
Synchrotron Radiation News

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

# AI@ALS Workshop Report: Machine Learning Needs at the Advanced Light Source

**Dilworth Y. Parkinson, Tanny Chavez, Monika Choudhary, Damon English, Guanhua Hao, Thorsten Hellert, Simon C. Leemann, Slavomir Nemsak, Eli Rotenberg, Andrea L. Taylor, Andreas Scholl, Ashley A. White, Antoine Islegen-Wojdyla, Petrus H. Zwart & Alexander Hexemer**

**To cite this article:** Dilworth Y. Parkinson, Tanny Chavez, Monika Choudhary, Damon English, Guanhua Hao, Thorsten Hellert, Simon C. Leemann, Slavomir Nemsak, Eli Rotenberg, Andrea L. Taylor, Andreas Scholl, Ashley A. White, Antoine Islegen-Wojdyla, Petrus H. Zwart & Alexander Hexemer (28 Aug 2024): AI@ALS Workshop Report: Machine Learning Needs at the Advanced Light Source, Synchrotron Radiation News, DOI: 10.1080/08940886.2024.2391258

**To link to this article:** https://doi.org/10.1080/08940886.2024.2391258

OPEN ACCESS

MEETING REPORT

# AI@ALS Workshop Report: Machine Learning Needs at the Advanced Light Source

## Introduction

A key component of the future vision for the Advanced Light Source (ALS) at Lawrence Berkeley National Laboratory is acceleration of scientific knowledge creation using synchrotron light through intuitive and transformative computational solutions. We believe that leveraging AI will be critical to achieving this vision, so to formulate our plans in this area, a focused AI@ALS workshop was held for ALS staff on February 28–29, 2024. The workshop started with a plenary with an introduction and flash talks on some of the work that has been done on AI at the ALS, then moved into two sets of breakouts: the first day of breakouts was based on synchrotron and scientific domains (e.g., accelerator, user office, controls, imaging, scattering, spectroscopy, and biology), while the second day was based on AI/ML domains (e.g., large language models, dimensionality reduction, autonomous, generative and digital twins, and AI-ready controls and data). The full charge to the workshop participants was as follows:

This workshop aims to survey the current use of machine learning (ML) at the Advanced Light Source (ALS), identify the main challenges faced by scientists in applying ML to accelerator work, data analysis, and autonomous data collection at beamlines and discuss potential future directions for ML support in these areas. It seeks to foster an exchange on the realities and possibilities of ML integration in scientific research workflow at ALS.

- **Current Applications Insight:** Obtain a clear picture of how ML is being utilized within the ALS community, highlighting existing practices, successes, and areas for improvement.
- **Challenges Identification:** Identify common challenges and bottlenecks in leveraging ML for enhancing data analysis and autonomous operations at beamlines.
- **Future Possibilities:** Outline realistic and aspirational goals for the advancement of ML support at beamlines, focusing on practical steps to achieve these objectives.
- **Strategic Recommendations:** Formulate practical recommendations for addressing current challenges and facilitating the adoption of ML technologies at beamline.

This report summarizes the outcomes of the workshop, and is divided into the following sections:

- FAIR Data for Synchrotrons
- Current and Future AI@ALS
  - AI@Accelerator
  - AI@Beamlines
  - AI@User Office
- Showcases of past AI/ML work at the ALS
- Conclusion/Executive Summary

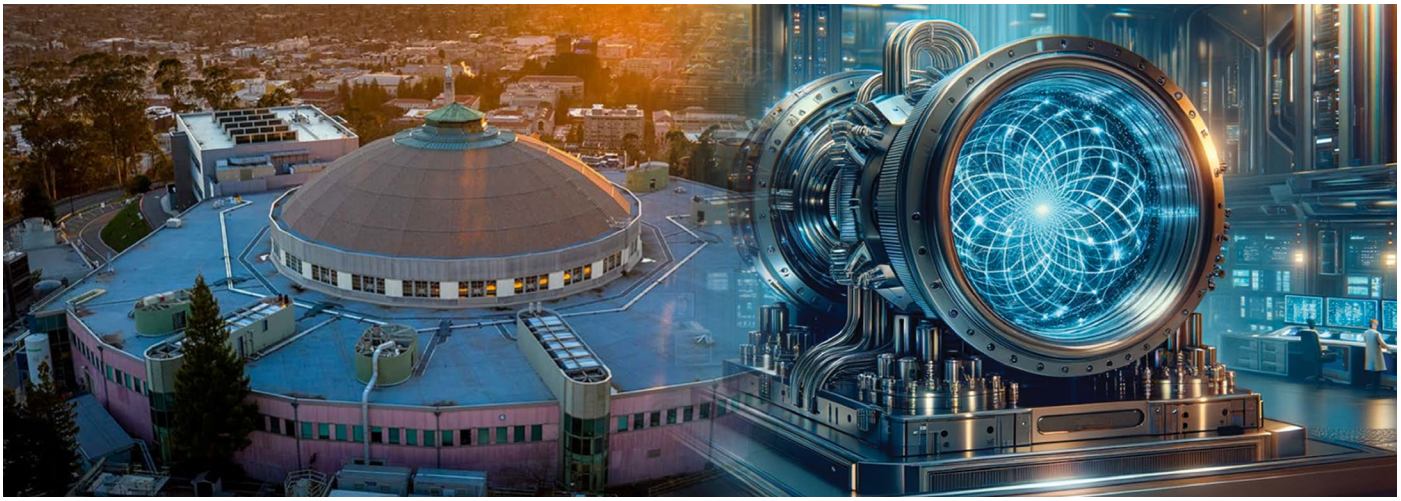
## FAIR data for synchrotrons

FAIR stands for Findability, Accessibility, Interoperability, and Reusability. During the workshop, the principles of FAIR data emerged as an important cornerstone of any AI and ML efforts [1]. By adopting these principles, the ALS can vastly improve the quality and usability of its datasets, fostering a collaborative environment that benefits every researcher. This section reviews the transformative benefits of the FAIR approach to data, the steps required to implement this approach, and how it can be integrated into the fabric of ALS operations.

### *Benefits of FAIR data*

There are a number of benefits of FAIR data, including the following.

- **Enhanced Data Quality:**
  - FAIR data principles ensure that datasets are thoroughly documented and standardized. This meticulous approach reduces errors and biases, providing a solid foundation for training robust AI/ML models.



- **Improved Accessibility:**
  - With data being easily findable and accessible, researchers can swiftly locate the datasets they need. This ease of access accelerates the research process and encourages collaboration, as data can be shared and reused across different studies and disciplines.
- **Interoperability:**
  - Standardized data formats and protocols allow integration of datasets from various sources. This is crucial for multi-modal data analysis, where combining different data types can lead to new insights and discoveries.
- **Reusability:**
  - When well-documented and preserved with comprehensive metadata, data becomes a valuable resource that can be reused for future research. This reusability ensures that datasets contribute to scientific progress long after creation.

***What's needed to make data FAIR***

Implementing FAIR data principles requires a concerted effort and several key components [2]:

- **(Meta)data Standards:**
  - Establishing robust metadata standards is vital. Metadata should provide detailed descriptions of the dataset, its origin, and the processing steps it has undergone, ensuring that anyone can understand and use the data correctly. The ALS works closely with other lightsources to establish and adopt standards. Many techniques used at the ALS do not have established standards and will require internal development. The ALS will curate, version, and publish the standards that it uses. (F, I, R)
- **Data Repositories:**
  - Centralized data repositories are essential. These repositories should be easily accessible, support standardized data formats, and provide robust data discovery and retrieval tools. The ALS is using SciCAT to accomplish this. Additionally, repositories will create globally unique identifiers for datasets, a key provision of FAIR principles. Integration with DOI minting services will enable swift exposure of published data sets. (F, A)
- **Data Governance Policies:**
  - Clear data governance policies are needed to outline data ownership, access rights, and usage guidelines. These policies ensure that data is managed responsibly and ethically. Some bottom-up effort to define this is useful, but it ultimately needs to be agreed upon by all stakeholders and the DOE.
- **Technological Infrastructure:**
  - Investing in the necessary technological infrastructure, such as high-performance computing resources and advanced data management systems, supports the efficient storage, processing, and analysis of large datasets. Infrastructure to curate and publish data standards and validate (meta)data will enable the future of FAIR data at the ALS.

***Bringing FAIR data to life at ALS***

Turning the vision of FAIR data into reality at ALS involves practical steps and strategic initiatives:

- **Developing Guidelines**:
  - Creating comprehensive guidelines that outline the steps to make data FAIR is essential. These guidelines should include templates and best practices for metadata documentation, data formatting, and repository submission. These guidelines must be developed collaboratively.
- **Centralized Repositories:**
  - The current ALS centralized repository (based on an implementation of SciCAT) needs to be extended to many more beamlines to foster a comprehensive centralized repository.
- **Automating Data Collection:**
  - Implementing automated systems for data collection that include metadata generation ensures that datasets are consistently and accurately documented from the outset.
- **Fostering a Culture of Sharing:**
  - Encouraging a culture of data sharing among researchers is essential. Highlighting the benefits of data reuse and collaboration and recognizing contributions to data repositories through incentives can promote this culture.
- **Collaborating Externally:**
  - Partnering with other research institutions and data standards organizations can help align ALS data practices with global FAIR data standards. Such collaboration enhances interoperability and facilitates broader data sharing and reuse.
- **Improve Training Tools**
  - To make the best use of available resources and effectively analyze and visualize their data using advanced computational resources, such as HPC, ALS users will require comprehensive training to. This improved training will enable users to fully leverage HPC capabilities, allowing them to continue their data analysis and visualization independently even after their experiments are completed.

Adopting FAIR data principles at ALS is not just about improving data management; it's about unlocking the full potential of AI and ML in scientific research. By ensuring that data is findable, accessible, interoperable, and reusable, ALS can provide better training datasets, leading to more accurate and reliable AI/ML models. Implementing these principles requires a strategic approach involving robust metadata standards, centralized repositories, clear governance policies, and continuous education and training. ALS can accelerate scientific discovery and innovation through these efforts, benefiting the entire research community.

## Current and future AI@ALS

***AI@accelerator***

Modern particle accelerators, such as 3rd- and 4th-generation synchrotrons, are large and

incredibly complex instruments that require tight control and optimization to render desired performance. State-of-the-art codes and parallel supercomputing is employed during the design phase to optimize the particle accelerator parameters and also to investigate the effect of imperfections in the as-built instruments along with possible mitigation strategies. Later, once commissioned, these machines require further optimization to exploit their full potential as well as various monitoring, data acquisition, and complex manipulation tools to ensure high operational reliability as well as quick recovery from both planned and unplanned outages.

It is not uncommon to find a modern particle accelerator with 100,000 process variables (PVs, which include measured values of sensors, motor positions, vacuum levels, or power supply settings), many of which update at high data rates often well in excess of what a human can process or react to. AI and ML are ideal tools to support accelerator physicists and engineers in all of these aforementioned tasks since they offer the potential to abstract from multidimensional parameter hyperspaces or implementation complexity and instead allow for the focus to remain on the physics involved in solving an issue or improving the performance. At the ALS we have pioneered AI/ML applied to synchrotrons: we have demonstrated the AI/ML approach both for the operational accelerator as well as in the design of future accelerators—see the showcases section below for more details on this work.

*Autonomous accelerator operation*

The pursuit of autonomous accelerator operation leverages AI and ML to create self-optimizing systems capable of adjusting operational parameters in real-time [3]. This autonomy aims to maximize accelerator performance, energy efficiency, and beam stability without continuous human intervention. Such advancements could significantly reduce downtime and enhance the consistency of beam quality, thereby supporting a wide range of scientific experiments with varying requirements.

*LLM-powered control room assistant*

Language Model (LLM) powered assistants in the control room represent a significant leap towards enhancing operational efficiency and decision-making [4]. These AI-driven tools can provide operators with instant access to a vast repository of documentation, operational logs, and diagnostic information, facilitating quicker resolution of issues and optimizing operational strategies. This intelligent assistant can transform the control room into a more responsive and informed environment. This development is particularly critical due to the enormous volume of information that is challenging to navigate using traditional search methodologies. By streamlining access to this wealth of data, LLM-powered assistants ensure that operators can swiftly locate and utilize the specific information needed, significantly accelerating operational workflows and decision-making processes.

*Auto-tuning with reinforcement learning*

Reinforcement learning [5] offers a promising avenue for auto-tuning accelerator parameters to achieve desired outcomes. This technique involves training models to make decisions that maximize some notion of cumulative reward. In the context of accelerators, it could automate the complex process of beam tuning, achieving optimal performance faster and more reliably than traditional manual tuning methods. This approach is expected to be particularly beneficial for the operation of ALS-U, given its significantly increased complexity. The application of reinforcement learning to ALS-U introduces a scalable method that is well-suited to navigating the enhanced intricacies of machine start up, beam tuning and operational optimization.

*Advanced feed forward correction algorithms*

The implementation of advanced feed-forward correction algorithms stands to significantly improve beam stability and quality. By anticipating and compensating for disturbances before they affect the beam, these algorithms ensure a higher level of precision in beam delivery, crucial for experiments that demand the utmost accuracy and stability.

*Anomaly detection and predictive maintenance*

AI-driven anomaly detection [6] and predictive maintenance frameworks are set to dramatically reduce downtime and maintenance costs. By identifying patterns indicative of potential failures [7], these systems can alert operators to issues before they escalate into critical failures. This approach not only ensures the longevity of the accelerator's components but also enhances safety and reliability.

*Comprehensive databases of operational records*

A pivotal aspect of these frameworks is their ability to sift through vast volumes of historical operational data. This deep analysis enables the identification of equipment failures and any deviations from normal operations, offering a comprehensive view of the accelerator's performance over time. By leveraging archived data, AI algorithms can detect subtle anomalies that might precede equipment failures, ensuring that preventative measures can be implemented in a timely manner. This is supported by our ongoing effort to streamline our control subsystems to ensure they are fully synchronized and systematically archived in the new archiver appliance.

*Machine learning for monitoring insertion device (ID) health*

ML algorithms are poised to revolutionize the monitoring of ID health through the analysis of data points such as motor forces, vibrations, and tracking errors. These metrics act as early indicators for identifying wear and tear in essential components. By foreseeing maintenance requirements before breakdowns occur, ML is set to prolong the service life of valuable machinery, reduce unexpected operational pauses, and refine maintenance planning.

*Virtual diagnostics and digital twins for accelerators*

The establishment of a digital twin [8–10], specifically for the injector complex, marks an important target for advancing diagnostic capabilities and optimization efforts. This virtual model, continuously updated with real-time operational data, can simulate and predict the behavior of the physical system under various conditions. Such a tool allows for risk-free experimentation on operational im-

provements and troubleshooting without impacting the facility's productivity. This is especially crucial because the key diagnostic instruments within the injector system are destructive and cannot be utilized during standard operational periods.

***AI@beamlines***

AI can affect every aspect of user and staff experience at beamlines—before, during, and after experiments. AI is applied to experiment planning, beamline setup, and alignment; to automate, accelerate, and improve data collection; and to analyze and extract information from data. In many cases, AI does not just accelerate experiments, it completely changes how experiments are done. To give one scientific example of why this is the case, consider the many-body degrees of freedom in real materials, which can result in multiple overlapping features in the spectroscopic data [11]. There is currently no general physics-based *a priori* interpretation of complex spectral features that could be used to automate data reduction, but ML-based approaches can help make progress in this area.

*Proposal phase, experiment planning*

While users prepare proposals, AI/ML algorithms and tools can streamline processes, offering enhancements and recommendations to ensure that proposers are equipped with the insights needed to refine their experimental designs and hypotheses. For example, sophisticated chatbots fine-tuned with information from previous open-access publications and technical beamline documentation, could offer invaluable assistance in experimental planning. These AI assistants can provide immediate, tailored advice on a range of topics—from selecting the most suitable beamline for a particular experiment to optimizing experimental parameters.

ML-augmented analysis that leverages existing data from previous experiments could offer proposers preliminary insights, suggesting potential outcomes and helping in refining hypotheses and experimental approaches. By analyzing data from similar studies, ML algorithms can pre-emptively identify then suggest the most relevant previous data and processing pipelines for a given proposal, enabling the development of bespoke analysis pipelines. This process ensures that by the time researchers arrive at the ALS, they are already equipped with ML models and analysis strategies finely tuned to their specific needs.

These tools could be pushed a step further, where users can use "digital twins", virtual surrogates of experimental setups, to conduct virtual preliminary experiments, exploring a wide range of scenarios and parameters before physically conducting their studies. This not only optimizes the use of beamline time but also significantly enhances the likelihood of experimental success.

*Beamline alignment and experiment setup*

This section explores how AI and ML technologies can be harnessed to streamline and accelerate beamline alignment and experiment setup, transforming these preliminary stages into more efficient and less labor-intensive tasks.

The concept of a digital twin [8–10] is particularly potent for beamline alignment and setup. By creating digital twins of every beamline, stage, and motor, AI algorithms can simulate and predict the optimal configurations for experiments. This virtual representation allows for extensive pre-experimentation planning and troubleshooting, significantly reducing the time required for physical adjustments.

The utilization of neural networks enables the precise prediction of beam position, size, and intensity, facilitating the fine-tuning of beamline configurations. These AI models can learn from historical data to predict the outcomes of different settings, guiding technicians and researchers in making informed adjustments that optimize the experimental setup.

AI-powered systems can also address mechanical imperfections, such as backlash and non-linearity, through automated calibration processes. Advanced Intelligent Systems (AIS) can be employed to dynamically adjust and calibrate beamline components, ensuring high precision and reliability without the need for constant human intervention.

The concept involves collecting and utilizing training data during non-user hours to refine AI models dedicated to beamline alignment. This approach leverages the downtime of facilities to enhance the capabilities of AI systems, ensuring that they are continually learning and improving. By performing fully autonomous beamline alignment, the initial coarse adjustments can be managed through ML techniques, such as Bayesian optimization. Following this, a hybrid approach that includes analytical refinement can fully optimize the beamline settings. A critical aspect of this process is the integration of digital twins with full physics modeling. This combination not only accelerates the learning process for deep learning approaches but also ensures that the AI's predictions and adjustments are grounded in the physical realities of the beamline operations. The ability of AI and ML to simulate and predict the complex interplay of factors affecting beamline performance represents a significant advancement in experimental setup processes [12].

*During beamtime*

During beamtime, AI and ML can enhance the efficiency, accuracy, and outcomes of scientific experiments through data preparation, autonomous operations, real-time analysis, and the implementation of digital twins.

At the core of AI-ready beamline controls is a robust data management infrastructure designed to handle the vast volumes of data generated during experiments, as demonstrated in Figure 1. This infrastructure encompasses data storage, formats, retrieval, and throughput—critical enabling technologies that support a wide range of AI@ALS capabilities. Centralized database management and automated data movement, coupled with strict adherence to data security and privacy policies, form the backbone of this system, ensuring that data is not only accessible but also protected.

The foundation of effective AI/ML application in beamline experiments is the preparation of "ML-ready" data. By ensuring compatibility among various systems and facilitating higher data throughput, this infrastructure paves the way for advanced data analysis and ML applications. Techniques to render scientific data digestible for ML algorithms are crucial. Equally important is the robust collecting and version control of analysis and ML solutions, linked

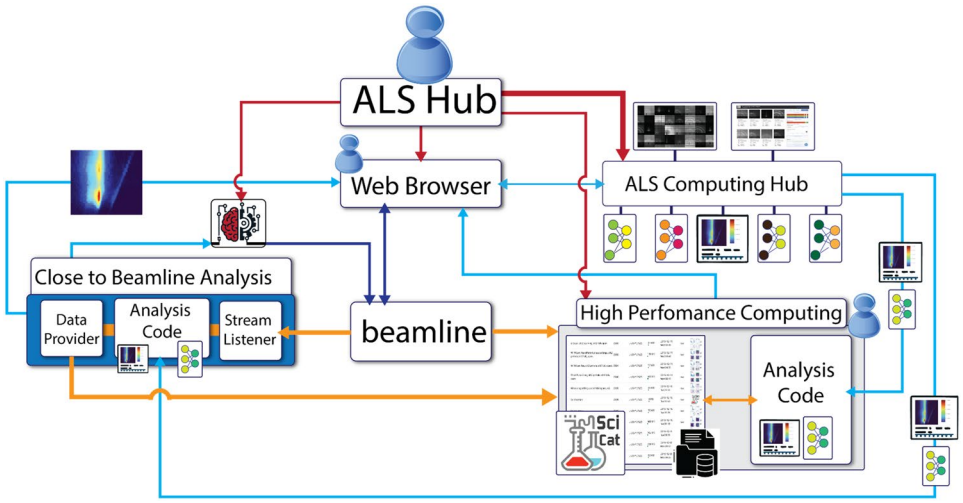



*Figure 1: This figure shows the interconnected infrastructure necessary to provide ML-as-a-service to an AI-ready beamline. This encompasses data storage, formats, retrieval, and throughput.*

seamlessly to the data as metadata. These steps ensure data integrity, reusability, and compatibility with advanced ML algorithms.

In addition to the traditional experimental data, data must be collected from beamline control logs and the other extensive parameters related to the experimental setups, including environmental conditions such as temperature, vibration, and slope, and sample information, as well as proposal information, to ensure comprehensive data documentation. This detailed capture of data enables the daily optimization and alignment of experimental systems, with recommendations driven by AI-powered analysis. By learning from user activities and needs, these AI-enhanced systems can significantly improve experimental outcomes and efficiency.

*Autonomous beamline operations*

Autonomous beamline operations, enabled by AI and ML, encompass a wide array of activities from phase space exploration and sample development to parameter optimization and adaptive controls [13, 14]. These advancements are driving a new era of experimental efficiency and precision at the Advanced Light Source (ALS).

Key to this autonomy is anomaly detection and phase space exploration. AI technologies also enable sophisticated anomaly detection by mining control log data and incorporating inputs from a variety of sensors for multi-modal analysis. This approach allows for the continuous monitoring of instrument health, anticipating potential failures before they occur. By identifying non-random patterns in the data, AI algorithms can notify relevant personnel of emerging issues, facilitating timely interventions.

Digital twins play a crucial role in closing-the-loop type experiments. By creating virtual surrogates of physical systems, digital twins can simulate and predict optimal configurations for experiments. This virtual representation allows for extensive pre-experiment planning and troubleshooting, significantly reducing the time required for physical adjustments and ensuring high precision and reliability in experimental setups.

Autonomous beamline operations are also advancing through adaptive scanning techniques. These techniques, supported by scalable and reusable elements, including common APIs and robust communication networks, exemplify the integration of AI in beamline control. Fast, high-performance controls capable of generating actionable information in real-time are crucial for the specific computational needs of each experiment.

AI-enhanced beamline controls incorporate the concept of the "human in the loop" [15, 16], allowing users to guide and influence automated behaviors. This approach fosters a collaborative interaction between researchers and AI systems, enhancing the user experience and experiment efficacy. Examples such as the Self-Driving Scanning Transmission X-ray Microscope (STXM) and the ARPES AARDVARK highlight how AI and ML are being deployed to increase scanning speeds and improve data acquisition and analysis in preparation for next-generation light sources like ALS-U. An example for IR and ARPES is presented in the AI@ALS Showcase section below.

Viewed holistically, an ALS experiment can be integrated into the search for new materials through the selection of compositional and structural properties, which is a proposed activity at Charter Hill (see https://research.lbl.gov/2021/03/11/lab-workshops-and-events-provide-input-for-charter-hill-materials-and-chemistry-campus-vision/). In some cases, ALS experiments already include synthesis and ancillary characterization capabilities, allowing for "complete" closed-loop experiments. In other instances, synthesis is conducted beforehand off-site, requiring judgment on which samples to prepare. This experimental need can be addressed by generative algorithms both before and during the experiments.

*Data analysis during beamtime*

ML-augmented analysis techniques, such as phasing, crystal structure determination, and image processing, are set to advance the

capabilities of beamline experiments significantly. ML-based tomographic and ptychographic reconstruction, image segmentation, peak and shape detection, and spectral fingerprinting are examples of how ML can transform data analysis, enabling deeper insights and discoveries. Dimensionality reduction techniques, such as PCA and NMF, further enhance the ability to distill and interpret complex datasets.

Often experiments are conducted at multiple conditions, where the goal is to optimize experimental information, optimize sample quality, or locate novel spectral features. These degrees of freedom include temperature, beamline settings, *in operando* parameters like applied voltage to control charge density or to scan along I-V trajectories. In other experiments the goal might be to survey these parameters to determine overall phase diagrams, e.g., in "library" samples with 2D compositional gradient structures. When scanning these libraries, there are large swaths of the search phase space where nothing happens, separated by interfaces where the phase and spectral information is most interesting, and time should be focussed on the latter, not wasted on the former. The extreme case of this is the "needle in a haystack" problem [17], where the interesting results occur for highly localized sets of conditions. Researchers are searching immense volumes of this high dimensional DOF at the Advanced Light Source. The wealth of information must be distilled to its essential elements without losing critical insights. Dimensionality reduction while focusing on the commonality and differences in the data set of techniques designed to simplify complex datasets, making them more manageable and interpretable [18, 19].

Principal Component Analysis (PCA) [20–22] helps identify the aspects of the data that matter most, those that carry the most significant patterns and trends in the data. PCA transforms the data into a set of orthogonal components, ranking them by the variance they explain. This method has become a staple in the ALS toolkit, especially useful in reducing the complexity of large datasets while retaining the core information.

Non-Negative Matrix Factorization (NMF) [23–25] offers a different approach. It decomposes the data into two smaller matrices with only non-negative elements, much like taking a colorful image and breaking it down into basic colors and their intensities. NMF shines in applications where the data naturally exhibit non-negativity, such as spectral imaging. By focusing on these fundamental components, researchers can isolate and study individual chemical compounds within complex mixtures.

Another powerful tool in the dimensionality reduction arsenal is Uniform Manifold Approximation and Projection (UMAP) [26]. UMAP excels at preserving the global structure of high-dimensional data while providing a clear, low-dimensional representation. It is particularly effective for visualizing complex relationships and identifying clusters within the data.

After UMAP has reduced the dimensionality of the data, clustering algorithms such as k-means [27, 28] or hierarchical clustering [29, 30] can be applied to identify distinct groups within the data. This approach is akin to mapping out different neighborhoods in a city after reducing the city's complexity into a manageable map. This technique is invaluable at ALS for categorizing various data for materials, biological samples, or experimental conditions based on their underlying characteristics.

Autoencoders [31–34] are a sophisticated type of neural network designed for efficient data representation. Think of them as skilled artisans who compress data into a simpler form without losing its essence and then reconstruct it almost perfectly. At ALS, autoencoders are pivotal in tasks such as denoising images and reconstructing high-quality tomographic images (Figure 2).

Dimensionality reduction can effectively reduce noise by focusing on the most critical components and filtering out irrelevant information, clarifying the data and making it more reliable. Computational efficiency is also improved, as reduced-dimensional data requires less processing power and time—a crucial factor in large-scale experiments. Moreover, machine learning models often perform better when trained on reduced-dimensional data, as this helps mitigate overfitting and enhances the models' ability to generalize to new data [35, 36].

Dimensionality reduction is more than just a data processing tool at ALS; it is a transformative approach that empowers researchers to unlock deeper insights and achieve more precise results. By simplifying the vast and complex data they work with, these techniques not only make the data more manageable but also significantly enhance the overall research process. As ALS continues to push the boundaries of scientific discovery, dimensionality reduction will remain a vital part of their analytical arsenal.

*Post beamtime data analysis*

While enhancing on-the-fly data analysis capabilities at the ALS remains a priority, it is equally important for users to have the ability to analyze and review results post-beamtime. This allows them to identify interesting outcomes or correlations from the entire collected dataset. To facilitate this analysis, datasets should be readily accessible and cataloged with corresponding experimental metadata. For example, integrating data repositories with searchable metadata tags can help users quickly locate specific datasets. Additionally, ML tools, such as dimension reduction combined with clustering algorithms to group similar results, should be capable of summarizing these findings and enabling users to explore the data efficiently.

Additionally, there is an opportunity to contribute training data to open-source community-wide datasets for the development of breakthrough tools, such as AlphaFold [37]. By providing diverse and high-quality datasets, scientists at the ALS can further improve pre-trained ML models for a variety of analysis pipelines. This collaborative approach accelerates innovation, as it allows for the validation and improvement of algorithms across techniques and user facilities. Moreover, open-access data repositories promote transparency and reproducibility in research, fostering a more robust scientific community.

ML techniques provide the ability to uncover valuable insights within complex data

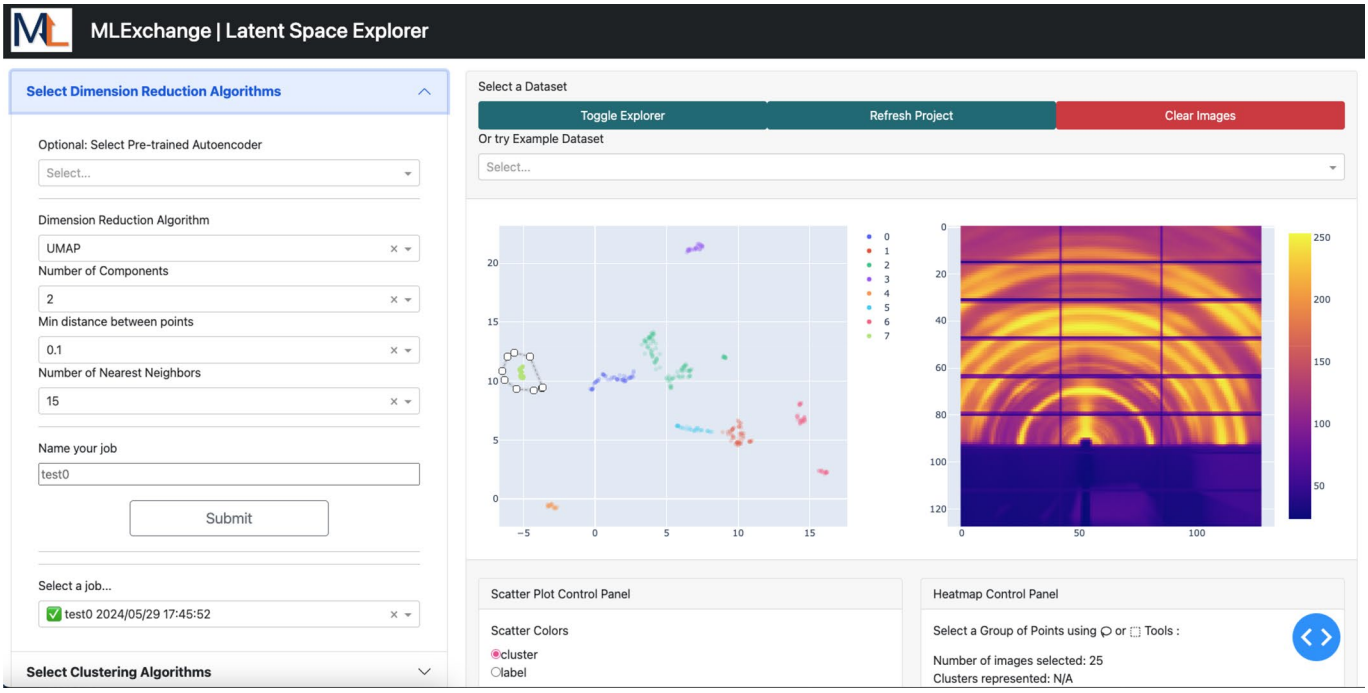



*Figure 2: Web-based interface for dimensionality reduction and visualization. This application makes use of autoencoders, dimension reduction techniques, and clustering algorithms for analysis.*

sets, such as those collected at the ALS. In the context of multi-modal experimentation, incorporating both in-situ and ex-situ characterization, ML-based analysis can help users process data from partial experiments and use these insights to guide future experimental designs. For instance, the analyzed results of partial multi-modal experiments can serve as valuable prior knowledge for planning and executing future ex-situ characterization (and vice versa). By integrating this prior knowledge into data-driven learning systems, such as reinforcement learning models [38, 39] or hybrid analytics frameworks, we can enhance the performance and execution of subsequent experiments. This approach informs decision-making and helps determine the optimal direction for future investigations.

Similarly, users can leverage past correlations with synchrotron data to "upscale" non-synchrotron data. An example of this is utilizing machine learning models trained on both lab-based microCT data and synchrotron scans to enhance the resolution of lab-based microCT images. The use of super-resolution ML-based enhancement methods, such as Super-Resolution Convolutional Neural Networks (SRCNN) [40] or Generative Adversarial Networks (GANs) [41], can further improve the correlation between synchrotron and non-synchrotron datasets, providing more detailed and accurate representations.

*Publications*

ML techniques enable the use of historical experimental data collected at the ALS, combined with published findings, to inform pre-trained models and analysis pipelines. By leveraging this wealth of past high-quality data, ML algorithms can identify patterns and insights that would be difficult to discern manually. This approach allows for the development of sophisticated, custom analysis pipelines tailored to specific experimental needs.

Additionally, these tailored pipelines can be made readily accessible to researchers prior to their beamtime. This preparation maximizes the efficiency and productivity of the beamtime, allowing researchers to derive more meaningful and actionable insights from their experimental data. By integrating historical data and cutting-edge ML techniques, this approach not only enhances the immediate research outcomes but also contributes to the continuous improvement and evolution of experimental methodologies at the ALS.

ML-based algorithms can integrate data from various sources to create comprehensive datasets for bibliometric analysis, which provides detailed information about users and their research groups. By analyzing co-authorship patterns in ALS-related publications, AI can map collaboration networks within the ALS user community. This information is valuable for targeted collaboration and outreach efforts, fostering stronger connections and more effective scientific partnerships.

***AI@ the user office***

The ALS User Office faces multifaceted challenges, from streamlining the affiliate processing and proposal review workflows to optimizing beamtime allocation and addressing language barriers. Similarly, the Communications team endeavors to identify high-impact publications, generate compelling content, and tailor communications to diverse audience

segments. In parallel, the complex domain of bibliometrics demands sophisticated tools for literature searches, citation analysis, and trend forecasting to gauge the impact of ALS-facilitated research on the broader scientific community.

*AI and ML in ALS user office*

Within the User Office at the Advanced Light Source (ALS), a series of innovative AI and ML applications are poised to transform the user experience and operational efficiency. One such application is the deployment of automated user support systems, such as chatbots and virtual assistants. These tools stand ready to deliver immediate and round-the-clock responses to common inquiries, significantly enhancing user interaction and satisfaction. This real-time assistance can streamline the resolution of queries related to the proposal process, training, safety requirements, and site access, thereby elevating the overall user experience [42].

In parallel, the utilization of Machine Learning (ML) in the initial screening of beamtime applications could represents a leap towards operational efficiency. By automating the review process based on predefined criteria, ML could swiftly identify applications that meet the ALS's stringent requirements, thereby streamlining the selection process and ensuring that beamtime is allocated to the most promising and impactful research projects.

Predictive analytics [43] further augments the User Office's capabilities by employing sophisticated models to forecast beamtime demand and identify emerging user trends. This forward-looking approach enables more effective planning and resource allocation, ensuring that the ALS can accommodate the evolving needs of its user community.

Scheduling optimization, another critical area of application, harnesses ML to consider a multitude of factors, including experiment complexity and user availability. This optimization ensures that beamtime is allocated in a manner that maximizes facility utilization while accommodating the diverse needs of researchers.

Sentiment analysis [44] of user satisfaction surveys offers insights into the user experience. By applying ML to analyze feedback, the ALS can proactively address concerns, refine services, and highlight areas of success, fostering a culture of continuous improvement and user-centric service.

Lastly, AI could play a role in enhancing the match between users and beamlines, as well as between proposals and reviewers. By analyzing the research needs and expertise available within its community, AI can ensure that users are directed to the beamlines that best match their experimental requirements. Simultaneously, proposals are assigned to reviewers whose expertise aligns with the research being proposed, ensuring a fair and informed review process [45].

These AI and ML applications collectively signify a paradigm shift in how the ALS User Office operates, promising a more responsive, efficient, and user-focused service model. Through these innovations, the ALS is set to redefine the standards of user support and operational excellence in the scientific research community.

*AI and ML in ALS communications*

The Communications team at the Advanced Light Source (ALS) is at the forefront of integrating Artificial Intelligence (AI) and Machine Learning (ML) to revolutionize how the facility engages with its diverse user base and the broader scientific community. These technologies can help in crafting targeted communications and segmenting audiences with unparalleled precision. By analyzing user engagement and preferences, AI can enable the Communications team to tailor messages and content that resonate with each segment of the ALS community, ensuring that information is relevant, engaging, and timely.

Website personalization can be facilitated by AI-driven tools capable of dynamically adjusting content in real-time based on user behavior and interests. This approach not only enhances the user experience but also fosters an interactive and personalized digital environment, encouraging deeper engagement with the ALS's resources and research opportunities.

The application of AI extends into the realm of impact analysis and sentiment analysis [44], providing the Communications team with a comprehensive understanding of the efficacy of their efforts and the public's perception of the ALS. These insights enable the team to refine their strategies, celebrate successes, and proactively address any areas of concern, ensuring that the ALS's communications are both effective and positively received.

Content generation, powered by AI, offers a solution to the often time-consuming task of drafting routine communications and summarizing complex technical topics for diverse audiences. AI tools can automate these processes, generating draft content that maintains the ALS's voice while making sophisticated scientific achievements accessible and understandable to non-specialists.

Finally, the enhancement of visual content through automated image and video processing showcases the potential of AI to revolutionize the creation and optimization of multimedia resources [46]. These tools can edit, caption, and optimize images and videos for various platforms, ensuring that visual content is both high-quality and tailored to the context in which it will be viewed.

Together, these AI and ML applications underscore a transformative shift in how the ALS Communications team operates, enabling more effective, personalized, and engaging interactions with the facility's community and enhancing the visibility and impact of ALS research on a global scale.

*AI and ML in bibliometrics for ALS*

The bibliometric analysis at the Advanced Light Source (ALS) is undergoing a transformative shift, thanks to the integration of Artificial Intelligence (AI) and Machine Learning (ML) technologies. These tools are not just enhancing the capabilities to track and analyze scientific outputs but also shaping the future of research impact assessment and strategic planning at the facility.

Automated literature searches stand at the forefront of this transformation. By employing AI, the ALS can efficiently streamline the search and retrieval process for publications associated with its facilities. This capability not only saves valuable time but also ensures

that even publications that may not explicitly mention the ALS in their acknowledgments are identified, thereby providing a more complete picture of the facility's scientific contributions.

The integration of diverse data sources and citation analysis further enriches bibliometric studies. By aggregating data from publication databases, citation indexes, and internal ALS records, AI algorithms can assess the research impact through citation patterns. This comprehensive dataset allows for a more nuanced analysis of how ALS research is utilized and referenced across the scientific community, highlighting its influence and reach.

Topic modeling and trend analysis, powered by AI, reveal emerging research areas and shifts in scientific inquiry [47]. This insight is invaluable for guiding strategic decisions at the ALS, ensuring that facility upgrades and new beamline developments are aligned with the future needs of the research community.

Collaboration networks [48] and future trends prediction are similarly enhanced through AI. By mapping out collaboration networks, the ALS can identify key researchers, institutions, and collaborations integral to its community. Predictive models offer foresight into future research trends, assisting in strategic planning and setting priorities that will keep the ALS at the cutting edge of scientific discovery.

Dynamic reporting tools and advanced visualization techniques, enabled by ML, automate the generation of metrics and foster sophisticated data visualization. This automation not only reduces the manual effort required in generating reports but also provides stakeholders with intuitive, interactive ways to understand research hotspots, collaboration networks, and temporal trends.

Lastly, personalized alerts and recommendations bring a new level of engagement to staff and users. AI systems can notify individuals of new publications in their areas of interest or alert them to emerging trends that might influence their research. This tailored approach ensures that the ALS community remains at the forefront of relevant scientific developments, fostering an informed and proactive research environment.

## AI@ALS showcases and activities

Integrating AI/ML at the Advanced Light Source represents a significant advancement in scientific research. By utilizing AI/ML, ALS improves how experiments are designed, conducted, and analyzed. This section presents key showcases and activities that demonstrate the impact of AI/ML on various aspects of the research process at ALS. From stabilizing photon beams to streamlining segmentation for tomography, here are a few examples that highlight the practical applications and benefits of AI/ML in enhancing scientific discovery.

### *Accelerator*

At the ALS we have pioneered AI/ML applied to synchrotrons: we have demonstrated the AI/ML approach both for the operational accelerator as well as in the design of future accelerators. In the former case, we used an ML-based approach to stabilize the ALS photon beams [49], thus delivering higher stable brightness in a feed-forward manner (including continuous online retraining) that does not rely on conventional feedback approaches with their associated tuning, stability, and latency issues (Figure 3).

In the latter case, we have shown how using AI/ML to develop surrogate models allows accelerator designers to perform multi-objective optimizations over many-dimensional input hyperspaces with large numbers of constraints using orders of magnitude less computational effort than conventional large cluster-based approaches [50]. Such an approach allows for physicists and engineers to effectively iterate much more heavily and closely on design optimization for most foreseeable sources of imperfection since the AI/ML approach brings performance evaluation down from the many-weeks to the few-hours regime (Figure 4).

### *Automated alignment of beamlines*

Beamlines typically contain a dozen of optical elements, from mirror to grating mono-

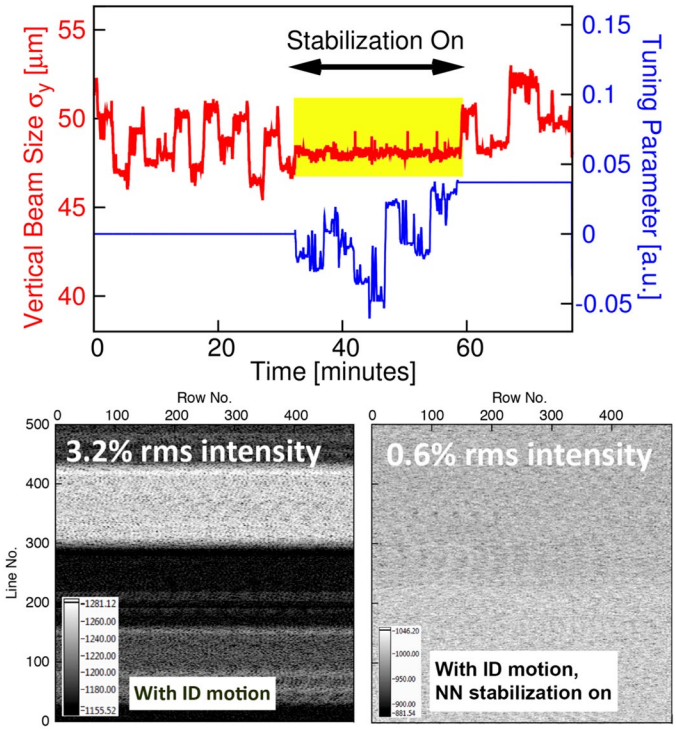



*Figure 3: Top: ALS vertical beam size (red) greatly improves when ML-based stabilization is on (yellow area). Blue trace is the parameter that an ML-based feed-forward tunes to cancel out fluctuations. Bottom: Banding that appears in scanning transmission x-ray microscopy (STXM) images at ALS beamline 5.3.2.2 during user ops ID motion (left) disappears after ML-based correction (right).*

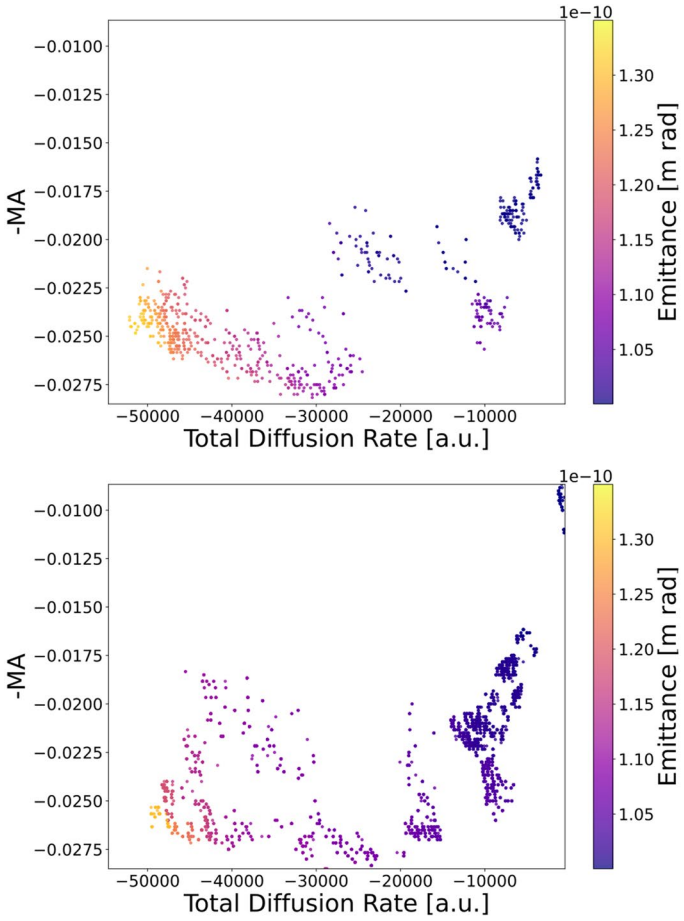



*Figure 4: ALS-U lattice optimization with an ML-enhanced multi-objective genetic algorithm in 11 dimensions. The ML-based optimization (top) achieves results similar to traditional processes (bottom) within much shorter time.*

chromators and slits, where many degrees of freedom are coupled. Machine learning provides a way to jointly optimize these multiple degrees of freedom and ensure optimal operation of the beamline – this will become increasingly critical as the next generation of coherent beamlines developed for ALS-U will be less tolerant to misalignment caused by environmental factors inducing drift.

We have demonstrated at the ALS R&D beamline 5.3.1 that automated alignment procedures based on Bayesian optimization and developed in collaboration with NSLS-II can be efficiently deployed and used to optimize the beam size and flux of the beam [51]. These procedures leverage the EPICS/bluesky framework [52] that we implemented at the beamline as part of a pilot program for instrument controls update and that will be the native framework for all new ALS-U beamlines (Figure 5).

The next generation of beamlines will feature adaptive optics, to provide exquisite control of the wavefront of the x-ray beam. Those adaptive optics are based on piezoelectric materials that exhibit drift and hysteresis, making them difficult to calibrate and limit their performance, thus making them hard to use in regular beamline operation. We have developed a procedure using neural networks to simplify the calibration and enable open-loop operation with diffraction limited performance [53]. This procedure was initially tested in a metrology lab (with visible optics), and then successfully implemented at a beamline (in collaboration with APS) [54] (Figure 6).

***Angle-resolved photoemission spectroscopy (ARPES)***

Nanoscale Angle-resolved photoemission spectroscopy (nanoARPES) is an information-rich spectromicroscopic technique that measures the electronic state energies as a function of their momentum. The spectra acquired are strongly modulated by matrix-element effects that are tied to the orbital character of the states. Finally, the spin of the electrons can also be measured. Thus, all the relevant quantum numbers are accessible in one probe, making it ideal to understand the origins of important phenomena such as superconductivity, magnetism, and topology.

Besides this feature, nanoARPES is also relatively fast, so that it is possible to screen many materials or phases to look for interesting spectral signatures. Our goals can vary: (i) to seek previously unobserved spectral signatures that can signal unknown or hidden orders; (ii) to evaluate different materials to understand their "electronic" phase diagram–e.g., Thermal- or charge-density-driven phases that are not accompanied by structural transitions; (iii) to perturb complex systems at the boundary between competing phases to understand the energy landscape. These goals can be accomplished by synthesizing multiple samples of different chemical or structural degrees of freedom–ideally assembled together in libraries on a single substrate platform, or it can be through probing heterogeneous assemblies to map out different phases–for example by inducing strong strain gradients in a sample to probe the strain-dependent phase diagram.

The challenges in this approach are as follows:

**Intrinsic complexity of the spectroscopic data.** The many-body degrees of freedom in real materials can result in multiple ARPES spectral features with complex lineshapes that reflect not only the ground states, but also complex excitation/deexcitation phenomena. The human eye can be drawn to "interesting" spectral features that may have nothing to do with the material function, while other, "boring" features like diffuse background might be overlooked. There is currently no general physics-based a priori interpretation of com-

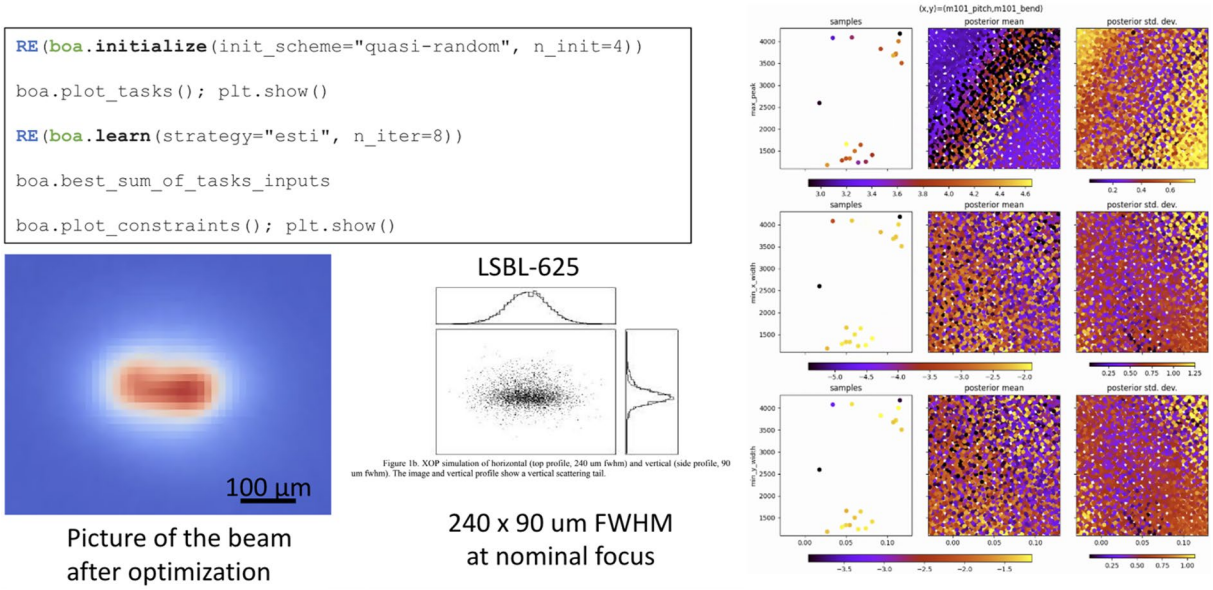



*Figure 5: Top left: sample code to start an automated alignment procedure (most of the logic is handled by bluesky); bottom left: beam at sample after optimization, compared with theoretical size; right: example of the underlying 4-dimensional logic of the scans used for Bayesian optimization.*

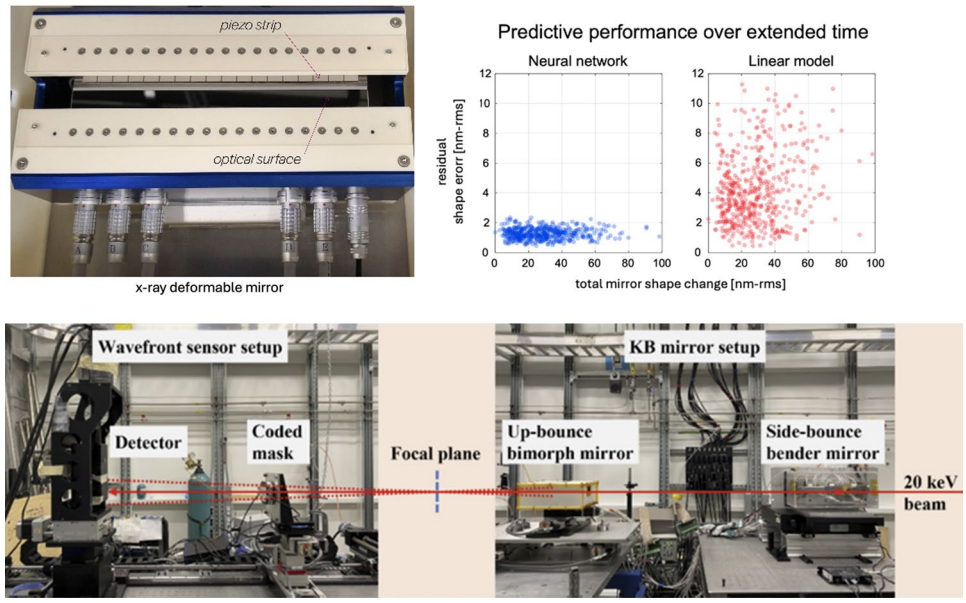



*Figure 6: Top left, X-ray deformable mirror. Top right, predictive performance over time using a linear model and a neural network. Bottom, beamline setup for a beamline ready for a beamline using automated alignment. Demonstration of machine-learning enabled adaptive optics at APS 28-ID.*

plex spectral features that could be used to automate data reduction. This is the rationale for ML-based approaches to data reduction.

**Dimensionality of the Spectroscopic Data.** The instantaneous acquired ARPES data unit is a 2D megapixel image (Energy vs momentum, acquired in ~seconds) but we frequently raster these over an additional momentum coordinate to form a 3D volume data set. Comparison of one experimental result to the next is complicated by differing energy or momentum calibrations or random instrument alignment errors. These experimental realities are an important constraint on the ML algorithms such as Gaussian processing and dimensionality reduction which have to be able to accurately assess what is "new" or "different"

about a spectrum without being hung up on extrinsic changes in the experiments.

*Dimensionality of Experimental Degrees of Freedom (DOF).* Often experiments are conducted at multiple conditions, where the goal is to optimize experimental information, optimize sample quality, or locate novel spectral features. These degrees of freedom can include combinations of temperature, beamline settings, in operando parameters like applied voltage to control charge density or to scan along I-V trajectories. When scanning these libraries, there are large swaths of the search phase space where nothing happens, separated by interfaces where the phase and spectral information is most interesting, and time should be focussed on the latter, not wasted on the former. The extreme case of this is the "needle in a haystack" problem, where the interesting spectra occur for highly localized sets of conditions. Searching this high dimensional DOF space efficiently is a common need of these approaches.

**Goals and Progress**. Our short-term goal is to scan a 2D spatial domain (x,y) coordinates, where each measurement consists of a 2D ARPES data map (E vs k). The aim is to segment the images, identifying all the regions of interest with their characteristic spectrum. The scanning should be done as efficiently as possible, so that the information obtained is comparable to what would be attained in a standard grid scan, but in the minimal number of measurements. The number of individual domains, and their spectral features are not known a priori. Results should be presented in real-time as the data are acquired, enabling "human-in-the-loop" interaction with the data collection program. Lastly, the results should be integrated into an active ARPES beamline so that the technique can be tested with real-life user samples.

To test different algorithms, we acquired a "ground truth" data set consisting of $N_0$~8200 gridded measurements. The sample is a library of twisted bilayer graphene composed of a polycrystalline graphene layer grown on Copper, then lifted off and deposited on a wafer-scale crystalline graphene layer [55], Fig. ARP1(a). The total spectral intensity is shown in Fig. ARP1(b). Probing a number of mea-

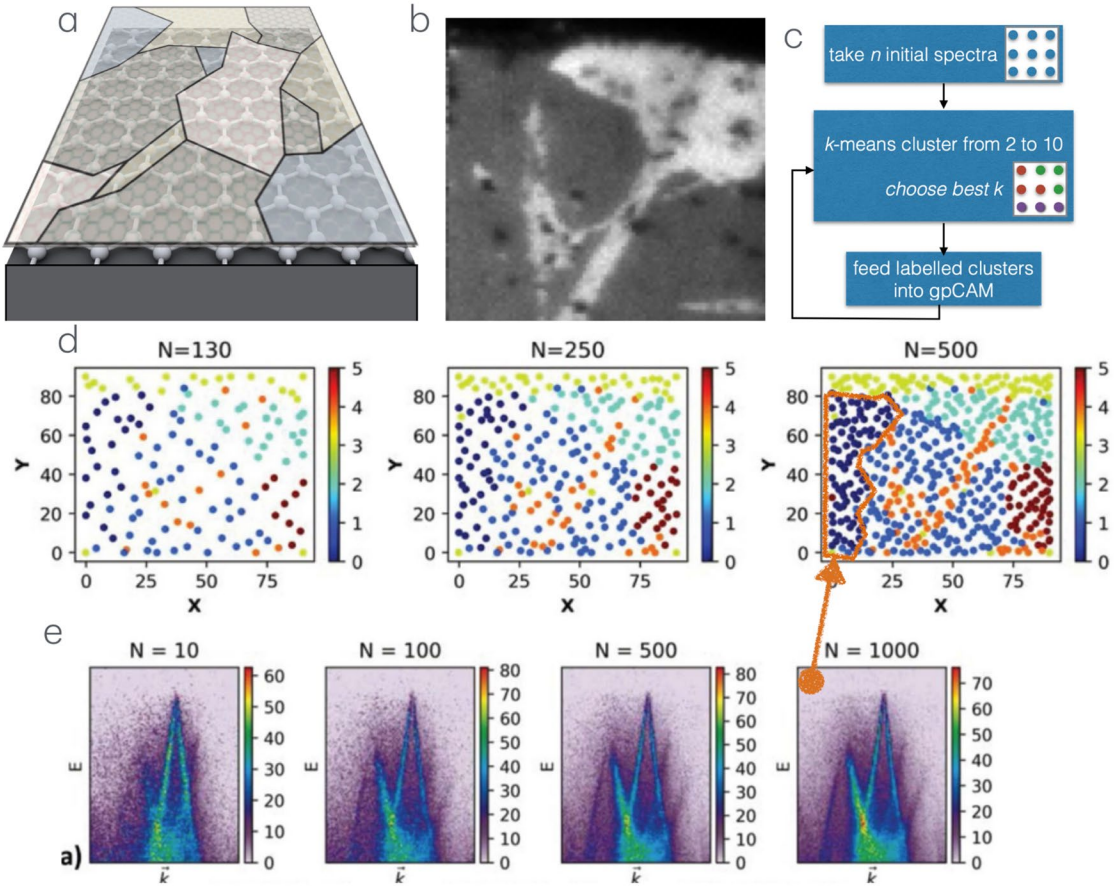



*Figure 7: ARP1 (a) Test sample consisting of polycrystal graphene on crystalline graphene, creating an array of bilayer graphene with a distribution of twist angles. (b) Ground truth image of sample, where each pixel represents the integrated section of the detected 2D ARPES image. Image consists of N0 ~ 8200 gridded measurements (c) block diagram of Gaussian Processing + K-means clustering algorithm. (d) sequence of clustered positions as a function of number of GP samples N. (e) The representation of the ARPES spectrum for cluster #0 (highlighted region in (b) as a function of N. From Ref. [55].*

surements $N < N_0$ of this data using a combination of K-means clustering and Gaussian Processing (GP) (Fig. ARP1(c)), the autonomous algorithm able to identify the unique domain elements (# clusters) (Fig. ARP1(d)) in about 10% of the data ($N$~500) [55]. The representative spectrum for each cluster (for example in Fig. 1ARP1(e) corresponding to the highlighted domain in (b)) similarly converged to its final value with only $N \sim N_0/10$.

The results above that 90% of the domain-defined data can be achieved using about 10% of the number of measurements. At the time of [55] the test code was run on a laptop so there was no time advantage, but in the interim we have moved the code to a GPU cluster which adds only a small overhead to the data collection, so that we can achieve approximately the same x10 performance improvement. In addition, the code was moved to a RESTful interface so that it could be accessed by users at BL7 using the existing labview code, or could be used by other beamlines (Figure 7).

### *Infrared spectromicroscopy*

Infrared spectroscopy (IR) measures the vibrational modes of chemical groups within molecules, providing a unique spectral "fingerprint" that identifies different chemical moieties. By detecting how molecules absorb infrared light at specific wavelengths, IR can reveal detailed information about the chemical structure and composition of a sample. This technique is highly effective for distinguishing between various functional groups and elucidating chemical abundance in a wide array of substances. The technique is label free and when performed at a synchrotron, it yields a sophisticated, label-free characterization technique that provides spatially resolved imaging modality.

Synchrotron Fourier Transform Infrared Spectromicroscopy (SFTIR) is used across a diverse range of fields, including materials science, chemistry, geobiology, and health sciences. The technique generates energy-resolved maps with a spatial resolution of 1 micron, covering areas of hundreds of microns. Despite the exceptional brightness of the Advanced Light Source, which significantly surpasses conventional benchtop IR instruments, the time required to map a sample size of 1.5 mm x 1.5 mm at 1 micron resolution can exceed a week. This extended duration makes extensive spatial spectral surveys challenging. This limitation is particularly significant given that IR spectromicroscopy is ideally suited for mapping and understanding complex spatio-chemical heterogeneity.

To address these challenges, the Berkeley Synchrotron Infrared Structural Biology (BSISB) program, in partnership with the Center for Advanced Mathematics for Energy Research Applications (CAMERA), has developed an innovative AI-based surrogate modeling method. This method leverages Gaussian Process Regression to adaptively adjust measurement strategies, enabling rapid and autonomous redirection of data collection efforts towards areas likely to contain interesting samples. Our approach has drastically reduced the time required to survey 1500 x 1500 micron samples to just two hours—a task that would otherwise be unfeasible [56]. This advancement significantly enhances the practicality and efficiency of utilizing SFTIR in extensive and detailed chemical imaging studies (Figure 8).

### *Tomography*

The tomography program at the ALS has a long history of applying AI/ML and computer

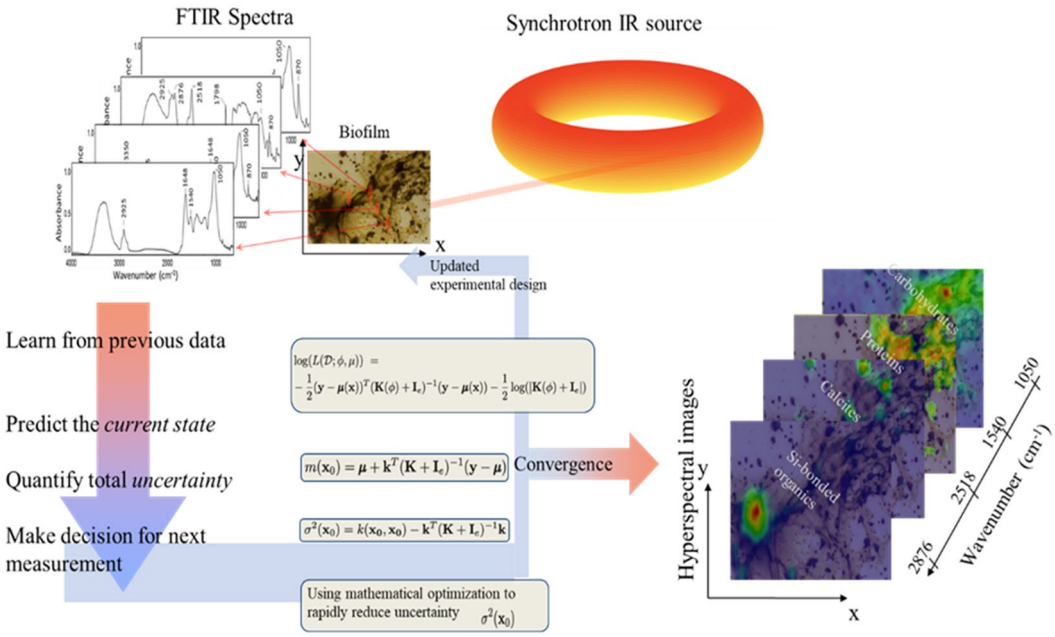



*Figure 8: Iterative feedback loop between observed data and underlying probabilistic surrogate model of the full experiment. By modeling the underlying spectra in the sample and their predicted uncertainty, informed decisions can be made on where to measure next.*

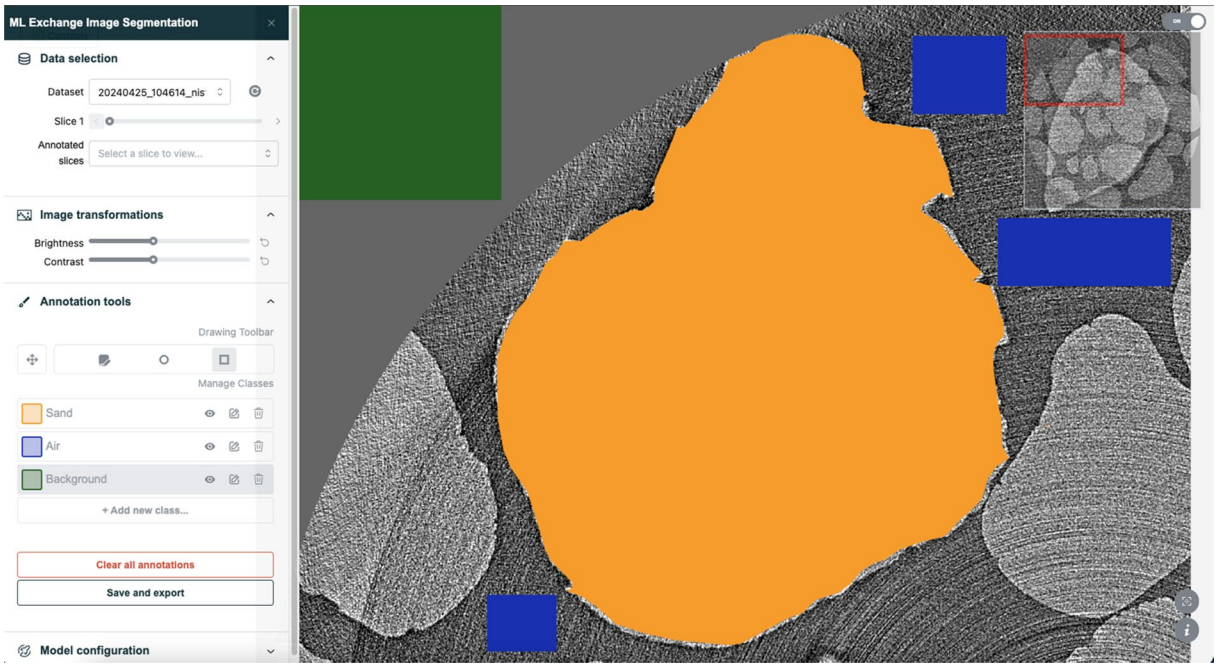



*Figure 9: Screenshot of the MLExchange image segmentation interface using DLSIA, developed at CAMERA, as a ML backend [62].*

vision tools to 3- and 4D imaging problems. This has included the invention of new types of convolution neural networks [57] that can be applied to image denoising and image, automatic detection of fibers in ceramics [58], a reverse image search platform for scientific images [59], a machine learning approach for detecting features in batteries [60].

Recently, a major focus has been to deploy segmentation capabilities within MLExchange, a web-based platform that aims to lower the introduction barrier of ML techniques for materials science applications. This platform is a cross-facility collaborative effort among five U.S. Department of Energy national laboratories such as Lawrence Berkeley National Laboratory (LBNL), Brookhaven National Laboratory (BNL), Argonne National Laboratory (ANL), Oak Ridge National Laboratory (ORNL), and SLAC National Accelerator Laboratory. Within this platform, MLExchange offers an assortment of web applications and ML algorithms for on-the-fly scientific data analysis. In particular, its high-resolution image-segmentation application [61] enables users to annotate high-resolution images on the web and use these annotations to train ML-based segmentation models with DLSIA, a Python library that provides customizable neural network architectures for various data analysis tasks, including segmentation [62]. With robust trained models, users can automatically segment large image data sets and visualize these results on the web. To achieve this, MLExchange makes use of Tiled from the Bluesky ecosystem for chunk-wise data access, and Prefect for workflow orchestration. These segmentation capabilities have been successfully tested for the segmentation of high-resolution reconstructed tomography data sets during experiments at a scheduled beam time (Figure 9).

## Conclusion and executive summary

The advent of machine learning (ML) is ushering in a transformative era across domains, making an important impact on science overall and at scientific user facilities such as the Advanced Light Source. The workshop on *Machine Learning Needs at the Advanced Light Source,* held on February 28–29, 2024, brought scientists together to discuss the current use, challenges, and future directions of ML in accelerator operations, data analysis, and autonomous data collection. The workshop identified key challenges, highlighted successful applications, and formulated strategic recommendations for integrating ML into ALS workflows. This workshop underscored the increasingly pivotal role of ML in revolutionizing the scientific discovery process, encompassing every stage: from initial proposal conception to autonomous data collection and near-real-time analysis to comprehensive analysis of complex multi-modal data.

## Current applications

ML is being effectively utilized at ALS in various domains:

1. **Accelerator Operations**: ML supports the optimization and control of complex systems, including autonomous operations, anomaly detection, and predictive maintenance. Notable advancements include ML-based beam stabilization and the development of surrogate models for accelerator design optimization.
2. **Beamline Operations**: AI and ML transform beamline setups, data collection, and analysis. Applications include automated beamline alignment, experiment planning, and ML-augmented real-time data analysis.

## Key challenges

The key challenges that must be addressed to realize the potential additional benefits from AI/ML include:

1. **Data Management**: Ensuring data is ML-ready involves standardized data formats, robust data management systems, and adherence to FAIR (Findable, Accessible, Interoperable, Reusable) principles.
2. **Integration Complexity**: The complexity of integrating AI and ML systems into existing workflows and the need for continuous model retraining to adapt to evolving data sets.
3. **User Training and Adoption**: Enhancing training tools and materials to enable users to leverage ML capabilities fully and independently.

## Future directions

Going forward, we plan to pursue these high impact directions in our AI/ML work:

1. **Advanced AI Systems**: Development of AI-driven control room assistants, reinforcement learning for auto-tuning, and advanced feed-forward correction algorithms to enhance accelerator performance.
2. **Enhanced Beamline Operations**: Expanding the use of digital twins for beamline alignment, autonomous operations, and adaptive scanning to improve experimental outcomes and efficiency.
3. **Comprehensive Data Analysis**: Implementing ML for multi-modal data analysis, predictive maintenance, and anomaly detection to ensure high operational reliability and quick recovery from outages.

## Strategic recommendations

1. **Data Infrastructure**: Invest in centralized data repositories, standardized metadata documentation, and automated data collection systems to support AI and ML applications.
2. **Machine Learning as a Service:** Implement a centralized system where ML inference is enabled through API calls, hosting models, algorithms, and GUIs accessible to the ALS community. This approach will streamline workflow setup, enable near real-time analysis, and support automation, thereby enhancing the efficiency of utilizing ML models.
3. **Collaborative Efforts**: Foster collaborations with other research institutions to align with global FAIR data standards and share datasets for broader scientific impact.
4. **Continuous Improvement**: Implement continuous education and training programs to ensure researchers can fully utilize AI and ML technologies, promoting a culture of innovation and collaboration.The integration of AI and ML at ALS promises to revolutionize research by enhancing experimental design, data collection, and analysis processes. This will ultimately lead to more efficient and impactful discoveries, breakthroughs, and advancements across various scientific disciplines.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research used resources of the Advanced Light Source, which is a DOE Office of Science User Facility under contract no. DE-AC02-05CH11231. ■

## References

1. M. D. Wilkinson et al., *Sci. Data.* **3** (1), 1 (2016). doi: 10.1038/sdata.2016.18
2. A. Jacobsen et al., *Data Intel.* **2** (1–2), 10 (2020). doi: 10.1162/dint_r_00024
3. A. S. Garcia et al., 15th International Particle Accelerator Conference (2024). doi: 10.18429/JACoW-IPAC2024-TUPS62
4. F. Mayet, *arXiv preprint arXiv:2405.01359* (2024).
5. J. Kaiser et al., *arXiv preprint arXiv:2306.03739* (2023).
6. J. P. Edelen and N. M. Cook, *arXiv preprint arXiv:2112.07793* (2021).
7. J. Branlard et al., Proceedings of ICALEPCS2013, 2013.
8. M. Shafto et al., *Technology Area.* **11**, 1 (2010).
9. E. Negri, L. Fumagalli, and M. Macchi, *Procedia Manuf.* **11**, 939 (2017). doi: 10.1016/j.promfg.2017.07.198
10. M. G. Juarez, V. J. Botti, and A. S. Giret, *J. Comput. Inf. Sci. Eng.* **21** (3), 030802 (2021). doi: 10.1115/1.4050244
11. F. Giustino et al., *J. Phys. Mater.* **3** (4), 042006 (2021). doi: 10.1088/2515-7639/abb74e
12. L. Rebuffi et al., *Opt. Express.* **31** (24), 39514 (2023). doi: 10.1364/oe.505289
13. A. Barbour et al., *Synch. Rad. News.* **35** (4), 44 (2022). doi: 10.1080/08940886.2022.2114716
14. C. Benmore et al., *Synch. Rad. News.* **35** (4), 28 (2022). doi: 10.1080/08940886.2022.2112500
15. E. Mosqueira-Rey et al., *Artif. Intell. Rev.* **56** (4), 3005 (2023). doi: 10.1007/s10462-022-10246-w
16. C. Wiethof and E. A. C. Bittner, ICIS (2021).
17. K. J. Craswell, *Two-Year Coll. Math. J.* **4** (3), 18 (1973). 10.2307/3026514
18. L. V. D. Maaten et al., *J. Mach. Learn. Res.* **10** (13), 66 (2009).
19. W. Jia et al., *Complex Intell. Syst.* **8** (3), 2663 (2022). doi: 10.1007/s40747-021-00637-x
20. A. Daffertshofer et al., *Clin. Biomech.* **19** (4), 415 (2004). doi: 10.1016/j.clinbiomech.2004.01.005
21. A. Maćkiewicz and W. Ratajczak, *Comput. Geosci.* **19** (3), 303 (1993). doi: 10.1016/0098-3004(93)90090-R
22. H. Abdi and L. J. Williams, *Wiley Interdiscip. Rev. Comput. Stat.* **2** (4), 433 (2010). doi: 10.1002/wics.101
23. D. D. Lee and H. S. Seung, *Nature.* **401** (6755), 788 (1999). doi: 10.1038/44565
24. D. Lee and H. S. Seung, *Adv. Neural Inform. Process. Syst.* **13**, 556 (2000).
25. Y. Wang and Y. Zhang, *IEEE Trans. Knowl. Data Eng.* **25** (6), 1336 (2013). doi: 10.1109/TKDE.2012.51
26. L. McInnes, J. Healy, and J. Melville, *arXiv preprint arXiv:1802.03426* (2018).
27. K. Krishna and M. N. Murty, *IEEE Trans. Syst. Man Cybernet. Part B (Cybernet.).* **29** (3), 433 (1999). doi: 10.1109/3477.764879
28. M. Ahmed, R. Seraj, and S. M. S. Islam, *Electronics.* **9** (8), 1295 (2020). doi: 10.3390/electronics9081295
29. F. Murtagh and P. Contreras, *Wiley Interdiscip. Rev. Data Min. Knowl. Disc.* **2** (1), 86 (2012). doi: 10.1002/widm.53
30. F. Nielsen, *Introduct. HPC MPI Data Sci.* (Springer, Cham, 2016), p. 195–211. doi: 10.1007/978-3-319-21903-5_8

31. W. H. L. Pinaya et al., *Machine Learning* (Academic Press, 2020), p. 193–208. doi: 10.1016/B978-0-12-815739-8.00011-0
32. U. Michelucci, *arXiv preprint arXiv:2201.03898* (2022).
33. D. Bank, N. Koenigstein, and R. Giryes, *Mach. Learn. Data Sci. Handbook: Data Min. Knowl. Disc. Handbook.*, (Springer, Cham, 2023), p. 353–374. doi: 10.1007/978-3-031-24628-9_16
34. P. Vincent et al., Proceedings of the 25th International Conference on Machine learning (2008), p. 1096–1103. doi: 10.1145/1390156.1390294
35. J. Clark and F. Provost, *Data Min. Knowl. Disc.* **33** (4), 871 (2019). doi: 10.1007/s10618-019-00616-4
36. S. Gyamerah et al., *Am. J. Elect. Comput. Eng.* **7** (2), 27 (2023). doi: 10.11648/j.ajece.20230702.12
37. J. Jumper et al., *Nature.* **596** (7873), 583 (2021). doi: 10.1038/s41586-021-03819-2
38. L. P. Kaelbling, M. L. Littman, and A. W. Moore, *Jair.* **4**, 237 (1996). doi: 10.1613/jair.301
39. Y. Li, *arXiv preprint arXiv:1701.07274* (2017).
40. C. Dong, C. C. Loy, and X. Tang, Computer vision–ECCV 2016: 14th European Conference (Springer¸ Amsterdam, The Netherlands, 2016), p. 391–407. doi: 10.1007/978-3-319-46475-6_25
41. Z. Pan et al., *IEEE Access.* **7**, 36322 (2019). doi: 10.1109/ACCESS.2019.2905015
42. C. M. Pham et al., arXiv preprint arXiv:2311.01449 (2024).
43. G. Shmueli and O. R. Koppius, *MIS Quarter.* **35** (3), 553–572 (2011). doi: 10.2307/23042796
44. M. Wankhade, A. Rao, and C. Kulkarni, *Artif. Intell. Rev.* **55** (7), 5731 (2022). doi:10.1007/s10462-022-10144-1
45. C. Yang et al., *Appl. Soft Comput.* **94**, 106483 (2020). doi: 10.1016/j.asoc.2020.106483
46. Y. He et al., *arXiv preprint arXiv:2405.19334* (2024).
47. C. M. Pham et al., *arXiv preprint arXiv:2311.01449* (2023).
48. M. E. J. Newman, *Phys. Rev. E.* **64** (1), 016131 (2001). doi:10.1103/PhysRevE.64.016131
49. S. C. Leemann et al., *Phys. Rev. Lett.* **123** (19), 194801 (2019). doi: 10.1103/PhysRevLett.123.194801
50. Y. Lu et al., *Nucl. Instrum. Methods Phys. Res, Sect. A.* **1050**, 168192 (2023). doi: 10.1016/j.nima.2023.168192
51. T. W. Morris et al., *Advances in computational methods for X-ray optics VI*; 126970B (2023). doi:10.1117/12.2677895
52. D. Allan et al., *Synch. Rad. News.* **32** (3), 19 (2019). doi: 10.1080/08940886.2019.1608121
53. G. Gunjala et al., *J. Synchrotron Rad.* **30** (1), 57 (2023). doi: 10.1107/S1600577522011080
54. L. Rebuffi et al., *Opt. Express.* **31** (13), 21264 (2023). doi: 10.1364/OE.48818 10.1364/OE.488189
55. C. N. Melton et al., *Mach. Learn. Sci. Technol.* **1** (4), 045015 (2020). doi: 10.1088/2632-2153/abab61
56. M. M. Noack et al., *Nat. Rev. Phys.* **3** (10), 685 (2021). doi: 10.1038/s42254-021-00345-y
57. D. M. Pelt and J. A. Sethian, *Proc. Natl. Acad. Sci. USA.* **115** (2), 254 (2018). doi: 10.1073/pnas.1715832114
58. T. Perciano et al., *Fibripy: a software environment for fiber analysis from 3d micro-computed tomography data* (2017). https://escholarship.org/uc/item/91n6h7t9
59. F. H. D. Araujo et al., *Expert Syst. Appl.* **109**, 35 (2018). doi: 10.1016/j.eswa.2018.05.015
60. Y. Huang et al., *Npj Comput. Mater.* **9** (1), 93 (2023). doi: 10.1038/s41524-023-01039-y
61. *MLExchange*. *Dash app for segmentation of high-resolution images* (2024). https://github.com/mlexchange/mlex_highres_segmentation.
62. E. J. Roberts et al., *J. Appl. Crystallogr.* **57** (2), 392 (2024). doi: 10.1107/S1600576724001390



**Dilworth Y. Parkinson, Tanny Chavez, Monika Choudhary, Damon English, Guanhua Hao, Thorsten Hellert, Simon C. Leemann, Slavomir Nemsak, Eli Rotenberg, Andrea L. Taylor, Andreas Scholl, Ashley A. White, Antoine Islegen-Wojdyla, Petrus H. Zwart, and Alexander Hexemer**

*Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA*