the most impactful deep learning techniques because it mimics the feature analysis of the NNs within animal brains. It has been applied to simulated and actual transmission x-ray microscopy scans of complex biological samples such as mouse brains. For this work, tomographic images were derived from raw low-dose projections using two conventional methods and CNN (see Figure 8). CNN demonstrated a tenfold improvement in x-ray signal quality compared with conventional image processing methods and was used to reconstruct myelinated mouse brain axons as shown in Figure 9.

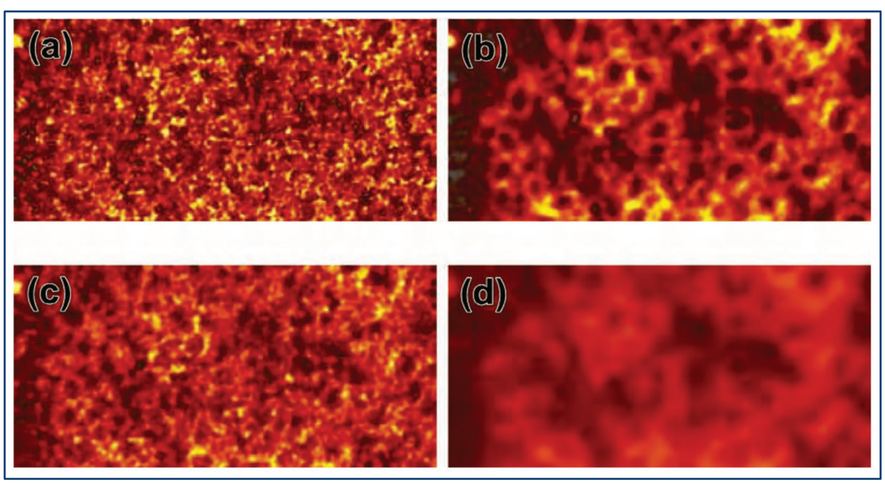


Figure 8. Low-dose transmission x-ray microscopy images from a section of a mouse brain at 30 nm resolution comparing (a) unenhanced reconstruction, (b) CNN, and (c, d) conventional enhancement. *Source: Figures adapted from X. Yang et al., Sci. Rep. 8, 2575 (2018). Licensed under Creative Commons Attribution 4.0 International License.*

Typical image recovery algorithms on which BCDI relies are iterative in nature and hence are time-consuming and computationally expensive, making real-time imaging a challenge. Furthermore, these algorithms struggle to converge to the correct solution, especially when there is a strong gradient in the phase of scattered radiation from the material, typically due to strain or material defects such as line dislocations. To address these algorithmic shortcomings that limit the scope and speed of BCDI, APS staff have built and trained coherent diffraction imaging NNs (CDI NN), deep convolutional encoder–decoder networks that have learned the mapping between raw diffraction data and corresponding object structure and strain (Figure 10). CDI NN is ~500× faster than traditional algorithms used to recover objects and is more robust against the presence of large strains in the material being imaged.

Figure 9. Individual myelin sheaths surrounding mouse brain axons are contrasted with different colors. *Source: Derived from CNN-analyzed data in Figure 8. Source: X. Yang et al., Sci. Rep. 8, 2575 (2018). Distributed under Creative Commons Attribution 4.0 International License.*

AI/ML for 4D image reconstruction

At the APS, x-ray imaging and scattering on 100 ps temporal scales is a revolutionary tool for the study of irreversible processes, such as turbulent flows, shock-induced structural transformations, and laser-assisted advanced manufacturing, in which materials are placed in states far from equilibrium. It is now possible to record time-resolved 2D images of highly transient systems, such as the dispersal of supersonic liquid flows as occurs in vehicle fuel injectors. An example is shown in Figure 11. For improved design of combustion engines, determining the actual 3D spray morphology from such images is crucial.

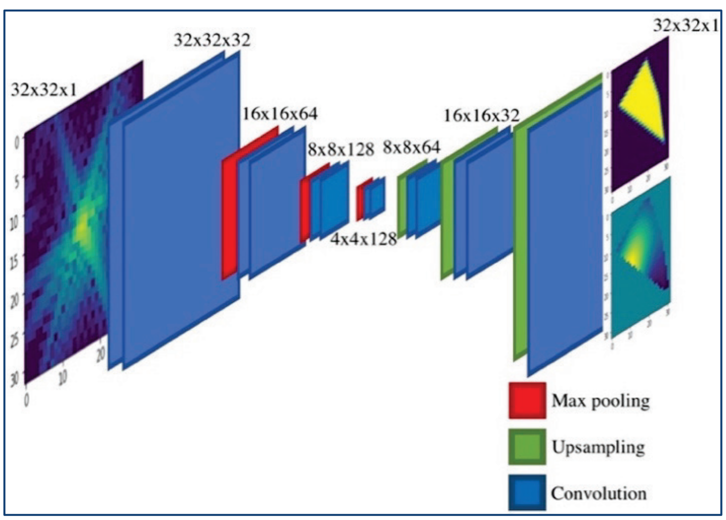


Figure 10. Structure of CDI NN, a deep convolutional encoder–decoder network trained to predict structure and strain from raw x-ray diffraction imaging data. *Source: Cherukara, M. J., Nashed, Y. S., and Harder, R. J. (2018). Sci. Rep. 8(1), 16520. Distributed under terms of Creative Commons CC BY License.*

A reconstruction process based on full-wave scattering physics is under development in which convergence is made possible by a rigorous AI/ML algorithm, and computational fluid dynamics is incorporated to train deep learning models. This approach is much needed for contemporary experiments, but after the APS-U comes on-line, ultrafast imaging experiments are expected to yield much more spatial information as a result of a hundred-to-thousandfold increase in beam coherence at the upgraded source.

Harnessing ML frameworks

The software behind NN training (e.g., Google's Tensorflow or Facebook's Pytorch), along with parallelized and highly optimized reverse-mode AD

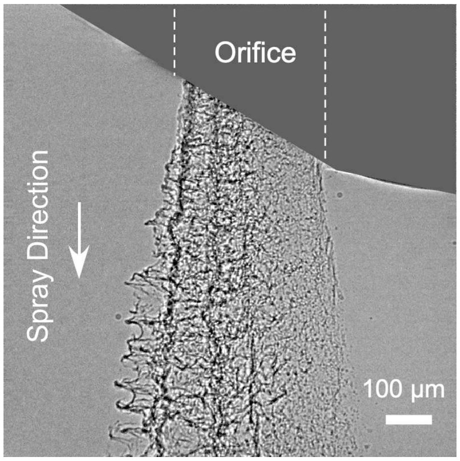


Figure 11. A 100 ps x-ray phase contrast image showing a fast-turbulent flow of liquid from a fuel injector. Combined simulation and reconstructive techniques that incorporate AI and deep learning are being developed to reveal 3D liquid complex morphology and fast dynamics. *Source: Image courtesy of J. Wang, Argonne National Laboratory.*

tools, provides an excellent toolkit for implementing iterative image reconstruction algorithms at scale. Furthermore, the use of ML to provide an estimate for the solution, followed by refinement with AD, maximizes speed while retaining the accuracy of conventional data inversion approaches.

AI/ML-guided experiments

Gaining transformative insight into dynamic materials processes requires identifying, tracking, and quantifying the most relevant volumes within a sample under various conditions of applied stimuli. In addition to choosing the most important volumes to sample, an imaging modality must also be chosen. This situation presents a vast measurement parameter space that is very difficult to navigate when concrete connections are sought between sparse local phenomena, such as dislocation motion and grain boundary stress concentration, and bulk irreversible processes. To this end, the APS is developing AI/ML approaches, shown schematically in Figure 12, aimed at real-time in-experiment decision-making to capture relevant microscale and nanoscale structural changes that govern macroscale response.

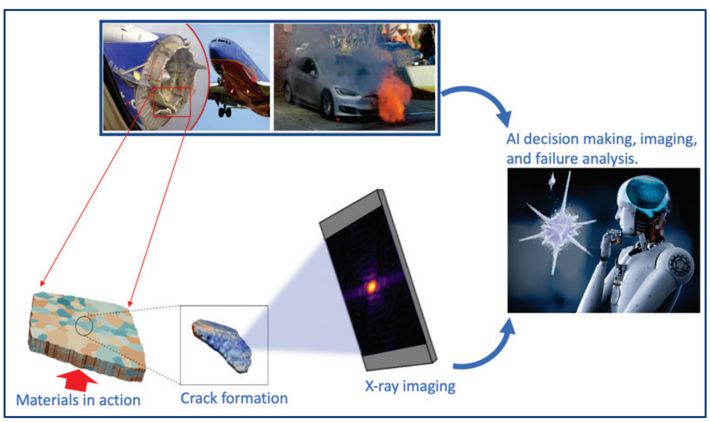


Figure 12. AI-guided workflow to automate experimental decision-making during beamtime experiments. A trained agent will make decisions regarding optimal resource allocation. *Source: Brian Toby, Argonne National Laboratory.*

For example, AI and ML promise to identify rare events, such as failures in materials under stress that occur on timescales too short for humans to observe. Adaptive control of experiments, implemented as real-time in-experiment decision-making, not only can identify regions of interest but also can save historical data from a circular buffer before they are overwritten once an event of interest has been noted. The introduction of AI/ML into instrument control systems also has the potential to allow instruments to detect when their alignment has drifted and then perform automated alignment and recalibration. The increased brightness afforded by new and upgraded sources such as the APS-U, coupled with advances in detector technologies, enables the study of interesting dynamic phenomena at timescales that were once inaccessible.

These advances in sources and detectors will result in the generation of orders of magnitude more data over exceedingly shorter timescales. Humans are not capable of processing such vast amounts of data on such short timescales, and as experiments progress to speeds at which humans are too slow to make control decisions, AI/ML-informed adaptive control becomes imperative.

Wider data challenges

With huge and expanding data streams, the APS is an obvious target for AI/ML, which has the potential to revolutionize nearly every aspect of APS operation and use. However, it should be recognized that obtaining labeled training data to build these ML models remains a challenge. Each of the dozen or so classes of APS experiments uses separate data reduction and analysis techniques, and at present nearly every instrument has its own local formats for data storage. Furthermore, data maintained within the APS are not uniformly archived and typically contain only the experimental observations and varying levels of data collection metadata. Datasets typically do not contain sample provenance information or user results, the former of which is crucial for data analysis and the latter for data mining. Nor do most beamlines typically retain data processing histories. For example, if the Beamline 11-BM: High Resolution Powder Diffraction processing pipeline had been written to record the observations from data quality review, these data would be an excellent AI/ML training set, but the value of doing so was not foreseen when the workflow was developed.

The process of recording provenance information and connecting user results to beamline data is best directed by BES, as all BES user facilities face similar problems and the problem is large in scope. The mechanisms for data archiving are being investigated, but each class of beamlines needs to develop an ontology for how data and results can be universally classified. Only a few fields, such as crystallography, have done so. Compelling software packages that implement these ontologies as data standards are needed so that users have uniformly coded electronic results. BES can help by incentivizing users to provide their results in electronic formats and to connect these results to archived datasets.

AI/ML at SLAC Accelerator Facilities

SLAC operates a number of accelerator user facilities, including LCLS, SPEAR3 (Stanford Positron Electron Asymmetric Ring), FACET (Facility for Advanced Accelerator Experimental Test), and the UED (Ultrafast Electron Diffraction) instrument. In the near future, LCLS-II and FACET-II will be commissioned and enter user operation. Ensuring stable, high-performance operation of these facilities is a primary goal of the SLAC accelerator teams. In response to the challenges in realizing and maintaining high machine performance and supporting user science, ML has been applied to a variety of accelerator applications by SLAC physicists. With successful demonstration of this approach, ML studies on accelerators have been gaining momentum. The accelerator ML work at SLAC falls into several categories: (1) online optimization of accelerator performance, (2) surrogate modeling of the accelerators, (3) data analysis and diagnostics, and (4) application in accelerator design optimization. This section briefly summarizes the current state of ML studies on the SLAC accelerators.

Online optimization

Online optimization is an effective method to discover the optimal operation setting when discrepancies exist between the actual machine and the design model, a common challenge in both storage rings and free electron lasers (FELs). SLAC accelerator physicists pioneered the development and implementation of online optimization algorithms. The robust conjugate direction search method was designed for the optimization of noisy functions with a complex terrain in the parameter space [1]. It has been applied in ~30 laboratories worldwide, including work that resulted in substantial improvement in the storage ring nonlinear beam dynamics at SPEAR3, the European Synchrotron Radiation Facility, MAX-IV, and NSLS-II, and improvement in FEL power at LCLS [2].

Application of Bayesian optimization to online tuning has been extensively studied on the LCLS [3]. The Gaussian process optimizer was developed in collaboration with Stanford University's Computer Science Department. Its application to the tuning of optics-matching quadrupoles in LCLS has led to a significant reduction in FEL tuning time.

Surrogate modeling

Modeling of accelerators is essential to accelerator design and control. High-fidelity modeling of complex systems such as FELs typically employs computationally expensive simulations. Evaluation of the computer model could take hours or days on large-scale computer clusters, making it impractical for many applications, such as real-time controls or diagnostics. NN-based models can provide a fast, accurate surrogate for the physics model in such circumstances. These models can also be trained using measured data. The surrogate modeling approach has been demonstrated for the LCLS and FACET machines with problems ranging from individual accelerator modules, such as injectors and bunch compressors, to start-to-end simulations (Figure 13) [4]. Surrogate models may be used in experiment planning, design optimization, virtual diagnostics, and machine tuning.

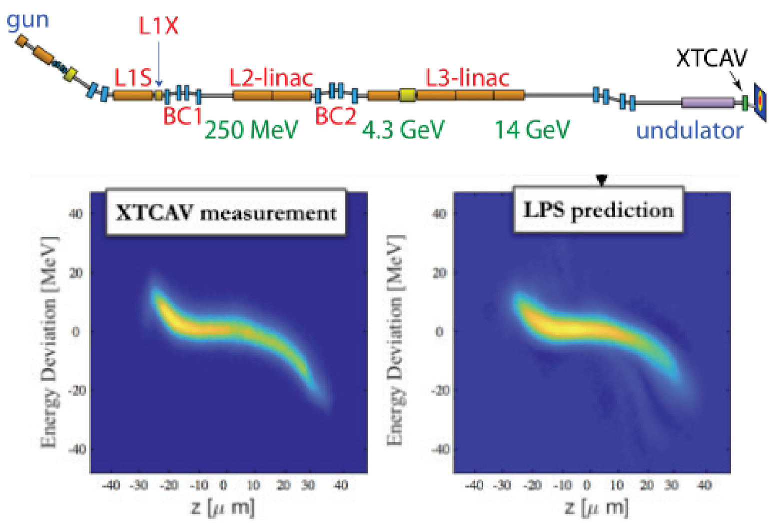


Figure 13. The NN-based surrogate model prediction (right) reproduces the electron beam distribution obtained through measurement (left) for the LCLS beamline (schematic at top). *Source: C. Emma, et al., Phys. Rev. Accel. Beams (2018), American Physical Society under terms of Creative Commons Attribution 4.0 License.*

Data analysis and diagnostics

An accelerator has many diagnostics that monitor and archive its state and performance. It is a challenge to extract useful information to benefit machine operation and user experiments. ML methods have been applied to data analysis in several scenarios. In an application on LCLS, a method using a CNN to analyze the images of electron beam $(z, \Delta E)$ distribution obtained with a deflecting cavity was shown to be faster and more accurate than the standard algorithm [5]. ML is being applied in the reconstruction of sample structure in ghost imaging, an experimental technique that employs the shot-to-shot jitter to extract information from the sample [6]. NNs have also been applied to analyze the operation history data for the discovery of hidden connections between environment variables and the machine performance on SPEAR3 [7].

Accelerator design optimization

Present-day accelerator design practice often requires a global search of the optimal solutions of a multidimensional parameter space. High efficiency of the optimization algorithm is critical, as the simulation involved in evaluating a solution is typically time-consuming. Surrogate models have been found to enable a significant speedup of the multiobjective optimization of an accelerator [8]. Recently, a multigeneration Gaussian process optimizer was proposed and shown to converge substantially faster than traditional algorithms (e.g., genetic, particle swarm [9]).

Although ML studies in these areas at SLAC have led to many exciting results, they are still in the early stages. Many opportunities exist to make breakthroughs in these areas. In the near term, the studies will result in improvements in accelerator performance and user support. These include improving both beam quality and tuning efficiency. Continuing ML research in accelerators may change philosophies regarding accelerator design, commissioning, and operation and open a new era in which ML and AI play a central role.

Growth of ML research requires sustained investment. In most cases, the ML activities in the SLAC accelerator community started spontaneously as the physicists saw opportunities to improve their work with ML techniques. Initial ML work has attracted internal and external funding under several projects. The Gaussian process optimizer study is supported by a 2-year Laboratory Directed Research and Development project. Two ML projects are funded jointly by BES and ASCR: surrogate modeling, and beam-based optimization and ML for synchrotrons. Continued funding support in these areas and new areas such as accelerator fault detection and prediction is essential.

AI/ML for LCLS X-ray Science

DOE hosts and collaborates on some of the most data-intensive scientific facilities in the world. From XFELs (e.g., the LCLS [10]) to international efforts to understand fundamental particles (e.g., the Deep Underground Neutrino Experiment [11] and the ATLAS experiment at CERN's Large Hadron Collider [12]), humanity's appetite for scientific insight is driving a change in how we must view information. The SLAC National Accelerator Laboratory (SLAC) operates the world-leading XFEL, the LCLS; a near-term upgrade, LCLS-II, will enable a ten-thousandfold increase in the LCLS data rate.

Examples of ML at LCLS include (1) high-throughput, low-latency experimental autonomous decision-making; (2) Bayesian signal separation; and (3) multiplex statistical analysis. These methods have wide application across experiments in photon science, including materials science; soft and hard x-ray spectroscopy; atomic, molecular, and optical (AMO) physics; and biological sciences.

Machine learning at the edge

A typical AMO experiment at LCLS is represented by Figure 14. This experiment used 10 fs x-ray pulses (shown as a magenta pulse) to eject the innermost electrons from specific atoms in optically aligned molecules. Exaggerated triatomic N_2O in the center of the ring of detectors, kicked into alignment by the sequence of eight optical laser pulses [13], is shown as red in the image. The Auger electron emission pattern then reveals not only the energy of the involved electronic levels but also the molecular frame electronic symmetries [14,15]. This paradigm serves as a motivation for one of the major efforts driving high-repetition-rate, angle-resolved soft x-ray spectroscopy at LCLS-II [16,17].

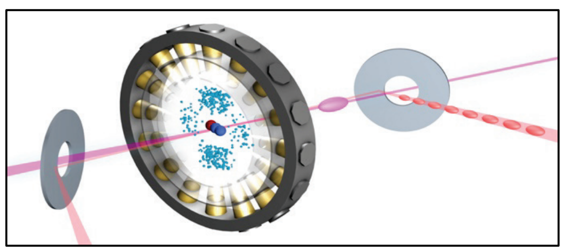


Figure 14. Typical LCLS soft x-ray experiment with high-resolution angle resolved electron spectroscopy.
Source: Image courtesy of Hegazy Kareen, Stanford University, and Ryan Coffee, SLAC National Accelerator Laboratory.

A new superconducting accelerator structure for LCLS-II (-HE) will enable repetition of measurements every microsecond, in a continuous stream of 1 MHz operation. This boost in frame rate portends a potential for a ten-thousandfold increase in the rate of data acquisition, moving the facility from the current ~100 MB/s to the scale of TB/s. This LCLS-II upgrade will give an incredible boost in source repetition rate and thus enable a transformative increase in the rate of "result acquisition"—a phrase used to distinguish the desire for results from the misconception that an increase in acquired data is desired. Beyond simply accelerating the acquisition of aggregate data, the biggest gains will be delivered by an ability to analyze data on the fly, intercepting data as they stream through sensor-local analysis and inference hardware such as EdgeTPUs and FPGAs.

Implementing analysis immediately at the detector is dubbed EdgeML, foreseeing the advantages of ultra-low latency and high-throughput machine-learned inference engines over conventional deterministic computation. EdgeML will provide autonomous data routing decisions that are performed as early in the process as the firmware on the sensor FPGAs. Decisions should be made locally, as near to the sensor as possible; as rapidly as possible; and with an awareness of the local data environment. Data that are destined for disposal should not be transferred.

Multiplex analysis and ghost imaging

Light sources traditionally measure quantities in space, time, or spectrum by raster scanning a probe across the parameter of interest, for example, sequentially probing each pixel of a spatial sample or scanning the x-ray energy for a spectral measurement. However, it is also possible to measure multiple points simultaneously. These "multiplex" schemes provide several benefits: Felgett's advantage improves the signal-to-noise ratio and compressive sensing reduces the time required to scan sparse samples. Finally, creating the sharp features required for a raster scan can be expensive, time-consuming, or impossible at light sources; multiplex schemes instead take advantage of the natural randomness of the standard source. For instance, an x-ray beamline could install a cheap spectrometer in lieu of an expensive monochromator to obtain the same scientific results. Modern implementations of multiplex schemes such as ghost imaging and adaptive illumination rely heavily on ML methods, including sparse fitting with priors, RL, and Bayesian optimization.

Examples of multiplex measurements are found widely across SLAC's light sources. A ghost imaging application uses the jitter of the cathode drive laser to passively measure the cathode quantum efficiency without interfering with operations. On the x-ray side, ghost imaging can exploit the randomness of the self-amplified spontaneous emission (SASE) process to replace pump–probe scans in the time domain [18]; spectral measurements sharpen resolution in the frequency domain for resonant inelastic x-ray scattering, self-seeding, and attosecond spectroscopy (e.g., see sections on Bayesian Signal Separation and Unmixing and Multimodal Attosecond Spectroscopy). At the Stanford Synchrotron Radiation Lightsource (SSRL), adaptive schemes dynamically change the illumination pattern to maximize the information contained in each measurement.

Bayesian signal separation and unmixing

LCLS is pursuing signal unmixing tasks with applications for linear and nonlinear x-ray spectroscopy, for example, isolating high-resolution monochromatic absorption from samples excited by a polychromatic XFEL beam or isolating the two photon signals in double core-hole photoemission. Unmixing, or signal separation, is an inverse problem that needs strong regularization and constraint to recover physically realistic results. Researchers at Stanford University use approximate Bayesian inference, specifically variational inference, to both regularize and quantify the uncertainty of the unmixed signals. These methods scale in a computationally similar way to NNs and enjoy similar hardware acceleration (e.g., on a GPU, which enables unmixing for large-scale datasets containing billions of measurements).

Multimodal attosecond spectroscopy

The new attosecond capabilities of LCLS open up the possibility of measurements of electronic motion on a natural timescale. A useful and well-developed method for making such measurements is attosecond transient absorption spectroscopy. In conventional transient absorption spectroscopy, a well-controlled central x-ray photon energy of narrow bandwidth is scanned while the absorption is recorded at consecutive points along the energy scan. The Fourier bandwidth theorem, however, precludes the application of this conventional technique in the attosecond regime at the LCLS; owing to the large bandwidth attosecond pulses, it is not possible to isolate a narrow bandwidth central photon energy. To overcome this limit, we rely rather on the inherent instabilities of SASE operation, measured for every individual shot, to make a statistical analysis rather than simply recording averages. Similar to ghost imaging and ptychography methods, this technique recovers sub-bandwidth spectral resolution in absorption spectra. A correlation between inherent fluctuations in the enhanced SASE spectrum and a single measured value such as total x-ray absorption forms a convex solution space that optimization methods such as alternating direction method of multipliers are efficient in solving. Thus, the spectral response of a sample is reconstructed at a higher resolution than the bandwidth of the incident pulse.

Machine learning in materials science

The LCLS-II is also targeting a demonstration of general tools for ML in fundamental materials science. This effort will bind together a partnership that includes the Stanford Institute for Materials and Energy Sciences, SSRL, LCLS, and the Stanford School of Earth, Energy and Environmental Sciences. It will also engage with the SETI Institute to extend its unique perspective on AI/ML for astrobiology into the energy science domain. The collaboration aims to extend established and emerging AI/ML tools developed by industry into new areas of materials research. Benchmarked with fundamental problems within SLAC's mission, the group will disseminate tools to the materials community as an open-source library.

AI/ML for SSRL X-ray Science

Major investments have been made at SSRL over the past decade to build brighter, undulator-based beamlines and acquire faster and larger detectors with higher dynamic range. The rate and complexity of data collected following these upgrades has risen dramatically; however, the emergence of new scientific breakthroughs and insights has lagged. For example, the rate of peer-reviewed publications from scientists at SSRL and other light sources has nearly doubled in the past decade, but the volume of data generated has increased by over an order of magnitude in the same time frame.

The rate of discoveries is slower than hoped because the rapid rise of data has not been accompanied by new tools for analysis and management; the volume of data produced by experiments has consequently made all three steps of the discovery cycle—hypothesis building (predicting), data collection (performing), and knowledge extraction (finding)—suboptimal. Discoveries are still extracted mostly by human curators using techniques from the days when data were scarce and researchers had time to decide which experiments to perform. That approach is increasingly challenged by the rate of data generation exceeding our human ability to analyze (or even visualize) data as they are generated and by the complexity (i.e., increasing dimensionality) of data, exceeding human perception to find interesting trends hidden in it.

We have, therefore, undertaken a pilot program that leverages emerging advances in ML and AI to build new data analytics tools to improve the productivity of the three stages of the discovery cycle (shown in the schematic in Figure 15).

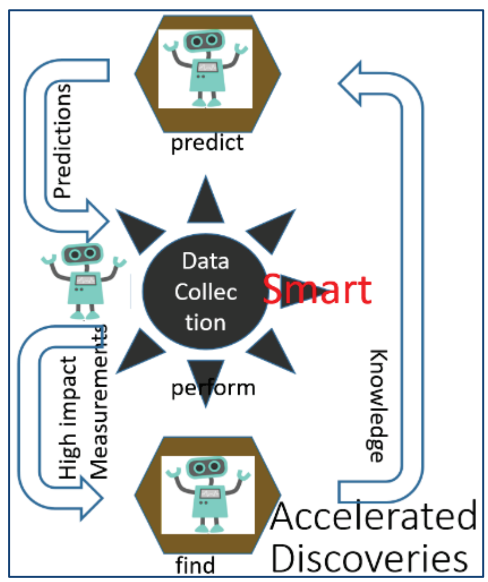


Figure 15. Schematic visualization of an AI/ML-driven discovery cycle that SSRL is beginning to develop and deploy at beamlines. *Source: Image courtesy of Apurva Mehta, SLAC National Accelerator Laboratory.*

Machine learning predictions

Increasingly large data-generating experiments at SSRL are either searches for new materials in complex composition-processing combinatorial spaces, such as high-entropy alloys with superior high-temperature structural properties, or reaction, transformation, and aging of materials in complex environments such as charging and discharging of lithium-ion batteries. These complex experimental spaces are too vast to search blindly. For example, there are easily over a billion compositions for possible high-entropy alloys that contain at least 5 elements from a pool of 25 cheap and earth-friendly elements. Over the past decade, large computational efforts such as the Materials Project and the Open Quantum Materials Database have begun to calculate the stability of a few of these complex alloys, but even the largest such database has fewer than a million calculations. Moreover, the computations for material stability are usually for 0K structures without defects, whereas real materials frequently deviate from these ideals. The inclusion of experimental observations complements computations. Even when the databases are incomplete, trends are present that can be exploited to predict the missing alloys, but finding these trends is beyond human perception. ML, on the other hand, is designed to find subtle trends in complex data. An ongoing collaboration among SSRL, Argonne National Laboratory, and Citrine Informatics is developing ML predictions based on computational and experimental databases to predict new compositionally complex alloys with a range of properties suitable for applications, from wear-resistant coatings to high-performance thermoelectrics. Figure 16 shows a comparison of ML predictions with experimental validation from that project for complex glass-forming alloys [19].

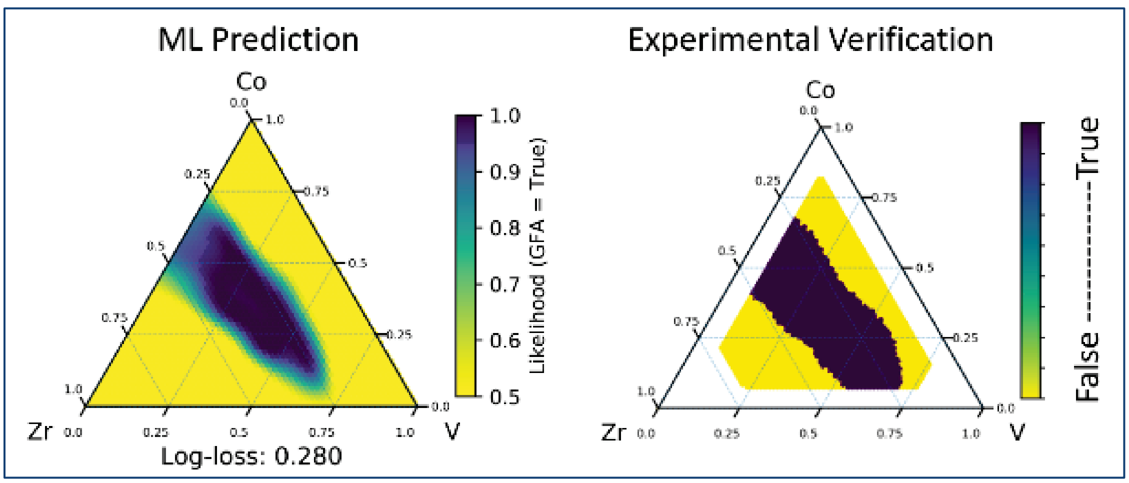


Figure 16. Machine–learned prediction of amorphous compositionally complex alloys and experimental verification. *Source: F. Ren et al., Sci. Adv. 4(4) (2018), eaaq1566. Distributed under terms of Creative Commons Attribution License.*

AI/ML-driven smart data collection

An unbiased experimental campaign must explore regions of the parameter space larger than those predicted by a model because, often, negative predictions and the boundaries between negative and positive predictions contain invaluable information. The traditional approach is a brute-force scan of the parameter space; however, as the parameter space becomes larger, even with increased throughput, such an approach becomes increasingly unsustainable. The brute-force approach also produces large amounts of data, only a small fraction of which is useful. A smarter search strategy is to continuously leverage the emerging new information as measurements are performed to construct a model of the experimental space and subsequently use that model to decide on the next measurement of maximum utility. Iterating such a Bayesian real-time decision-making process may locate the information-rich regions of the experimental space with far fewer measurements. The Bayesian approach is particularly powerful for searching a parameter space that is very poorly predicted. This is often the case at SSRL for experiments in which process/operating conditions play a significant role.

Recently, such an approach was developed to discover synthesis conditions for a specific target size, phase, and polydispersity for a class of nanocrystal catalyst materials for which very few reliable prior predictions existed. To accomplish this task, a suite of ML classification and regression algorithms, which collectively automates the interpretation of small-angle x-ray scattering data, was developed. These algorithms enabled machine interpretation of the results of a nanocrystal synthesis measured in situ using x-ray scattering methods. The output of the ML classification was incorporated in a Bayesian algorithm, which automated the experimental design to efficiently search the enormous parameter space associated with colloidal synthesis. The design algorithm attempts to predict a synthesis for a specified target or class of targets. Based upon its acquisition function value, it either attempts that synthesis or prioritizes gathering more training data. After each subsequent experiment, the algorithm is retrained and queried again for new results. This collection of algorithms is contained within a software package designed to control the autonomous synthesis, which includes clients to interface with the beamline data acquisition, flow reactor, data-handling workflows, and third-party analytics clients. The autonomous synthesis workflow is shown in Figure 17.

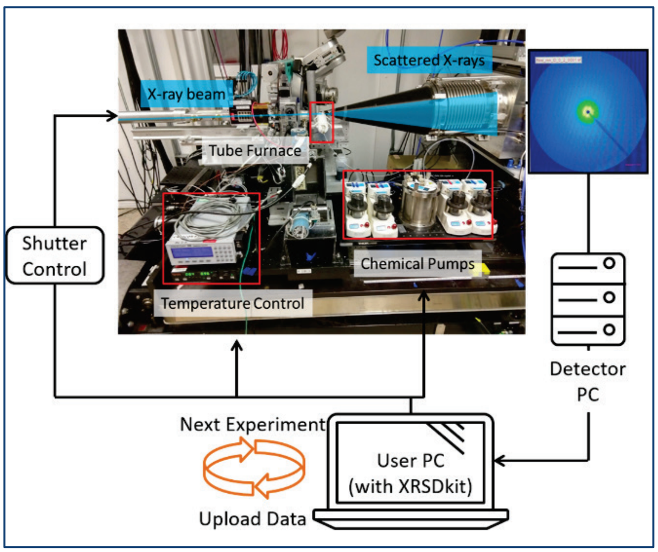


Figure 17. Data collection and computation workflow for an autonomous experiment designed to discover process conditions for a nanocrystal catalyst material of specific size and polydispersity. *Source: Apurva Mehta, SLAC National Accelerator Laboratory, Stanford University.*

An alternative approach uses RL to optimize data acquisition. To train an RL "agent," a simulation environment provides the agent with sequential opportunities to change acquisition parameters.

After each measurement, the environment rewards the agent based on the measurement error, teaching the agent to focus on the highest-value measurements. In a first test on x-ray fluorescence images, the adaptive RL agent significantly outperformed a standard raster scan [20].

Rapid information extraction

Extracting information rapidly and with minimal human intervention—but with performance equal to and in many cases surpassing that of a human—is essential for a more productive AI/ML-driven discovery cycle. Without on-the-fly information extraction, the kind of autonomous experimentation described above cannot happen. Moreover, many other experiments at SSRL produce large amounts of highly multidimensional data very quickly. Often, a human researcher does not have the throughput or even perception to find discoveries buried in such databases. Therefore, multiple efforts at SSRL use unsupervised factorizations and supervised ML to rapidly extract information from large databases. There is a concerted push to push to increase such efforts because they have the potential not only to increase the rate of discoveries but also to produce the types of discoveries that often are missed with the current analysis approach.

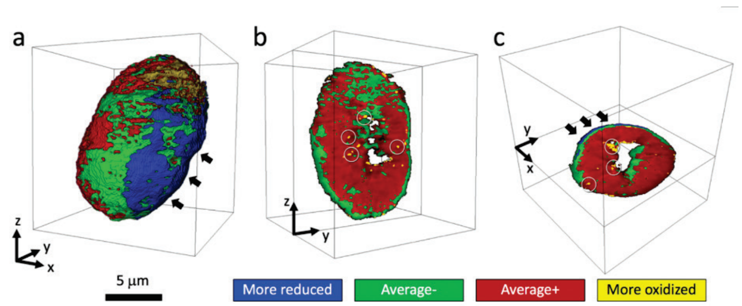


Figure 18. 3D distribution of four cobalt phases discovered from a hybrid supervised and unsupervised analysis of a collection of x-ray absorption spectra collected on a multiply cycled and partially charged battery cathode particle. The red and green phases were anticipated, but the blue and the yellow phases were unexpected. *Source: Y. Mao, et al. Adv. Func. Mater. 29, 1900247 (2019) Permission granted by Wiley.*

Figure 18 illustrates an example of such an effort applied to understanding modification of battery cathode particles under operating conditions (i.e., charging/discharging of lithium-ion batteries). The datasets contained x-ray absorption spectra produced by nanoresolution full-field x-ray absorption near edge structure (XANES) microscopy. A full-field XANES microscope produces over 4 million spectra in 10 minutes. A hybrid supervised and unsupervised classification approach was developed to search for unanticipated and undesired chemical phases in the particles that have gone through more harsh conditions. The researchers found that, in addition to the anticipated chemical phases, two different types of chemical outliers coexist in the aggressively cycled particle (labeled "more reduced" and "more oxidized" in Figure 18.) These two different types of chemical outliers have different chemical fingerprints and different spatial distribution. They were, therefore, attributed to different side reactions that need to be mitigated for improvement in the overall device (i.e., battery) performance [21].

Denoising

Deep learning methods for removing noise from images may be useful for measurements for which diffraction intensity is limited, such as radiation-sensitive samples or diffuse scattering measurements. A CNN can find correlations in the intensity distributions between a pair of images and thus predict an image representing the shared structure with any uncorrelated noise removed. Using this technique, we analyzed diffraction images from a time-resolved pump–probe measurement at SSRL and found an improvement in the signal-to-noise in diffraction images in Figure 19.

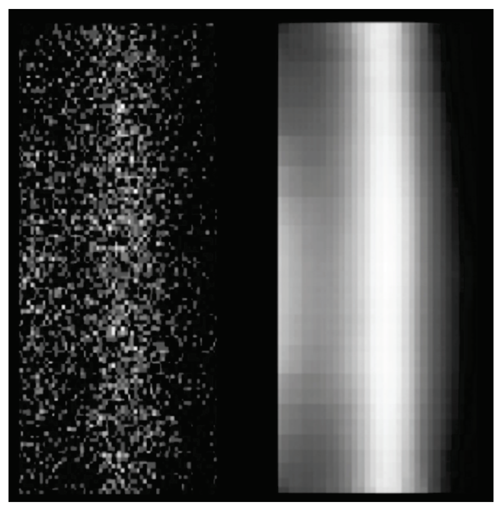


Figure 19. (Left) Diffraction from a polycrystalline gold sample. (Right) Output of a neural network. *Source: Clara Nyby, Stanford University.*

AI/ML at NSLS-II

ML and AI are becoming increasingly critical tools at NSLS-II. Increased photon intensity makes many experiments, even complex ones, high-throughput. This in turn demands intelligent execution of experiments and frequently requires near-instantaneous analysis and decision-making. It is becoming less and less possible for humans to play a real-time role in this loop. At the same time, as the newest light source in the United States, NSLS-II has developed a modern data acquisition architecture that has recently been adopted as the DOE–complex-wide standard. By design, this architecture is the perfect foundation to build the AI and ML solutions that will transform synchrotron science.

The NSLS-II is investing in the development and adaptation of AI and ML algorithms and packages in the following areas:

1. Analysis of synchrotron data of various types, both to produce physically meaningful information and to identify and remove experimental artifacts;
2. Experiment automation including the coupling of such algorithms to the beamline data acquisition system;
3. The combination of the first two areas to facilitate scientific investigations that are only possible through the application of such techniques; and
4. Accelerator operations, to enable fault prevention and fast fault diagnosis for accelerator issues.

Within these areas, we think it is *essential* that any proposed solutions using AI and ML techniques develop an integrated architecture for the nonexpert to enable the use of the solution at the beamline or within the accelerator control room.

Current AI/ML projects at NSLS-II

Autonomous experiments

The NSLS-II, in collaboration with Brookhaven National Laboratory's Computational Science Initiative and CAMERA, has been pioneering autonomous x-ray scattering experiments at Beamline 11-BM (CMS) Complex Materials Scattering beamline, based on a workflow consisting of automated data acquisition, real-time data processing, and CAMERA algorithms for autonomous decision-making. This is being applied to the field of additive manufacturing to *autonomously make or modify materials*. For example, this work will investigate the particle dynamics, dispersion, and alignment of anisotropic "nanofillers" (e.g., metallic nanorods, clay nanosheets) in a printable polymer–nanofiller composite ink as it is extruded from a nozzle, is deposited, and solidifies. The AI/ML-guided exploration of the vast material and process parameter space will be used to unravel the correlations among control parameters, material dynamics, and final structure to understand the underlying physical mechanisms for material evolution and formation during the out-of-equilibrium additive manufacturing process.

Spectroscopic data analysis

The NSLS-II and the CFN are currently applying the use of AI and ML techniques to x-ray absorption spectroscopy (XAS) data. XAS is a powerful tool to probe electronic and structural features in a wide variety of energy materials. Currently, synchrotron sources can collect XAS data in a high-throughput mode faster than data can be analyzed with physics-based models, which are key to determining the underlying structures of materials. Recent research at NSLS-II has prototyped AI/ML engines to extract structural descriptors from XAS spectra using a first-principles simulation-guided ML approach, thereby allowing the researcher rapid on-line data interpretation. This approach is being developed as a future robust, general tool for XAS data analysis, enabling efficient data analysis to match the high-throughput beamline capability and the implementation of real-time feedback for self-guided experimentation.

Proposed future AI/ML projects at NSLS-II

Accelerator operations

Machine reliability is vitally important for the productivity of a synchrotron facility and is a key metric of the accelerator. Since FY 2017, NSLS-II has kept its annual reliability above 96%. As the accelerator and its components age, common points of failure can change. If not anticipated and adapted to, faults and downtime can increase, and reliability will suffer. Because of the modern controls environment at NSLS-II, all machine parameters are actively archived (>106,000 individual variables). The first 5 years of operation have resulted in an unparalleled dataset that can be used for training ML algorithms.

Currently, alarm thresholds are established on parameters to alert the operator to a fault. While this arrangement prevents some beam dumps, it does not account for trends and patterns that do not approach a threshold. ML and AI techniques can be engaged to learn from the archived data to characterize patterns that can predict accelerator faults and therefore alert the operator, or the algorithm, to a fault *before it occurs* so corrective action can be taken. Such a system would increase the reliability and therefore the productivity of the NSLS-II.

Materials synthesis

To dramatically speed up the feedback loop between experiment and theory, leading to accelerated materials discovery, NNs can be used to encode simulations of the shape distribution of nanocrystals by physics-based simulation. These simulations are sufficiently complex and time-consuming that they cannot be performed in real time. If NNs are trained on many simulation runs, they can then be used in real time to determine the shape distribution from the experimentally measured total scattering pattern, allowing the possibility of real-time materials discovery.

Artifact detection

One important possible application for AI and ML techniques is the detection and quantification of systematic errors in data collection. For example, coherent x-ray diffraction patterns collected for photon correlation spectroscopy are taken successively over many hours to measure dynamics. Currently, any systematic errors (e.g., beam or sample drifts) in such datasets are not found until the whole dataset is analyzed; the delay can result in highly inefficient use of beamtime. Furthermore, the sources of such errors are often difficult to identify and hence difficult to remove. We will train ML algorithms to identify and classify such artifacts, training on both simulated coherent diffraction patterns and actual collected data. This approach will provide the experimenter with a real-time measure of data quality and, if systematic errors are present, indicators of their origins.

Complex beamline optimization

Diffraction-limited synchrotrons such as NSLS-II have enabled such techniques as ultra-high-resolution inelastic x-ray scattering. The SIX (2-ID) beamline at NSLS-II currently leads the world in experimental resolution, and a key component of that resolution is the spectrometer. To obtain such resolution, however, requires careful alignment of the spectrometer's many optical elements by an expert scientist. AI/ML techniques can be used to develop an autonomous system for aligning such spectrometers, leveraging NSLS-II expertise in beamline optical simulations using coherent wave-front propagation. Such forward simulations can be performed to explore the whole parameter space of the optical component alignment to generate a dataset on which NNs can be trained. Initially, the resulting network will be used to more quickly guide a scientist to the perfect alignment. Ultimately, it will serve as a basis for an autonomous alignment system that will not only align the beamline but also maintain it at an optimum alignment throughout the course of the experiment. Once developed, this approach should be easily generalizable to other beamlines across the DOE complex.

Additional points

For AI/ML techniques to be successful, there needs to be a sociological change in how experimenters approach experiments. Scientists developing AI/ML techniques will need to educate the synchrotron community to build confidence that these techniques offer accurate and robust results. Doing so will require careful determination of the true uncertainties present in such analysis and the applicability of these techniques to particular experimental problems. To build confidence in their use, AI/ML algorithms will have to be benchmarked against trusted datasets.

Finally, the use of AI/ML techniques, with their inherent need for large training datasets, should and will drive a need for uniformity of software environments across the DOE user facility complex, for both data-sharing standards and policies and metadata standards. Initial steps are being taken in this regard with the adoption of the open-source, python-based BlueSky data acquisition software stack. This effort ties into existing large, highly sophisticated python-based ML packages. Development of AI/ML for synchrotron science should be carried out with agreed-upon programming standards and environments and ideally with a coherent strategy across the light sources so that the development work can be leveraged to the greatest extent possible.

AI/ML at the Neutron Facilities

Along with SNS, the High Flux Isotope Reactor (HFIR) at ORNL are world-class neutron scattering user facilities supported by DOE BES. There are many benefits to the application of AI and ML techniques to neutron scattering experiments at these facilities, many of which are only beginning to be understood. This section briefly describes several areas in which the application of AI/ML is expected to be most beneficial to neutron scattering facility users.

Although neutron scattering experiments produce fewer data than experiments at light sources, the volume of data is still significantly larger than what can be digested and understood easily by humans. The data are also rich in physics information that may not be easily analyzed in coupled analyses in a straightforward way. Thus, AI and ML applied to neutron scattering experiments can significantly improve the rate at which user data can be analyzed while simultaneously revealing new physics. For example, unsupervised ML techniques (e.g., DBSCAN and other clustering algorithms) can be used to extract common features in the data without having to define what these features are a priori. One common approach is to apply these unsupervised ML techniques to extract features such as phase transitions or data outliers. Such algorithms have recently been shown to reveal the presence of different phases of materials and can also be used to alert the experiment team or facility staff of unexpected, changing conditions during an experiment.

Supervised ML techniques can also be used. In this case, labeled training and test data are needed to train a network for a given task. This approach has been used in a DOE Data Demo to categorize single-crystal diffuse scattering trained on simulated diffraction patterns. The same approach was used to categorize images on an x-ray detector as isotropic or background. This opens up the possibility of automatically detecting “bad” background data and using the tags as metadata. Using these approaches, it would in principle be possible to create a live diagnostics engine that assesses data quality and lets users know when there are problems with the instrument. By training the ML process with standard-looking data projections, it would be possible to determine whether the acquisition is proceeding appropriately.

ML can also be used to speed up computationally expensive first-principles calculations by training and using NNs to evaluate larger models more rapidly. This will give users access to the needed model sizes, for example, to explain diffuse neutron scattering data and eventually use these techniques in near real time. Neutron reflectivity and quasielastic neutron scattering are two areas in which AI/ML tools have been used by ORNL staff and where they show significant promise. The ICEMAN (Integrated Computational Environment for Modeling and Analysis) project provides high-order, nonlinear, regression-based ML tools for quasielastic neutron scattering model fitting in its QClimax package, which allows users to consider the deeper physical meaning hiding in their data more quickly and in highly natural ways. Figure 20 shows a similar application of regression-based ML techniques for neutron reflectivity, in which the k-nearest neighbors (k-NN) algorithm was able to successfully model the input data with one-, two-, and three-layer thin-film models. A similar approach for classifying small-angle neutron scattering (SANS) data for the most appropriate model to use for data analysis has also been developed.

The application of super-resolution methods to spectroscopy and SANS is also promising and currently under way for multiple instruments at the SNS FTS. Super-resolution methods form a class of techniques that can improve signal quality by better exploiting the information in the signal through various statistical means. Super-resolution techniques have advanced scientific fields such as fluorescence microscopy and biology; the 2014 Nobel Prize in Chemistry was awarded for super-resolution imaging. Rare attempts at “deconvolution” for neutron spectroscopy earlier in the 1990s have not yet found wide application. Pilot studies for direct-geometry spectrometers at the FTS have shown that traditional fusion–deblurring–denoising techniques from super-resolution imagery are applicable to obtain sharper dispersions and phonon densities of states with a resolution improvement of 5× greater than normal. It was also demonstrated that super-resolution imaging techniques can be used to refine theoretical material models such as spin-coupling coefficients in spin-wave models. In SANS, super-resolution deep CNNs have been shown to reproduce intensity curves as well as a bicubic method, but super-resolution performed better than the bicubic methods when the intensity reflected physical features such as a beam stop. Thus, it is clear that additional research and application of such techniques across multiple instrument types will provide significant resolution improvements, especially for those instruments with expensive or otherwise impractical hardware upgrades.

The utility of super-resolution methods to neutron spin-echo measurements has not been established but may be especially beneficial because of the flux constraints encountered by those instruments. Super-resolution methods are also promising for instruments planned for the STS. For engineering and powder diffractometers, super-resolution methods will make it possible to better utilize the information from higher- and lower-angle detector banks to improve diffraction patterns. Similarly, in single-crystal diffractometers, super-resolution will help improve the accuracy of modeling of peak profiles, making it easier to obtain accurate integrated intensities of diffraction peaks. Similar and related improvements are expected for other STS instruments.

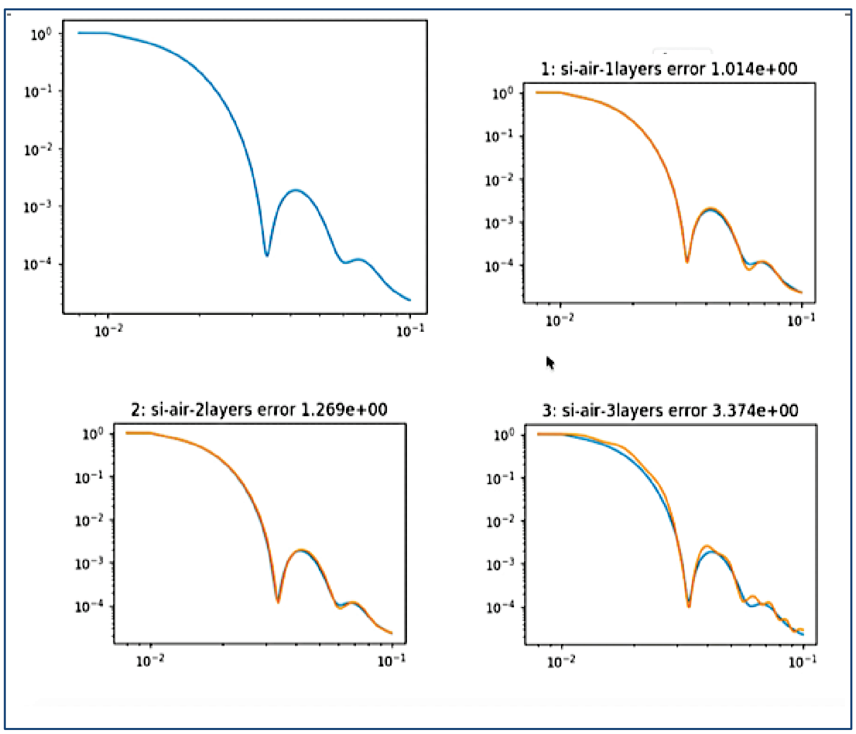


Figure 20. An application of regression-based ML for neutron reflectivity. The k-NN algorithm successfully modeled input data with thin-film models. *Source: Mathieu Doucet, Oak Ridge National Laboratory.*

Understanding neutron measurements of complex emergent materials such as quantum magnets requires the integration of multimodal datasets and application of computation and modeling that go well beyond traditional fitting approaches. As a result, current experiments at SNS and HFIR take from weeks to months to analyze. Advances in ML combined with high-performance simulations of models could provide a new way to meet this need. By training on wide ranges of realistic simulations, ML resources capable of interpreting data, quantifying the uncertainty of model parameters, and guiding experimental strategies can be prepared before experiments are started. They can be integrated into the data acquisition process to provide real-time feedback and steering, as well as removal of experimental artifacts and noise. ML combined with modeling simplifies the complexity of materials experiments and high experimental data volumes by providing the researcher the essential science with a direct visualization of how the material behaves. It can do so by integrating and distilling the information content of dissimilar datasets (e.g., different temperatures, applied fields, heat capacity, and magnetic susceptibility) and working directly on the difficult-to-visualize high-dimensional data collected in scattering experiments (e.g., 3D diffuse scattering and 4D single-crystal inelastic datasets).

The first demonstrations of potential applications for diffuse and inelastic scattering datasets have been completed. Figure 21 shows a 3D diffuse scattering dataset collected on a spin-liquid material at the CORELLI diffractometer at SNS. ML was used to train an NN to interpret the data, as well as to identify the experimental features and phases over five independent dimensions of model interactions/parameters. This approach enables the identification and automatic removal of experimental artifacts and noise. Validation of the model was undertaken by comparing its results with data taken in applied fields and at temperatures different from that used to extract the model. Dissimilar datasets, here involving heat capacity and susceptibility, were used to provide a unique solution. Figure 21(d) shows a visualization of the formation of a complex glass phase in the material determined from the combined data modeling. This demonstrates how new science that goes beyond current theory and analysis can be realized by the uniquely close combination of modeling and data possible with such new approaches. It provides the means to solve long-standing problems previously considered too complex.

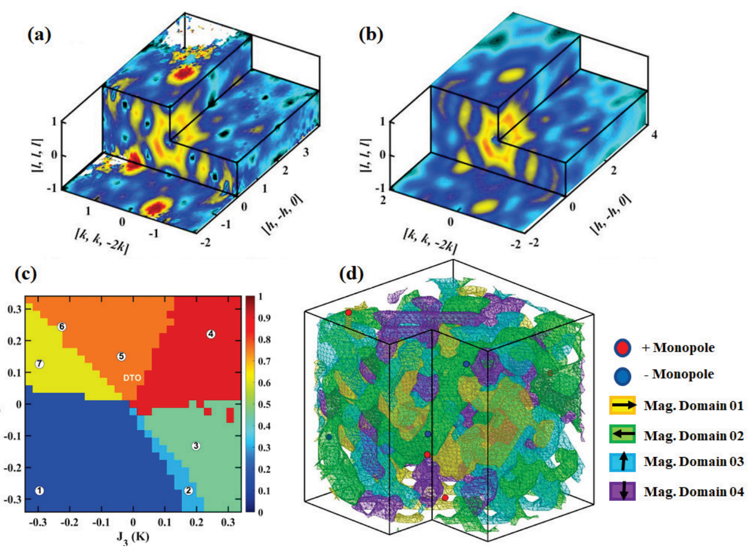


Figure 21. Machine learning is used to train autoencoders resulting in visualization of glass formation in a cooled spin liquid. (a) A rendering of a 3D diffuse scattering measurement of the spin correlations in the gauge spin liquid $Dy_2Ti_2O_7$ measured as 1 of 40 datasets on CORELLI. (b) The result of denoising and removal of experimental artifacts by the autoencoder, which provides optimized model parameters from the 3D dataset. (c) A slice through the model space of interaction parameters where ML has identified distinct physical behaviors (i.e., phases) and learned features. (d) Visualization of the model in real space showing monopole localization. *Source: Alan Tennant and Mathieu Doucet, Oak Ridge National Laboratory.*

Next-generation neutron imaging instruments, such as VENUS at the SNS, will produce large datasets that may be impossible to analyze and understand without advanced AI/ML tools. Traditional attenuation-based imaging systems that use neutrons from reactors, such as the imaging instrument at HFIR, produce manageable amounts of scalar data. However, VENUS, which will provide TOF 3D Bragg-edge imaging, will produce data with six unknowns per voxel. Important imaging information is therefore expected to be sparse or nearly nonexistent in such huge datasets. In addition to data reduction and analysis, it is expected that AI/ML will help in planning imaging experiments. For example, during computed tomography scans, AI/ML can determine which angles are more important than others to image a material's microstructures. Real-time feedback to AI/ML "observers" could also be used to guide experiments in progress, which would significantly boost both experiment and data quality. One major challenge that AI/ML will be critical for addressing is the mapping of applied and residual stress in 3D at micron-scale resolution with 3D Bragg-edge imaging. Presently, both limited boundary conditions and partial information lead to multiple answers and a complex (and incomplete) inverse problem. Using advanced AI/ML techniques that are currently applied in the broader imaging community, along with both modeled and synthetic data, could have significant impacts on materials research because Bragg-edge imaging is nondestructive, fast, and inexpensive compared with more traditional methods.

Efforts to integrate more closely with HPC capabilities and computer science research at ORNL are under way, including the successful Ugly Data Days and AI Expo events, hosted annually to connect staff experiencing problems with AI/ML with AI/ML experts on staff.

The opportunity for AI/ML to impact user productivity and scientific discovery in neutron experiments is substantial; perhaps the greatest gains will be in the integration of AI/ML technology into the full user workflow, from experiment to analysis. Such tight integration with workflow systems would make it possible to scan data archives for missed or previously unrevealed science, to suggest or automatically conduct new data reduction and analysis based on the suggestions of AI/ML engines, and to highlight the most important parts of a user's data by referencing large physics-based models.

3. Cross-Facilities Integration

There are ongoing efforts to integrate capabilities across the SUFs via AI/ML, networking, and advanced math. Coordination and execution are highly collaboration-based and mostly fall under guidance from the Energy Sciences Network (ESnet) and the CAMERA project.

AI/ML at ESnet and NERSC

ESnet and NERSC provide high-speed networking and HPC capabilities, respectively, to BES facility users and their complex workflows. Both facilities are constantly engaged in improving workflow performance engineering, for example, through efforts of science engagement teams working with BES scientists to provide support from design to execution phases for seamless experiment-to-result. Current efforts also include the superfacility project, in which facilities are collaborating to enable near-real-time, end-to-end computing workflows for beamlines, supported by predictable bandwidth across high-speed networks. Going forward, as BES experiments generate increasing data quantities with upgraded instruments, real-time feedback for experimental control and advanced computing and network capabilities will need to support the exponential data growth. AI/ML efforts can play key roles to seamlessly support BES workflows, potentially accelerating the game-changing BES science results. The following sections address questions and answers related to use of AI/ML at ESnet and NERSC.

How can AI/ML change facility operations?

Upgrading to advanced detectors can lead to a data explosion needing quick analysis to discover patterns. For computing and network resources, supporting these *high-speed/high-volume* workflows will require optimized resource allocation based on workflow demands. AI/ML efforts can help find and optimize these distributed resources and provide on-demand availability, storage, and bandwidth. In addition to monitoring workflow health, the AI/ML controllers can help recommend changes to promote energy-efficient, low-power computing and reduce traffic hot spots across the network. Facilities are also investigating the use of AI/ML to ward off potential threats and predict anomalies and component failures that may impact science experiments, proactively mitigating and minimizing downtime. Various ML algorithms designed for real-time and streaming analysis can be developed for these support structures.

Are there any limitations regarding where AI/ML can help?

ESnet and NERSC provide operational support to BES science. At the operation layers, we see more data being generated at high speeds, faster data transfers, and the need for end-to-end performance improvements. Oftentimes, these advances require a few trial-and-error cycles to learn optimal parameters and experimental settings. ESnet is investigating the provision of high-speed telemetry services—packet-level transport characteristics such as packet sizes and timing patterns—that can be used to infer workflow health. Researchers will be able to deduce application and system characteristics and determine if certain parts of the workflow need fine-tuning. Eventually, this analysis will be crucial to support the sciences.

Can AI/ML change the way users approach data acquisition, analysis, and adaptive control?

AI/ML techniques can augment current manual, heuristic-driven data acquisition and adaptive control techniques employed by BES facilities. ESnet and NERSC can exploit AI/ML techniques to improve workflow performance and spot problem areas before they jeopardize workflows. Using learning and control, facilities can learn optimal configurations and "adjust" workflows accordingly to support high-impact science.

Are there limitations to successful progress in AI/ML for data production and analysis at ESnet and NERSC?

Current AI/ML methods are most effective in the regime of supervised learning, for which access to training datasets is a critical requirement. BES, ESnet, and NERSC facilities currently lack standardized tools to capture, label, and share such datasets broadly within their respective user communities or the wider research community.

Are there opportunities to better integrate with ASCR HPC and high-speed networking capabilities for data-intensive experimental and theoretical problems?

NERSC has deployed a relatively new set of AI/ML capabilities on HPC hardware, and BES facilities can engage with NERSC staff to integrate these modern capabilities in their workflows. Besides system-level capabilities, NERSC staff have expertise in AI/ML methods and can assist with pointing BES users to appropriate techniques. ESnet data transfer and workflow scheduling services can currently be leveraged by BES facilities; we expect that the superfacility project will provide a more standardized, integrated interface for end-to-end services. In addition to the points discussed above, ESnet and NERSC are already exploiting AI/ML capabilities to understand their users better as part of their own research agendas.

What aspects of AI/ML are exciting?

AI/ML offers exciting capabilities to provide time-sensitive experiments and reliability among results, along with the potential for resource optimization and cost savings. These will lead to breakthrough research in applying novel AI/ML approaches for time series, improving data processing in streaming ML applications, and controlling infrastructure in real time. From the ESnet perspective, it will be exciting to see how the high-speed telemetry service will provide knowledge inferences regarding what scientists and the corresponding data are doing on the network and whether potential problems can be isolated before they affect the science mission. From the NERSC perspective, AI/ML is being used to solve a number of data analytics problems. A new breed of applications in surrogate modeling and inverse modeling are now emerging at the intersection of classical applied math and modern data science. AI/ML controls applications are very exciting, with applications in both BES device control and improved efficiency of data center operations.

4. CAMERA: Machine Learning and Mathematics across Facilities

Fundamental computational methods across light sources and the NSRCs are needed to extract information from datasets of all sizes, interpret experimental results, and provide near-real-time analysis as data are generated. Advanced algorithms combined with ML can screen candidate materials that are expensive and time-consuming to manufacture, rapidly find optimal solutions to energy-related challenges, and suggest new experiments for scientific discovery. To meet these scientific challenges, BES and ASCR have jointly established CAMERA to build cross-facility cooperation to meet these critical challenges. CAMERA's mission is to develop and couple the required applied mathematics, statistical, and ML techniques to build shared algorithms and software tools to analyze and interpret the data coming out of complex experiments, optimize experiments, and accelerate scientific understanding. The CAMERA project has had a significant impact on the development and adoption of new mathematical technologies combined with ML approaches for DOE light sources, including the ALS, APS, LCLS, NSLS-II, and SSRL, with additional expanding connections with the NSRCs. Within CAMERA, cross-disciplinary teams of applied mathematicians, ML experts, software engineers, and facility scientists formulate models, derive appropriate equations, develop algorithms, build and test prototype codes and ML applications, and deliver usable software. CAMERA has undertaken several AI/ML projects related to light sources, including scattering, imaging, tomography, and autonomous experimental design, even under low–signal-to-noise scenarios. The following sections discuss a few selected examples.

Autonomous Optimization of Experiments

X-ray scattering experiments are often lengthy procedures in which the light source user attempts to find the structure–property relationships of materials, subject to parameters including pressure and temperature. As the number of these parameters grows, the scientist faces the challenge of visualizing and processing the parameter space and related data with the hope of making informed decisions for the next experiment. A common solution is to perform experiments randomly or at discrete predetermined points, an approach that does not take advantage of the information collected in previous experiments. Additionally, frequent monitoring of the experiment allows scientists to react to changes when necessary. An exciting new approach is to steer the experiment through ML optimization, exploiting mathematical optimization to make autonomous decisions based on past experiments and without human interaction. This new approach is encapsulated in the "SMART algorithm" (Figure 22) built by scientists at CAMERA, NSLS-II, and CFN for full autonomous control of the CMS beamline at the synchrotron. The introduction of the SMART algorithm at the NSLS-II CMS beamline has increased the beamline utilization during the autonomous runtime from 20% to 80%. This new capability to choose an optimal path through parameter space using near-real-time analysis and ML will have a significant impact on the scientific output of DOE facilities.

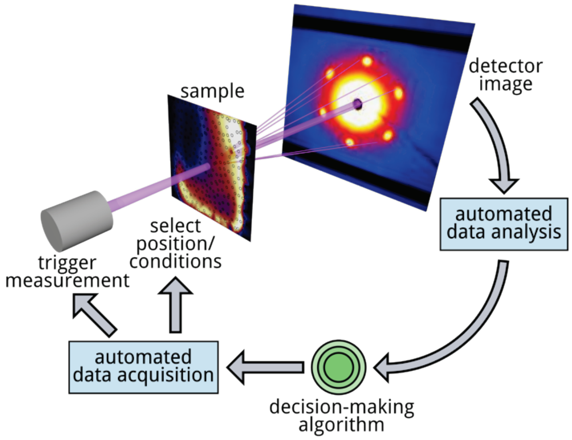


Figure 22. CAMERA schematic of an autonomous x-ray scattering experiment. When the measurement is performed, the data acquisition and processing occur automatically. From the processed data, the SMART algorithm selects the next measurement parameters. *Source: Kevin Yager, Brookhaven National Laboratory.*

Machine Learning for 3D Reconstruction and Inverse Problems

The light sources allow for ever-more-complex sample environments, and users require ever-faster detection and measurements. The development of high-frame-rate time-resolved tomography measurements allows deeper insights into dynamic processes such as fracture in extreme environments or manufacturing of complex materials. High-frame-rate detectors, together with high-flux beamlines, have profoundly increased the possible frame rates and hence the time resolution of traditional tomography. To extract the most information possible from these new capabilities, CAMERA has recently developed ML mixed-scale dense (MS-D) NN schemes that significantly increase the time resolution of traditional tomography beamlines by using existing beamline technology combined with deep learning networks. The goal is to minimize damage to samples and enable advanced dynamic experiments by acquiring tomographic scans at very low x-ray doses. While these scans can be taken quickly, resulting images are typically noisy; hence, a method is needed to take noisy input data and reconstruct higher-resolution images. Rather than follow the traditional scheme of downscaling and upscaling to capture features at different scales, CAMERA's MS-D architecture uses dilated convolutions to capture and preserve additional features. The training requires a low-resolution reconstruction as the input to the MS-D and high-resolution reconstruction as the output to train. The NN training is required only for the first time step; sequential time steps need only the low-resolution input and therefore fewer projections. Figure 23 shows the CAMERA MS-D, which is now being used by scientists across fields such as biology, pattern recognition, EM, tomography for metallic composites, MRI scans, segmentation of satellite images, and sonar imagery.

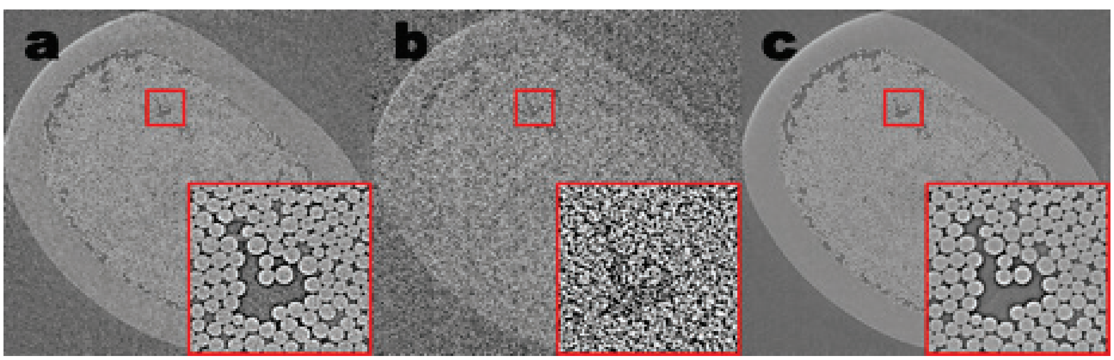


Figure 23. Tomographic images of a fiber-reinforced minicomposite, reconstructed using (a) 1,024 projections and (b) 128 projections and the result from the (c) MS-D training. *Source: Jamie Sethian, University of California–Berkeley.*

Machine Learning for Scientific Pattern Recognition

Images in real space or reciprocal space are a common occurrence in scientific data collection. ML provides exceptionally powerful tools for extracting and categorizing natural image data. However, for scientific data, some common approaches must be reexamined and developed further. Given images produced from experiments using x-rays, neutrons, or electrons, a major task is to detect and extract characteristics of imaged structures. This is typically done through painstaking manual segmentation by a domain expert; it is a costly and time-consuming procedure that cannot scale to high-throughput experiments or high frame rates.

In collaboration with the ALS, the National Center for Electron Microscopy at TMF, and the Berkeley Institute for Data Sciences, CAMERA has built automatic algorithms to quickly extract and measure patterns from scientific image data. These algorithmic tools, which include classical image processing, geometric priors, partial differential equations, deep learning, and generalized physics-specific ML, are in use across a wide field of applications. They include the following:

Analyzing micro-CT images

A major task in processing micro-CT data is to detect and quantify properties of imaged solids, as a step toward assessing the quality of materials and measuring microstructures. Challenges include dealing with corrupted scans, reconstructing artifacts, and multiphase volumes. Much of this metrology requires tools that offer flexibility, a variety of algorithms, and efficient implementation to allow for fast iterations and scalability to data streams.

To address materials metrology through micro-CT experiments, CAMERA has built tools to automatically extract structure. In the context of analyzing CMCs for microstructural damage, these tools can process large numbers of images to assess:

- number of components, including their absence as part of detected defects;
- deformation and failure under tension and high temperature; and
- damage due to preparation and/or loads of samples.

Image quality has been enhanced with the design of scalable 3D filtering algorithms based on anisotropic diffusion and mathematical morphology to emphasize contrast and edge maps. These algorithms handle data streaming and can load from out-of-core sources, and the resulting software tool enables parallel processing of large datasets, removing random access memory–based constraints.

Separation of the dense material from the background involves volume partitioning into solid-phase and interstitial regions, using graph-based models based on adaptive statistical merging predicated on intensity levels and voxel vicinity that runs in linear time. These methods are combined with unsupervised algorithms (e.g., fast clustering approaches including SLIC [simple linear iterative clustering] and Felzenszwalb efficient graph-based image segmentation) and supervised algorithms (e.g., random forest, support vector machine, multilayer perceptron, CNN, and recurrent NN).

Extraction of target microstructures has been conducted using priors and geometric constraints to reduce the size of the search space with regard to the pattern to be detected. For example, to identify fibers from high-resolution CMC images, it is possible to model fiber cross sections as an ellipse and define the fiber detection as a search problem. Because the fiber cross section is consistent, a variant of template matching can be used to search for fibers in some cases. It depends on two main steps: first, to define similarity metrics between prototypes and local regions, and second, to determine the best matches. When the fiber contrast (coating) is low and/or the voxel resolution is poor, more advanced approaches using CNNs are necessary. They often drastically reduce processing time for a material to a few minutes. Recent investigations and accomplishments also include characterization of carbon textiles in collaboration with ALS and the National Aeronautics and Space Administration.

PyCBIR: Content-based image retrieval

CAMERA has built a new recommendation system for content-based image retrieval (CBIR), pyCBIR, which allows for scientific image retrieval based on pictorial similarity. This open-source tool is capable of retrieving relevant images using datasets across science domains. CAMERA's package has been used to find the closest matches of scattering data to curated libraries of stored images, from GISAXS and micro-CT to optical microscopy and photography (Figure 24). PyCBIR provides real-time image retrieval using a compact data representation that leverages historical data tagged by domain experts and presents an associated confidence metric (e.g., "class membership") for each image. The latest version of pyCBIR is available at http://bit.ly/aimagesearch.

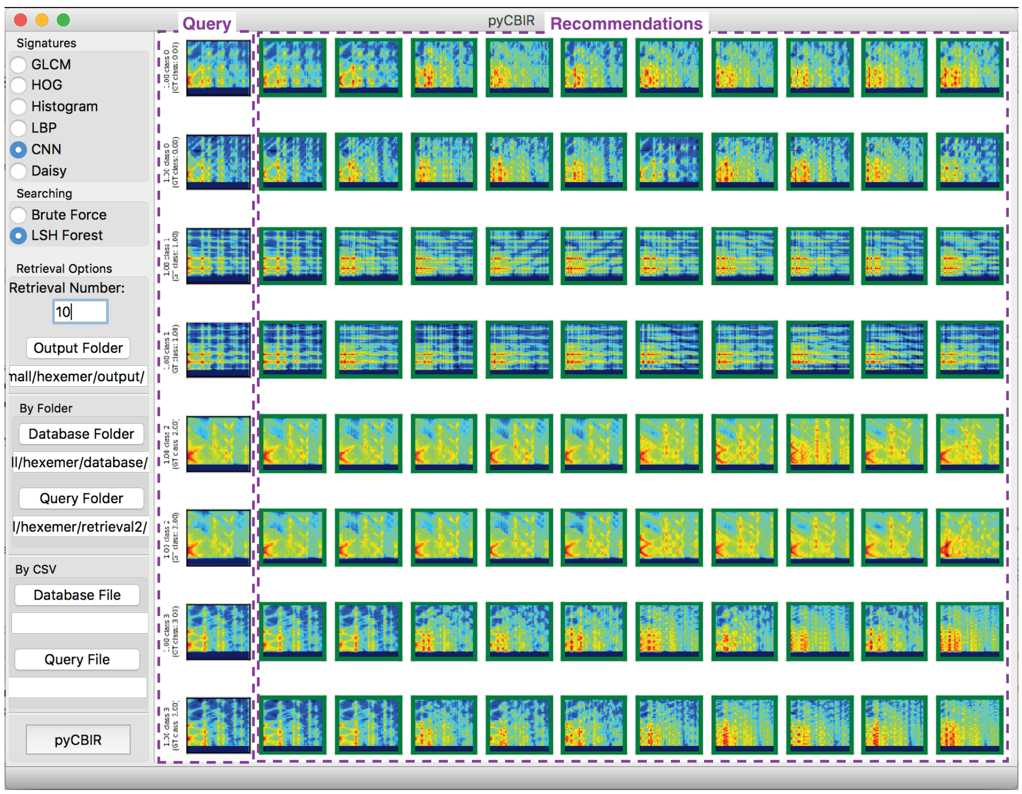


Figure 24. Developing algorithms to categorize millions of GISAXS patterns. | *Distributed under the MDPI Open Access Information and Policy*

5. Summary and Conclusions

The ability of SUFs to reach their full potential in terms of user experience, resilience, and optimized operation and performance will require innovations to solve a variety of technical challenges in critical areas including data acquisition, simulations, control, analysis, and curation for AI/ML applications. The activities/opportunities and current gaps identified in this Facilities' Current Status and Projections document will require implementation of pilot projects in some cases and scale-up of successful pilot projects in other cases. It will also depend on continued development of AI/ML methods for science and will require dedicated staff. Future work in AI/ML and facility needs will require access to increasingly large computing resources. For example, simultaneous access to powerful computing at the edge and in burst (i.e., on-demand) modes will become necessary; federation of computing and data analysis will be important to minimize layers of different strategies (see Figure 25). Opportunities clearly exist for strong collaboration with ASCR to help develop new AI/ML algorithms and software stacks, given the large datasets generated and the robust control systems used by the SUFs.

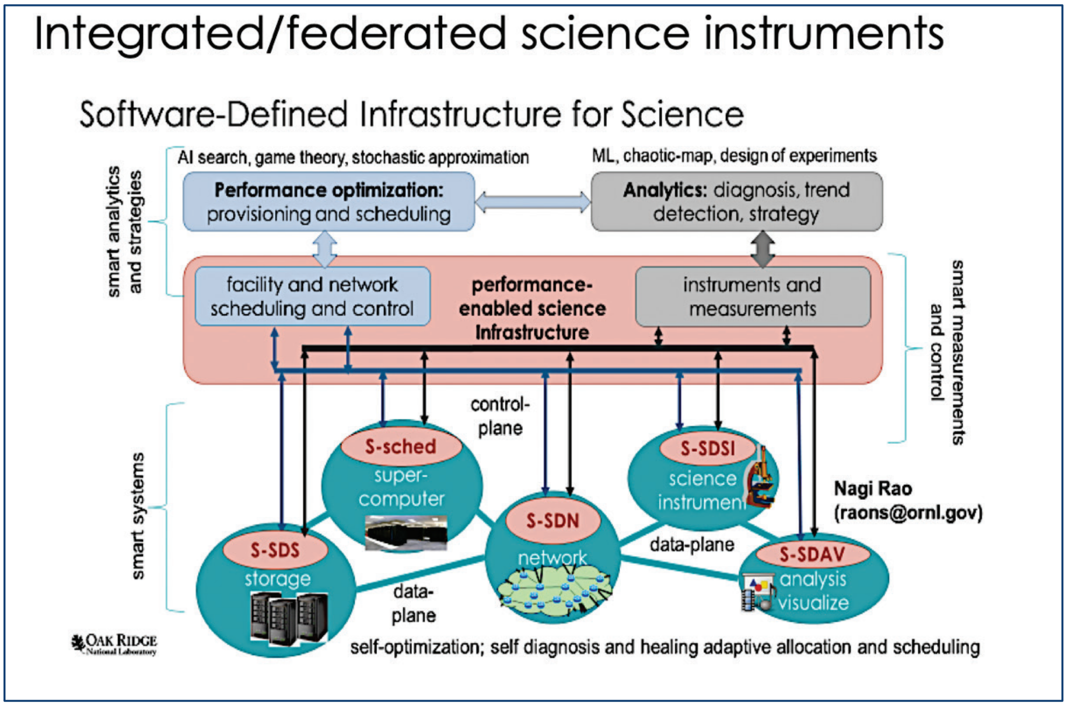


Figure 25. Schematic diagram illustrating different levels of "self-optimizing" computing and the concept of a federated software infrastructure for science. *Source: Nagi Rao, Oak Ridge National Laboratory.*

References

[1] X. Huang et al., *Nucl. Instrum. Meth. A*, 726, 77–83 (2013); X. Huang, *Proceedings of NAPAC'16*, p. 1287 (2016).

[2] X. Huang et al., *Phys. Rev. ST Accel. Beams*, 18, 084001 (2015); S. Liuzzo et al., *Proceedings of IPAC'16*, p. 3420 (2016); D. K. Olsson, *Proceedings of IPAC'18*, p. 2281 (2018); X. Yang et al., IPAC'19 (Melbourne, Australia, May 19–24, 2019); J. Wu et al., FEL'17 (Los Alamos National Laboratory, August 20–25, 2017).

[3] M. McIntire et al., *Proceedings of IPAC'16*, p. 2972 (2016); J. Duris et al., *PRL*, 124, 124801, 2020.

[4] C. Emma, A. Edelen et al., *Phys. Rev. Accel. Beams*, 21,112802 (2018).

[5] X. Ren et al., 23, 040701 (2020).

[6] S. Li et al., *Proceedings of PRL 121*, 114801 (2018); D. Ratner et al., *Proceedings of PRX 9*, 011045 (2019).

[7] F. Wang et al., 2nd ICFA Machine Learning for Accelerators Workshop, February 2019.

[8] A. Edelen, et al., https://arxiv.org/pdf/1903.07759.pdf, March 2019.

[9] X. Huang, https://arxiv.org/pdf/1907.00250.pdf, June 2019.

[10] LCLS News Center. https://lcls.slac.stanford.edu/news.

[11] DUNE–Deep Underground Neutrino Experiment. https://www.dunescience.org/.

[12] CERN: Accelerating Science. https://atlas.cern/.

[13] J. P. Cryan et al., *Phys. Rev. A*, 80 (2009), 063412.
[14] J. P. Cryan et al., *Phys. Rev. Lett.*, 105 (2010), 083004.
[15] J. P. Cryan et al., *J. Phys. B: At., Mol. Opt. Phys.*, 45 (2012), 055601.
[16] R. Coffee, Enabling Long Wavelength Streaking for Attosecond X-ray Science. Field Work Proposal no. 100498.
[17] J. M. Glownia et al., *Opt. Express*, 18 (2010), 17620.
[18] D. Ratner, et al., *PRX* 9 (2019), 011045.
[19] F. Ren et al., *Sci. Adv.* 4, no. 4 (2018), eaaq1566. DOI: 10.1126/sciadv.aaq1566.
[20] J. Betterton et al., Submitted to ICRA 2020 (2019).
[21] Y. Mao et al., *Adv. Funct. Mater.* 1900247 (2019).

U.S. DEPARTMENT OF ENERGY | Office of Science

DISCLAIMER: This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government.