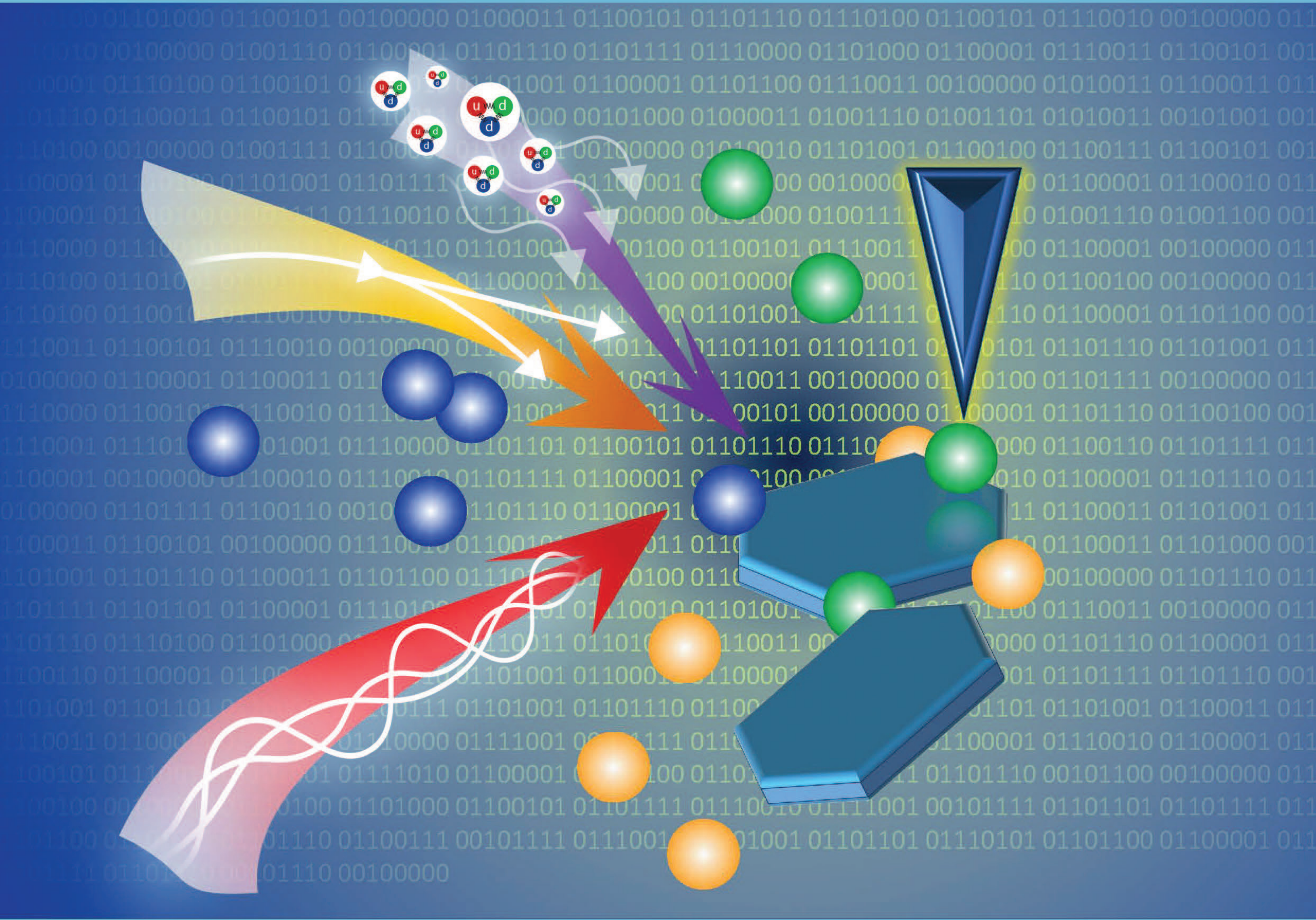
Roundtable on

Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning



Accelerating experimental and computational discovery through artificial intelligence and machine learning

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

BES Roundtable on

Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning

October 22–23, 2019
Bethesda, Maryland

Chairs
Daniel Ratner, SLAC National Accelerator Laboratory
Bobby Sumpter, Oak Ridge National Laboratory

Roundtable Participants:
Frank Alexander, Brookhaven National Laboratory
Jay Jay Billings, Oak Ridge National Laboratory
Ryan Coffee, SLAC National Accelerator Laboratory
Sarah Cousineau, Oak Ridge National Laboratory
Peter Denes, Lawrence Berkeley National Laboratory
Mathieu Doucet, Oak Ridge National Laboratory
Ian Foster, Argonne National Laboratory
Alex Hexemer, Lawrence Berkeley National Laboratory
Dean Hidas, Brookhaven National Laboratory
Xiaobiao Huang, SLAC National Accelerator Laboratory
Sergei Kalinin, Oak Ridge National Laboratory
Mariam Kiran, Lawrence Berkeley National Laboratory
A. Gilad Kusne, National Institute of Standards and Technology
Apurva Mehta, SLAC National Accelerator Laboratory
Anibal (Timmy) Ramirez-Cuesta, Oak Ridge National Laboratory
Subramanian Sankaranarayanan, Argonne National Laboratory
Mary Scott, Lawrence Berkeley National Laboratory
Mark Stevens, Sandia National Laboratories
Yipeng Sun, Argonne National Laboratory
Jana Thayer, SLAC National Accelerator Laboratory
Brian Toby, Argonne National Laboratory
Daniela Ushizima, Lawrence Berkeley National Laboratory
Rama Vasudevan, Oak Ridge National Laboratory
Stuart Wilkins, Brookhaven National Laboratory
Kevin Yager Brookhaven National Laboratory

Basic Energy Sciences team:
Beliane Lessner, James Murphy, and Katie Runkles

Oak Ridge National Laboratory Publications team: Brenda Wyatt, Deborah Counce, Laurie Varma, Laddy Fields, Kathy Jones, Jim Kidder

DOI: 10.2172/1630823

Table of Contents

List of Figures

Acronyms

AI/ML	artificial intelligence/machine learning
ASCR	Advanced Scientific Computing Research program
ASIC	application-specific integrated circuits
BES	Office of Basic Energy Sciences
CAMERA	Center for Advanced Mathematics for Energy Research Applications
CNN	convolutional neural network
DFT	density functional theory
DNN	deep neural network
DOE	US Department of Energy
DRAM	dynamic random access memory
ESnet	Energy Sciences Network
FAIR	findability, accessibility, interoperability, and reusability
FPGA	field-programmable gate array
GPU	graphics processing unit
HPC	high-performance computing
I/O	input/output
LCLS	Linac Coherent Light Source
NERSC	National Energy Research Scientific Computing Center
NSRC	Nanoscale Science Research Center
PRO	Priority Research Opportunity
R&D	research and development
RAM	random access memory
RL	reinforcement learning
SC	DOE Office of Science
SciDAC	Scientific Discovery through Advanced Computing
SUF	scientific user facility
XFEL	x-ray free electron laser

Executive Summary

The US Department of Energy's (DOE's) scientific user facilities provide access to the world's most advanced research instruments and produce increasingly larger quantities of data. DOE's Basic Energy Sciences (BES) scientific user facility instruments for x-ray, neutron, and nanoscale science are among the world's most productive, serving over 16,000 users per year with impact reported in nearly 7,000 publications and resulting in unprecedented quantities of scientific data. While this record is impressive, using the rapidly growing data stream to its full potential will require new innovations to solve a variety of technical challenges in data acquisition, control, modeling, and analysis. Artificial intelligence and machine learning (AI/ML) have opened corresponding new avenues in optimization, efficient surrogate models, data analytics, and inverse problems. These intriguing capabilities suggest that AI/ML can greatly accelerate the quest to probe and understand fundamental phenomena across a vast range of length, time and energy scales, potentially leading to transformative advances across scientific disciplines.

Both industry and the scientific community already use AI/ML approaches for data analysis. User facilities, however, crucially require AI/ML tools throughout the lifetime of an experiment: not just for data analysis, but also for data creation, acquisition, and storage. In the next 10 years, AI/ML are expected to go beyond traditional data analysis to aid the design and control of complex facilities, enable real-time capabilities to acquire and analyze large data volumes, automatically steer data collection for in-the-loop experiments, and support experimentalists' use of exascale computing. These advances will in turn open new avenues of scientific research in energy sciences and many other fields. For example, AI/ML can help the scientific community transition from relatively simple performance and properties measurements of materials and molecules to complex intertwined functionalities in batteries, information technology, chemical and biological systems, and quantum-based devices and sensors, where classical serendipitous materials discovery and sequential optimization paradigms are impractical. We envision a future of AI/ML-enabled scientific user facilities that maximize the DOE's scientific impact.

To identify specific Priority Research Opportunities (PROs) for AI/ML at the user facilities, BES convened a roundtable of facility experts encompassing the fields of physics, chemistry, materials synthesis science, computational science, detector and accelerator technology, theory, modeling, simulation, and atomic-scale characterization techniques. The roundtable met on October 22–23, 2019, to identify coordinated, long-term AI/ML research efforts that will drive major advances in neutron, photon, and nanoscale sciences.

This report describes the four PROs identified at the roundtable: PRO 1 on how AI/ML can extract high-value information from the large datasets; PRO 2 on how AI/ML can use such information in real time to maximize the facilities' scientific output; PRO 3 on using AI/ML virtual laboratories (i.e., computational models of experimental facilities) to aid the facilities and user community in design and control of machine parameters and the design and execution of experiments, including training AI/ML models for PROs 1 and 2; and PRO 4 on how a shared scientific data infrastructure can provide tools to assemble and analyze the totality of data coming from user facilities. A section at the end of the report provides a summary on computer science and mathematics that highlights where enhanced AI/ML capabilities could be particularly impactful for BES user facilities.

PRO 1: Efficiently extract critical and strategic information from large, complex datasets

Key question: *How do we extract robust and meaningful information from the increasingly vast and complex data now being produced at BES's scientific user facilities?*

Advances in the tools and techniques at BES's x-ray, neutron, and nanoscale scientific user facilities allow capture of increasingly larger datasets, often taken in a variety of experimental modalities. Paradoxically, the explosion of data can make it harder to arrive at desired scientific insights, because of the monumental level of effort needed to process and analyze the data. AI/ML techniques have the potential to significantly reduce that effort while allowing rapid, real-time information extraction of properties from noisy, imperfect measurements. Additionally, AI/ML can help unmask the complexity hidden in problems in high-dimensional spaces (e.g., multimodal measurements, many experimental variables) by finding connections elusive to human observation.

PRO 2: Address the challenges of autonomous control of scientific systems

Key question: *How do we address the challenges inherent in real-time operation of large, complex scientific user facilities?*

Realizing the full potential of current and next-generation measurement capabilities will require advanced methods in order to develop and maintain optimal performance as well as automated experimental approaches to guide scientific discovery. AI/ML-based methods are needed to efficiently search large, complex parameter spaces in real time and to predict the health and failure of instruments at high-power sources and the experiments that are run on those instruments. Such capabilities will dramatically reduce facility tuning time and downtime, improve facility performance, and maximize the productivity of BES SUFs.

PRO 3: Enable offline design and optimization of facilities and experiments

Key question: *How do we enable virtual laboratories—offline design and optimization of facility operation—to achieve new scientific goals?*

Physically accurate, virtual laboratory environments of experimental facilities (i.e., a lab in the computational cloud) will help in guiding in silico experiments from conception to synthesis and measurements. Digital twins that faithfully mimic facilities, including shared workflows and continuous updates from real experiments, can enable the design of new facility capabilities and execution of optimal experimental strategies to drive physics knowledge acquisition for user facilities. These digital twins could also facilitate development of AI/ML methods for the other Priority Research Opportunities.

PRO 4: Use shared scientific data for machine learning–driven discovery

Key question: *How can we catalyze scientific discovery by leveraging the wealth of diverse and complementary data recorded across the BES scientific user facilities?*

Radical improvement in data sharing, curation, and analysis is needed to catalyze scientific discovery across all facilities. Through the application of new AI/ML platforms to integrate diverse scientific data resources, extensive new datasets could be created from heterogeneous experimental and simulated data, leading to new opportunities for scientific discovery. Coordinated development of workflows on a shared facility–based data repository could catalyze development of data standards, formats, and priorities. These curated datasets could, in turn, serve as training sets for developing new AI/ML methods.

Introduction

The US Department of Energy (DOE) operates a wide range of scientific user facilities (SUFs) that provide access to the world's most advanced research instruments. The world-leading Basic Energy Sciences (BES) x-ray, neutron and nanoscale SUFs serve over 16,000 users per year and produce petabytes—the equivalent of a million gigabytes—of data that deliver high-impact science. Current and upgraded user facilities face a variety of technical challenges related to data acquisition, control, modeling, and analysis. Improvements to instruments will enable more sophisticated studies by providing a greater quantity and quality of probe particles (i.e., photons, electrons, neutrons), while concomitant advances in detection and data volumes demand new techniques to obtain the scientific results. The synchrotron light sources, neutron sources, and Nanoscale Science Research Centers (NSRCs) enable ambitious new experiments that can weave together complex, multimodal datasets. NSRCs further require active control to synthesize new materials. Therefore, advances in our ability to handle large quantities of data, rapidly extract useful information, and use it to guide experiments and simulations, promise to open new avenues of research across the physical, biological, and engineering sciences.

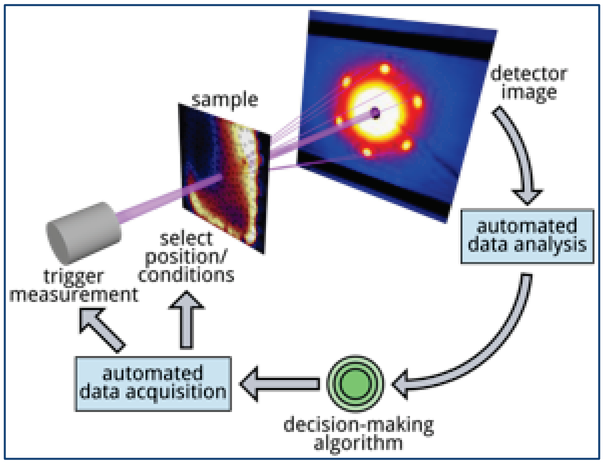

sample
detector image
automated data analysis
select position/ conditions
trigger measurement
automated data acquisition
decision-making algorithm

Figure 1. Autonomous control of experimental systems promises to open the study of problems previously considered impossible. Automating the entire experimental workflow—instrument setup and tuning, sample selection and synthesis, measurement, data analysis and model-driven data interpretation, and follow-up experimental decision-making—will bring about revolutionary efficiencies and research outcomes.| *Distributed under a Creative Commons Attribution International License 4.0*

As an example of an emerging challenge, coherent x-ray diffraction imaging (or "lensless" imaging) is growing rapidly in usage at both storage ring–based synchrotrons and x-ray free electron lasers (XFELs) as new and upgraded sources provide a higher degree of coherence. Because the source properties are critical to the experiment, sophisticated prediction and feedback are required to maintain source quality, and high-fidelity simulations are needed both for designing new capabilities and guiding online control. To reach peak performance, accelerators require frequent optimization in high-dimensional spaces as well as anomaly/breakout detection to protect the high-power, high-repetition-rate machines. On the experimental side, lensless imaging is both data- and computationally intensive. Sophisticated compression/rejection data pipeline tools operating at the "edge" (i.e., next to the detector or experiment) are needed to extract and save information "on the fly." Active control is needed to automatically steer experiments and synthesis through a high-dimensional parameter space. Figure 1 depicts an autonomous control process for experimental systems.

Even after datasets are recorded, new tools are needed to share and analyze the enormous, multimodal datasets that span the SUFs, including data merges and simulations. Large-scale computation will require the development of automated science workflows and novel data science approaches. Example applications include molecular dynamics simulations for comparison to neutron scattering data, density functional theory (DFT) for comparison to neutron spectroscopy data, Monte Carlo ray tracing for simulating instrument and complex sample effects, and diffuse scattering modeling for investigating defects in solids and large-scale tomographic reconstructions. At the NSRCs, the ability to discover new materials and chemical compounds with desired properties relevant for societal applications is primarily driven by a relatively slow process of intuition, design rules, models, and theories derived from scientific data generated by experimentation and simulation. The number of materials and chemical compounds that can be derived is astronomical. Finding desired properties through random experimentation is like looking for a needle in a haystack.

Computational and data science challenges exist throughout the facility operations life cycle, but there is an expectation that artificial intelligence (AI) and machine learning (ML) methods will have a transformative impact on SUF science. AI/ML methods for analysis, control, and modeling will drastically accelerate experimental and computational discovery. AI involves machines that can perform tasks characteristic of human intelligence such as planning, understanding language, recognizing objects and sounds, learning, and problem-solving. ML is a means of achieving AI; ML refers to systems that can learn from data without being explicitly programmed. We envision that, in the next 10 years, AI/ML will be an integral part of the DOE's discovery and design arsenal, just as experimental, theoretical, and computational tools are today. Scientists at the SUFs will work in synergy with AI/ML experts at the DOE to operate facilities and generate scientific data, formulating new physical models and theoretical insights that drive scientific discovery and open new paths of design of materials and chemicals.

While AI/ML is widely recognized as a set of tools for data analysis, opportunities at the SUFs extend across facility operations, spanning everything from the design of new machines to the inference of new science. For example, AI/ML can merge physics, simulations, and data to guide online optimization of accelerators, enabling design of challenging configurations that will deliver new capabilities to users. Autonomous control of experimental systems could revolutionize the way experimentalists work, allowing them to explore high-dimensional problems previously considered impossible. Such advances could, for example, enable the discovery of targeted materials and chemicals a thousand times faster than current methods; resolve conformational landscapes of proteins; and reveal complex hierarchical correlations, from molecular-scale interactions to transport phenomena, to mapping the energy landscapes of chemical and materials transformations.

To identify Priority Research Opportunities (PROs), BES convened a roundtable of experts on October 22–23, 2019, from the SUFs and user communities encompassing a wide range of disciplines and cross-cutting experimental sciences, computational science, detector and accelerator technology, theory, modeling, simulation, and atomic-scale characterization techniques. This roundtable generated future research opportunities that could form the basis of a coordinated, long-term research effort enabling major advances for neutron, photon, and nanoscale sciences. See Appendix A for a list of participants and their affiliations and Appendix B for the roundtable agenda.

The roundtable participants were asked to provide insight on how big data and AI/ML techniques could be used to reach the full operational potential and scientific impact of the SUFs. Technical challenges are expected for simulations, control, analysis, data acquisition, and deep data analysis. Participants considered new technologies for speeding up high-fidelity simulations for online models, fast-tuning in high-dimensional spaces, anomaly/breakout detection, "virtual diagnostics" that can operate at high-repetition rates, and sophisticated compression/rejection data reduction workflows operating at the edge to capture high-value data and steer experiments in real time.

Prior to the roundtable meeting, members of the community were asked to provide a two-page summary of past and current work on AI/ML for machine control, data manipulation, and analysis at their facilities. This companion to the present report was compiled into the *Facilities' Current Status and Projections for Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning*, (to be published in https://science.osti.gov/bes/Community-Resources/Reports) which set the stage for the roundtable by addressing six questions that described the current AI/ML applications, developments, and opportunities at the SUFs:

1. How can AI/ML change or improve the way your lab operates its facilities?
2. What are the detectors' limitations, and how can AI/ML help?
3. Can AI/ML improve DOE facility users' experiences during data acquisition through novel experimental methods, data analysis, adaptive control, etc.?
4. Do you feel there are particular limitation(s) to successful progress in AI/ML for data production and analysis at your facility (please elaborate)?
5. Are there opportunities to better integrate with Advanced Scientific Computing Research (ASCR) data analytics, high-performance computing (HPC), and high-speed networking capabilities for data-intensive experimental and theoretical problems?
6. What aspect of AI/ML is exciting to you? Explain how this could be enabling for user facilities.

During the roundtable, an interactive discussion session of all participants identified potential themes. Four breakout sessions identified dominant themes in the areas of online control, data acquisition, multimodal analysis, and models/simulations. From the results of the first day's discussions, the writing team identified four PROs, which were finalized on the second morning along with example "killer applications" to highlight the potential impact of each PRO.

This report summarizes the key challenges and future research directions identified by the AI/ML roundtable. These are articulated through the four PROs, with each described in a detailed section of the report.

PRO 1: Efficiently extract critical and strategic information from large, complex datasets

PRO 2: Address the challenges of autonomous control of scientific systems

PRO 3: Enable offline design and optimization of facilities and experiments

PRO 4: Use shared scientific data for machine learning–driven discovery

The report ends with a section highlighting the cross-cutting opportunities for collaboration with ASCR to help enable enhanced AI/ML capabilities relevant to the four PROs.

Successfully addressing the PROs will establish coordinated, long-term research efforts to enable major advances for neutron, photon, and nanoscale sciences that propel these user facilities to next generation capabilities for users and for scientific discovery.

PRO 1. Efficiently Extract Critical and Strategic Information from Large, Complex Datasets

Key question: *How do we extract robust and meaningful information from the increasingly vast and complex data now being produced at BES's scientific user facilities?*

Introduction

The BES user facilities' array of x-ray, electron, neutron, atom, and optical probes are producing increasingly larger and more complex data streams at a faster pace than current analysis methods can handle [1–10]. Scientific understanding requires the extraction of physical and chemical information from these data, in the form of underlying electronic, atomistic, nanoscale, and mesoscale structures and dynamics. AI/ML approaches, constrained and informed by physical models, are needed to enable and accelerate sampling of the underlying structure and dynamics spaces; efficient forward modeling; and pattern matching from these large, high-throughput data streams [11].

As tools and techniques (e.g., x-ray, neutron, optical, and electron probes, along with other microscopies) to study matter at the nanoscale dramatically improve, our challenges are harnessing the quantity of the data produced and connecting different pieces of scientific information generated from that data. Overcoming these challenges will lead to three dramatic advances: (1) faster time for sample characterization and understanding, (2) real-time analysis for online control and autonomous experiments, and (3) the ability to address higher levels of complexity in experiments by elucidating connections hidden in high-dimensional spaces. If the scientific community does not overcome these challenges, the scientific output realized by the SUFs may not keep pace with the facilities' capabilities.

Research Directions

This PRO has three underlying themes. AI/ML techniques offer the potential to enhance the scientific productivity of the BES SUFs by:

1. Accelerating the transformation of data into scientific information (i.e., taking raw, noisy, imperfect snapshots of observations and extracting physical quantities and useful information).
2. Enabling rapid information extraction to provide real-time feedback to experimenters, allowing the modification of the course of an experiment as it is being performed; and, more broadly, as a foundation for autonomous smart control of the experiment (PRO 2).
3. Delivering enhanced analytical techniques (via AI/ML) of large, complex datasets, including joint analysis of experimental data and simulations, to allow scientists to observe hidden connections across experimental modalities in complex, high-dimensional spaces.

Each of these themes is described in more detail below.

Accelerating the transformation of data into scientific information

The explosion of data volumes, and in some cases data generation rate (i.e., how fast data are produced), poses a challenge for effective data analysis at the SUFs. Data reduction techniques such as experiment-specific vetoes, lossy or lossless compression, and feature extraction must be employed to achieve the dual goal of reducing the data volume *and* extracting physical information from raw measurement data. These techniques must be capable of adapting to frequently changing experiments, must be scalable up to the maximum detector input/output (I/O) capabilities and easily configurable to accommodate rapidly changing experimental conditions. AI/ML methods have potential to address these challenges, with large datasets enabling the use of deep learning techniques [12].

Extracting information from the data is likely to require a layered approach. For example, in particle physics, various levels of “trigger” are used, each one successively more complex, in order to determine whether the data from an individual event is worth recording (e.g., [13]). In particle physics today, an experiment may have decades of simulation studies to build confidence in the trigger algorithm. While such selective saving of data is also beneficial to the SUFs (e.g., shot-by-shot measurements at an XFEL), designing triggers is more challenging due to the short duration of experiments, which can change day to day (see PRO 3). An alternative data reduction approach would not discard data by a trigger but rather begin to extract information at a low level (i.e., as close to the detector as possible). An example is “clusterization,” determining the impact point of a probe particle on the detector with subpixel accuracy. Conventional techniques are computationally expensive and sensitive to noise and calibration errors. AI/ML methods may improve both the speed and spatial resolution of such tasks. As one moves further from the detector and more computing power is available, more sophisticated techniques can be employed. For example, artifacts and distortions that arise in x-ray scattering data collection might be healed computationally [14–15], thereby revealing the true underlying structural motifs. Data reconstruction methods that have been avoided historically due to the high computational cost could be replaced by fast-executing AI/ML methods [16]. At the broadest level, AI/ML methods could extract physical insights directly from raw experimental signals without any information loss from intermediate steps [17].

Enabling rapid information extraction to provide real-time feedback

Driving an experiment in real time requires both rapid and sophisticated data analysis so that each measurement can inform future experiments (see PRO 2). Current analysis methods are inadequate for this task. For example, new and upgraded light sources dramatically increase both source brightness and coherence. Coherence can be exploited, for example, for lensless imaging, but with a concomitant increase in computational complexity. It is estimated [18–19] that a single coherent imaging beamline will generate approximately 130 petabytes of raw data per year and that over 30 petaflops of continuous computing power will be needed to keep up with this anticipated data generation rate using current inversion algorithms.

Recent preliminary results [20–21] suggest deep neural networks (DNNs) can be used to learn a wide range of inverse problems; for example, the inversion of raw x-ray (and electron data) from the NSRCs to real-space coordinates. Once trained, these networks could be deployed on the edge to enable real-time experimental feedback. Incorporating the physics of both the sample being studied and the model connecting the raw data to a real-space image could further constrain the optimization space and improve the results of experiments that use AI/ML. Further research challenges in optimizing the tuning of massive DNNs and active learning will need to be explored to optimize these techniques.

Delivering enhanced analytical techniques

Multimodal characterization tools that provide complementary information are indispensable for BES’s SUFs. The challenge is to properly connect often disparate information, similar to multiscale problems in biological and physical sciences. For example, precise knowledge of atomistic and electronic structure of materials during synthesis and dynamics is needed for discovery of new materials, but combining scattering, microscopy, and spectroscopy data to recover structures remains a challenging inverse problem. Whether due to projection of a 3D structure onto one or two dimensions—as in pair distribution functions and transmission electron microscopy, or the reduction of a large number of matrix elements into an overall energy-dependent amplitude, as in x-ray absorption spectroscopy, or the interference of scattered coherent x-ray, as in coherent diffractive imaging and x-ray photon correlation spectroscopy—the result is that inversion of this mapping is time-consuming, imprecise, and sometimes even wildly inaccurate, which limits and delays new knowledge, despite the availability of the impressive array of multimodal in situ/operando instrumentation at the SUFs.

Solving inverse problems such as these requires (1) a large number of measurements, (2) simulations and predictions of how to incorporate signals from different modalities, and (3) physical constraints (i.e., solutions need to be driven towards optimal matching with experiments while ensuring adequate physical representation). Fortunately, inverse problems are particularly well-served by the confluence of AI/ML, which can handle under-determined mappings, and atomistic and first principles modeling, which allows high-throughput configurational sampling, forward modeling of multimodal characterization data, and, most importantly, severe constraints of the solution space.

Enabling Capabilities

To achieve faster time to sample characterization and a deeper level of understanding of complex experiments, several corresponding advances in infrastructure and facilities are required.

- **Sufficient network bandwidth availability:** The DOE's Office of Science (SC) Energy Sciences Network (ESnet) provides high-bandwidth interconnection between national laboratories and universities, and it is imperative to ensure sufficient ESnet shared backbone capacity is available to allow data flows that adequately connect data and compute resources between HPC facilities, neutron and light sources, and NSRCs.

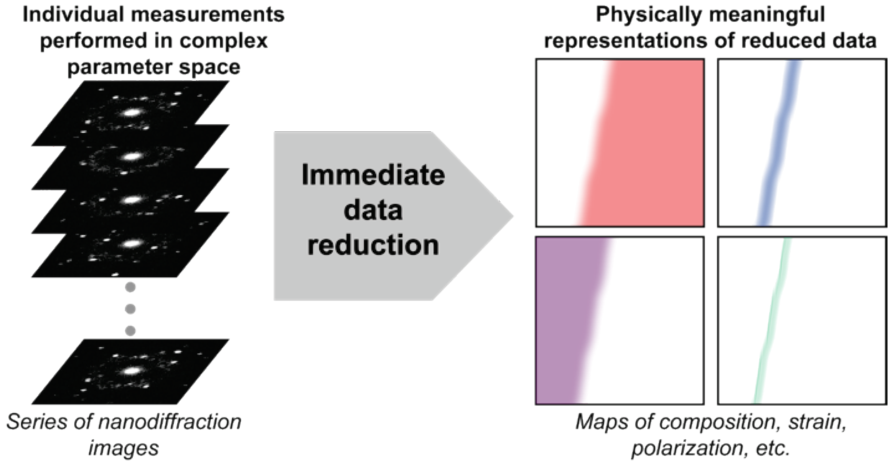
- **Analyzing data at its natural production rate:** AI/ML will facilitate several key elements in accomplishing on-the-fly information extraction at accelerating data generation rates: (1) solving inverse problems, such as those listed above, which may combine measurements from different facilities and modalities; (2) finding surrogate models covering the transitions between discrete measurements; and (3) parameter-space learning, which enables more efficient experimental searches through that parameter space. A wide range of classification and regression ML approaches, together with stochastic, decision tree, evolutionary, active learning, and Bayesian optimization methods, are applicable to these problems. Methods for training models on sparsely labeled data will be critical.

- **Ability to leverage domain knowledge in analysis:** Extraction of physical and chemical information from large and fast data streams using AI/ML approaches must be constrained and informed by physical models to efficiently enable and accelerate the sampling of the underlying parameter space and allow for pattern matching and forward modeling.

- **Simultaneous analysis of all the data, no matter the source, machine, or format:** A comprehensive dataset representative of the SUFs' science interests must be generally available to train and validate AI/ML models. To support such a dataset, centralized storage capabilities and policies that encourage data sharing across facilities will need to be implemented. It will be critical to devise metadata standards because experimental metadata are not systematically collected across facilities and most metadata are not standardized and are recorded in user logbooks [5]. A uniform metadata tagging process would make it easier for users and developers of AI/ML methods to locate and use relevant data by enabling data searchability. Both PRO 4 and the report titled *Data and Models: A Framework for Advancing AI in Science* cover this topic in more detail [22].

- **Validation and verification (trust but verify):** How does one provide evidence that a model is sufficiently accurate for its intended use? Each AI/ML data reduction method that is developed must work reliably and robustly. Standards must be developed for validation and verification of AI/ML methods to convince users that these methods are accurate and do not systematically bias results.

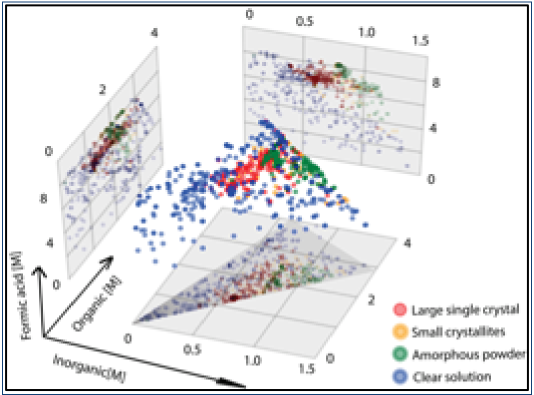
Harnessing Complexity for Multicomponent, Multifunctional Materials Design


Individual measurements performed in complex parameter space
Immediate data reduction
Physically meaningful representations of reduced data
Series of nanodiffraction images
Maps of composition, strain, polarization, etc.

Measured values of interest often are not apparent to the researcher during data collection. For example, many material properties are encoded in scanning electron nanodiffraction datasets. Properties such as crystalline phase, strain, and polarization and any correlations between them are typically discovered only after an experiment. Immediate data reduction into relevant physical properties alters the way an experimentalist interacts with the system, allowing direct navigation through complex experimental parameter spaces on the time scale of the experiment. | *Image courtesy of Mary Scott, National Center for Electron Microscopy, Lawrence Berkeley National Laboratory*

Many material systems comprise complex, heterogeneous phases. Damage-tolerant [23] and self-healing [24] structural materials or multi-component catalysts for cascade reactions [25] all rely on heterogeneity. The exact nature of these phases, their distribution, and any correlation between them is often not apparent until after an experiment is performed. To design and optimize multicomponent materials, where direct observation of underlying structures may be impossible, experiments must navigate the resulting relevant functional properties. This goal necessitates high-throughput experimental tools that employ immediate data reduction to navigate a complex experimental parameter space. For example, an experimentalist may need to classify and interpret raw electron microscopy images or x-ray ptychography data during an experiment, as illustrated in the figure above. AI/ML approaches, which automatically classify features, detect patterns and correlations, and interpret data can therefore significantly impact the development of heterogeneous material systems. To enable these capabilities, the advances outlined in PRO 1 will be required.

A practical example involves identifying crystallization conditions and crystal structure in organic–inorganic hybrid perovskite materials. This is a time-consuming, "needle in a haystack" search through thousands of combinations of reaction parameters, even when those parameters are known for materials with similar organic components. Robotic workflows at the Molecular Foundry nanoscience user facility at Lawrence Berkeley National Laboratory recently performed over 9,000 perovskite reactions and screened over 50 different organic precursors for single-crystal formation in perovskites [26]. ML algorithms classified reaction outcomes such as crystal size, crystal structure, dimensionality, and material properties (see figure at right). A network of ML experts, acting as "virtual users," used a software pipeline [27] developed by Molecular Foundry users to propose new experiments using the robots. These data are uploaded automatically into the software database and then used to train learning algorithms using transfer learning or Bayesian optimization. A major challenge for such ML modeling of materials, and thus an ultimate goal, is to extrapolate outcomes from one chemical system to untested systems.



Nine hundred sixty-eight perovskite reactions were performed by a robot to explore conditions for producing single crystals (red dots) of ethylammonium lead iodide. | *Distributed under a Creative Commons Attribution Noncommercial License 4.0 (https://creativecommons.org/licenses/by-nc/4.0/)*

- **Provenance preservation:** Provenance in computational science is the record of data lineage and software processes that operate on these data to enable the interpretation, validation, and reproducibility of results. In experimental science, provenance includes experimental conditions, calibrations, and notes that contain the record of how the data were produced and analyzed. Similar to metadata, provenance for data and software is crucial to enable transparency and trust in experimental and computational results. Provenance provides a record for the numerous transformations in the scientific process of discovery and in the design of new materials. An accurate provenance record provides a measure of quality for results. Such a record should include references to the software code and initialization conditions used to produce particular datasets, samples, and experimental conditions such as motor positions at a beamline and the names of researchers and facilities involved in a particular project. With AI/ML and computational algorithms increasingly guiding more steps in the discovery process, detailed and comprehensive provenance is required to trace how results are obtained, especially in dynamic settings, when computer-driven decisions guide autonomous experiments.

Potential Impact

The ability to extract key features of data generated at SUFs confers a number of advantages. Overall throughput and storage requirements are reduced, and the ability to stream data to DOE's ASCR computing facilities becomes possible. Information extraction allows for crucial real-time feedback to drive experiments toward the highest value measurements and reduce the time from measurement to scientific insight (see PRO 2). Furthermore, compact information in relevant physical units can be more easily shared across facilities, enabling multimodal data analysis and synthesis. AI/ML methods for data reduction and feature extraction will enable BES facilities to process higher rates of streaming data to characterize heterogeneous ensembles, capture rare transient events, and map spontaneous dynamics in operando. AI/ML techniques make it possible to address new levels of complexity such as mapping reaction landscapes or capturing rare events via automatic pattern recognition and to probe high-dimensional space.

If a real-time data reduction capability is not delivered, the consequences for the BES SUF mission are significant. Facilities would be forced to artificially limit the delivered particle flux or the readout rates of future detectors, constraining both the quality and number of experiments that can be performed at the SUFs, which in turn limits the science output of the facilities. Many experiments requiring high statistics would not be feasible without some form of on-the-fly feature extraction to handle the throughput. Experiments that require complex multimodal analysis would not be possible, reducing the potential scientific impact of the SUFs. The inability of SUF users to efficiently acquire, manage, and analyze their data would increase the time to understanding and publication.

Current research in AI/ML has been shown to be effective in feature identification and processing of large datasets, helping data analysis and visualization keep pace with the explosion of experimental data. In moving from capturing *data* to recording *information*, fast extraction of information from experimental measurements underpins the autonomous control and experimentation discussed in PRO 2.

PRO 2. Address the Challenges of Autonomous Control of Scientific Systems

Key question: *How do we address challenges inherent in real-time operation of large, complex scientific user facilities?*

Introduction

Advances in AI/ML approaches are fueling a new paradigm wherein any and all automatable tasks might be ceded to machine control and human experts are liberated to work on the challenging high-level problems of *understanding* the underlying science. For example, AI/ML-driven autonomous control of scientific systems promises to deliver scientific systems that self-regulate to yield ultra-high performance and experimental platforms that can autonomously explore scientific problems, optimally selecting experimental sequences and synthesizing the accumulating datasets into human-readable physical insights. This could radically improve facility operation efficiencies and provide a means to discover and understand new scientific phenomena, ultimately accelerating delivery of new scientific discoveries and next-generation technologies for energy production, storage, utilization, and national security.

Within the SUFs, a pertinent case is that of accelerators, which underlay the large photon, electron, and neutron science communities. For an accelerator to operate efficiently, thousands of component systems must work together within tight tolerances, providing nonlinear and highly coupled responses. Conventional control approaches, such as relying on static design models and manual tuning cannot hope to contend with the real-world complexity of these systems, especially as the SUFs move toward physics-limited sources. Future online control systems will require AI/ML methods that leverage known device physics (via detailed modeling) and real-world operational knowledge (via data mining the system's archive). Similarly, modern experimental tools—including synchrotron and neutron endstations, electron microscopes, scanning probe instruments, and advanced optical systems—are becoming more complex and require precisely controlled, interconnected hardware systems that must handle a high-volume of data generated at a high rate.

AI/ML methods also hold promise in controlling user experiments through autonomous selection of measurement conditions. By leveraging fast, real-time data analytics, this approach will increase the quality of experimental datasets, reduce wasted instrument time, and accelerate experimental studies. Such improvements would have immediate impact across the entire SUF program.

Research addressing modern materials, chemistry, and biosciences faces a similar problem, with frontier experiments probing vast and complex parameter spaces. The study of multicomponent, hierarchical, and nonequilibrium materials requires exploring the enormous spaces defined by material composition and processing history. Searches to identify target functional materials, or to uncover salient trends, are very difficult to achieve using traditional approaches. Instead, the field requires the ability to efficiently predict, explore, and navigate materials and processing parameters. This necessitates the development of autonomous experiment control to adaptively update data gathering. AI/ML autonomous experimentation will similarly help enable real-time material synthesis, allowing potential access to metastable and nonequilibrium materials that can only be realized via active control of the synthesis pathway. Steered synthesis will also allow study of additive manufacturing technologies that rely on active computation and controlled processing to yield desired materials and structures.

Research Directions

Autonomous control of experimental systems promises to open study of problems previously considered impossible to address. The research goal is to automate the entire experimental workflow, from instrument setup and tuning to sample selection and synthesis, measurement and data analysis, model-driven data interpretation, and follow-up experimental decision-making. As such, a coordinated set of advances is required across a range of systems.

The PRO identified two broad areas of research that could add value to BES research programs:

1. Automating facility control, enabling higher reliability, increased efficiency from self-regulation, and ability to reach physics-limited capabilities. Examples are given for both accelerators and beamlines.

2. Automating the experimental process such as automated measurement and synthesis platforms coupled with AI/ML algorithms allow intelligent exploration of complex problems. Examples are given for both scientific discovery and synthesis of new materials.

Automating facility control

With each successive generation of the SUFs, the complexity of the operational and experimental challenges increases. A successful experiment at an SUF requires real-time control and tuning in a high-dimensional space, where response is nonlinear and parameters are strongly coupled. For example, achieving high coherent flux in a focused spot in a modern synchrotron beamline currently relies on simple feedback loops that maintain beam intensity; ideally, they would directly guarantee stability of the coherent wavefront at the sample position. Advanced AI/ML-driven control that leverages physical simulation of beamline systems would enable previously impossible performance and stability. Existing AI/ML methods will need to be adapted to the specific challenges associated with distinct experimental tools. Simultaneously, an opportunity exists to define a common toolset of AI/ML methods useful across a variety of control problems.

The SUFs face a looming challenge in the need to control and tune in a truly end-to-end manner. For example, a synchrotron beamline experiment can be viewed as a distinct set of systems that must be optimized individually, or it can be viewed as a large coupled problem, where accelerator performance, beamline optics, endstation measurement systems, and the experiment itself must all be co-optimized to maximize a target scientific objective. Similarly within an electron microscope, the source, optics, detector, and sample environment can be viewed as a coupled system that must be tuned for a particular objective, whether that is high-resolution imaging or in situ atomic-scale synthesis [28].

A critical example of the need for online control comes from SUF accelerators, which deliver photon, electron, and neutron beams to a large community of researchers. Modern accelerators are enormously complex, with thousands of components, each having dozens or even hundreds of control parameters that must be modulated in a concerted fashion. The impact of a parameter on the performance of any of these complex systems is often realized through complicated, nonlinear physical processes. For example, in a storage ring, nonlinear beam dynamics determine the ring's injection efficiency and beam lifetime; in a self-amplified spontaneous emission XFEL, nonlinear beam dynamics determine the self-bunching of the electron beam. The control parameters may be coupled, and the optimal configuration may drift as the environment changes. The traditional control approach consists of setting parameters according to a static design model, and manually tuning subsystems. This approach has many limitations that ultimately stunt scientific productivity. Real-world performance typically falls short of simulated predictions, owing to environmental variables omitted from design models. Manual tuning can improve performance but is time-consuming and depends strongly on the training and experience of the operator. For complex machines such as the Linac Coherent Light Source (LCLS), preparation of the machine for some special operation modes can take hours of tuning—time that would be better spent conducting user science. Some exotic beam characteristics may not even be offered, owing to the underlying tuning challenge.

After achieving the desired accelerator setup, it is equally important to maintain the conditions during user operation. At present, feedback loops are used to stabilize subsystems, typically assuming simple linear relationships; orbit feedback is one such example. However, in many cases, machine performance is affected by the environment through unknown connections, which require continual compensation by adjusting control parameters [29–30]. Typical tuning methods may not be suitable for this purpose because they would exert large changes to the control parameters and perturb user experiments. In recent years, automated tuning has become increasingly popular on machines ranging from colliders to light sources [31–41], with solutions addressing complications of noise, drift, and outliers. While examples of accelerator control using AI/ML methods already exist (e.g., Gaussian process optimization) [34, 38], effectively searching large, complex parameter spaces remains a substantial challenge. Smart control methods that can rapidly and smoothly tune a coupled set of nonlinear control parameters are needed.

High reliability and availability are important for an SUF, which may serve thousands of users a year on a strict schedule. While each individual component in a facility is expected to operate reliably for a long time, it is not uncommon to have component failures in a large system. Because one failure can bring down the entire system, and the recovery time from a failure is typically much longer than replacing components during scheduled maintenance, it is critical to understand the health and failure of the accelerator components and subsystems. For example, knowledge of failure patterns enables quick identification of the root cause of failures, which helps expedite recovery. The ability to predict failures is even more important, as it can prevent failures through preemptive maintenance or reduce downtime by initiating protective procedure before failures occur. The failure prediction capability is especially important for superconducting systems as failure-induced quenches can cause substantial loss of operation time.

AI/ML provides a unique opportunity to address challenges in the operation of large, complex SUFs. For example, while traditional tuning methods treat the target system as a black box, AI/ML-based methods can learn a model that approximates the physical behavior of the complex machine (see PRO 3). An online learning model can be continuously updated and refined using new machine measurements. The ability to make accurate predictions with a model opens up the possibility of dramatically increasing the efficiency of optimization algorithms for high-dimensional parameter space.

AI/ML control methods can also be applied to compensate for environmental drift; minimizing perturbation to ongoing user experiments can naturally be included as part of the ML target–reward function. Such methods could enable previously impossible maximization of both performance and stability. Integrating the advanced tuning, control, and prognostics methods made possible by AI/ML into operations will make it feasible to operate an SUF largely through an intelligent, autonomous program, minimizing the need for human intervention while maximizing key metrics. Advances are required in tuning algorithms, parameter-space searching methods, fast modeling of components, and integration of these advanced methods into real-world hardware systems. Properly implemented, these methods will deliver novel beam capabilities and system reliability to the scientific user community, empowering a new generation of leading-edge BES research.

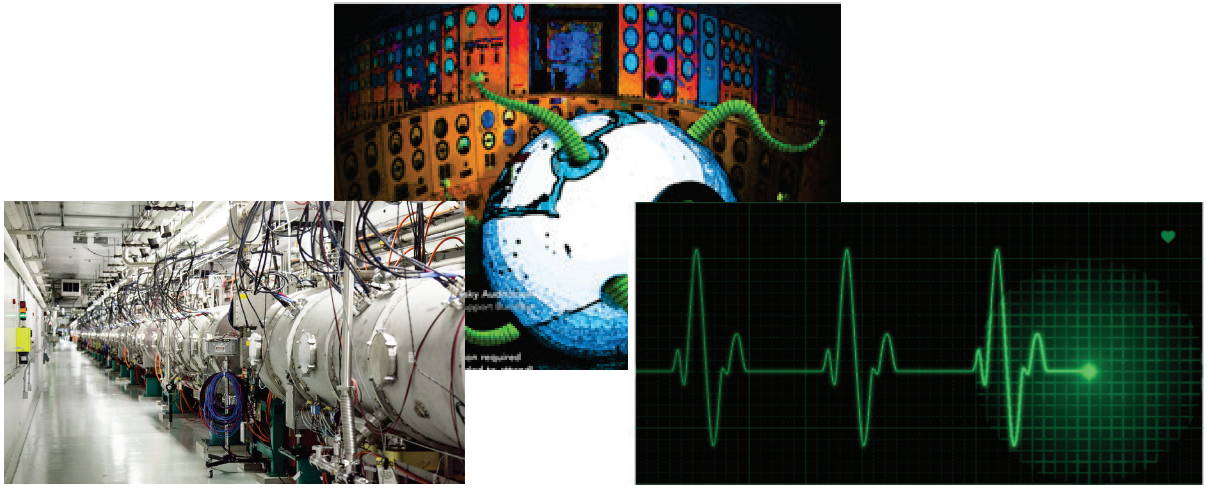
Automating the experimental process

In addition to instrument tuning, AI/ML methods could revolutionize experimental platforms by automating selection of measurement conditions, experimental conditions, sample measurement sequence, and overall experimental execution. Such automation—necessarily leveraging accelerated real-time data analytics—would dramatically increase the quality of experimental datasets, reduce wasted instrument time, minimize sample damage from probes, and accelerate experimental studies.

Modern experimental measurements are high-dimensional and multimodal; the traditional approach of exhaustively probing a sample will be impossible as complexity and resolution increase. For example, imaging of dynamic materials implies a 4D space, while multimodal acquisitions that combine rich spectra with scattering/diffraction patterns further broadens signal complexity. Autonomous control of experiments will allow the parameters of each measurement to be informed by previous measurements, focusing the SUFs' resources to always capture the highest value data. For example, studying dynamic processes under operando conditions requires the identification, tracking, and quantification of the most relevant volumes within the sample as a function of the applied stimuli. In addition to the choice of the most important volumes to sample, researchers can also choose the imaging modality to apply to this subvolume. This situation presents a vast measurement parameter space that is very difficult to navigate when seeking concrete connections between sparse local phenomena (e.g., dislocation motion and grain boundary stress concentration) and bulk irreversible processes. AI/ML agents that can make real-time decisions are needed to navigate these parameter spaces. The higher brightness afforded by new and upgraded light sources and the development of ultra-fast electron microscopy methods coupled with advances in detector technologies enables the study of interesting dynamic phenomena at time scales that were previously inaccessible. These advances in sources and detectors will result in the generation of orders of magnitude more data over exceedingly shorter time scales. As experiments progress beyond speeds at which humans can make real-time decisions, AI/ML-informed adaptive control becomes imperative.

Monitoring the Heartbeat of an Accelerator

A self-healing accelerator would achieve record-setting reliability.



Modern accelerators rely on the operation and high-precision tuning of hundreds of thousands of parameters simultaneously. Traditional human-driven control of these complex, nonlinear, and coupled systems does not scale when one desires physics-limited performance and uninterrupted operation. Using AI/ML, one can build a "self-driving" accelerator that is able to monitor its own health through AI/ML analysis of its operation, predict failures, avoid downtime, and automatically retune in real time using physical models to maintain stable high performance. This would enable customizable shot-by-shot configurations for XFEL experiments, reductions in reconfiguration time between experiments from days to minutes, and orders of magnitude increases in source-to-detector beam stability.

Left image courtesy of Christopher Smith, SLAC National Accelerator Laboratory | Middle image courtesy of Terry Anderson, SLAC National Accelerator Laboratory | Right image courtesy of Genevieve Martin/Oak Ridge National Laboratory, US Department of Energy.

Similar opportunities exist in the autonomous guidance of materials synthesis. Modern materials are inherently complex, owing to the compositional complexity of formulations, blends, and composites; the structural complexity of hierarchical materials exhibiting order at multiple length scales; and the processing complexity of nonequilibrium materials exhibiting pathway-dependent ordering. While the search spaces are exponentially large, the subset of materials exhibiting desired characteristics is extremely small, thus defining an exceptionally challenging "needle in the haystack" search problem. In contrast to stability control problems, where anomalous events are generally to be avoided, searches in material physics must emphasize variability to identify the interesting anomalies that represent radically new materials with record-setting material properties. Traditional correlative searches will generally fail to find important outliers, as they emphasize interpolation, perform poorly at extrapolation, and tend to average out potentially relevant variations. AI/ML methods hold promise to accelerate the searches of these spaces [42], because they can handle the scale of data as well as the search for subtle correlations. AI/ML methods can further accelerate discovery through physics-informed search, constraining the search space to physically reasonable regimes while also guiding scientific examination toward areas of predicted novelty or experimentally identified "surprise," such as mismatches with established theories. Physics-informed search methods are poised to significantly enhance materials discovery [41–53], by providing scientists with the ability to rapidly explore materials problems, uncovering the underlying physics and identifying target materials.

Moreover, the simultaneous or sequential use of multiple probes (e.g., optical, electron, x-ray, neutron, scanning probes) represents an opportunity to more deeply interrogate a material because the probes provide complementary information about material makeup. Autonomous experimentation would benefit enormously from control methods tailored to take full advantage of these rich datasets. For example, sample measurements in one modality should leverage any preexisting measurements in different modalities to identify optimal measurement strategies (e.g., concentrating points to leverage the new modality and resolve ambiguities associated with previous measurement methods). Moreover, the real-time data reconstruction associated with autonomous data-taking must leverage all available multimodal signals. For example, in tomographic experiments, the reconstruction should yield the actual material composition, structure, and subvoxel ordering based on constraint satisfaction of all available signals, rather than reconstructing a parallel set of tomograms for the distinct imaging modes. Here, AI/ML approaches can offer excellent performance because their architecture lends itself natively to handling and reconciling multiple data channels. Artificial neural networks enable complex information processing by combining nonlinear response nodes through a dense set of interconnects; the connecting network weights are tuned to yield the desired input–output response and thus encode the desired complex computation. These networks can be physics-constrained through a variety of methods, including pretraining on physically constrained synthetic data, sophisticated constraints on the loss function, or tailoring the network architecture by enforcing physically meaningful output to certain intermediate layers.

To realize the full potential of autonomous experimentation, new decision-making algorithms that allow integration of material physics must be developed. Existing work in Bayesian frameworks can be adapted to allow for arbitrary physics "priors" to constrain models. Such priors can guide experimentation by focusing results on a target or parts of the parameter space where models are uncertain. Moreover, such systems can be used for hypothesis testing where multiple competing models are available because they can localize measurements in regions that distinguish between model predictions. More sophisticated methods should also be sought by which surrogate models can be dynamically composed by smoothly transitioning among different physical models over different parts of the space. An outstanding challenge in the field is to combine input knowledge that spans the full gamut of experimental reality—from rigorous analytic theories, to parametric simulation studies, to coarse-grained models that may capture relevant trends but miss the absolute scale, to fuzzy heuristics and experimenter intuition.

Figure 2 depicts an optimal autonomous experimental design. Autonomous experimentation will also benefit from AI/ML approaches that can handle sparse data and finite time-horizon predictions. A promising approach is reinforcement learning (RL), a type of ML capable of dealing with unlabeled and sparse data and learning from "experience" in dynamic environments with limited foresight [54–60]. RL is based on goal-oriented algorithms, where suitable actions are taken to maximize reward and to identify a policy the algorithm should apply in a specific situation. Unlike in supervised learning, in RL there is no single correct answer; instead, an agent decides how to best perform the given task, learning in a trial-and-error manner. The learning agent formulates optimal policies while simultaneously maximizing its reward with respect to the situation it currently faces. Current RL approaches have been largely developed for small-scale problems, with enormous success. They have been enormously successful in navigating tasks like playing games such as Go (AlphaGo) [55, 59]. Further research efforts could lead to RL approaches adaptable for SUFs' needs, in particular handling large state spaces and continuous reward problems.

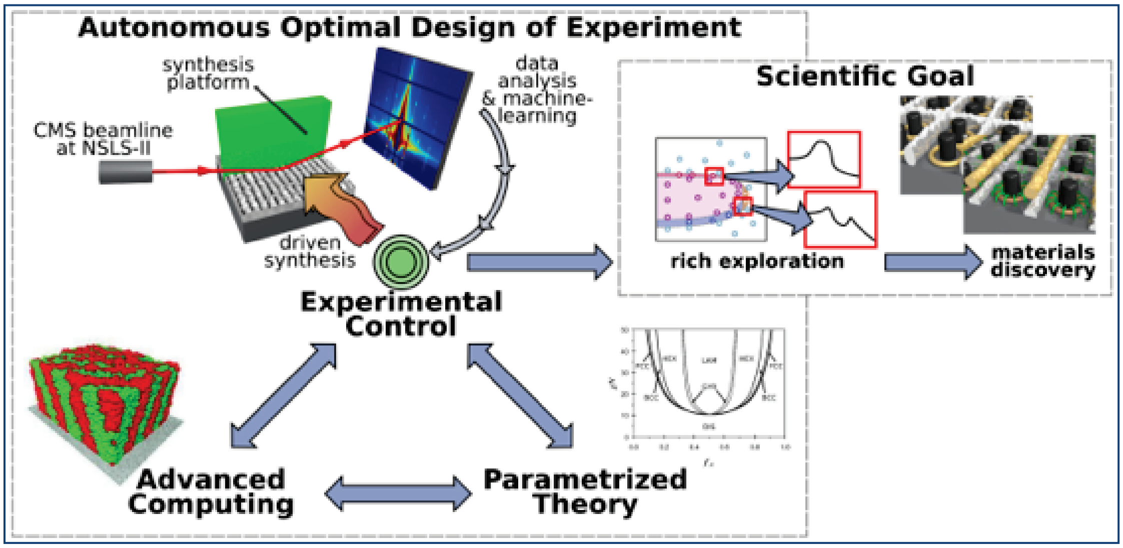


Figure 2. Autonomous experimental workflow being developed for the Complex Materials Scattering beamline at the National Synchrotron Light Source II; similar motifs can be envisioned for a wide range of experimental tools. By incorporating input from theory and fast real-time modeling into the decision-making algorithm (here, a method developed by the CAMERA [Center for Advanced Mathematics for Energy Research Applications] project [41]), material synthesis can be controlled and driven in real time. This enables access to previously impossible classes of materials, especially metastable states that appear during nonequilibrium ordering. | *Image courtesy of K. G. Yager, Brookhaven National Laboratory*

Enabling Capabilities

Online control of accelerators, measurement instruments, and experimental platforms requires corresponding advances in infrastructure and facilities.

- **Computing infrastructure:** Real-time control coupled with material modeling requires fast and flexible access to databases of precomputed results, as well as the ability to trigger model computations. This requires the development of new computing infrastructure that can handle the intense and unsteady workload that will arise in response to dynamic experimental execution. Ideal solutions will need to combine edge computing, elastic access to centralized HPC resources, and cloud computing integration. Moreover, new access modes to DOE compute infrastructure for on-demand HPC should be investigated.
- **Edge computing:** Ultra-fast recomputation of large and complex models will require computing at the edge, exploiting specialized hardware (including graphics processing units [GPUs] and field-programmable gate arrays [FPGAs]) where appropriate.
- **Data practices:** Integration of experiment, theory, simulation, and AI/ML will require significant changes in the scientific community with respect to data practices. Community members will need to

share data more broadly while being mindful of credit and incentives and will need to establish standards for data curation, annotation, and aggregation. Mixing theory and experimental results will require planning with respect to consistent tagging, nomenclature, and data representation formats, such that results from different models can be compared and seamlessly integrated into experiment.

- **Workflow infrastructure:** Infrastructure development will need to integrate existing materials and chemistry databases into autonomous workflows. Simple, flexible standards should be agreed upon such that new databases from collaborators and industry partners can be easily adapted to online control environments. Autonomous experimental control should leverage the multiple probes and imaging modalities available across the DOE complex (e.g., x-ray, neutron, electron, optical, and scanning probes). Development of multimodal science will require collaboration across user facilities, amplifying the needs discussed above with respect to data curation and sharing.

Autonomous, smart control attempts to automate a complex control loop, requiring integration of improved data collection, data analysis, system modeling, and decision-making. Development in this area thus requires corresponding advances in the other PROs.

- **PRO 1:** Advanced analytic and collection strategies are required to intake and process data, delivering derived insights suitable for driving online control.

- **PRO 3:** Decision-making algorithms should be pretrained based on synthetic data obtained by running virtual experiments and continually updated based on advanced physical models.

- **PRO 4:** The training sets required for AI/ML algorithms must have a shared data infrastructure.

The most crucial required advances are the AI/ML method developments that have been noted that will enable unprecedented computational efficiency and complexity (see section on Enabling Capabilities in Computer Sciences and Mathematics)

- **Search/optimization:** Advances in data mining and search/optimization methods are required to handle the complex, high-dimensional spaces inherent to scientific problems.

- **Correlation analysis:** Advanced correlation detection would enable fault prediction for accelerators and self-calibration for experimental tools. Identifying correlations between datasets would also empower new sets of multimodal measurement schemes.

- **Uncertainty quantification:** Online control algorithms that correctly incorporate experimental uncertainties and costs are required. The variety of use cases requires a range of strategies, including Bayesian methods, RL, and active learning.

- **Approximants:** Fast approximants for both data analysis and system or material modeling are critical for the autonomous experimental loop to run in real time.

- **Physics from AI/ML:** Equation-learning methods can contribute to theory building in physics and chemistry, as these methods enable the direct determination of physical equations from the data [61–62]. From a theoretical perspective, purely numerical solutions, while valuable, limit further development that can be done with analytic solutions. While ML tends to produce numerical solutions, developments are ongoing in learning equations. Obtaining equations from the data provides a more easily, thoroughly interpretable result that is more easily converted into other forms and a compact formulation of the result.

Potential Impact

Intelligent automation has the potential to revolutionize science, allowing scientists to tackle more challenging problems while also liberating them to think about science at a higher level. Conversely, it is becoming increasingly clear that the current approach does not fully leverage the capabilities of modern scientific tools. The rapid growth in brightness of synchrotrons [63] and similar trends for other high-end experimental tools (e.g., modern electron microscopes that can achieve enormous frame rates approaching 100,000 images per second) may not be fully utilized owing to the currently limited analysis pipelines. Advanced automation of experimental workflows will allow scientists to take full advantage of the capabilities of modern tools, while also allowing them to investigate problems of a complexity previously considered too daunting. Figure 3 shows the expansion of synchrotron publications output and improved brightness over time.

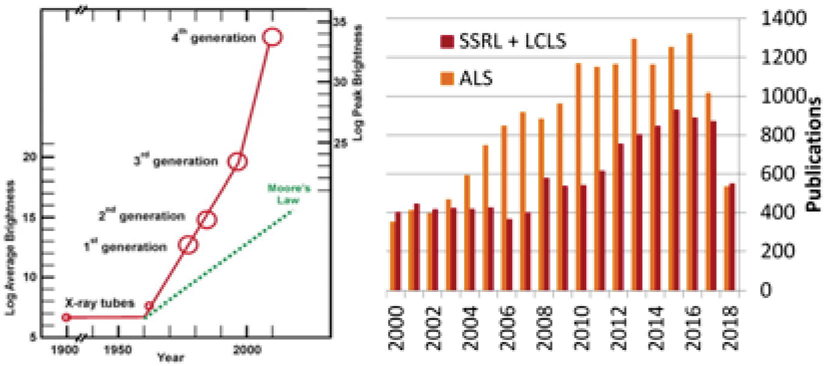


Figure 3. The brightness of synchrotron light sources has grown enormously over time (left), even outpacing the rapid scaling observed for microelectronics (Moore's law). The publication output (right) from synchrotrons has also increased, but not as dramatically as source characteristics. This implies that existing light sources have untapped potential (i.e., more efficient use of existing resources could lead to dramatic improvements in scientific productivity). | *Left image reprinted with permission from J. Stohr and H. C. Siegmann,* Magnetism: From Fundamentals to Nanoscale Dynamics *(Springer, 2006). | Right image courtesy Apurva Mehta, SLAC National Accelerator Laboratory*

New accelerator capabilities are generally associated with more difficult setup and operation. For example, a storage ring with ultra-high brightness often has an ultra-small stable operating space known as the dynamic aperture. Dynamic aperture is especially small during commissioning, when a host of errors have not yet been compensated. The performance of a future storage ring design may also be limited by the need to reserve a dynamic aperture overhead, which could be eliminated by advanced tuning methods. Similarly, fast implementation of challenging XFEL operation modes will enable new types of scientific experiments by delivering exotic beam configurations to users. The development of autonomous accelerator operation will revolutionize the design and operation of future accelerators as well as large SUFs in general: machine control will be largely automated; accelerator tuning will be carried out by efficient, consistent computer programs; the central control program will be fully aware of the status of accelerator components and subsystems and thus able to make tuning and maintenance decisions. The ability to ensure design performance through advanced tuning methods will have a substantial impact on accelerator design. AI/ML methods have the potential to achieve unprecedented capabilities and availability for future accelerators.

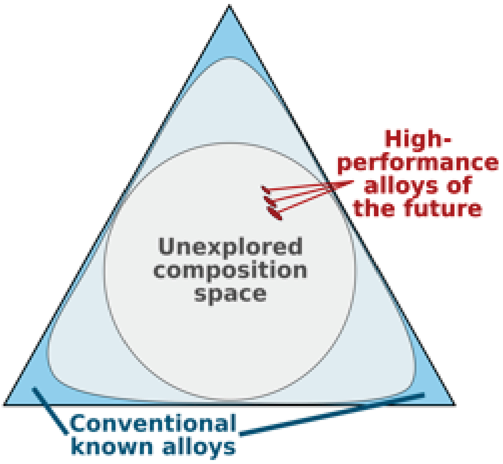
Advanced autonomous experimentation holds promise for revolutionizing materials, chemistry, and bioscience discovery by providing means of identifying exotic and high-performance materials previously hidden in a sea of complexity. It is difficult to overstate the potential impact and the breadth of experiments that will benefit. One can anticipate significant impact to problems currently hindered by material compositional complexity, including formulations and blends, biomaterials and biomimetic systems, and alloys. For example, certain classes of metallic glasses may hold promise for yielding ultra-high strength-to-weight ratios [64]. The ideal "steel of the future" material that would yield transformative improvements in demanding applications (e.g., aerospace) is currently hidden in the enormous space of different alloys that can be envisioned and exacerbated by the enormity of the processing space one must consider to correctly form and quench glassy metastable states. More generally, the study of pathway-dependent phenomena would be revolutionized by autonomous exploration. Self-assembling materials exhibit a set of nonequilibrium states that can only be accessed using the correct processing history [65–66]. In narrow cases, researchers have been able to perform "pathway engineering" wherein a desired target that cannot be achieved with equilibrium processing methods is selected and enforced by using the correct sequence [67]. Online control of synthesis platforms would generalize upon these early successes, allowing researchers to navigate complex assembly landscapes and guide the full spectrum of complex self-assembling materials, including block copolymers [68–69], liquid crystals [70], supramolecular structures [71], nanoparticle superlattices [72–75], and DNA [76–78] into vital structural motifs. The exploration of many classes of functional materials could be greatly accelerated by tight coupling with appropriate material modeling. For example, design of advanced thermoelectrics would benefit from experimental searches with coupling of structural/spectroscopic probes, operando functional measurements, and structure–property modeling. Similarly, studies of quantum heterostructures already benefit greatly from detailed physical simulations. Integration of these predictive theories into the measurement loop would optimize experimental searches for unique materials geared towards quantum information science applications.

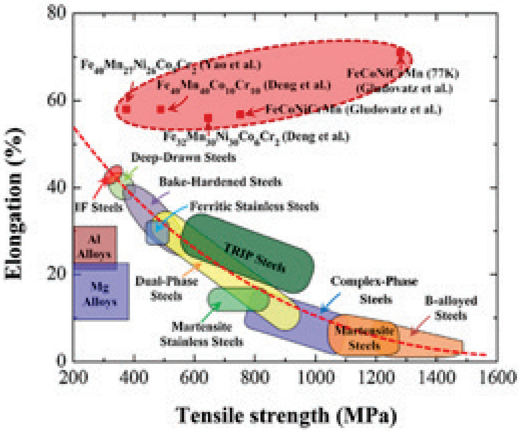
A broadened research program in autonomous experimentation would be expected to yield both near-term and long-term impacts. In the near term (3–5 years), dedicated research should yield a set of specialized tools including models, AI/ML methods, and hardware systems for autonomous exploration of samples. In the long term (10 years), it should be possible to deliver robust generalized autonomous synthesis platforms, which can tackle a wide range of materials, chemistry, and bioscience problems while simultaneously revealing new physics. Overall, the goal of autonomous experimentation is to liberate scientists from the task of micromanaging the execution of experiments, including optimizing experimental conditions, which will allow them to tackle scientific problems at a higher level.

Many of the experimental tools developed within the DOE complex could benefit from advanced AI/ML control methods. Proposed AI/ML methods will improve performance and stability, benefitting all user experiments through enhanced uptime and reliability, while also increasing the sophistication of experiments that can be executed, benefitting the most ambitious and cutting-edge research programs. As these advanced experimental tools underlie a wide variety of modern scientific studies—from geosciences chemistry to biosciences to energy research—improvements would have broad benefits throughout the BES research program.

“Steel of the Future”

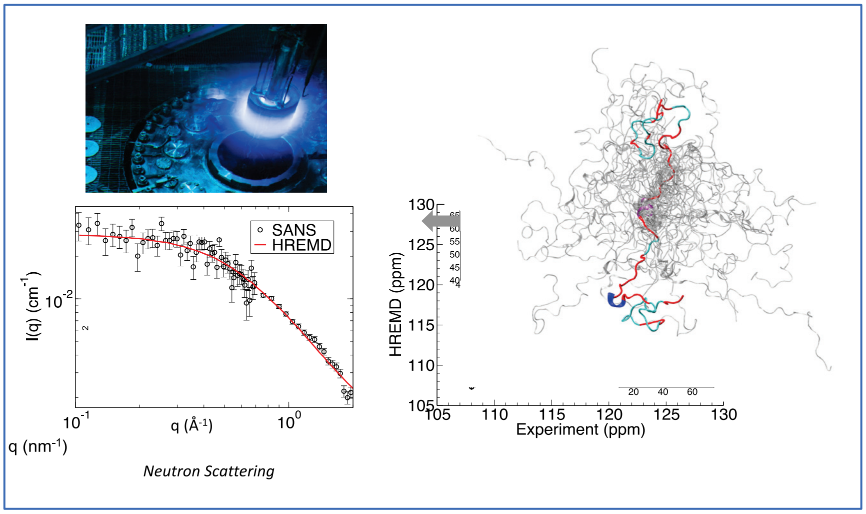
Amazing new materials are waiting to be discovered. How can we find them?


High-performance alloys of the future
Unexplored composition space
Conventional known alloys



Conventional alloys are prepared by mixing together a small number of metals. Conceptually, such materials access only a tiny fraction of the total possible space of alloy materials. High-entropy alloys contain a higher number of elements than conventional alloys, enormously expanding the parameter space of possible compositions. Such materials hold promise for record-setting mechanical properties (e.g., strength-to-weight ratio), especially if frustrated and metastable states such as found in metallic glasses can be accessed. However, the enormity of these parameter spaces cannot be explored using conventional methods, or even naïve high-throughput searching, because high-performance materials represent an infinitesimal island in an enormous ocean of uninteresting materials. Autonomous experimental modes, leveraging input from accelerated physical modeling, can efficiently search such spaces, identifying interesting outliers and guiding further studies in meaningful directions. Properly implemented, such methods could yield the high-performance alloys of the future that would have important applications in transportation, aerospace, and energy harvesting. | *Left image courtesy K. G. Yager, Brookhaven National Laboratory.* | *Right image distributed under a Creative Commons Attribution Noncommercial License*

Decoding the Structure of Intrinsically Disordered Proteins



Small-angle neutron scattering and Hamiltonian replica exchange molecular dynamics simulation can synergistically generate the configurational ensemble of a flexible protein.

The structural characterization of flexible biosystems poses a major challenge in biology because of intrinsic protein disorder. It is critical to gain structural information to understand protein function. To help decode the complexity of a disordered biomolecule, a combination of neutron scattering and high-performance molecular simulations can be used to generate the configurational ensemble (i.e., the collection of 3D structures the biomolecule adopts) [79]. However, this combination currently can require weeks to gather the appropriate information for a single biomolecule. AI/ML approaches could be used to learn how to optimally couple neutron scattering data with high-performance molecular simulations in real time. This would enable AI/ML-driven steering of simulations toward experimental neutron scattering results that in turn would significantly reduce the time to solution. Such a tremendous acceleration in the ability to decode structural characterization could have tremendous impact on the structural biology of flexible biosystems. Similar challenges are feasible in x-ray and electron microscopy; while current approaches reconstruct the average structure, new methods in AI/ML are starting to reveal conformational changes [80–81].

Top left image reprinted from https://www.energy.gov/ne/articles/7-fast-facts-about-high-flux-isotope-reactor-oak-ridge-national-laboratory. Courtesy Oak Ridge National Laboratory. /Bottom left image reprinted from https://www.ornl.gov/news/supercomputing-neutrons-unite-unravel-structures-intrinsically-disordered-protein. /Right image reprinted from Shrestha, U. R.; Juneja, P.; Zhang, Q.; Gurumoorthy, V.; Borreguero, J. M.; Urban, V.; Cheng, X.; Pingali, S. V.; Smith, J. C.; O'Neill, H. M.; and Petridis, L. "Generation of the Configurational Ensemble of an Intrinsically Disordered Protein from Unbiased Molecular Dynamics Simulation." Proc. Natl. Acad. Sci. U.S.A. *116 (2016): 20446–20452. doi: 10.1073/pnas.1907251116.*

PRO 3. Enable Offline Design and Optimization of Facilities and Experiments

Key question: *How do we enable virtual laboratories—offline design and optimization of facility operation and experiments—to achieve new scientific goals?*

Introduction

A key challenge within the SUFs lies in the design and optimization of facilities and experiments to achieve scientific goals. On the facility side, modern SUFs house expensive and complex accelerators, which are challenging to design, build, and operate. Experiments are time-consuming and require carefully planned sequences of steps, including formulating the scientific hypothesis, performing the experiment(s), modeling the results, and theory–experiment matching with data analytics to draw conclusions. A key challenge is to optimize experimental planning at both the individual and facility levels to reduce time to discovery of scientific knowledge, minimize redundancy, and maximize physics knowledge from each experiment.

Due to the complexity and high cost of experiments conducted at the SUFs, and the difficulty in simulating all parts of the experiment and facility, lengthy iterations of trial and error to achieve optimal experiments are rarely feasible. This can substantially reduce the scope of possible experiments. Moreover, complete knowledge of the probing signal (e.g., at a beamline) can potentially be brought to bear on postexperiment data analysis, but this is rarely done due to unavailability or invasiveness of measurements and incompatibility with experiments. For example, the wavefront of the x-ray pulses from an XFEL can be useful information for users, but its measurement precludes delivery to the sample on the same pulse. Thus, some combination of simulations and virtual diagnostics is necessary.

Examples of planning and optimization challenges at modern SUFs include determining optimal synthesis conditions for a new material; selecting the correct set of multimodal experiments to solve a structural inverse problem; optimizing the parameters of specific tools to achieve computational and experimental end goals; and generating precise, continuously calibrated models of accelerators for data analysis and interpretation. Recently, the use of AI/ML approaches has shown promise in situations involving path planning and optimization under uncertainty [82]. At the same time, experimental design and optimization is best performed in a simulated environment to allow for exploration of the parameter space in silico, as only a small number of experiments can ever be conducted in real laboratories and experimental time is costly.

Key to achieving this objective is the need for a digital twin of each SUF that enables users to design, operate, and optimize experiments in a safe, virtual environment guided by AI/ML so they can seamlessly transition to the real facility, reducing the time to scientific discovery [83]. Such virtual laboratory environments (figure 4) will need to be closely coupled with the experimental facilities to continuously update their simulations to reflect objective reality (e.g., to run virtual experiments off-line that reflect what would happen at the SUFs). Furthermore, these virtual environments require high fidelity as well as a mix of accurate and fast simulations optimized for efficient reproduction of the essential physics of the laboratory measurement or synthesis process.

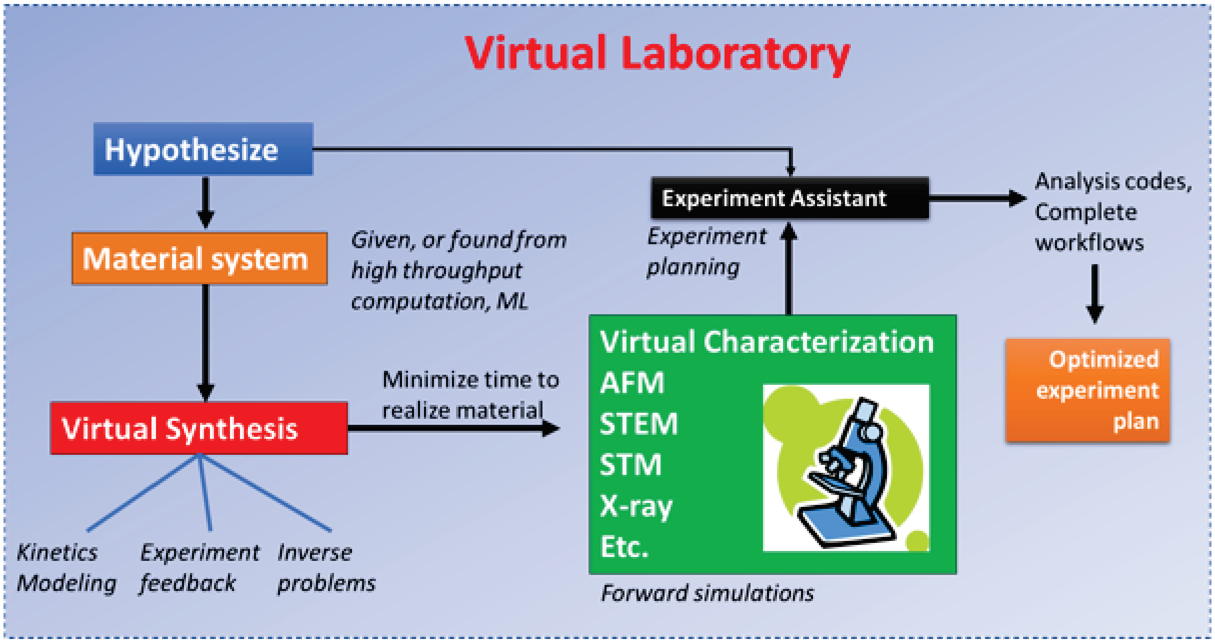


Figure 4. Virtual laboratories can enable optimization of experimental plans, accelerate training, provide starting points, and enable automated generation of analysis codes and workflows for end-to-end scientific experiments, from hypothesis to execution to analysis. | *Image courtesy Rama Vasudevan, Oak Ridge National Laboratory*

Research Directions

The primary research objective is to create physically accurate, virtual laboratory environments of experimental facilities that guide conception to synthesis and/or characterization in silico, closely coupled to the actual facilities and continuously updated based on real experiments to ensure faithful reproduction. This will enable automated and/or AI/ML-assisted design of optimal experimental strategies and analysis workflows for physics knowledge acquisition.

These virtual laboratory environments should include:

- Rapid simulations
  - Rapid on-the-fly methods (including surrogate modeling) within the virtual environment for simulating results
  - Acceleration of simulations via both hardware methods (e.g., FPGA, GPUs) as well as more efficient numerical approximators, such as DNNs
- Accurate simulations and theory–experiment matching
  - First-principles modeling (e.g., of predictions from heterostructures)
  - Calibration of physics models to observations in a continuous fashion, and appropriate theory–experiment matching routines
- User interfaces
  - Immersive, interactive 360° environments delivered to train and enable users to design and conduct virtual experiments, analyze data, and modify scientific hypotheses

Once built, these digital twins would then enable:

- Experimental planning and design
  - AI/ML-enabled context-specific experimental design, along with ML-generated analysis workflows
  - AI/ML-guided planning using, for example, reinforcement learning [84] or genetic algorithms to discover the optimal sequence of measurements to answer a scientific question
  - Virtual diagnostics that provide real-time input to both accelerator operation and user experiment data analysis
- Facility optimization
  - Expedited accelerator design and precise accelerator control
  - Statistics on the common use patterns of instruments and potential opportunities for optimization (e.g., co-location of different instruments and staffing)

More generally, the digital twins are expected to expand the scope of possibility for experiments, given that advanced planning can enable tackling of more ambitious projects.

One example of the envisioned utility of digital twins and the challenges involved arises from modern accelerators. Creating a high-fidelity digital twin implies an accurate model that can predict the accelerator performance and the particle beam characteristics. Such an accelerator model is critical to the design, analysis, and interpretation of user experiments, as well as the continuous improvement of the machine, as the model can be used in accelerator tuning and control and upgrade studies.

Virtual laboratories could additionally assist with user onboarding and training procedures, making facility operations more efficient, and aid in planning and design of existing and new facilities.

An accelerator is always built with a design model, which is based on simulation of the physics processes involved. The design model is the basis for machine operation, for example, setting the working point of accelerator components. However, the actual machine often deviates significantly from the design model. Owing to the differences between the model and the machine, the accelerator typically does not reach the desired performance without extensive efforts to adjust experimentally the control parameters.

Calibration of a physics model with measurements can bridge the gap between the model and the machine. It would enable the discovery and compensation of errors in the machine and precise prediction of the machine performance. Presently, model calibration is typically based on minimizing model predictions and measurements with least-square fitting. It is applicable only to some limited subsystems with strong measurable signals, such as storage ring and linac linear optics [85–88] and may suffer from significant under- or over-fitting [89]. New AI/ML methods, such as Bayesian inference techniques [90], can enable precise and comprehensive model calibration, with coverage wider and deeper than traditional techniques. Examples include the calibration of storage ring nonlinear beam dynamics or start-to-end models of XFELs. AI/ML may also be deployed to model accelerator subcomponents, especially with regard to predicting anomalies or imminent failure. Predictive models will enable preemptive maintenance so facilities avoid unscheduled downtime. Such models will also enable rapid identification of the root causes of faults, expediting recovery and minimizing recurrence of the fault. Rapid tuning and fault prediction both require fast execution of modeling and control algorithms, which can be AI/ML-accelerated.

It should be noted that complete physics modeling of a complex system may involve intensive computer simulation. The complexity of the accelerator and experimental systems, and the level of accuracy required for simulations, demand costly computing resources. This limitation precludes digital twin applications that require frequent and fast model evaluations. For example, experiments on the LCLS-II could benefit from knowledge of photon beam characteristics, which could be predicted with the model, but only through hours of computer simulation; thus, this knowledge would not be available in real time. AI/ML can enable modeling that is millions of times faster than physics-based modeling by employing flexible neural networks or other models that are trained from simulation or experimental data to serve as substitutes for the first-principles simulations [91]. The surrogate models can be constantly updated and refined, enabling high-fidelity predictions of machine performance with extremely rapid predictions.

Another example of the envisioned utility of digital twins is to reduce the time to synthesis of materials and chemicals with desired properties. Assuming that a target structure is predicted from a first-principles model, a series of syntheses could be carried out within a virtual environment, with a small subset of real experiments serving to ground and tune the synthesis models. Subsequently, optimal characterization schemes could also be found within the digital twin to determine whether the structure of interest was generated; ideally, the analysis routines could be self-generated within the digital twin to further reduce the time spent. The latter is a particularly important aspect of facility operations in addition to experimental design that enables both to be optimized in the same step.

For first-principles models, fast approximants for materials' interaction potentials used in a wide variety of simulation codebases would enable a considerable expansion in simulation size, as well as improved coupling to real-time experimental platforms (PRO 2). One of the major challenges in performing molecular simulations is sampling the complex phase space, which puts many phenomena outside the capability not only of today's computers but also those planned in the near future. Recent advances have been made in sampling with ML [92]. For example, Boltzmann generators are an ML algorithm that determines the invertible transformation between the Boltzmann distribution and a Gaussian distribution. Conceptually, this work shows a path to addressing a major bottleneck in using molecular simulations to determine material behavior from the molecular constituents. Such developments need to be part of the capability development for the SUF user community as it appears to be a paradigm shift. Additionally, developments for other aspects of molecular simulation are still needed. For example, nonequilibrium dynamics are not currently handled by Boltzmann generators.

To satisfy the diversity of SUFs problems, a wide variety of modeling approaches should be evaluated and implemented to achieve a digital twin. AI/ML models would ideally identify important physical parameters, such that these could be modified without retraining the entire network. This would allow a co-design approach, where the physics model parameters are identified and refined alongside data collection. Even more exciting would be the development of AI/ML models that are predictive for a wide range of analogous but distinct physical problems. Convolutional neural networks (CNNs) combine the computational richness of networks with nodes that perform local convolution operations on datasets. The hierarchy of convolution operations inherently aggregates features and can thus be used to model hierarchical physical phenomena. For example, the assembly of colloids, nanoparticles, proteins, block co-polymers, and liquid crystals could all be captured by a generalized CNN model that is trained on the underlying physics of interacting anisotropic building blocks, with each specific system being represented only by slightly different network weights. Rigorously mapping known physics parameters to the CNN weights would be of added value, thereby yielding an interpretable ML model where the meaning of retrained weights can be inferred. More generally, what is sought by scientific researchers are AI/ML models that return meaningful physical insights; in this sense, studies in interpretable AI/ML applied to specific scientific problems would be fruitful.

Enabling Capabilities

To achieve the level of accuracy and speed in facility modeling required by the digital twin paradigm, development of new capabilities is needed, including:

- Adoption of a unified data management system that provides coverage from facility status monitoring to diagnostics readbacks. This system will achieve facility operation data in a consistent, accessible manner to facilitate the application of ML methods for the creation of a digital twin. The data record across the facility needs to be synchronized.
- Development of AI/ML methods that can train facility-scale large models on heterogeneous input data of different sources and format and are able to impose physics-principle constraints to the AI/ML models.
- Development of facility control systems that can accommodate AI/ML data flow requirements (e.g., allowing local GPU integration or low-latency remote GPU access).
- Development of AI/ML methods to reliably evaluate the uncertainty of the AI/ML models on a large scale and with diverse input/output data types to ensure the digital twin's validity.

AI/ML advances

AI/ML advances needed to realize the virtual lab environments include:

- Rapid inverse structure predictions from imaging or spectral data, with uncertainty quantification.
- AI/ML-assisted speedups for dynamical simulations optimized for specific hardware.
- RL and Bayesian learning algorithms for efficient exploration of large multidimensional parameter spaces under uncertainty.
- Feature learning under realistic experimental constraints for theory–experiment matching.

Potential Impact

The use of digital twins of SUFs can open new paradigms for experimental science. The potential impacts include:

- Order of magnitude decrease in time from theory to practical realization of a new materials phenomenon (e.g., of new electric topological order to a synthesis "recipe" for creating the material itself).
- Optimal experimental design for specific mechanistic questions such as mechanism of long carrier lifetimes in semiconductors for photovoltaics or high electromechanical response in ferroelectric relaxors.
- Expedited accelerator design cycles and thorough search of parameter space, which will maximize facility performance.
- Noninvasive virtual diagnostics that provide real-time information to facilitate automation of user experiments and accelerator operation.
- Enabling of new experiments that were previously infeasible due to perceived risk from a lack of analysis routines or complexity of the steps.
- Environments to train autonomous AI/ML-driven agents for scientific exploration.

PRO 4. Use Shared Scientific Data for Machine Learning–driven Discovery

Key question: *How can we catalyze scientific discovery by leveraging the wealth of diverse and complementary data recorded across the BES scientific user facilities?*

Introduction

Although scientific knowledge and operational insights are shared across the existing SUFs, most data analytics, infrastructure, and workflows are siloed. Operation in isolation and a lack of tools to search and analyze datasets leads to repeated work, unnecessary experiments, and missed opportunities to leverage the vast amount of data collected at facilities. Radical improvement in data sharing, analysis, and curation can catalyze scientific discovery across facilities, resulting in transformative tools for multimodal, multiuser science and creating a test bed to develop the next generation of AI/ML tools for the BES and SUFs communities [93].

This section describes an opportunity to build a common facilities data repository to house the collective output of the BES SUFs. To facilitate data sharing, the repository would need to include infrastructure to support the full data lifecycle: AI/ML tools to aid automated recording and structuring of metadata; annotated, curated, high-quality datasets to guide future use; tools to format, search, and analyze both data and metadata; and finally, benchmark datasets to help train new AI/ML models and advance research across the SUFs.

Development of a searchable, common repository of scientific data will accelerate experimental design and enable hypothesis creation and observation comparisons. Integrating diverse scientific data resources would enable automatic development of benchmark datasets built from heterogeneous experimental and simulated data; these training sets could both speed up development of AI/ML methods described throughout this report and contribute to development of scientific AI/ML capabilities across the DOE complex. A byproduct of the repository would be scientific domain-specific schemas and abstractions. This would expand *search* from simple metadata exploration to *investigation* driven by scientific motifs such as crack formation in composite materials or phase transition in simulations that can catalyze coordinated efforts toward defining standards, formats, and priorities across SUFs.

Research Directions

By 2025, the BES SUFs could generate thousands of petabytes of data per year. While individual user groups may extract science from their own data, at present the scientific community is missing the opportunity to leverage the totality of acquired data to improve the SUFs and accelerate discovery. This PRO provides the vision for a shared data repository that spans facilities and scientific domains. The repository would include infrastructure throughout the data lifecycle, with critical capabilities during acquisition of data and metadata; curation of high-value datasets; search; and multimodal, multiexperiment analysis. AI/ML could be harnessed to improve this process, with autonomous data curation working to capture provenance, context, and data quality and tools to enable large-scale, multimodal search and analysis. The ultimate goal is to coordinate continual creation, curation, and application of large quantities of data and knowledge as well as associated models, workflows, computations, and experiments. Finally, byproducts are discussed, including creation of benchmark datasets and coordinated efforts centered on emergent scientific motifs. Many topics in PRO 4 are discussed in greater detail in the ASCR Data and Models for AI Workshop report [22], the *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence* [94], and the international call for FAIR (findability, accessibility, interoperability, and reusability) datasets [95].

This PRO identifies four research themes associated with a shared data infrastructure:

1. Automated capture of data and metadata to ensure all SUFs data are stored with high-quality metadata from every experiment and calculation.
2. Data search functions that can locate relevant, high-value datasets.
3. Meta-analysis for simultaneous analysis of diverse, multimodal datasets.
4. Benchmark datasets created from the shared data that can train new AI/ML models and support general research and development (R&D) in scientific AI/ML.

In brief, the research directions examine how data should be stored, how the data should be accessed once stored, how relevant datasets should be analyzed, and one particular application of using the data to develop new AI/ML models and methods. This report did not consider potentially significant issues on data privacy, embargo, and culture of sharing across research groups.

- **Automated capture of data and metadata:** Regardless of how data are generated, automating manual data capture and curation is key to increasing both the quantity and quality of data collected and their usability. A major hurdle to sharing datasets is the acquisition and structuring of metadata to describe experimental conditions, samples, acquisition parameters, anomalies, and data quality. Current facility users may lack the time, tools, and/or incentive to collect and structure metadata themselves; therefore, lowering the barrier to generating complete data provenance will be critical to the repository's success. A major research direction will involve automating the collection and structuring of both data and metadata as part of typical SUF operation and experiment. AI/ML should be harnessed to improve this process, with autonomous data curation working to capture provenance and context information and to encode associated uncertainty ranges. AI/ML methods (e.g., natural language processing to read logbooks) are expected to play a role in this challenging task.

- **Data search functions:** A shared repository assumes users are able to collect high-value data relevant to their study (i.e., access to a robust, sophisticated search function). To locate a relevant experiment or data collection, the search function should provide access to keywords from specific scientific domains. The need for keywords also presumes that datasets are well labeled and that facilities should enable automated data tagging of relevant descriptive keywords.

  A key aspect of a search is the ability to identify data quality to guide future selection and use. For example, metadata should highlight anomalous behavior to warn scientists of problematic content, or it could be used to identify datasets of particularly high quality (e.g., stable experimental setups, high signal-to-noise ratio). This general area of data curation will also require automation to be applied to large-scale SUF datasets.

A shared data repository enabled by AI/ML will leverage the yearly work of 16,000 users at scientific user facilities to accelerate discoveries. An easy-to-use search for all the data collected at the facilities will unearth new research targets, ideas, and previously undiscovered correlations.

- **Meta-analysis:** The repository should enable new scientific approaches for meta-analysis across a large and diverse set of experimental data. The heterogeneity of datasets collected by different experimental groups at multiple facilities using a range of methods presents a host of new challenges. Datasets will come in different formats, use different samples, and suffer from different experimental anomalies. Moreover, meta-analysis methods will have to be efficient and scalable to allow application to real-time control tasks (see

PRO 2). AI/ML methods are expected to play a critical role incorporating metadata into the combined analysis.

- **Benchmark datasets:** Though the datasets in the repository are expected to serve specific domain science goals, automated development of large structured, labeled datasets presents an opportunity for the scientific AI/ML community as well. Just as benchmark datasets such as MNIST and ImageNet [96, 97] played a critical role in the development of AI/ML methods in industry, benchmark scientific datasets can open new avenues for AI/ML in science. Given the differences in scale, problems, data types, and questions asked in science versus industry, it is expected that scientific datasets will be necessary for AI/ML to reach its potential. Development of benchmark datasets will require significant effort in curation; datasets should be cleaned of anomalies and missing data; and accuracy of labels will determine accuracy of the final model. Some datasets will require uncertainties on both the data and labels. Datasets should also be divided into train, test, and validation sets; division can be a surprisingly challenging task in complex scientific datasets where data leakage may introduce correlations between different samples. Finally, there should be a diversity of datasets to target the wide range of tasks encountered at the SUFs.

Enabling Capabilities

A shared data repository will require development of a range of enabling tools. Some of these will be directly related to AI/ML capabilities, and some will support knowledge representation, data curation, data integration, and mechanisms to access the data for a variety of purposes.

- **Workflows:** Analysis requires integrated tools to sort, rank, compare, and guide scientific discovery at different levels of fidelity determined by different computational capabilities. Whether running models at ASCR facilities, leveraging edge computing resources, or using resources at an SUF, users require data, models, and workflows that are matched with available computational capabilities and different expectations of accuracy. For example, depending on the needs and resources, tested models may be needed for shallow neural network feature extraction for low-dimensional descriptors or DNN feature extraction for high-dimensional, high-fidelity description.

- **Standardized file formats:** One of the main challenges in any repository is the need to standardize file formats to enable search and analysis to take place. In particular, data models need to be capable of representing most commonly observed data regardless of its size; dimensionality; or lack of N-dimensional form, modality, precision, or even instrument of origin [98]. Most SUFs produce open data file formats, although some instrument vendors still output formats that are difficult to integrate into modern ML workflows. As a result, repositories will need to accommodate a subset of common file formats that can easily be addressed via appropriate translators that can ingest data from many formats and convert them into the repository's allowable formats.

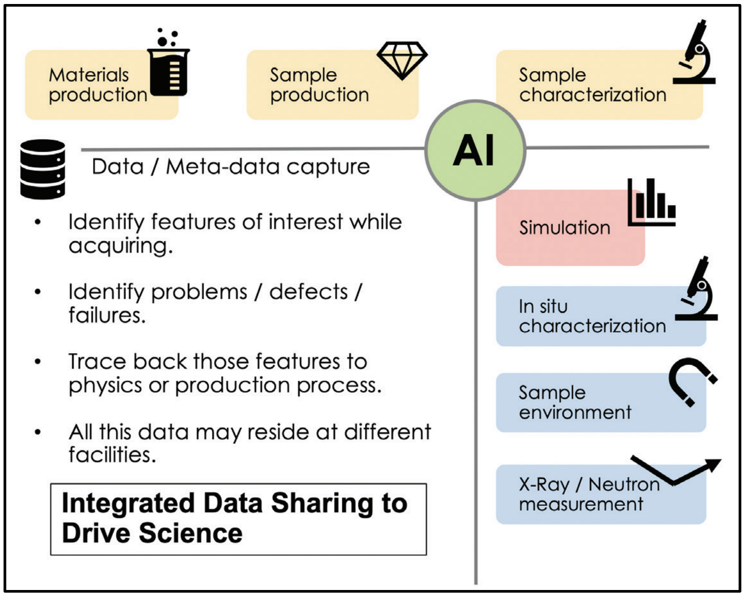
- **Catalogs:** Search functions will require catalogs of linked data, metadata, analysis workflows, and results (e.g., scientific motifs that can be searched, studied, and discovered across facilities). The challenge with this task is to enable open and interoperable access patterns for facilities, given that they may be producing data from different instrument types, stored in different formats, some proprietary, and described using a variety of metadata without a common ontology. Adding to this challenge is the fact that some of that data may be incompletely described or of poor quality. A system to assess data quality should also be considered to address those points.

- **Assembly tools:** Creating benchmark datasets will require tools to assemble training sets from heterogeneous experimental and simulated data. Having a common repository of tagged data will be useful for creating training sets for AI/ML models. Combining multifacility and/or multimodal data to form a cohesive collection that can be processed by AI/ML algorithms is a challenge that will require new tools, such as registration methods.

- **Integration with existing databases:** Many important repositories already exist from the perspective of SUFs. For example, theory databases with force-field or DFT calculations, experimental libraries of materials synthesis, and crystal structures [99]. Providing appropriate links to external libraries where possible (e.g., to link a particular electron microscopy image to a crystal structure) will expand the possibilities of ML by providing access to a richer feature space over which to mine, categorize, and correlate data.
- **Recommendation tools:** Deployment of recommendation tools should be tied to catalogs that leverage metadata entered by the user in a previous interaction with the system, characterizing an intelligent interface that can auto-complete metadata upon request and by which the auto-complete function learns with each user individually, creating models that are customized for each user. By having such services, users may be incentivized to curate and share more of their experimental data.
- **Labeling benchmark datasets:** The benchmark datasets need to consist of simulated and experimentally measured sets that are fully tagged and cross-verified by domain experts. The datasets will need to be accompanied by a set of basic AI/ML models and results to provide a baseline for new developments and competition. In addition, the sets will bolster new instrumentation with a baseline for compute, accuracy, and time requirements to provide scientists with AI/ML-driven results.

Potential Impact

A shared scientific data infrastructure is the cornerstone in development of AI/ML capabilities in the coming decades (figure 5). PROs 1 and 2 described how AI/ML can help push scientific discoveries at the BES SUFs, and PRO 3 described how those capabilities can be brought to the user community and facility operators. All these advances need a shared scientific data infrastructure that provides the tools to access experimental data and algorithms from different instrument types and from all the SUFs. Having a shared data infrastructure linking all the SUFs will significantly impact science output in several ways: shared knowledge and models that accelerate analysis and understanding, shared data for validation and forensics, benchmark training data for AI/ML models that address analysis, control, digital twins, and validation/anomaly detection.



- Identify features of interest while acquiring.
- Identify problems / defects / failures.
- Trace back those features to physics or production process.
- All this data may reside at different facilities.

Figure 5. AI-supported experiment lifecycle. | *Image courtesy Daniela Ushizima, Lawrence Berkeley National Laboratory*

Knowledge/model sharing

A major impact of creating a linked and easily searchable data system across SUFs is the possibilities it opens up for sharing knowledge. Using a scientific query language, users will be able to search through vast amounts of tagged data produced by the16,000 yearly users across BES's SUFs. With such a tool at their disposal, users will be able to assemble relevant data about their system of interest, opening the door to questions that are beyond the scope of a single experiment. For example, while user experiments typically probe one sample or a small number of samples, metastudies could enable the probing of material families to look for overarching patterns. The combination of data across multiple and varied sources could then enable better, more targeted experiments because a more holistic picture of the sample is created, for example using data from synchrotrons or neutrons to guide active learning via scanning probe microscopy and spectroscopy.

In addition to being used for specific domain questions, integrated datasets can be used as training sets to develop the AI/ML methods described in the other report PROs. Information extraction (PRO 1) presumes analysis methods have been validated on known systems, and any technique must be understood prior to application to a new domain science question. Pretraining is particularly important for online control (PRO 2) when there may not be time to train new algorithms on the current task. Likewise, the digital twins (PRO 3) presume access to both simulation and experimental datasets. The shared data can be seen as an enabling capability for each of the other PROs described in this report.

Knowledge sharing doesn't have to stop at instrument data. Trained models and analysis workflows can also be stored, described, automatically tagged, and made available for others to leverage. That knowledge sharing will greatly speed up the time from idea to publication across all fields of science the DOE supports.

Data validation and forensics

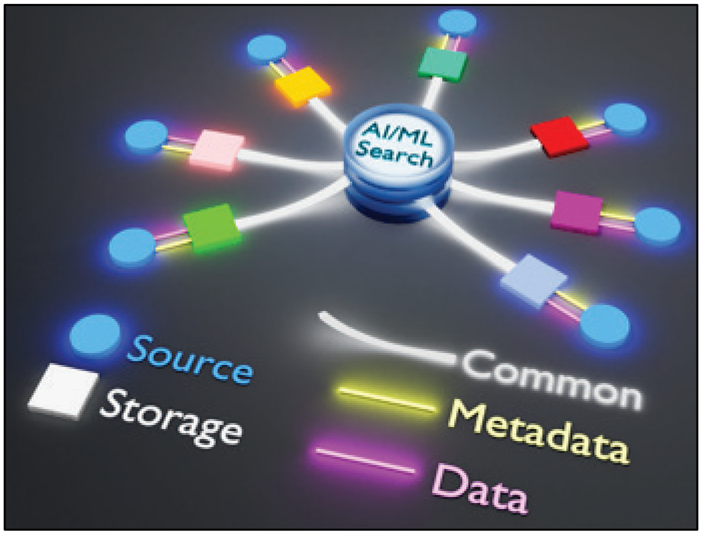
Having access to a wealth of existing data will greatly impact data quality. Assessing data quality during acquisition will ensure that the use of BES facilities is optimal. The scientific community should, however, be aware of the ways acquired data can diverge from expectations. An anomaly can identify new science, or it can point to an instrumental problem. Key to ensuring reproducibility when performing experiments is the ability to validate new measurements against previous ones. A common data infrastructure will allow scientists to tap into existing data for this purpose. A well-curated and labeled data repository will allow users to quickly retrieve the right data for the right task and explore the cause of anomalous measurements. Tracking data provenance, for example, can help identify the experimental differences that led to incompatible data. Whether sample misalignment or unrecognized differences in sample preparation caused the incompatibilities, being able to fully understand the nature of observed features will improve the quality of research produced at SUFs.

Benchmark datasets for AI/ML R&D

A shared data infrastructure will also have an impact on DOE science beyond the SUFs. Benchmark datasets have played a key role in the recent AI/ML revolution, providing both data for training and a framework for rigorous comparison of methods. Given the scale, problems, data types, need for uncertainty/robustness, and differences in questions asked in science versus industry [94], it is expected that datasets specifically designed for scientific questions will be necessary for AI/ML to reach its potential in the sciences [22]. For example, while common applications of AI/ML in industry (e.g., the digit recognition task of MNIST) assume new examples will be drawn from an identical distribution to the training set, scientific problems often involve looking for novel phenomena, explicitly or implicitly outside the training set. The SUF tasks discussed throughout this report will require special attention to robustness, uncertainty, and interpretability, reaching beyond the current state of art in industrial AI/ML. The creation of benchmark datasets specifically designed for scientific AI/ML would not only spur development of new AI/ML methods but could have a significant impact on the safe, reliable application of AI/ML to the SUFs.

Overall, an integrated AI/ML mechanism would enable searching for scientific motifs across all the SUFs' collected data and augment the analysis and decision-making with the wealth and knowledge distilled by the shared data platform. This would allow users not only to accelerate their analysis and data exploration, but to also leverage the combined knowledge of SUFs to build more accurate and comprehensive understanding of their science. Analysis workflows would not have to be remade given shared knowledge and analytics, as the integrated facility would allow searching for similar, past analysis and provide the necessary codes and references to analyze and interpret the experimental data.

Example Applications



Different sources (shown as blue circles) of data and metadata write and store data at different storage places (different color squares) using different access patterns and formats. To apply AI/ML, search and correlate data from different sources it will require building a common access platform (light blue cylinder in the center) that communicates with the individual storage locations using a "Common" access pattern. | *Image courtesy Alex Hexemer, Lawrence Berkeley National Laboratory*

By capturing data and provenance throughout the data lifecycle of an experiment or simulation, researchers will be able to use AI/ML to incorporate a broad spectrum of insight into their systems to make decisions about how best to approach a scientific question and optimize the process. Such decisions could be related to the measurement process during the experiment itself, for example, determining regions of interest in moment transfer space or determining a better force field where a simulation matches better with experiments. Similarly, this information can be used by AI/ML models to help determine modifications to the sample-making process or to develop better models. By leveraging previous scientific knowledge, such models would inform scientists about modifications to those processes that would home in on the materials characteristics of interest. With such capability in place, one can easily imagine having an integrated system where synthesis, sample-making, and simulations can be brought closer to the experiment so that scientists can make significantly greater progress during their visit to a facility. In addition to merely helping drive the experiment, such a data infrastructure would be the core building block to accelerate and drive science during experiments.

Opportunities and Challenges in Computer Science and Mathematics

The four PROs discussed in this report lay out a vision for AI/ML to transform SUF operations, creating new facility capabilities, maximizing performance, and opening new avenues of research for the scientific user community. However, in addition to investment in the SUFs' traditional research areas, bringing the PROs to fruition will require substantial advances in both fundamental and applied computational sciences. This section describes the computational capabilities needed to enable the PROs to reach their full potential.

At the heart of many of the SUFs' challenges is the explosion in data from the latest generation of facilities and detectors. Advances in data acquisition have led 90 percent of the total volume of data being created in just the past few years [100], with current estimates of a daily output of 2.5 quintillion bytes of data [101]. While the ability to record data has increased exponentially, reliance on manual/visual inspection continues to be a roadblock in many data analytics: it delays science discovery across DOE SUFs, and often precludes the full utilization of data acquired at high cost with sophisticated instruments. Manual inspection is especially problematic at the SUFs, where real-time analysis is a critical component of machine control, fault prediction and recovery, and autonomous guidance of in-the-loop experiments.

A key outcome from the BES roundtable was the identification of computational capabilities necessary to support each of the PROs. First, tools should be available to convert big datasets from the SUFs into usable and accessible forms (PROs 1 and 4). Moreover, this extraction of information should be fast enough to aid real-time, autonomous facility operation (PRO 2), which will itself make extensive use of AI/ML methods. Both information extraction and autonomous control will require AI/ML-enabled fast, accurate models trained on both simulations and data (PRO 3). Finally, the AI/ML tools in each application should be sufficiently robust and interpretable to be deployed online at a major research facility.

While many of the AI/ML needs can leverage existing solutions developed by industry, the challenges faced by the DOE are sufficiently distinct to necessitate new research in AI/ML techniques. Examples include extremely high throughput (TB/s) and low latency (microseconds), extremely large (PB) or small (single example) datasets, and rigorous statistical analysis of uncertainties and interpretability. For more details on applying AI/ML to scientific discovery, see the Priority Research Directions discussed in the *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence* [94].

Tackling these challenges will require multidisciplinary teams [93] involving both domain scientists and applied mathematicians, computer scientists, data scientists, and skilled software engineers to ensure that constructed algorithms and tools are widely applicable. In this respect, there may be opportunities for synergy with other DOE SC programs, especially within ASCR. One example of a coordinated effort between BES and ASCR deploying a multidisciplinary team is the CAMERA project. The effort has impacted the facilities in areas such as x-ray scattering reconstruction, image analysis, computer vision, and autonomous self-steering experiments. Other examples have found success in accelerator physics and computational chemistry such as the SciDAC (Scientific Discovery through Advanced Computing) program.

These efforts show that coordinated ASCR/BES efforts can have a transformative impact on the SUFs. Indeed, new AI/ML techniques will need to be customized and advanced to support BES emerging missions and enable a full utilization of next-generation facilities. For example, dependency on algorithms and software might need to come with some guarantees:

- **Transparency:** physics-informed strategies, software documentation, and organized data repositories for benchmarking, including persistent and unique identifiers, will be vital to understanding AI/ML tools.
- **Reproducibility:** AI/ML algorithms will require measures of reliability, uncertainty quantification, trustworthiness, and data ethics.
- **Instrument experience improvement:** Automation must be accompanied by user-friendly systems for better human–machine interface and accessibility.
- **Maintenance services:** Human-based teams for transition and support to new modes of operation.
- **Extensibility and modularity for software integration:** Automation must allow for inclusion of new modules, mechanisms for interoperability, and compatibility.
- **Faster and I/O-aware:** Multiscale data representations for fast access, given diverse SUF computational infrastructure, and for different scientific questions.
- **Portability to diverse computational platforms:** From edge to leadership-class computers, including the ability to handle terabytes on millisecond scales across computing facilities.

The cross-cutting issues identified by the PROs are discussed below: AI/ML algorithms, data management and infrastructure, HPC, and data networks, although these themes themselves are highly interrelated.

AI/ML cross-cutting issues

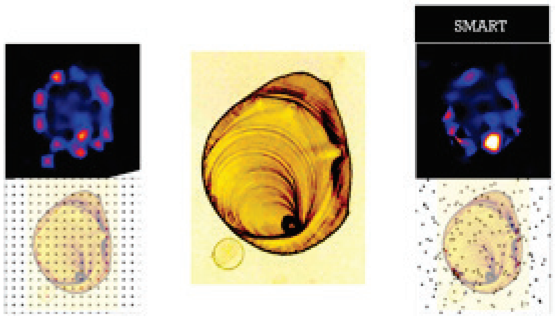


Figure 6. Autonomous imaging experiments using Gaussian processes. Here, an optical image of a nanoparticle coating (middle) having a "coffee ring" pattern and the automatically reconstructed image from a dense sampling (left) versus a sparse sample (right) [102]. | *Distributed under a Creative Commons Attribution Noncommercial License 4.0*

Successful execution of the PRO research directions will require both expertise and innovation in AI/ML techniques. Methods will extend beyond the neural network and deep learning approaches most commonly associated with ML to include Gaussian processes (figure 6) [102]; decision trees (e.g., Monte Carlo tree search) [103]; reinforcement learning; Boltzmann generators applied to fundamental problems in statistical physics [104]; Bayesian optimization [38]; and dimensionality reduction methods such as variational auto-encoders. While many of these innovations are driven by industry, the SUFs will also require AI/ML advancements that are specific to DOE science challenges. Examples include:

1. **Physics-based constraints:** New ML algorithms are needed that employ physics-based constraints in understanding data both to ensure that models produce relevant information and to greatly accelerate convergence to reasonable models. This will require exploiting advances in underlying mathematics in areas such as physics-appropriate projection operators.
2. **Robustness:** ML-based methods must deal with experimental conditions such as noise, jitter, drift, dropout, and alignment, exploiting the mathematics of multiobjective energy minimizers and deep convolutional denoising of Poisson noise.

3. **Scaling:** Existing ML solutions must scale to the high dimensional parameter spaces, continuous variables, and extreme data sizes common to SUF applications. Real-time applications (e.g., data reduction) must handle both high data rates (terabytes/second) with microsecond latency. While high-performance computing will be essential, scaling will also require innovations in the ML algorithms.

4. **Super resolution:** New approaches are needed to extract subgrid resolution from coarse sampling in space or time, aided by ML models that learn resolution capabilities from coupled resolved or under-resolved training data [106].

5. **Multimodal analysis:** Analysis methods should handle multimodal comparisons across length scales, techniques, and users, enabling intelligent learning of similarities and linkages across different experimental modalities that allow information fusion during data acquisition to converge on physics-appropriate models. This will require developing ML models that combine multiobjective descriptions across disparate sources.

6. **Automated labeling:** Diverse scientific datasets require automatic ML techniques to tag and annotate data, making use of mathematically based networks [106] specifically designed to work with limited data and to determine appropriate features. This will require developing techniques that maximally exploit computationally expensive, complex scientific data, rather than count on vast databases of simplified objects, to build and determine appropriate feature vectors for reduced, efficient, and sparse representations (Figure 7).

7. **Fast-executing approximations:** Fast-executing approximations are needed, including reduced coarse reconstruction methods, optimized inversion techniques, surrogate models, and data-driven approximation models to quickly perform “data triage” to determine if an experiment is on track and generating important data, as well as to extract critical features and compression opportunities to pinpoint key information as an experiment progresses. This will require exploiting advances in underlying mathematics in areas such as search and optimization methods, Bayesian experimental design, dimensional reduction methods to efficiently explore high-dimensional parameterization spaces, parameter estimations, and reduced-order models.

8. **Data reduction:** AI/ML methods are needed for streaming, data reduction, and storage protocols for heterogeneous experiments at high acquisition rates, exploiting computer science research on fast network transfer, optimal ways to load-balance computer resources across detectors, local compute facilities, HPC, and edge services. Figure 8 shows automated image search output.

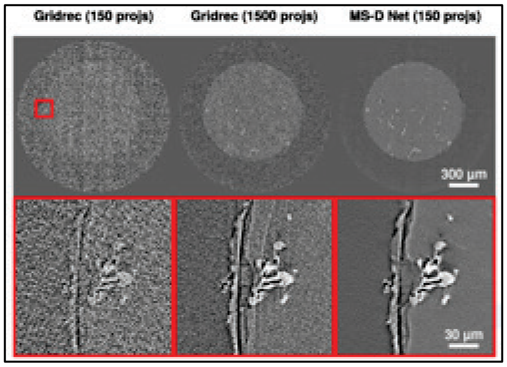


Figure 7. Deep neural network using limited labeled samples to classify tomographic images of a fiber-reinforced minicomposite. The top pannel shows SEM images of a minifiber and the bottom left shows zoomed in images of the red region marked at the left in the top pannel. The bottom middle and right panels show reconstructed images using sparse and noisy data [105]. | *Distributed under the MDPI Open Access Information and Policy*

9. **Data mining:** The shared data repository will require new mathematics and computer science to exploit fast indexing, such as locality sensitivity hashing, clever feature vectors [107], ontologies, and inferential engines. For example, materials researchers will need key data services to promote open data sharing and data reuse, simplified data publication and curation workflows, and powerful data discovery interfaces for data of all sizes and sources.

10. **User friendly:** A form of automated AI/ML selection or recommendation system will be useful to encourage a broader range of users with limited AI/ML expertise. For example, automatic selection methods for ML algorithms and/or hyper-parameter values for a given ML approach.

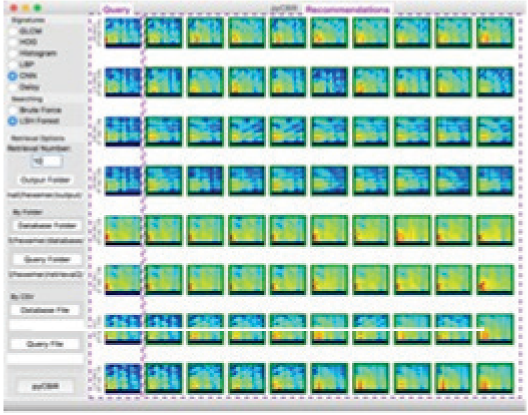


Figure 8. Automated image search by content-based image retrieval of millions of grazing-incidence small-angle scattering patterns [107]. | *Distributed under the MDPI Open Access Information and Policy*

Data management infrastructure

AI/ML models are fundamentally linked to the datasets on which they are trained, and data infrastructure needs are ubiquitous in AI/ML workflows. These needs are highlighted by PRO 4, which discusses the opportunity to create a shared data repository to host the totality of data generated at BES SUFs. PRO 4 identifies a range of enabling capabilities, including standardized file formats, search functions, data catalogs, recommendation tools, autonomous data labeling, and challenges relating to capture of data and metadata. Data mining in the repository will require new mathematics and computer science to exploit fast indexing, such as locality sensitivity hashing, clever feature vectors [107], ontologies and inferential engines. Though less central to the other PROs, nearly every topic covered during the roundtable will face challenges relating to data workflows for training, testing, and deployment of models. The recent ASCR workshops on data and models for AI/ML covered many of these topics in depth [22, 94].

High-performance computing cross-cutting issues

The AI/ML approaches delineated under the PROs will require access to extreme computation to process data, run high-fidelity simulations to generate or augment measured data, and train models. The DOE is well-positioned to address these challenges, with plans to deploy NERSC-9 (Perlmutter) and the first generation of exascale computers: Aurora at the Argonne Leadership Computing Facility and Frontier at the Oak Ridge Leadership Computing Facility. Already, these ASCR facilities support the most popular AI/ML frameworks. The research being conducted at the National Energy Research Scientific Computing Center (NERSC) includes examples of running extreme-scale training and optimizing DNNs for massive climate data, as well as computational modeling for energy-efficient industrial applications [108]. It is expected that additional HPC-focused AI/ML frameworks will be developed in the next decade, as highlighted in recent AI for Science Town Halls (e.g., see [109]). For example, AI/ML methods will eventually support rapid data processing at HPC facilities to enable quasi-real-time feedback on experiments and observations. These advances are fundamental to the PROs identified in this BES AI/ML roundtable.

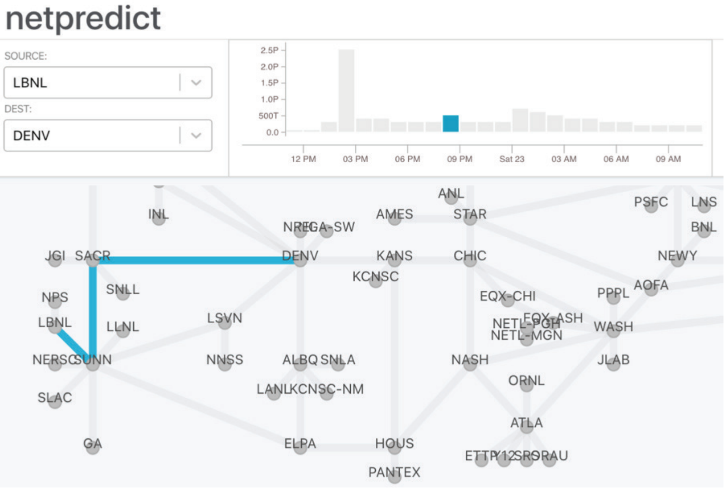
The scientific community has an opportunity to define a common toolset of AI/ML methods that are useful across a variety of control problems and available for HPC. In addition to instrument tuning, AI/ML methods in HPC could revolutionize experimental platforms by automating the selection of measurement conditions, experimental conditions, sample measurement sequence, and overall experimental execution. Such automation—necessarily leveraging accelerated real-time data analytics—would dramatically increase the quality of experimental datasets, reduce wasted instrument time, minimize sample damage from probes, and accelerate experimental study.

However, several research developments are needed to enable HPC AI/ML models for experimental data, such as verifying accuracy of high performance codes, because many discoveries result from serendipitous events, and it is essential to avoid confusing signal with noise. ML is often not easily adapted to current and emerging heterogeneous instrument hardware. Experimental facilities place unique requirements on AI/ML systems. For example, the data streams at experimental facilities may be very high volume, approaching the TB/s range. This data may need to be processed by ML algorithms on the fly (e.g., at the edge) while meeting power requirements, and any configuration or interface to algorithms should be approachable by domain scientists who do not have AI/ML expertise. Future computing environments that can address these challenges will likely be heterogeneous, consisting of GPU accelerators, possibly in conjunction with FPGAs, application-specific integrated circuits (ASICs), and emerging hardware custom designed for deep learning workloads. These computing systems may also have novel memory hierarchies, involving traditional dynamic random access memory (DRAM) alongside technologies like nonvolatile random access memory (RAM), 3D stacked memory, and chips with processing-in-memory capabilities.

Further research in edge computing can help advance solutions to these AI/ML challenges as well as HPC capabilities. While these systems are highly customizable and can potentially deliver orders of magnitude improvement over traditional systems—even those with GPUs—achieving this speedup will be challenging. It may require hardware modeling expertise and a large investment in porting codes to the new computing architectures. Therefore, intuitive programming interfaces are needed, including software that can automatically transfer AI/ML models developed in standard GPU/central processing unit-friendly frameworks onto FPGAs/ASICs and emerging deep learning accelerators. Furthermore, in cases where systems have multiple accelerators, it is desirable to have a turnkey solution for the device placement problem, which involves mapping the different operations describing an AI/ML model onto available hardware resources to maximize parallelism. In systems with hybrid memory hierarchies, the device placement problem will include determining storage locations for large arrays, be they on traditional DRAM, nonvolatile RAM, or other specialized memory modules. Being able to solve these problems automatically or online is especially important in experimental facilities, as researchers often have limited time with the equipment and the experimental setup changes from one user to another.

Network cross-cutting issues

The HPC applications discussed above presume the enormous datasets generated at the SUFs can be shipped between SUFs and HPC facilities at speeds that permit real-time analysis. The data movement problem will itself require advances to meet the needs of the PROs, and the DOE is already working on deploying ESnet6, the next generation of high-speed networks for science use cases. Figure 9 shows an example of real-time network traffic prediction output. With the new networking capabilities, the network will enable novel new research that can support BES workflows. Examples include:



netpredict

SOURCE: LBNL

DEST: DENV

Figure 9. Real-time prediction using deep learning for network traffic. | *Image courtesy Mariam Kiran, Energy Sciences Network*

- **Intelligent protocols that enable querying target subfields of an experiment:** Using intent and named networking, new network schemes will deliver new protocols that allow scientists to query the exact datasets needed for their data analysis. This will streamline experiments, prioritizing the relevant data being generated by the facilities.

- **Data reduction for accelerated I/O:** Current efforts at NERSC and ESnet are exploring how the speed of I/O correlates with network transfer speed and how it impacts experiments overall. Current projects such as SENSE (SDN for End-to-End Network Science at the Exascale) [110] and the ASCR Early Career Project 2017, called DAPHNE (Large-scale Deep Learning for High-Performance Networks) [111] have been exploring how this end-to-end workflow can be optimized to improve science. Current efforts will also require exploring the use of AI/ML to allow the network to make intelligent decisions on managing the data transfer rate.

- **Improving network utilization and enabling higher bandwidth for on-demand experiments:** ESnet research efforts have been exploring the use of deep RL techniques to improve network utilization, making science transfers faster and optimizing bandwidth for experiments [112].

- **Predicting performance multiple hours ahead:** Developing advanced time-series prediction libraries can help networks predict how they will be utilized in the future, including aspects related to their power consumption and required utilization. Advance knowledge will help engineers optimize the infrastructure use, such as diverting flows to underutilized links [113] and powering machines down when not needed.

Additionally, further research is needed to develop new AI/ML algorithms that enable fast stream data processing leading to clustering and classification in unlabeled datasets. This capability is particularly important in network and compute facilities for learning behavior with fast streaming operational data.

Summary

Strategic findings from a recently published *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence* [94] summarized some of the key points in this section as well as key thrust areas to be worked in synergy.

1. **Incorporating domain-aware knowledge:** Research must develop supervised, unsupervised, and feature selection methods to incorporate domain qualities into the models. There are currently research constraints that need further work, such as unconstrained optimization and loss function calculations. Further work in Bayesian approaches, surrogate models, and read-only memory will be highly relevant.

2. **Interpretable scientific AI/ML:** Effort must be made to develop methods that help organize and explore datasets, and building of optimized models, including comparison methods and probabilistic approaches that supports optimization toward the scientific problems being explored.

3. **Robust scientific AI/ML:** Developments must seek out reproducible solutions in certain conditions and to investigate the limitations of the models. This is an important area to identify the current challenges of how the AI/ML models will behave under certain conditions and how methods can be generalized to a larger set of domains. Additionally, methods to measure the correctness of the model are being explored here.

4. **Working with complex datasets:** Efficient sampling is required to analyze high-dimensional, noisy datasets. Innovative solutions that use Monte Carlo, Bayesian, and active learning methods are needed to make progress in this research thrust.

5. **Intelligent automation and decision support:** Developments must steer experiments using AI/ML-informed decision-making. Work on uncertainty quantification and sensitivity analysis will be developed further to enhance this research thrust.

These research efforts will be essential to the success of the four PROs discussed during the roundtable. Scientific AI/ML faces challenges of different scales, data types, and goals compared to the challenges inherent in commercial AI/ML efforts. Academic and national laboratory research communities will need to develop their own computational tools. For this purpose, the benchmark datasets described in PRO 4 could provide AI/ML researchers with large, well-labeled, realistic datasets for AI/ML algorithm R&D. Historically, scientific challenges have at times spurred development of new computational paradigms, notably the World Wide Web. A close collaboration between computer scientists and the SUFs to create AI/ML tools could have a transformative impact on both fields.

Closing Remarks

As detailed in the Executive Summary, this roundtable has identified four key Priority Research Opportunities (PROs) that establish a vision for accelerating scientific discovery at the BES Scientific User Facilities. It is clear that AI/ML can integrate with modeling, simulation and data analysis to become an essential part of our scientific laboratory activities, from engineering to operations to scientific analysis. AI/ML-enabled research will help power automated experiments, control of complex systems, discovery of new materials and processes, and application of exascale computing to maximize scientific output and discovery. In the next ten years it is fully expected that AI/ML will enable the DOE to attack and solve new problems for energy sciences.

References

[1] White, W., et al. “The Linac Coherent Light Source.” *J. Synchrotron Radiat.* 22 (2015): 472–476.

[2] Liang, M., et al. “The Coherent X-ray Imaging Instrument at the Linac Coherent Light Source.” *J. Synchrotron Radiat.* 22 (2015): 514–519.

[3] Boutet, S., et al. “The New Macromolecular Femtosecond Crystallography (MFX) Instrument at LCLS.” *Synchrotron Radiat. News* 29 (2016): 1.

[4] Thayer, J., et al. “Data Systems for the Linac Coherent Light Source.” *J. Synchrotron Radiat.* 49 (2016): 1363–1369.

[5] US Department of Energy, Office of Science, Basic Energy Sciences. *The 5 BES Light Source Directors’ Data Working Group (DWG): Report on Data Retention*, April 17, 2018.

[6] US Department of Energy, Office of Science. BES User Facilities Data Management and Analysis Resource Needs, BES/ASCR Data Call: ALS Response, July 30, 2018.

[7] US Department of Energy, Office of Science. BES User Facilities Data Management and Analysis Resource Needs, BES/ASCR Data Call: APS Response, July 30, 2018.

[8] US Department of Energy, Office of Science. BES User Facilities Data Management and Analysis Resource Needs, BES/ASCR Data Call: SLAC Response, July 30, 2018.

[9] US Department of Energy, Office of Science. BES User Facilities Data Management and Analysis Resource Needs, BES/ASCR Data Call: NSLS-II Response, July 30, 2018.

[10] US Department of Energy, Office of Science. BES User Facilities Data Management and Analysis Resource Needs, BES/ASCR Data Call: SSRL Response, July 30, 2018.

[11] US Department of Energy, Office of Science, Basic Energy Sciences. *Gap Analysis: Materials Discovery through Data Science at Advanced User Light Sources Workshop Report*, October 3–5, 2018

[12] LeCun, Y., Bengio, Y., and Hinton, G. “Deep Learning.” *Nature* 521 (2015): 436–444.

[13] Khachatryan, V., et al., “The CMS Trigger System.” *J. Instrum.* (2018). https://doi.org/10.1088.

[14] Liu, J., and Yager, K. G. “Unwarping GISAXS Data.” *IUCrJ* 5 (6) (2018): 737–752.

[15] Liu, J., Lhermitte, J., Tian, Y., Zhang, Z., Yu, D., and Yager, K. G. “Healing X-ray Scattering Images. *IUCrJ* 4 (4) (2017): 455–465.

[16] Hezaveh, Y., Levasseur, L., and Marshall, P. “Fast Automated Analysis of Strong Gravitational Lenses with Convolutional Neural Networks.” *Nature* 548 (2017): 555–557. https://doi:10.1038/nature23463.

[17] Pelt. D. M., and Sethian, J. A. “A Mixed-scale Dense Convolutional Neural Network for Image Analysis.” *Proc. Nat. Accad. Sci.* 115, no. 2 (2018): 254–259.

[18] *Proceedings of SC19, The International Conference for High Performance Computing, Networking, Storage, and Analysis. Data Processing at the Linac Coherent Light Source*, November 2019.

[19] “Exafel.” Linac Coherent Light Source. SLAC National Accelerator Laboratory Source. https://lcls.slac.stanford.edu/exafel.

[20] Cherukara, M. “AI CDI: Deep Convolutional Neural Networks for Real-time Inversion of Coherent X-ray Diffraction Data.” https://press3.mcs.anl.gov/jlse/projects/ai-cdi-deep-convolutional-neural-networks-for-real-time-inversion-of-coherent-x-ray-diffraction-data/.

[21] Cherukara, M., Nashed, Y. S. G., and Harder, R. J. “Real-time Coherent Diffraction Inversion Using Deep Generative Networks.” *Sci. Rep.* 8, no. 1 (2018): 16520. DOI: 10.1038/s41598-018-34525-1.

[22] Fagnan, K., Nashd, Y., Perdue, G., Ratner, D., Shankar, A., and Yoo, S. *Data and Models: A Framework for Advancing AI in Science*. DOI:10.2172/1579323.

[23] Ardell, A. J. “Precipitation Hardening.” *Metall. Trans. A* 16, no. 12 (1985): 2131–2165.

[24] Hager, M. D., et al. “Self-healing Materials.” *Adv. Mat.* 22, no. 47 (2010): 5424–5430.

[25] Ramachary, D. B., and Jain, S. “Sequential One-pot Combination of Multi-component and Multi-catalysis Cascade Reactions: An Emerging Technology in Organic Synthesis.” *Org. Biomol. Chem.* 9.5 (2011): 1277–1300.

[26] Li, Z., Najeeb, M. A., Alves, L., Sherman, A., Parrilla, P. C., Pendleton, I. M., Zeller, M., Schrier, J., Norquist, A. J., and Chan, E. “Robot-accelerated Perovskite Investigation and Discovery (RAPID): 1. Inverse Temperature Crystallization.” ChemRxiv. https://doi.org/10.26434/chemrxiv.10013090.v1.

[27] Pendleton, I. M., Cattabriga, G., Li, Z., Najeeb, M. A., Friedler, S. A., Norquist, A. J., Chan, E., and Schrier, J. “Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): A Software Pipeline for Automated Chemical Experimentation and Data Management.” *MRS Communications* 9, no. 3 (September 2019): 846–859. https://doi.org/10.1557/mrc.2019.72.

[28] Kalinin, S. V., Borisevich, A., and Jesse, S. “Fire Up the Atom Forge.” *Nature* 539 (2016): 485–487.

[29] Hanuka, A., Duris, J., Shtalenkova, J., Kennedy, D., Edelen, A., Ratner, D., and Huang, X. “Online Tuning and Light Source Control Using a Physics-informed Gaussian Process.” arXiv: 1911.01538v1.

[30] Huang, X. *Beam-Based Correction and Optimization for Accelerators*. Boca Raton, FL: Taylor & Francis, 2020.

[31] Huang, X., Corbett, J., Safranek, J., and Wu, J. “An Algorithm for Online Optimization of Accelerators.” *Nucl. Instrum. Meth. Phys. Res. Section A: Accelerators, Spectrometers, Detectors Assoc. Equip.* 726 (2013): 77–83.

[32] Scheinker, A., and Krstić, M. “Minimum-seeking for CLFs: Universal Semiglobally Stabilizing Feedback under Unknown Control Directions.” *IEEE Trans. Automat. Contr.* 58 (2013): 1107–1122.

[33] Huang, X., and Safranek, J. “Online Optimization of Storage Ring Nonlinear Beam Dynamics.” *Phys. Rev. Accel. Beams* 18 (2015): 084001.

[34] McIntire, M. W., Cope, T. M., Ratner, D. F., and Ermon, S. In *Bayesian Optimization of Fel Performance at LCLS*, p. WEPOW055, International Particle Accelerator Conference, Busan, Korea, May 11, 2016.

[35] Huang, X. “Robust Simplex Algorithm for Online Optimization.” *Phys. Rev. Accel. Beams* 21 (2018): 104601.

[36] Scheinker, A., Edelen, A., Bohler, D., Emma, C., and Lutman, A. “Demonstration of Model-Independent Control of the Longitudinal Phase Space of Electron Beams in the Linac-Coherent Light Source with Femtosecond Resolution.” *Phys. Rev. Lett.* 121 (2018): 044801.

[37] Scheinker, A., Huang, X., and Wu, J. "Minimization of Betatron Oscillations of Electron Beam Injected into a Time-Varying Lattice Via Extremum Seeking." *IEEE Trans. Control Sys. Technol.* 26 (2018): 336–343.

[38] Duris, J., Kennedy, D., Hanuka, A., Shtalenkova, J., Edelen, A., Egger, A., Cope, T., and Ratner, D. "Bayesian Optimization of a Free-electron Laser." *Phys. Rev. Lett.* 124 (2020): 124801.

[39] Bergan, W. F., Bazarov, I. V., Duncan, C. J. R., Liarte, D. B., Rubin, D. L., and Sethna, J. P. "Online Storage Ring Optimization Using Dimension-reduction and Genetic Algorithms." *Phys. Rev. Accel. Beams* 22 (2019): 054601.

[40] Scheinker, A., Bohler, D., Tomin, S., Kammering, R., Zagorodnov, I., Schlarb, H., Scholz, M., Beutner, B., and Decking, W. "Model-independent Tuning for Maximizing Free Electron Laser Pulse Energy." *Phys. Rev. Accel. Beams* 22 (2019): 082802.

[41] Noack, M. M., Yager, K. G., Fukuto, M., Doerk, G. S., Li, R., and Sethian, J. A. "A Kriging-Based Approach to Autonomous Experimentation with Applications to X-ray Scattering." *Sci. Rep.* 9 (2019): 11809.

[42] Reyes, K. G., and Maruyama, B. "The Machine Learning Revolution in Materials?" *MRS Bull.* 44 (2019): 530–537.

[43] Casciato, M. J., Kim, S., Lu, J. C., Hess, D. W., and Grover, M. A. "Optimization of a Carbon Dioxide-assisted Nanoparticle Deposition Process Using Sequential Experimental Design with Adaptive Design Space." *Ind. Eng. Chem. Res.* 51 (2012): 4363–4370.

[44] Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., and Ramprasad, R. "Accelerating Materials Property Predictions Using Machine Learning." *Sci. Rep.* 3 (2013): 2810.

[45] Kusne, A. G., Gao, T., Mehta, A., Ke, L., Nguyen, M. C., Ho, K.-M., Antropov, V., Wang, C.-Z., Kramer, M. J., Long, C., and Takeuchi, I. "On-the-Fly Machine-learning for High-throughput Experiments: Search for Rare-Earth-free Permanent Magnets." *Sci. Rep.* 4 (2014): 6367.

[46] Reyes, K., Chen, S., Li, Y., and Powell, W. B. In *Quantifying Experimental Characterization Choices in Optimal Learning and Materials Design*, TMS 2015 144th Annual Meeting & Exhibition. (Springer International Publishing: Cham, 2016), pp. 697–704.

[47] Godaliyadda, G. M. D., Ye, D. H., Uchic, M. D., Groeber, M. A., Buzzard, G. T., and Bouman, C. A. "A Supervised Learning Approach for Dynamic Sampling." *Electron. Imaging* 2016 (2016): 1–8.

[48] Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., and Lookman, T. "Adaptive Strategies for Materials Design Using Uncertainties." *Sci. Rep.* 6 (2016): 19660.

[49] Nikolaev, P., Hooper, D., Webber, F., Rao, R., Decker, K., Krein, M., Poleski, J., Barto, R., and Maruyama, B. "Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth." *npj Comp. Mater.* 2 (2016): 16031.

[50] Lookman, T., Alexander, F. J., and Bishop, A. R. "Perspective: Codesign for Materials Science: An Optimal Learning Approach." *APL Mater.* 4 (2016): 053501.

[51] Scarborough, N. M., Godaliyadda, G. M. D. P., Ye, D. H., Kissick, D. J., Zhang, S., Newman, J. A., Sheedlo, M. J., Chowdhury, A. U., Fischetti, R. F., Das, C., Buzzard, G. T., Bouman, C. A., and Simpson, G. J. "Dynamic X-ray Diffraction Sampling for Protein Crystal Positioning." *J. Synchrotron Radiat.* 24 (2017): 188–195.

[52] Lookman, T., Balachandran, P. V., Xue, D., Hogden, J., and Theiler, J. "Statistical Inference and Adaptive Design for Materials Discovery." *Curr. Opin. Solid State Mater. Sci.* 21 (2017): 121–128.

[53] Ren, F., Ward, L., Williams, T., Laws, K. J., Wolverton, C., Hattrick-Simpers, J., and Mehta, A. "Accelerated Discovery of Metallic Glasses through Iteration of Machine Learning and High-throughput Experiments." *Sci. Advances* 4 (2018): eaaq1566.

[54] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. "Continuous Control with Deep Reinforcement Learning." arXiv: 1509.02971 (2015), cs.LG.

[55] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. "Human-level Control through Deep Reinforcement Learning." *Nature* 518 (2015): 529–533.

[56] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. In *Asynchronous Methods for Deep Reinforcement Learning*, International Conference on Machine Learning, 2016, pp. 1928–1937.

[57] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. "Proximal Policy Optimization Algorithms." arXiv: 1707.06347 (2017), cs.LG.

[58] Liu, Y., Ramachandran, P., Liu, Q., and Peng, J., "Stein Variational Policy Gradient." arXiv: 1704.02399 (2017), cs.LG.

[59] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. "A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-play." *Science* 362 (2018): 1140–1144.

[60] Sutton, R. S., and Barto, A. G. *Reinforcement Learning*, 2d ed. (Cambridge, MA: MIT Press, 2018), p. 552.

[61] Martius, G., and Lampert, C. H. "Extrapolation and Learning Equations." arXiv: 1610.02995 (2016), cs.LG.

[62] Sahoo, S. S., Lampert, C. H., and Martius, G. "Learning Equations for Extrapolation and Control." arXiv: 1806.07259 (2018), cs.LG.

[63] Parise, J. B., and Brown, G. E. "New Opportunities at Emerging Facilities." *Elements* 2 (2006): 37–42.

[64] Ye, Y. F., Wang, Q., Lu, J., Liu, C. T.; Yang, Y. "High-Entropy Alloy: Challenges and Prospects." *Mater. Today* 19 (2016): 349–362.

[65] Majewski, P. W., and Yager, K. G., "Latent Alignment in Pathway-Dependent Ordering of Block Copolymer Thin Films." *Nano Lett.* 15 (2015): 5221–5228.

[66] Steinschulte, A. A., Scotti, A., Rahimi, K., Nevskyi, O., Oppermann, A., Schneider, S., Bochenek, S., Schulte, M. F., Geisel, K., Jansen, F., Jung, A., Mallmann, S., Winter, R., Richtering, W., Wöll, D., Schweins, R., Warren, N. J., and Plamper, F. A. "Stimulated Transitions of Directed Nonequilibrium Self-Assemblies." *Adv. Mater.* 29 (2017): 1703495.

[67] Choo, Y., Majewski, P. W., Fukuto, M., Osuji, C. O., and Yager, K. G. "Pathway-engineering for Highly-Aligned Block Copolymer Arrays." *Nanoscale* 10 (2018): 416–427.

[68] Hu, H., Gopinadhan, M., and Osuji, C. O. “Directed Self-assembly of Block Copolymers: A Tutorial Review of Strategies for Enabling Nanotechnology with Soft Matter.” *Soft Matter* 10 (2014): 3867–3889.

[69] Doerk, G. S., and Yager, K. G. “Beyond Native Block Copolymer Morphologies.” *Mol. Sys. Des. Eng.* 2 (2017): 518–538.

[70] Tschierske, C. “Development of Structural Complexity by Liquid-crystal Self-Assembly.” *Angew. Chem. Int. Edit.* 52 (2013): 8828–8878.

[71] Chakrabarty, R., Mukherjee, P. S., and Stang, P. J. “Supramolecular Coordination: Self-assembly of Finite Two- and Three-dimensional Ensembles.” *Chem. Rev.* 111 (2011): 6810–6918.

[72] Murray, C. B., Kagan, C. R., and Bawendi, M. G. “Synthesis and Characterization of Monodisperse Nanocrystals and Close-packed Nanocrystal Assemblies.” *Ann. Rev. Mater. Sci.* 30 (2000): 545–610.

[73] Glotzer, S. C., and Solomon, M. J. “Anisotropy of Building Blocks and Their Assembly into Complex Structures.” *Nat. Mat.* 6 (2007): 557–562.

[74] Boles, M. A., Engel, M., and Talapin, D. V. “Self-assembly of Colloidal Nanocrystals: From Intricate Structures to Functional Materials.” *Chem. Rev.* 116 (2016): 11220–11289.

[75] Kumar, S. K.; Kumaraswamy, G.; Prasad, B. L. V.; Bandyopadhyaya, R.; Granick, S.; Gang, O.; Manoharan, V. N.; Frenkel, D.; Kotov, N. A., Nanoparticle Assembly: A Perspective and Some Unanswered Questions. *Curr. Sci.* 112 (2017): 1635-1641.

[76] Knorowski, C., and Travesset, A. “Materials Design by DNA Programmed Self-assembly.” *Curr. Opin. Solid State Mater. Sci.* 15 (2011): 262–270.

[77] Tan, L. H., Xing, H., and Lu, Y. “DNA as a Powerful Tool for Morphology Control, Spatial Positioning, and Dynamic Assembly of Nanoparticles.” *Accounts Chem. Res.* 47 (2014): 1881–1890.

[78] Zhang, X., Wang, R., and Xue, G. “Programming Macro-materials from DNA-directed Self-assembly.” *Soft Matter* 11 (2015): 1862–1870.

[79] Shrestha, U. R., Juneja, P., Zhang, Q., Gurumoorthy, V., Borreguero, J. M., Urban, V., Cheng, X., Pingali, S. V., Smith, J. C., O’Neill, H. M., and Petridis, L. “Generation of the Configurational Ensemble of an Intrinsically Disordered Protein from Unbiased Molecular Dynamics Simulation.” *Proc. Natl. Acad. Sci. U.S.A.* 116 (2016): 20446–20452.

[80] Ourmazd, A. “Cryo-EM, XFELs and the Structure Conundrum in Structural Biology.” *Nature Methods* 16 (2019): 941–944.

[81] Zhong, E., Bepler, T., Davis, J., and Berger, B. “Reconstructing Continuous Distributions of 3D Protein Structure from Cryo-EM Images.” arXiv:1909.05215, 2019.

[82] Mousavi, S. S., Schukat, M., and Howley, E. In *Proceedings of SAI Intelligent Systems Conference.* Springer, pp. 426–440.

[83] Er, S., Suh, C., Marshak, M. P., and Aspuru-Guzik, A. “Computational Design of Molecules for an All-quinone Redox Flow Battery.” *Chem. Sci.* 6 (2015): 885-893.

[84] Sutton, R. S., and Barto, A. G. *Reinforcement Learning: An Introduction* (Cambridge, MA: MIT Press, 2018).

[85] Safranek, J. “Experimental Determination of Storage Ring Optics Using Orbit Response Measurements.” *Nucl. Instrum. Meth. Phys. Res. Section A: Accelerators, Spectrometers, Detectors Assoc. Equip.* 388 (1997): 27–36.

[86] Tomás, R., et al. "Record Low β Beating in the LHC." *Phys. Rev. Accel. Beams* 15 (2012): 091001.

[87] Yang, X., and Huang, X. "A Method for Simultaneous Linear Optics and Coupling Correction for Storage Rings with Turn-by-turn Beam Position Monitor Data." *Nucl. Instrum. Meth. Physics Res. Section A: Accelerators, Spectrometers, Detectors and Assoc. Equip.* 828 (2016): 97–104.

[88] Zhang, T., Huang, X., and Maxwell, T. "Linear Optics Correction for Linacs and Free Electron Lasers." *Phys. Rev. Accel. Beams* 21 (2018): 092801.

[89] Huang, X., Safranek, J., and Portmann, G. "LOCO with Constraints and Improved Fitting Technique." *ICFA Beam Dyn. Newslett.* 44 (2009): 60–69.

[90] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. "Taking the Human Out of the Loop: A Review of Bayesian Optimization." *Proc. IEEE* 104 (2015): 148–175.

[91] Emma, C., et al. "Machine Learning-based Longitudinal Phase Space Prediction of Particle Accelerators." *Phys. Rev. Accel. Beams* 21 (2018): 112802.

[92] Noé, F., Olsson, S., Köhler, J., and Wu, H. "Boltzmann Generators: Sampling Equilibrium States of Many-Body Systems with Deep Learning." *Science* 365 (2019): eaaw1147.

[93] *Basic Research Needs for Innovation and Discovery of Transformative Experimental Tools* (DOE BES, 2017). https://science.osti.gov/-/media/bes/pdf/reports/2017/BRNIDTET_rpt_print.pdf?la=en&hash=C64F64A94E41D8E5F0239C52709.64CB4C9AA6412.

[94] *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. https://www.osti.gov/servlets/purl/1478744; https://www.osti.gov/biblio/1478744.

[95] Wilkinson, M., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Sci. Data* 3 (2016): 160018. DOI:10.1038/sdata.2016.18.

[96] LeCun, Y., Cortes, C., and Burges, C. J. C. "The MNIST Database of Handwritten Digits." http://yann.lecun.com/exdb/mnist/.

[97] ImageNet. http://www.image-net.org/.

[98] Somnath, S., Smith, C. R., Laanait, N., Vasudevan, R. K., Ievlev, A., Belianinov, A., Lupini, A., Shankar, M., Kalinin, S. V., and Jesse, S. "USID and Pycroscopy: Open Frameworks for Storing and Analyzing Spectroscopic and Imaging Data." arXiv:1903.09515. https://arxiv.org/abs/1903.09515.

[99] Himanen, L., et al. "Data-driven Materials Science: Status, Challenges, and Perspectives." *Adv. Sci.* (2019): 1900808.

[100] Marr, B. "How Much Data Do We Create Every Day? The Mind-blowing Stats Everyone Should Read." *Forbes*, May 21, 2018. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#565db9ed60ba.

[101] Bradshaw, L. "Big Data and What It Means." US Chamber of Commerce Foundation. https://www.uschamberfoundation.org/bhq/big-data-and-what-it-means.

[102] Noack, M. M., Yager, K. G., Fukuto, M., Doerk, G. S., Li, R., and Sethian, J. A. A Kriging-based Approach to Autonomous Experimentation with Applications to X-ray Scattering. *Sci. Rep.* 9, no. 1 (2019): 1–19.

[103] Silver, D., et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (2016): 484–489.

[104] Noé, F., et al. “Boltzmann Generators: Sampling Equilibrium States of Many-body Systems with Deep Learning.” *Science* (2019). arXiv:1812.01729.

[105] Pelt. D.M., and Sethian, J.A., “A Mixed-scale Dense Convolutional Neural Network for Image Analysis.” *PNAS* 115, no. 2 (2018): 254–259.

[106] Harken, R. “From Simulation to Automation in a Data-rich World.” https://www.olcf.ornl.gov/2019/08/29/from-simulation-to-automation-in-a-data-rich-world/.

[107] Araujo, F., Silva, R. R. V., Medeiros, F. N. S., Parkinson, D. D., Hexemer, A., Carneiro, C. M., Ushizima, D. M. “Reverse Image Search for Scientific Data within and beyond the Visible Spectrum” *Expert Sys. Appl.* 109 (November 2018): 35–48.

[108] Troutman, K. “Industry and CRD Partner to Build Energy Efficiency into Paint Design: Computational Modeling Will Drive Efficiency in Automotive Painting Process.” Computational Research Division, Lawrence Berkeley National Laboratory. https://crd.lbl.gov/news-and-publications/news/2019/industry-and-berkeley-lab-partner-to-improve-paint-design-processes/.

[109] AI for Science: Report on the Department of Energy Town Halls on Artificial Intelligence for Science. https://anl.app.box.com/s/f7m53y8beml6hs270h4yzh9l6cnmukph.

[110] Monga, I. SENSE (SDN for End-to-End Networking @ Exascale) Project. Presentation. http://es.net/assets/pubs_presos/SENSE-Thomas-20160217-on-Web.pdf.

[111] Lawrence Berkeley National Laboratory, ESNet, Computational Research Division. DAPHNE Project Website. https://sites.google.com/lbl.gov/daphne/home.

[112] Kiran, M., Mohammed, B., and Krishnaswamy, N. “DeepRoute: Herding Elephant and Mice Flows with Reinforcement Learning.” Presented at 2nd IFIP International Conference on Machine Learning for Networking (MLN'2019), Paris, France, December 3–5, 2019.

[113] Mohammed, B., Kiran, M., and Krishnaswamy, N. “DeepRoute on Chameleon: Experimenting with Large-scale Reinforcement Learning and SDN on Chameleon Testbed.” Presented at 2019 IEEE 27th International Conference on Network Protocols (ICNP), Chicago, October 7–10.

Appendix A: Roundtable Participants

BES Roundtable on AI/ML
Hilton Washington, DC/Rockville Hotel & Executive Meeting Center
1750 Rockville Pike, Rockville, Maryland
October 22–23, 2019

Invited Participants

Frank Alexander, Brookhaven National Laboratory
Jay Jay Billings, Oak Ridge National Laboratory
Ryan Coffee, SLAC National Accelerator Laboratory
Sarah Cousineau, Oak Ridge National Laboratory
Peter Denes, Lawrence Berkeley National Laboratory
Mathieu Doucet, Oak Ridge National Laboratory
Ian Foster, Argonne National Laboratory
Alex Hexemer, Lawrence Berkeley National Laboratory
Dean Hidas, Brookhaven National Laboratory
Xiaobiao Huang, SLAC National Accelerator Laboratory
Sergei Kalinin, Oak Ridge National Laboratory
Mariam Kiran, Lawrence Berkeley National Laboratory
A. Gilad Kusne, National Institute of Science and Technology
Apurva Mehta, SLAC National Accelerator Laboratory
Anibal (Timmy) Ramirez-Cuesta, Oak Ridge National Laboratory
Daniel Ratner, SLAC National Accelerator Laboratory
Subramanian Sankaranarayanan, Argonne National Laboratory
Mary Scott, Lawrence Berkeley National Laboratory
Mark Stevens, Sandia National Laboratories
Bobby Sumpter, Oak Ridge National Laboratory
Yipeng Sun, Argonne National Laboratory
Jana Thayer, SLAC National Accelerator Laboratory
Brian Toby, Argonne National Laboratory
Daniela Ushizima, Lawrence Berkeley National Laboratory
Rama Vasudevan, Oak Ridge National Laboratory
Stuart Wilkins, Brookhaven National Laboratory
Kevin Yager Brookhaven National Laboratory

Invited Observers

Manouchehr Farkhondeh, Nuclear Physics
Tim Fitzsimmons, Basic Energy Sciences
Robin Hayes, Basic Energy Sciences
Linda Horton, Basic Energy Sciences
Harriet Kung, Basic Energy Sciences
Jeffrey Krause, Basic Energy Sciences
Peter Lee, Basic Energy Sciences
LK Len, High Energy Physics
Eliane Lessner, Basic Energy Sciences
George Maracas, Basic Energy Sciences
Gail McLean, Basic Energy Sciences
Raul Miranda, Basic Energy Sciences
James Murphy, Basic Energy Sciences
Van Nguyen, Basic Energy Sciences
Katie Runkles, Basic Energy Sciences
James Rustad, Basic Energy Sciences
Andrew Schwartz, Basic Energy Sciences
Wade Sisk, Basic Energy Sciences
Emily Smith, Basic Energy Sciences
Ed Stevens, Basic Energy Sciences
Pappannan Thiyagarajan, Basic Energy Sciences
Brenda Wyatt, Oak Ridge National Laboratory
Jane Zhu, Basic Energy Sciences

Appendix B: Roundtable Agenda

BES Roundtable on AI/ML
Hilton Washington, DC/Rockville Hotel & Executive Meeting Center
1750 Rockville Pike, Rockville, MD
October 22-23, 2019

Tuesday, October 22, 2019

7:30–8:30 AM	**Registration/Continental Breakfast–Roosevelt**
8:30–8:45 AM	Welcome Harriet Kung, Associate Director of Science, Basic Energy Sciences
8:45–9:00 AM	Roundtable Charge Eliane Lessner, Scientific User Facilities Division
9:00–9:30 AM	Roundtable Goals and Logistics and Introductions Daniel Ratner, SLAC National Accelerator Laboratory Bobby Sumpter, Oak Ridge National Laboratory
9:30–10:15 AM	*A Materials Opportunity: Machine Learning, Artificial Intelligence, and Autonomous Experimentation in Microscopy for Materials Design* Sergei Kalinin, Oak Ridge National Lab
10:15–10:45 AM	**Break**
10:45 AM–12:15 PM	Brainstorming Session All Participants
12:15–1:30 PM	**Working lunch** (Writing Team to meet in **Adams**)
1:30–3:00 PM	Breakout Session 1: Development of Key Science Challenges and Draft PROs Panel I–Data Acquisition–**Madison** Frank Alexander, Brookhaven National Lab Peter Denes, Lawrence Berkeley National Lab Panel II–Online Control–**Regency** Kevin Yager, Brookhaven National Laboratory Xiaobiao Huang, SLAC National Accelerator Laboratory *Break items will be available from 3:00–3:30*
3:00–4:30 PM	Breakout Session 2: Development of Key Science Challenges and Draft PROs Panel III–Multimodal Learning**–Madison** Alex Hexemer, Lawrence Berkeley National Laboratory

	Rama Vasudevan, Oak Ridge National Laboratory
	Panel IV–Models Simulations–**Regency** Subramanian Sankaranarayanan, Argonne National Laboratory Mathieu Doucet, Oak Ridge National Laboratory
4:30–5:00 PM	Report-outs from each Panel by Panel Leads–**Roosevelt** Key ideas/challenges/approaches; first draft of PROs from panels
5:00 PM	Adjourn for Day 1
5:30 PM	Writing Team Dinner–**Adams**

Wednesday, October 23, 2019

7:30–8:30 AM	**Continental Breakfast**
8:30–8:45 AM	FYI: Where We Are Now?–**Roosevelt**
8:30–10:30 AM	Breakout Session: Development and Refinement of Down-selected PROs *(note: may require adjustment of panels to align with PROs)*
10:30–10:45 AM	**Break**
10:45–11:00 AM	Breakout Session (cont'd): Panelists develop potential side bars and other assignments Panel leads work on PROs with chairs for Presentation
11:00–11:45 AM	Report-outs from each Panel by Panel Leads**–Roosevelt**
11:45 AM–12:00 PM	Closing Remarks Roundtable chairs, BES
12:00 PM	Roundtable Adjourns
12:00–5:00 PM	**Writing Team assembles (by invitation only)**

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.

DOI 10.2172/1630823

U.S. DEPARTMENT OF ENERGY | Office of Science