Check for updates

NIST Special Publication 1500-25
NIST SP 1500-25

Workshop Report on Autonomous Methodologies for Accelerating X-ray Measurements

Zachary T. Trautt
Brian L. DeCost
Howie Joress
Austin S. McDannald
A. Gilad Kusne
Francesca Tavazza
Thomas N. Blanton

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-25

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication 1500-25
NIST SP 1500-25

Workshop Report on Autonomous Methodologies for Accelerating X-ray Measurements

Zachary T. Trautt
Brian L. DeCost
Howie Joress
Austin S. McDannald
A. Gilad Kusne
Francesca Tavazza
Data and AI-Driven Materials Science Group
Material Measurement Laboratory

Thomas N. Blanton
International Centre for Diffraction Data

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-25

November 2024

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Publication History
Approved by the NIST Editorial Review Board on 2024-09-16

How to cite this NIST Technical Series Publication:
Trautt ZT, DeCost BL, Joress HL, McDannald AS, Kusne AG, Tavazza F, Blanton TN (2024) Workshop Report on Autonomous Methodologies for Accelerating X-ray Measurements. (National Institute of Standards and Technology, Gaithersburg, MD), NIST SP 1500-25. https://doi.org/10.6028/NIST.SP.1500-25

Brian L. DeCost, Howie Joress, and Austin S. McDannald contributed equally to this report.

Author ORCID IDs
Zachary T. Trautt: 0000-0001-5929-0354
Brian L. DeCost: 0000-0002-3767-926X
Howie Joress: 0000-0002-3459-5888
Austin S. McDannald: 0000-0002-6552-2972
A. Gilad Kusne: 0000-0001-8904-2087
Francesca Tavazza: 0000-0002-5602-180X
Thomas N. Blanton: 0000-0002-9506-7237

Contact Information
zachary.trautt@nist.gov

Abstract

The National Institute of Standards and Technology and the International Centre for Diffraction Data co-hosted a workshop on 17-18 October 2023 to identify and prioritize the goals, challenges, and opportunities for critical and emerging technology needs within industry, with an emphasis on leveraging artificial intelligence, data-driven methodologies, and high-throughput and automated workflows for accelerating x-ray-based structural analysis for materials development and manufacturing. Participants, predominantly from industry, gathered in-person at ICDD® headquarters in Newtown Square, Pennsylvania. The findings of this workshop report provide critical input for strategic planning and the convening activities serve as a kickoff for future public-private cooperation.

Keywords

X-ray diffraction, X-ray analysis, Artificial Intelligence, Machine Learning, Autonomous Laboratories, Materials Synthesis and Characterization, Robotics

Table of Contents

List of Tables

List of Figures

Author Contributions

Zachary T. Trautt, Brian L. DeCost, Howie Joress, and Austin S. McDannald: Conceptualization, workshop planning, workshop facilitation, writing, reviewing, and editing. **A. Gilad Kusne**: Reviewing and editing. **Francesca Tavazza**: Supervision, workshop planning, workshop participation, writing, reviewing, and editing. **Thomas N. Blanton**: Workshop planning, workshop participation, community engagement, reviewing, and editing.

Executive Summary

> *"[X-ray analysis] is a solved problem, depending on your definition of* ***solved*** *and your definition of the* ***problem****."*
>
> -Workshop Attendee

The National Institute of Standards and Technology and the International Centre for Diffraction Data co-hosted a workshop on 17-18 October 2023. Its purpose was to identify and prioritize the goals, challenges, and opportunities for critical and emerging technology needs within industry, with an emphasis on leveraging artificial intelligence (AI), data-driven methodologies, and high-throughput and automated workflows for accelerating X-ray-based structural analysis for materials development and manufacturing. Participants, predominantly from industry, gathered in person at ICDD® headquarters in Newtown Square, Pennsylvania. The findings of the workshop provide critical input for strategic planning and the convening activities serve as a kickoff for future public-private cooperation.

The rapid development of data-driven materials methodologies and automated materials experimentation (predominately in academia and national labs) presents a significant opportunity for industry. However, the maturation and adoption of this technology is limited by several challenges, which were identified and prioritized during the workshop. The workshop methodology gathered input from participants on notecards, which were later prioritized by participants with stickers. The facilitated sessions generated 322 cards, which consisted of 83 goals, 164 challenges, and 75 possible solutions.

Summarized Findings

- **Data and Metadata** — Data and metadata are key enablers of data-driven methodologies. Workshop participants authored a number of cards, which can be categorized within the well-defined concept of FAIR (findable, accessible, interoperable, reusable) data principles or within the emerging concept of AI-ready data. Within the category of FAIR data, workshop participants highlighted the need to revive conversations about file formats for diffraction data. Furthermore, there were considerable discussions surrounding machine actionability and quality of data and metadata. Within the category of AI-ready data, we first make a clear distinction from FAIR data. Furthermore, a major theme was the need to publish more benchmark diffraction data, with sufficient variety, to train data-driven models. Workshop participants emphasized null data and data of varying quality. Workshop participants emphasized that data traditionally unworthy of publication is important for training data-driven models.
    - **FAIR Data Principles** — Achieving the aspirational goals within the FAIR data principles is: (i) critical to the seamless adoption of data-driven methodologies, (ii) a long-term challenge within the materials community, and (iii) often perceived by researchers to be a "costly distraction" that diverts time and effort from productive research. Workshop participants identified a clear need for the structural analysis community to de-

velop and adopt standardized file formats that better support diffractograms and metadata, supporting phase identification and quantitative analysis. Other opportunities were also identified in relation to the findability and interoperability of models and tools, which warrant consideration.

  - **AI-Ready Data** — Data-driven methodologies such as machine learning require systematic, fit-for-purpose data for the development and training of new algorithms and models. The past centuries of scientific research have resulted in the publication of scholarly data that was novel or interesting enough to warrant a peer-reviewed publication. However, data from "failed" or null experiments, non-ideal data (e.g., poor resolution), or data from laboratory mistakes (e.g., poor sample preparation or misalignment) is rarely published, which makes it impossible to develop robust algorithms and models. Workshop participants highlighted the need for more data that represents real-world complexity to be generated and well annotated with experimental metadata.

- **Physical Infrastructure** — An emphasis of the workshop was next-generation hardware needs. Workshop participants noted the need for increased automation for sample preparation and measurement, which is a task often performed by humans. Workshop participants also indicated the need to have the capability to better detect when problems occur with sample preparation or operation of the instrument. Workshop participants noted cases of needing to perform X-ray-based analysis under non-ambient conditions, which may be inaccessible to current hardware. Workshop participants indicated the need for programmatic access to equipment such that they can be driven by independent computer systems. Finally, research and development into standards for modular equipment may be a promising opportunity for public-private cooperation.

- **Algorithm and Model Development** — Data-driven algorithms and models are at the heart of data-driven science and autonomous laboratories. Workshop participants discussed a broad range of issues related to the development of autonomous laboratories and also provided a focused perspective on phase identification and quantitative analysis. Workshop participants also reaffirmed the importance of uncertainty quantification.

  - **Phase Identification and Quantitative Analysis** — Workshop participants identified a specific need for algorithms for phase identification and quantitative phase analysis. Existing innovations in data-driven methodologies tend to be applied to simple and idealized data and lack the depth to be later applied to complex (real-world) problems. With an in-depth focus on phase identification and quantitative analysis during the workshop, a number of specific opportunities emerged including supporting complexity in the physical samples, complexity in the measurement methodology such as low data quality from benchtop systems, multimodal data streams, and high throughput methodologies.

  - **Broad AI for Materials Development** — Rapid innovation in algorithm and model development, broadly speaking, presents a significant opportunity for industry to introduce targeted autonomous solutions for repetitive problems that typically require human expertise and experience. These technologies can be developed and adopted most efficiently if industry and other end-users prioritizes which tasks ML methods will have the largest impact on their productivity. These can include methods for combining multimodal data streams, automated experiment planning, metadata assessment, and automated selection of analysis tools.

- **Uncertainty Quantification (UQ)** — As ML methods are intrinsically statistical in nature, workshop participants noted that UQ is of critical importance if data-driven methodologies are to be implemented in robust and reliable production applications. Strengthening UQ evaluation requires accurate evaluation of standard statistical quantities, like mean average error (MAE), as well as the determination of the confidence of each individual ML prediction (error bounds), as it is individual predictions that are used in industrial applications. Obviously, no amount of model UQ can reveal or mitigate bad quality in the training data, as systematic errors in the generation of the data set will propagate through a machine learning model. Careful curation of the training and testing dataset is, therefore, a crucial step in developing trustworthy ML models, and there is a need for developing an accuracy measure to communicate the quality of such a dataset.

- **Community Engagement** — Workshop participants authored a number of cards centered around community engagement. One major theme was workforce development, with an emphasis on cross-disciplinary education and considering both the existing workforce and the next-generation workforce. Other cards discussed culture change, intellectual property, and economic considerations. Data-driven methodologies are ultimately implemented as software, and workshop participants highlighted a number of issues related to the sustainability of both open-source and proprietary software. Finally, workshop participants highlighted the need for new and ongoing convening and coordination efforts.
  - **Workforce Development** — Workforce development was identified as an important need where the depth of knowledge in both experimental methodology and data-driven methodologies is critical for success. It is important to have a depth of understanding of the physical processes and geometries of the diffraction experiments to enable correct interpretation of measurement results. However, it is becoming increasingly important to have additional depth of knowledge in the machine learning algorithms to understand their implications and biases for use as analysis tools. Having a workforce trained in both disciplines is needed for the full realization of machine learning and autonomous workflows of diffraction experiments. Even non-expert practitioners need to be educated in framing their problems as ML tasks and in identifying and deploying existing solutions.
  - **Software Development and Sustainability** — Workshop participants highlighted a number of issues related to the development and sustainability of software. Much of the algorithm development is occurring in an academic research environment and the pathways to production software are plagued with many challenges. Research software is highly decentralized and often poorly documented, making it challenging for industrial users to discover and reuse. Further, industrial developers have little incentive to assume the responsibility of maintenance and support for open-source software they might otherwise incorporate into or enable integration with their products.

- **Culture Change, Intellectual Property, and Economic Considerations** — The economic and intellectual property landscape of the community makes the development of a sustainable ecosystem of interoperable solutions challenging. The use of intellectual property rights associated with software and data formats generated by commercial instrument vendors makes implementation of these tools and data into autonomous pipelines challenging. Conversely, end-users are reluctant to share their proprietary data to improve analysis models. To overcome these barriers workshop participants highlighted the need for a clear value proposition for instrument vendors and other service providers to move towards more interoperable and open models. Workshop participants highlighted a divide between industry and academia in the incentives and constraints around adopting new methodologies and the publication of data.
- **Convening and Coordination** — Workshop participants highlighted the need for new convening and coordinating efforts, such as working groups and consortia. Expanded participation in working groups will be an important mechanism to make progress on specific opportunities identified during the workshop.

Future Desired Outcomes

1. **FAIRness of XRD and other Experimental Data** — The community achieves wide adoption of a consensus-based specification for data and metadata. This consensus-based specification provides a rich plurality of metadata fields to enable wide reuse of data and metadata. This consensus-based specification enjoys wide adoption across instrument vendors and software providers.

2. **Programmatic Control and Administration of XRD and other Experimental Equipment** — There is wide adoption of a consensus-based specification for protocols and semantics for the programmatic control and administration of experimental instruments. This consensus-based specification provides an extendable mechanism for instruments to enumerate and describe the available operations and parameters for controlling and administering the instrument. This consensus-based specification enjoys wide adoption across instrument vendors and software providers.

3. **Comprehensive Datasets for Phase Identification and Quantitative Analysis** — Available data for training data-driven methodologies covers far more real-world use cases. For example, labeled data would cover situations of diverse measurement quality, sample preparation quality, and diverse chemistry and processing conditions.

4. **Robust Tools for Phase Identification, Quantitative Analysis, and Autonomous Laboratories** — Operators of autonomous laboratories and other laboratory equipment have the ability to leverage a plethora of robust and well-documented tools for autonomous laboratories, phase identification, quantitative analysis, and other laboratory activities. Open-source and for-profit providers coexist in delivering and supporting high-quality software.

5. **Vibrant Marketplace for Autonomous Laboratory Equipment and Services** — It is possible to affordably procure a wide variety of equipment to build an autonomous laboratory. There are a number of industry-lead standards that enable diverse components from different vendors to easily plug and play.

6. **Workforce Equipped for the Autonomous Laboratory** — Next-generation and current-generation workforce have a plethora of mechanisms to acquire skills and expertise in leveraging autonomous labs and data-driven methodologies in the laboratory.

A Call to Action

Materials and manufacturing are essential sectors of the U.S. economy. This report presents a clear consensus on specific challenges and opportunities within the emerging field of autonomous materials science and within the established field of X-ray-based methodologies. The community has an opportunity to unite around this consensus, develop action plans around desired end states, and ultimately push the US and its allies as the world leaders in autonomous methodologies. We hope that the community finds this report informative and valuable. Furthermore, we hope that this report stimulates action within the community, with a specific emphasis on industry stakeholders building out the foundational cyber-physical infrastructure components of autonomous laboratories and industrial stakeholders adopting them within their organizations.

1. Introduction

The field of materials science and engineering revolves around the discovery and manipulation of the processing-structure-property-performance relationships. X-ray-based scattering methods are a common and powerful approach for characterizing the structure of materials on the atomic and near-atomic scale. Technological advances in AI and automation have great promise for accelerating the field of materials science. Yet, to reach their full potential, these techniques must be specialized for the materials science domain. This work has started in the field as a whole, but given the importance of X-ray methods to the field, there is still much work to be done on adapting these methods for X-ray-based structural analysis.

Materials science and engineering has undergone rapid innovation over the past decade as a result of community-wide efforts such as ICME[1] and MGI[2]. The rapid adoption of modeling and simulation foreshadowed the more recent rapid adoption of data-driven methodologies in materials science and engineering such as: ML[3], LLMs[4], and autonomous experimentation (AE) systems[5–8]. The rapid pace of technological innovation has predominantly occurred in the setting of academia or national laboratories. This innovation presents a significant opportunity for industry. However, the maturation of this technology likely faces many challenges, which can be addressed by organizations that support materials and manufacturing industries.

1.1. NIST and ICDD Role

ICDD and NIST are both industry-serving organizations. ICDD is a non-profit scientific organization dedicated to collecting, editing, publishing, and distributing powder diffraction data for the identification of materials. NIST is a U.S. Government agency that promotes U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life. NIST is a larger organization with many different groups and this workshop was co-organized by the Data and AI-Driven Materials Science Group. The Data and AI-Driven Materials Science Group[9] was established in the Summer of 2023 as a successor to the former Materials for Energy and Sustainable Development Group, which had an extensive publication history on related topics including: HTE[10–17], ML[18–26], AE[27–29], UQ[19, 30], and MDI[31–34]. The Data and AI-Driven Materials Science Group develops methods, algorithms, data, and tools, to accelerate the discovery, development, commercialization, and circularity of industrially-relevant materials. The group enables the trustworthy use of data and AI-driven methodologies within both experimental and computational materials science and engineering workflows. This workshop was organized to help inform the strategic directions of both ICDD and the Data and AI-Driven Materials Science Group as well as the group's parents in the organizational structure of NIST, which includes the Materials Measurement Science Division, and the Material Measurement Laboratory.

2. Workshop Methodology

The primary objective of the workshop was to identify and prioritize goals, challenges, and possible solutions within the intersecting topics of autonomous methodologies and X-ray-based structural analysis. The workshop was divided into four, half-day units, with each half-day unit including both invited talks and a facilitated discussion. The facilitated session was structured to gather and organize input from participants. Input was gathered via adhesive note cards on topic-focused easel boards and 8 easel boards were used over the duration of the workshop and had the following titles:

1. Goals: Self-Driving Laboratory
2. Goals: XRD of the Future
3. Challenges: Sample Characterization/Analysis
4. Solutions: Sample Characterization/Analysis
5. Challenges: AI-Driven Decisions/Experiments
6. Solutions: AI-Driven Decisions/Experiments
7. Challenges: Next-Gen Hardware/Methods
8. Solutions: Next-Gen Hardware/Methods

Each easel board had a built-in actor/role-based categorization system to help promote sufficient input for each type of actor/role within a laboratory where input was requested. These actors/roles are as follows:

1. Human Scientists, Engineers, Etc.
2. Synthesis Tools plus AI/ML
3. XRD Tools plus AI/ML
4. Other Tools plus AI/ML
5. Data, Protocols, Automation
6. AI/ML-Driven Experiments

There was an additional “Parking Lot” category for adhesive cards that did not fit into the selected categories. These categories can be viewed at the top of Figure 1. Participants split into two independent groups with identical boards for logistical reasons. NIST staff facilitated a discussion that would evoke responses from participants. Participants would either write their own responses on the adhesive note card, or NIST staff would write down verbal responses for them and select a category. Approximately 30 minutes was dedicated to note card authoring before participants would rank the cards within their group. Each participant was given the same number of adhesive circular stickers, which they may place upon note cards or note card clusters. Participants were not required to use all of their stickers. An example completed easel board is illustrated in Figure 1.

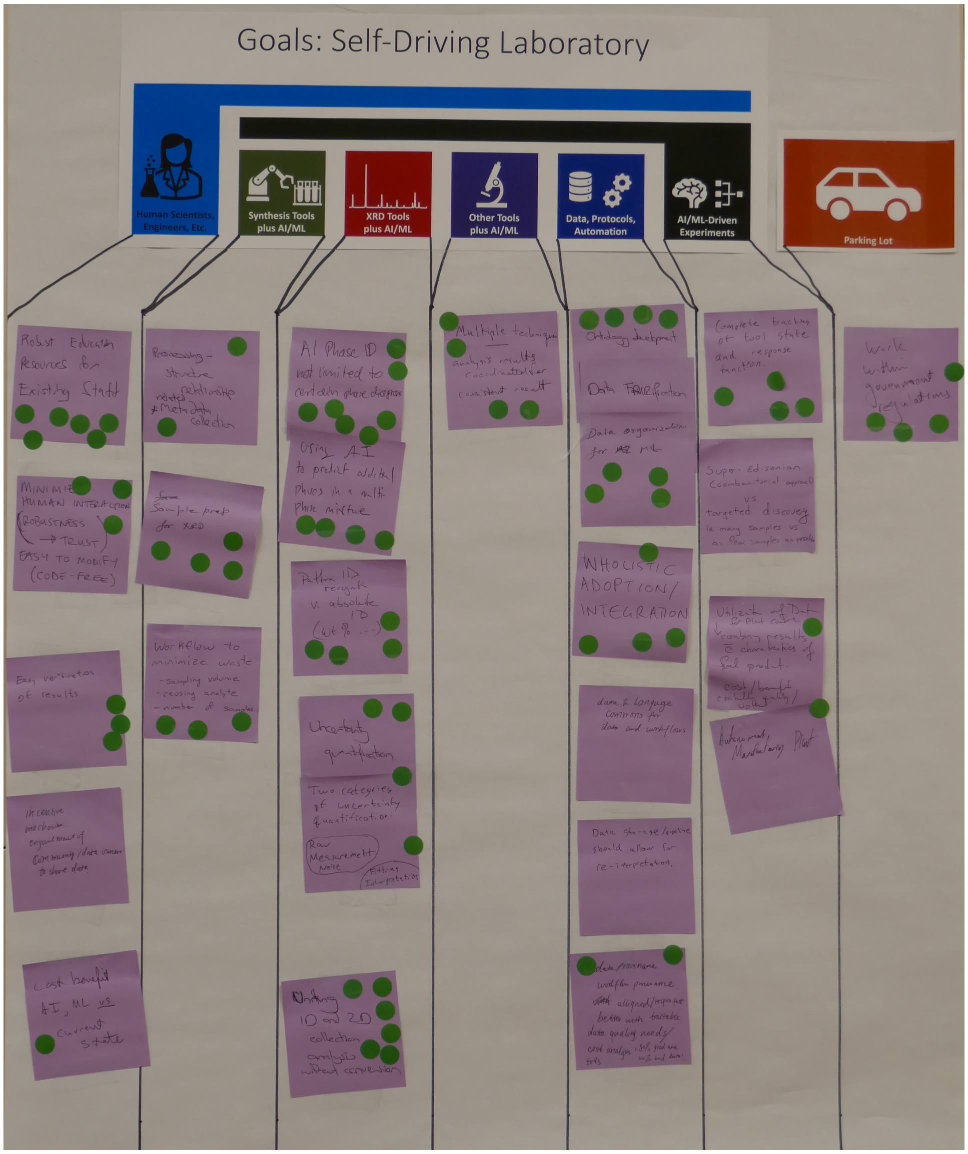

Fig. 1. An example of a completed facilitated discussion board. Adhesive cards were authored by participants and placed within a categorized column at the time of the discussion. After approximately 30 minutes of card authoring discussion, each participant was given an equal number of circular stickers, which they may place on cards based on priority.

3. Workshop Findings

As discussed in the previous section, NIST facilitators led discussions in two groups completing a total of 8 easel boards in each group (16 total boards). This resulted in the authoring of 322 individual adhesive note cards. Easel boards had the title prefixes of Goals, Challenges, and Possible Solutions. The distribution of cards is as follows:

- 83 Goals
- 164 Challenges
- 75 Possible Solutions

It is important to note that this distribution is based on the boards being used when the card was authored. Thus, it is possible that some cards are miscategorized. The raw images of the easel board and the digitized data is available as a data publication as listed in Appendix B. The note card data listed in the report may have undergone minor editorial revisions for spelling and clarity.

3.1. High Priority Goals, Challenges, and Possible Solutions

As discussed in the previous section, workshop participants prioritized cards with adhesive dots. The distribution of votes is shown in Figure 2. The most upvoted 47 cards (15 %) received 5 votes or higher and the remaining 275 cards (85 %) received 4 or fewer votes. The most upvoted cards are listed in Tables 1-3. These upvoted cards help demonstrate a consensus regarding the existing challenges and potential paths forward. These upvoted cards help inform the topics discussed in the following subsections. It is noted that cards with similar answers may have caused votes to be diluted in some instances. These topics were selected based on human expert interpretation and clustering of the cards created during the workshop.

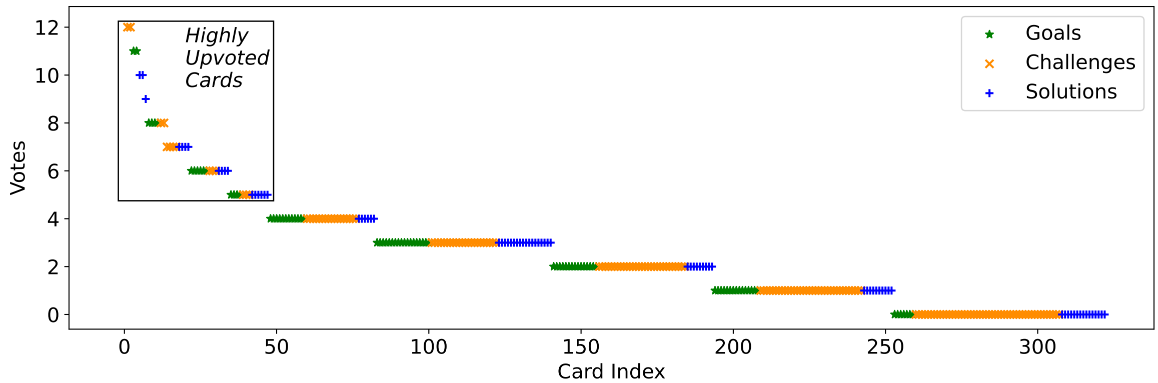

Fig. 2. Illustration of the distribution of the number of cards vs. the votes per card for the 322 cards collected. The boxed region of highly up-voted cards highlights likely consensus perspectives that are shared within the community. Cards with 5 or more votes are shown in Tables 1-3.

Table 1. Goal note cards with 5 or more votes.

Votes	Content
11	AI Phase ID not limited to certain phase diagrams
11	Using AI to predict additional phases in a multi-phase mixture
8	AI vs. existing methods? Are they known and /or already exploited?
8	Ontology development
8	Data organization for ML
8	Data FAIRification
6	Cross-disciplinary education: Statistics, experimental design, materials, physics, economics
6	Robust education for existing staff
6	Should not be a black box. Even with automation, expertise is needed.
6	How to make automated refinement coding resistant to software changes and updates or changing industry standards / updates
6	Automated analysis has a long way to go to robustness for challenging problems, but it could accelerate a lot
6	Uniting 1D and 2D collection analysis without conversion
5	Data transformation and preconditions scheme to improve ML interpretation
5	Uniform metadata file across all diffractometer manufacturers
5	Can we layer multiple complementary techniques to help QPA / XRD automation on?
5	Raw data + metadata should be published with results

Table 2. Challenge note cards with 5 or more votes.

Votes	Content
12	CIF is problematic for many use cases
12	No complete XRD file format exists
8	Need for new community members to learn what pitfalls to look out for – these are in the literature but hard to find without knowing already
8	Education! Users / practitioners need to know: – what factors matter – why they are important – where to learn more to be effective with next-generation tools (and current ones).
7	Need to build new technologies on existing programs rather than write more new programs
7	Few materials experts in general. Even fewer with knowledge of AI / ML / automation
7	Huge lack of knowledge XRPD + AI
7	Access to training datasets
6	Lack of Metadata Quality Score
6	Too many programming languages
6	Reproducible computation: – deployment – version control – metadata provenance – backward compatibility
5	How to maximize current generation methods?
5	What quality standards should apply to data used to train ML algorithms
5	Lack of interoperability (that can evolve)

Table 3. Potential Solution note cards with 5 or more votes.

Votes	Content
10	Sustainability Models for Open-Source
10	Consortia for databases + models
9	IP (intellectual property) rights are a major blocker for industry collaboration with universities.
7	XRD + XRF
7	AI or ML to identify poor sample prep or other artifacts
7	Generative AI for data search and queries
7	Exploratory querying/match making how to help users finding/uniting questions
6	Encourage funding for translational research collaborating with industry and academia
6	Quantitative data quality assessments
6	Push on IUCR & ICDD on past efforts PowdCIF
6	Limited personnel with necessary expertise: – work on career pipeline – collaborate routinely with industry
5	Academia lacks ideas about business basics, and a sense of what topics are of general interest.
5	User Facility Model for Autonomous Systems
5	Independent organization (ICDD) Champions a new Data File (with a Metadata Structure that can Evolve) Replace CIF
5	Simulate bad data using good data. Test ML on the bad data also.
5	Publish more poor/null data
5	Adding artifacts to high quality data – Synchrotron -> benchtop – Long count -> short times – Re-binning high fidelity -> low

As shown in Tables 2-3, the highest-ranking goals focus on the application of data-driven methodologies to realistic problems experienced in the field, where many published papers focus on simple or idealized problems. The highest-ranking challenges focused on issues related to poor data and system interoperability within experimental materials synthesis and characterization. The highest-ranking potential solutions focused on finding a sustainable model for open-source software and suggested a consortium for pressing problems. The following subsections discuss workshop findings through the perspective of Data, Physical Infrastructure, Algorithm and Model Development, and Community Engagement.

3.2. Data and Metadata

Data and metadata are key enablers of data-driven methodologies. Workshop participants authored a number of cards, which can be categorized within the well-defined concept of FAIR (findable, accessible, interoperable, reusable) data principles[35] or within the emerging concept of AI-ready data. Within the category of FAIR data, workshop participants highlighted the need to revive conversations about file formats for diffraction data. Furthermore, there were considerable discussions surrounding machine actionability and quality of data and metadata. Within the category of AI-ready data, we first make a clear distinction from FAIR data. Furthermore, a major theme was the need to publish more benchmark diffraction data, with sufficient variety, to train data-driven models. Workshop participants emphasized null data and data of varying quality. Workshop participants emphasized that data traditionally unworthy of publication is important for training data-driven models.

3.2.1. FAIR Data Principles

Achieving the aspirational goals within the FAIR data principles is: (i) critical to the seamless adoption of data-driven methodologies, (ii) a long-term challenge within the materials community, and (iii) often perceived by researchers to be a “costly distraction” that diverts time and effort from productive research. The FAIR data principles, which were published in 2016 have rapidly become a framework for continuous improvement in other areas beyond data publications. For example, the FAIR data principles have been considered for research software[36], research hardware[37], and machine learning[38]. Workshop participants identified a clear need for the structural analysis community to develop and adopt standardized file formats that better support diffractograms and metadata, supporting phase identification and quantitative analysis. Other opportunities were also identified in relation to the findability and interoperability of models and tools, which warrant consideration. The cards within this category are shown in Tables 4-6.

Table 4. Goal note cards for FAIR Data Principles

Votes	Content
8	Data FAIRification
8	Ontology development
5	Uniform metadata file across all diffractometer manufacturers
5	Raw data + metadata should be published with results
3	Streaming Curation: ETL Pipeline between data resources
3	Open-source data cleaning / normalization software – depends on standard database APIs
2	Adoption of FAIR principles for data storage – community standards API for data access
2	Processing-Structure relationship related + metadata collection
1	pdCIF standardizes powder diffraction data & analysis results & metadata. (No current work further dev work done in U.S.)
0	Data & language common for data and workflows

Table 5. Challenge note cards for FAIR Data Principles

Votes	Content
12	No complete XRD file format exists
12	CIF is problematic for many use cases
6	Lack of Metadata Quality Score
6	Reproducible computation: – deployment – version control – metadata provenance – backward compatibility
5	Lack of interoperability (that can evolve)
4	Data is too big to move to compute
3	Journals don't require sufficient metadata
3	Data curation problem + licensing
2	Too many data protocols
2	Hordes of dark data
1	Vendors have different file formats
1	Data Management: Standards – premature, if possible. Adoption is difficult, as is adoption that does not break standards.
1	Schema, Ontology, Vocabulary: Searchable, extensible, shared schema/ontology, and vocabulary – something that can be readily incorporated into commercial and open-source software
1	Search – The macroscopic property associated to the problem to solve is important
0	API software analysis paralysis
0	Technical Barriers: Disparity in computational resources (e.g., AWS, Azure, GCP, HCP, hybrid)

Table 6. Potential Solution for FAIR Data Principles

Votes	Content
6	Push on IUCR & ICDD on past efforts PowdCIF
5	Independent organization (ICDD) Champions a new Data File (with a Metadata Structure that can Evolve) Replace CIF
3	Awesome List of Papers
2	Repository of best practices already implemented

Data Formats for Diffraction Data — The most upvoted cards for the entire workshop indicated that community standard sufficient for X-ray diffraction data and metadata does not exist and that CIF[39] is problematic for many use cases. However, discussions with workshop participants didn't provide detailed insight into the actual problem. However, it wasn't clear if there is a technical problem with the CIF format or if there is a cultural problem in understanding and adopting an existing specification. Workshop participants upvoted a card suggesting reinvigorated community efforts surrounding pdCIF[40]. This report makes no statement on the validity of existing file formats such as CIF, NeXus[41], or others, but there exists a clear need to convene interested parties to better understand the issues.

Machine Actionable Data — Workshop participants once again highlighted the need for common machine-actionable data formats, vocabularies/ontologies, as well as improved or automated metadata collection. This represents a long-standing issue within the community. While community consensus has emerged for narrow classes of data with efforts like OPTIMADE[42], many challenges have persisted for the border materials data ecosystem. In fact, cards collected during this workshop describing challenges with interoperability and machine actionability of data are consistent with statements made in 2008 within the National Academies report on Integrated Computational Engineering[1], which mentioned widely varying levels of maturity for materials databases. The persistent nature of these issues suggests the existence of a sociotechnical gap that has not been adequately addressed with past incentives and community efforts. Workshop participants highlighted potential roles for IUCR and ICDD in charting a path forward.

Metadata Quality — Workshop participants highlighted the lack of a metadata quality score and indicated a goal of having uniform metadata across instrument vendors. Workshop participants emphasized that data and metadata should be published with derived data and manuscripts. However, most journals don't place significant requirements on the publication of data and metadata.

Search Tools — Workshop participants highlighted the need for improved search capabilities for the broad ecosystem of resources for data-driven materials science and engineering. Much focus has been placed on search capabilities for data (which are still underdeveloped), but significant opportunities for improvement exist for finding the knowledge contained in papers and books, models and algorithms contained in research papers and code, and tools and utilities in software packages. Some potential solutions include a repository, clearinghouse, or ”awesome list” (e.g., Awesome Materials Informatics[43]) of resources.

Prior Workshop Reports — Many of the issues discussed in this 2023 workshop were also discussed in detail in the 2017 Materials Data Infrastructure Study[44] organized by The Minerals, Metals & Materials Society. That study identified 36 challenges, which were mapped into four quadrants based on probability of success (y-axis) and potential impact (x-axis), which was intended to help prioritize the challenges. This report defined the four quadrants as follows:

- Quadrant I (Higher Potential Impact, Higher Probability for Success): these "no-brainers" are expected to be particularly appealing to most readers of this report since they have the highest likelihood for strongly impacting the materials data infrastructure.
- Quadrant II (Lower Potential Impact, Higher Probability for Success): this "low-hanging fruit" may not have as high an impact as quadrant I, but the likelihood of success of overcoming these barriers is expected to make them a relatively high priority in developing the materials data infrastructure.
- Quadrant III (Lower Potential Impact, Lower Probability for Success): these "tough sells" are expected to be some of the lowest priority challenges, but would nonetheless have a worthwhile impact on the MDI if successfully addressed.
- Quadrant IV (Higher Potential Impact, Lower Probability for Success): these "heavy hitters" have a lower likelihood of success, but they can have a notable impact on the MDI if challenges can be overcome.

Some of the topics identified in our workshop tend to align with challenges from the 2017 report that were classified as having a lower probability of success. For example, a "lack of robust APIs of connected systems and instrumentation" was listed in Quadrant III and a "lack of developed, agreed-upon ontologies for materials domain" was listed in Quadrant IV. This highlights that a changing landscape towards automated and autonomous systems has transformed the "tough sells" and "heavy hitters" into major bottlenecks limiting progress within the community.

3.2.2. AI-Ready Data

Data-driven methodologies such as machine learning require systematic, fit-for-purpose data for the development and training of new algorithms and models. The past centuries of scientific research have resulted in the publication of scholarly data that was novel or interesting enough to warrant a peer-reviewed publication. However, data from "failed" or null experiments, non-ideal data (e.g., poor resolution), or data from laboratory mistakes (e.g., poor sample preparation or misalignment) is rarely published, which makes it impossible to develop robust algorithms and models. Workshop participants highlighted the need for more data that represents real-world complexity to be generated and well annotated with experimental metadata. Tables 7-9 list Goal, Challenge, and Potential Solution cards that are relevant to an overarching theme of AI-ready data. While this term was not used in the workshop, it is rapidly gaining mainstream use. Thus, we attempt to conceptualize workshop input under the umbrella of AI-Ready Data.

Table 7. Goal note cards for AI-Ready Data

Votes	Content
8	Data organization for ML
3	Development of digital twins for automated & semi-automated laboratories
2	Data provenance workflow provenance aligned / respond fast better with traceable data quality needs / cost analysis tools 80 % good data 20 % bad data

Table 8. Challenge note cards for AI-Ready Data

Votes	Content
7	Access to training datasets
4	Training/testing data that resembles real life data
4	Dealing with unbalanced datasets
4	Lab quality data, limited time, etc.
1	Characterization challenge – ill-defined requirements that can be tested in application tests only
0	Real benchmark data / SRMs for method validation
0	Training data should account for different instrument configurations, measurement parameters, sample prep, etc.

Table 9. Potential Solution for AI-Ready Data

Votes	Content
5	Adding artifacts to high quality data – Synchrotron -> benchtop – Long count -> short times – Re-binning high fidelity -> low
5	Simulate bad data using good data. Test ML on the bad data also.
5	Publish more poor/null data
3	Challenge Dataset
3	Round Robin Studies
3	Round Robin on single dataset just testing the analysis of the measurement
3	Defining/Establishing a ground truth
2	Troubling datasets with known answers. Mixtures of materials amorphous organics
1	One Standard Dataset + Materials for each industry – known contaminates
1	Reference Materials for AI/ML
0	Different Datasets for Different areas
0	Internal Sample Variation?
0	See generative AI card
0	Synthetic Slag (e.g.,)

What is AI-Ready Data — Many of the cards in Tables 7-9 point to a desire to reduce the burdens associated with identifying, accessing, structuring, and cleaning a balanced dataset that is capable of addressing the problem at hand using data-driven methodologies. For the context of this report, we consider concepts associated with "AI-Ready" to be the concepts that remain after data has achieved high FAIR maturity[45, 46]. It is important to note that the FAIR data principles have already described the concept of machine actionable “to indicate a continuum of possible states wherein a digital object provides increasingly more detailed information to an autonomously-acting, computational data explorer”[35]. In this report, we define ”AI-Ready” to describe the overall suitability of a dataset to support the parameterization, training, and evaluation of computational agents with robust and trustable performance within the scope of their intended mission. We emphasize that AI-readiness is not an intrinsic property of the data, but rather is a context-specific property of the data shaped by the specific AI task.

Opportunities to Publish More Data — Workshop participants signaled a strong desire to stimulate the publication of additional data and highlighted a number of mechanisms to achieve such an objective. For example, workshop participants suggested round-robin studies and adding simulated features to known data sources. Workshop participants also indicated an interest in challenge datasets.

Null Data — Workshop participants noted that the discipline of materials science and engineering has a systemic problem of unbalanced and biased data publication, where only ”interesting” data is published (other causes of unbalanced data are described below in the section ”Unbalanced Data”). One card mentioned the need to publish more “poor/null data”, which was also mentioned in the report of a 2017 DOE Workshop on Artificial Intelligence Applied to Materials Discovery and Design[47]. One context for null data is for autonomizing instrument configuration. In order to develop ML algorithms for tasks such as detecting when samples are misaligned or instrument configuration is otherwise incorrect, there would need to be a database that included examples of those results. Another context for null data is in material synthesis. If a synthesis recipe is attempted for a particular material and the XRD pattern of the result shows something other than the target material, that measurement is still useful. Such XRD patterns, when contextualized with the details of the experiment (e.g. the attempted recipe), are valuable data for training ML models to predict synthesis routes. In general, in order to accurately make predictions ML models need both positive and negative examples: when the recipe worked and when it did not, the exciting data and the null data. An ML model shown only examples when the recipe worked will never learn when the recipe doesn’t work.

Data of Diverse Quality and Context — Workshop participants noted there is a critical need to increase the data available to train data-driven models, if they are to be used in real-world applications, which can include diverse quality, multiple phases, material defects, etc. Workshop participants noted that some of these problems could be addressed with more robustly representative synthetic data.

Many ML workflows analyzing diffraction data start with the assumption that the training dataset is of high quality. This typically implies accurate annotations for experimentally collected training data, and that synthetic datasets capture a realistic range of the relevant measurement conditions: representative signal-to-noise ratios, resolutions, background contributions, peak shapes, and distributions of sample parameters. Often, simulation-based datasets focus on the very high quality data: low signal-to-noise, high resolution, and low background data from a highly monochromatic source, with a well-prepared sample measured on a recently calibrated and aligned instrument. While this high data quality is ideal, especially if the task is to extract detailed, high-precision quantitative descriptions of the phases (e.g., grain size, strain, phase fraction), it is not representative of the data used for other tasks. Most diffraction measurements do not meet all of those criteria. A lab might only have very limited time on benchtop diffractometers to measure very complex samples. Often the purpose of performing diffraction measurements is to identify all the phases present in a sample, rather than determining a detailed quantitative description. The phase ID tasks can be very challenging with samples that might consist of powders with grains having a wide variety of sizes and non-spherical shapes and include many phases, some of which might be amorphous. A ML-based approach could be used to address these types of challenging phase ID tasks – provided the algorithms have a database to train on that is representative of the realistic data quality. A ML model that has only been trained on synchrotron data will likely not perform well on data from benchtop diffractometers, and vice versa. Without a dataset that reflects the realistic diversity of data quality, any ML tools will not be able to generalize to these common scenarios seen in many labs. More details about the ways in which the data could be diverse, and the type of tasks that ML models trained on such data could be developed for can be found in the discussion in Section 3.4.1.

Unbalanced Data — Once a particular task is identified, then the suitability of a dataset for that task can be evaluated. One factor is the balance of the dataset. Much of the crystallography data sets are clustered around materials that were interesting to study (for one reason or another), rather than evenly distributed across material systems. This type of clustered data is a form of dataset imbalance that can cause issues with many ML workflows. For example, if the goal is to distinguish two classes, if the dataset has very few examples of the second class it will be difficult for any ML algorithms to learn the appropriate distinguishing features from that data. In this case, accuracy can be a misleading evaluation metric. Consider a dataset where 98 % of the examples belong to class 1. A naive model could achieve 98 % accuracy by predicting class 1 unconditionally, completely ignoring the input data. Metrics like the mean average precision try to account for class imbalance by weighing the performance of the model on each class relative to the size of that class. To deal with class imbalance, a dataset might be sub-sampled or pruned to make the class distribution more balanced. Alternatively, oversampling, data augmentation, or active learning schemes might be employed to inflate the minority class. Finally, optimization techniques, such as the use of robust loss functions, like the focal loss, can help models automatically focus on the most informative examples. For classification tasks, imbalanced datasets can be clearly understood in terms of the relative populations of the classes. However, the balance of datasets for other tasks, like regression, is less well-studied or established. New metrics are needed to describe the balance of X-ray datasets with respect to tasks specific to X-ray analysis.

3.3. Physical Infrastructure

An emphasis of the workshop was on next-generation hardware needs. Workshop participants noted the need for increased automation for sample preparation and measurement, which is a task often performed by humans. They also indicated the need to have the capability to better detect when problems occur with sample preparation or operation of the instrument. They noted cases of needing to perform X-ray-based analysis under non-ambient conditions, which may be inaccessible to current hardware. Workshop participants indicated the need for programmatic access to equipment such that they can be driven by independent computer systems. Finally, research and development into standards for modular equipment may be a promising opportunity for public-private cooperation. The cards within this category are shown in Tables 10-12.

Table 10. Goal note cards for Physical Infrastructure

Votes	Content
4	Lab tools with standard APIs
4	Automating Collection: – metadata – raw data – contextual / domain
4	Complete tracking of tool state and response function.
4	Extreme condition XRD: – High temperature – High pressure – High speed (transient)
4	Sample prep for XRD
3	Workflow to minimize waste: – sampling volume – reusing analyte – number of samples
3	Tracking ambient condition of sample at time of measurement (temperature, humidity, barometric pressure)
3	Tool drift should be captured and tracked
2	All tool / prep metadata should be captured
2	Automatic manufacturing plant
2	Fully automated XRPD sample prep measurement analysis of toxic/potent materials
2	Make advanced analysis more accessible through automatic instrument function
1	What interface is required for the instruments working in the lab?
1	Automation to improve quality of life for technicians: – Sample prep – Safety
1	How to automate data collection under conditions well outside of ambient (T< 200 C, or > 300 C, P, etc.)?

Table 11. Challenge note cards for Physical Infrastructure

Votes	Content
4	Connectivity protocols with instrumentation
3	Use cameras or other sensors to verify data and sample quality (temperature, humidity, vibrations, etc.)
3	How do we tell the instruments that some of the data they are generating is poor quality and link poor quality data to prep issues with machines?
3	Automation interfaces are not standardized
3	How do you track development of a characterization technique in a unique or complex environment?
2	Controlling instrument will always be a Tower of Babel issue
2	Self Auditing Instrument
2	What standards can or should apply to data format coming from the instrument
2	To accelerate developments of AI-hardware integration we need to learn from pharma industry
1	How to standardize physical interface to equipment
1	How do we move powder samples from instrument to instrument without introducing sample errors?
1	Technical: Best practices: materials characterization software development & maintenance infrastructure (compute, equipment)
1	How do we train machines to prepare large numbers of powder samples without introducing granularity, z-errors, and preferred orientations
1	Need to combine / include XRD and NIR or FTNIR, CEC, technologies for clay quantification
1	Dedicated high throughput experimental approaches could be better to implement high fidelity measurements instead of trying to fully automate general gear
0	Bespoke hardware without APIs
0	What is the interface to the instrument? What standards should apply?
0	Lack of instrument metadata
0	Standardized sample loading
0	But instruments need to be designed around sample design
0	Globe detectors? (360 degree) Can the be implemented for faster data collection?
0	How do we reduce noise levels and scan faster?
0	Much faster if multiple probes and synthesis in one tool / robotic station
0	Dedicated instruments to autonomize
0	How do we resolve the need to collect data quickly (like with off-geometry models, detectors, and off-normal scanning) with background modeling corrections needed for non-FP refinements.
0	Sample prep and synthesis may need to adapt to geometry more amenable to measurement systems
0	Experiments may need to be tailored specifically to capture emergent or surprising phenomena – like first order phase transformation

Table 12. Potential Solution for Physical Infrastructure

Votes	Content
4	Self-driving lab for sample prep
3	Self-Auditing Instrument, Tool Health Monitor
2	Standard API for diffraction instruments with end-to-end pipeline
0	Adding optical microscope to XRD to verify sample prep

Reducing Repetitive Human Tasks — A common factor in all of these issues appears to be a desire to reduce the level of human involvement in sample manipulation through the instrument because it is repetitive and time-consuming and, in some cases, can involve significant safety hazards. These issues could be addressed with an increased level of modularity and standardized physical interfaces for sample management and other forms of measurement. Workshop participants noted that other disciplines have benefited from standardized sample management, such as robotic sample preparation systems designed for the 96-well plate format in the life sciences community.

Sample Preparation — A specific but important case of the need for automation is in sample preparation, which was identified frequently as a goal by workshop participants. Participants both discussed automation of the task itself, along with a desire for robust methodologies (hardware and software) for detecting sample preparation and instrument issues; this is discussed in the section below. This sample preparation is sample and instrument-dependent but may involve planarization of bulk samples or transferring powdered material to an appropriate sample holder, including adding binders, dilutants, or internal standards.

Detecting Sample Preparation and Instrumentation Issues — In both human-based and automated X-ray analysis there are always possibilities (and maybe even frequently) of experimental issues that lead to data degradation. These issues may stem, for example, from problems in sample preparation or from instrumentation issues including diffractometer misalignment or detector problems. Particularly as automation increases and the act of data collection becomes separated from the act of data analysis, it becomes even more crucial to detect these issues before time and resources are wasted, or even worse errors from misanalyzed faulty data create real-world issues outside of the lab. Detecting issues must handle a range of failure modes including one-off problems such as sample holder loading problems or emergent issues such as wear on mechanical components leading to misalignment. Conceivable, these issues can be detected either through sensors and self-diagnostic routines on the instrument or through anomaly detection algorithms that operate on the raw data. The latter approach is attractive as it requires less specialized hardware, but achieving this will require the collection of datasets specifically for this purpose.

Programmatic Access to Instruments — Workshop participants highlighted the need for programmatic access to instruments, both for control of the instrument and access to measurement data and diagnostic information. Participants noted that instruments should be able to automatically store machine state and configuration information while enabling machine-actionable auditing. Attendees representing commercial vendors stated that to this point, there was not sufficient demand for this type of feature. They also expressed reluctance on behalf of their companies due to demand for a wide range of types of interfaces requested by their customers, requiring a large engineering load. Further, they felt there was a risk that changes in software that interacted with their instrument but was not written or controlled by them could be changed, resulting in a failure of their programmatic interface that they would be held responsible for fixing. From the end-user perspective, workshop participants noted that many independent efforts exist and it is unclear how best to proceed. Efforts thus far have explored existing protocols including SiLA[48] within ChemOS[8] and MQTT[49] within MDML[50], explored streaming software like Apache Kafka[51] within OpenMSIStream[52], or built out bespoke REST APIs in HELAO7. Workshop participants noted that each approach likely has strengths and weaknesses, which are not sufficiently quantified in the context of an autonomous materials laboratory. We note that workshop participants view that maintaining and integrating these technologies is expensive and may distract from the broader mission goals. We note that interchange protocols are a small part of the broader cyberinfrastructure of the modern software-enabled laboratory.

Opportunities for Modular Equipment — Workshop participants highlighted the desire for improved physical interoperability of laboratory equipment. There is a clear opportunity to develop new standards for laboratory equipment. However, there are currently two approaches being actively used within the community. In one approach, researchers are developing new systems from the ground up with a sample-centric approach. In the other approach, researchers are developing human-scale robotics that are capable of operating existing laboratory equipment. As a follow-up to this workshop, to meet the need for modular equipment with hardware and software interfaces (as discussed above), we have generated a report proposing a national center for autonomous materials research, with the goal of enabling off-the-shelf modular autonomous infrastructure[53].

Multimodal Data Generation Another topic brought up by participants was the need for collecting multiple data streams to support more advanced analysis. This may be compositional data, spectroscopy techniques for generating chemical and additional structural data, or potentially measurements of functional properties. In order to collect this data more efficiently, it would be beneficial to allow for this data to be collected simultaneously. This could be achieved through instrument vendors developing these multimodal instruments or through the development of modular equipment as described above.

3.4. Algorithm and Model Development

Data-driven algorithms and models are at the heart of data-driven science and autonomous laboratories. These algorithms and models are often specialized to a specific type of data, and should thus not be considered monolithic. This subsection is divided further into three subsections. This workshop had a specific emphasis on X-ray-based methodologies. Therefore, the first subsection reviews issues related to phase identification and quantitative analysis of diffraction data. The next subsection focuses on broader issues relating to AI for materials science and engineering and its intersection with X-ray-based. The final subsection reviews issues related to uncertainty quantification.

3.4.1. Phase Identification and Quantitative Analysis

The workshop participants agreed that there was a need for algorithms for phase identification and quantitative phase analysis. In discussion, some participants felt that these are largely solved problems. Upon further probing, it was generally agreed that existing capabilities are sufficient in some cases, but for many instances with more complexity in the analysis goals and data these existing models fall short. There is, therefore, a need for a systematic evaluation of the strengths and limitations of the many competing traditional and ML approaches for different XRD-related analysis and prediction tasks. Existing innovations in data-driven methodologies tend to be applied to simple and idealized data and lack the depth to be later applied to complex (real-world) problems. With an in-depth focus on phase identification and quantitative analysis during the workshop, a number of specific opportunities emerged including supporting complexity in the physical samples, complexity in the measurement methodology such as low data quality from benchtop systems, multimodal data streams, and high throughput methodologies. The cards within this category are shown in Tables 13-15.

Table 13. Goal note cards for Phase Identification and Quantitative Analysis

Votes	Content
11	AI Phase ID not limited to certain phase diagrams
11	Using AI to predict additional phases in a multi-phase mixture
6	Automated analysis has a long way to go to robustness for challenging problems, but it could accelerate a lot
6	Uniting 1D and 2D collection analysis without conversion
5	Can we layer multiple complementary techniques to help QPA / XRD automation on?
4	Pattern ID recognition vs absolute ID (wt %)
3	Do we want our labs to be statistically perfect (use self-referential XRD analysis) or reference external tools such as XRF, analytical chemistry, SEM, to establish a standard?
3	Can we pair cluster analysis (automated) with automated QPA for more accurate analysis
3	XRD tools for many different complex materials that can be used by non-experts
2	Combination of PONCKS + QPA via IS for phases with distinct amorphous hump but needs IS. For normalization of d-spacing for phase ID
2	Automated XRD on benchtop data for highly overlapped samples -> What is the low benchmark for QPA & can the AI / ML intelligently decide
2	How to make measurements almost instantaneous? (without synchrotron)
2	Data screening granularity preferred orientation crystallinity instrument
2	XRD, XRR, XRF, should be able to be simultaneously fit
1	Can automated XRD parse apart a sample for QPA with > 10 solid solutions, polymorphs, etc. Extreme thermodynamic variability
1	Data quality specification of XRD missing outdated multi-resolution data repairing (AI-based)
1	XRD automation that can catch multiple errors

Table 14. Challenge note cards for Phase Identification and Quantitative Analysis

Votes	Content
3	General Phase ID
3	How do we operate over the best XRD technologies in dusty and noisy environments?
1	How to standardize QPA methodology to allow for many types of analyses, including sample prep. PONCKS / IS / ES / Rietveldt / intelligent RIR
1	Multiphase QPA combined with many cross validation techniques.
1	How to train machines to correctly choose overlapped phases?
1	SRMs appropriate for synchrotrons – Sharp lines – Wide Q range – 0.01 % mass impurity level
1	Correct for oversampling / undersampling (representation) of certain methods (XRD, Raman, NMR) & properties because the are easier or more directly performance related
1	Grain boundary orientation conventions round robin analogy – can we capture sufficient meta-data to recreate an analysis?
0	Really need quantitative phase analysis complementary cluster analysis + qualitative
0	Initially AI can be used to correlate XRD patterns to a macroscopy property
0	XRD is characterized to many ambiguities which require complementary data to move forward XRF, Raman, NMR
0	Texture and d-space shift complicate analysis but scanning is slow
0	Working with small sample sizes for XRD from autonomous labs – bad particle statistics

Table 15. Potential Solution for Phase Identification and Quantitative Analysis

Votes	Content
7	XRD + XRF
4	Take advantage of Chi information in 2D for textured powders
4	Use Bob He’s Chi Integration or Azimuth to Automate poor texture identification
2	Reynold’s Cup for clay minerals phase ID
2	More XRD/XCT/domain specific (XML) validation

This workshop placed a major focus on X-ray and neutron based scattering methods. Overall, workshop participants highlighted either the need to solve repetitive problems quickly, or the need for data-driven methodologies to accommodate complexity in the samples and data being analyzed. We illustrate this complexity schematically in Figure 3.

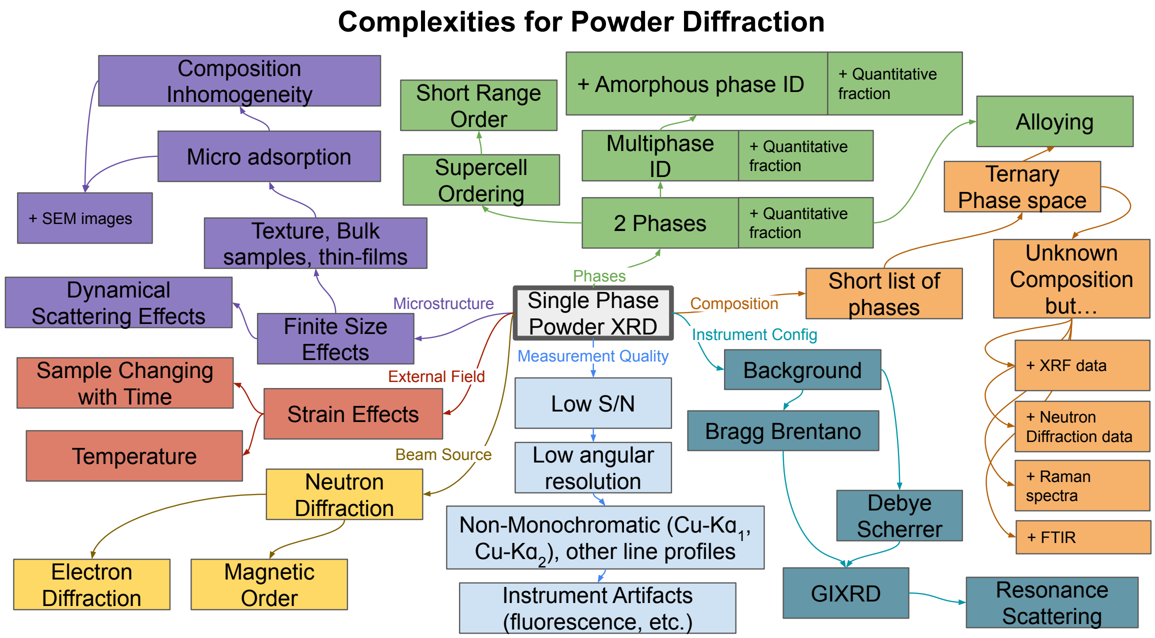

Fig. 3. Illustration of factors that increase the complexity of powder diffraction data, starting from the center box and roughly increasing in complexity outward. Each of the boxes represents a complicating factor for the analysis, color-coded by a theme of complexity. These complexities should be thought of as dimensions along which the analysis can be complicated, and can therefore be combined. This figure is meant to be illustrative and is by no means exhaustive.

Analysis of powder diffraction patterns can be made easier as justifiable simplifying assumptions are introduced. In the simplest case, if the sample is known (or assumed) to be a single phase randomly oriented powder with equiaxed grains, then the conventional phase identification approach of directly matching expected peak locations and intensities to a set of structural prototypes. This is represented in the gray box in the center of Figure 4. This phase identification task becomes more difficult as assumptions of the data quality are relaxed, as represented by the light blue boxes in Figure 4. Most current ML algorithms are designed to handle this phase identification task at this level – or with only a few complexities considered as nuisance factors. Relatively little work has been done explicitly targeting quantitative analysis of quantities such as phase fraction, grain size, and lattice parameter directly with ML systems.

Measurement Quality — Real-world measurements are non-ideal, and those complexities often break assumptions used to construct the ML models. We could consider measurements with low signal-to-noise ratios (S/N), lower angular or $\mathbf{q}$resolution, the effects of non-monochromatic X-ray sources, or other instrumental artifacts like X-ray fluorescence. Lowering the S/N ratio should lower our confidence of the phase identification and increases the uncertainty of quantitative modeling. Lowering the instrumental resolution makes it more difficult to distinguish between peaks that are nearby, and to obtain precise estimates of peak locations. Real X-ray sources have some distribution of photon wavelengths and angular divergence, which will be superimposed on the diffraction signal. Instrumental artifacts, such as X-ray fluorescence, are typically treated as background in an XRD measurement, but these artifacts make it more difficult to compare measurements across different instruments.

Instrument Configuration — One of the major instrument-to-instrument (or even sample-to-sample) differences in X-ray diffraction is the background. It would be difficult to enumerate all the potential sources of background signal and explicitly account for each of them. Even measuring the signal from a blank or empty sample holder has its challenges, particularly in a reflection geometry. This is because there are some contributions to the background, like the diffraction off of the sample holder material – which will be partially blocked by a sample – while other contributions, like air scattering, will not be. Therefore the signal from a blank sample holder will not simply add to the signal from the sample. Further complicating, background fitting is scattering from amorphous material or other sources of diffuse scatter within the sample itself can often not be separated trivially. It is common, then, to seek a heuristic approach that generalizes across several instruments. In order to evaluate the generalizability of these heuristic approaches, data is needed from several instruments with different background signals. Furthermore, the background signals are different for different instrument geometries. The Bragg-Brentano geometry is popular for PXRD measurements in laboratory-based diffractometers. However, Debye-Scherrer geometry is popular especially when the beam can pass through the sample in transmission mode.

Phases — If the assumption that the sample is single phase is relaxed to allow for the possibility of two or more phases, then analysis becomes much more challenging. Due to the overlapping of peak positions, it can be hard to simply identify minority phases, especially at low signal-to-noise ratios or resolution. Quantitative phase analysis will provide more reliable results not only at higher data quality, but also when the phases are more distinct from each other. These tasks are combinatorially more challenging as the number of possible phases increases. Of course, in many real-world applications, the total number of phases is not known a priori – which may, in fact, be the motivation for a particular PXRD analysis in the first place. Real-world samples may also include amorphous materials whose signals are more difficult to treat quantitatively. Depending on the processing history of the potentially multi-phase sample, other effects like alloying which causes peak shifting, or supercell reconstruction, which causes additional peaks, may be possible. Short-range ordering has a complex contribution to the scattering signal that is typically analyzed in the Fourier transform of the scattering pattern known as the pair-distribution function.

Composition — Knowing the composition of the sample can help populate a short list of likely candidate material structures. For phase mapping in composition space, the scattering patterns are related by how similar the samples are in composition. This means that adjacent compositions should have similar short lists of likely phases, which can aid in identifying phase regions in composition space. But if the composition is unknown, then combining the scattering pattern with another measurement may be necessary to provide enough information to identify the phase. Such joint inference is useful when the other measurement has a different operational mechanism and, therefore, provides a complementary signal. For example: X-ray photoemission spectroscopy (XPS) measures the energy required to produce photoelectrons, neutrons have a different scattering cross-section than X-rays providing different contrast, Raman spectroscopy measures how phonons couple to visible light.

Beam Source — X-ray scattering factors primarily depend on the electron cloud density of the atoms of the material. Neutrons, however, primarily interact with the nucleus of the atoms. Neutron scattering, therefore, provides a very different atomic contrast than X-ray scattering. In neutron scattering, it is possible to get a scattering signal from the low atomic number elements that X-rays are not very sensitive to. Further, neutrons will interact with the magnetic moments that are present in the material. Neutron scattering is, therefore, widely used to study magnetic ordering, where the scattering of the magnetic order manifests as additional scattering peaks, which may or may not overlap with the peaks from the atomic structure. Electrons have a much stronger interaction with the atomic electron cloud than X-rays do. That means that strong scattering signals can be obtained from comparatively less dense electron clouds, or smaller amounts of material, but it also means that artifacts – such as multiple scattering events – are more prevalent, making analysis more difficult.

External Field — Materials structures distort when subjected to external fields, which, in turn, distorts the diffraction patterns. Perhaps the simplest external field to account for is an applied stress which acts directly on the lattice parameters of the structure. This has the effect of shifting peak positions, which can manifest as a splitting of peaks if the symmetries are broken by the distortion. Strain gradients can cause broadening of diffraction peaks. Temperature not only distorts the structure through the effects of thermal expansion/contraction, but also has effects on the scattering parameters. As the temperature increases, the phonon modes in the structure are further excited, and therefore the average atomic positions are less localized. The beam itself can be thought of as an external field. Some materials, especially organic materials, can degrade in strong X-ray (or neutron, or electron) radiation. This can have the effect of altering the atomic structure. This is perhaps one of the most difficult external fields to account for since the transformation is typically irreversible. The type of damage experienced by a sample affects the operational mode to account for it. Some beam damage might be mitigated at low temperatures or low times. A signal can be acquired over many positions as the irradiated positions are damaged. In some cases, the beam damage might not be avoidable, in which case the signal could be acquired in a single high-intensity pulse. These acquisition modes affect how the data will be analyzed.

Microstructure Effects — In addition to the average crystal structure of the sample, the microscopic spatial distribution (the microstructure) of the material can also have substantial effects on relative peak intensities and broadening. As the crystallite size decreases, the diffraction peaks will broaden. Conversely, very large crystallite sizes can violate the powder assumption as individual Bragg reflections become resolvable on a 2D detector – therefore leading to incomplete statistics with respect to orientation when integrated to a 1D diffraction pattern. Many materials, including those in thin films, not only show finite-size broadening effects but also can have a preferred crystal orientation. Because of this preferred orientation, some atomic planes are normal to the scattering vector more often than other planes, increasing the diffracted intensity of those peaks relative to the others. Bulk samples can also often have a preferred orientation of the crystals due to processing history.

Micro-absorption is an effect seen in materials with heterogeneous microstructures that is not accounted for in most diffraction analysis software. This effect can be envisioned in the extreme, for example, as layered structures where the top layer is a strong scatterer and the bottom layer is a weaker scatterer. The diffraction signal from the bottom layer is much lower than would be expected from a simple sum of the signals from both layers. This is because the beam must pass through the top layer first which lowers the intensity that reaches the bottom layer. Furthermore, the scattered signal from the bottom layer must pass through the top layer, which additionally lowers the intensity that reaches the detector. Composition gradients can make this effect much more complex to handle since the heterogeneity goes from a discrete ordering of layers to a continuous function describing the composition. The analysis of samples that need to account for these microstructural effects can be aided by complementary characterization to directly quantify microstructural features through microscopy or other imaging techniques. The kinematic theory of diffraction, which neglects multiple-scattering, is a good approximation to use for most common X-ray diffraction measurements. However, as the crystal size and quality increases dramatically or as the interaction strength increases going from an X-ray beam to electron beam, then the effects like multiple scattering become increasingly important. These effects require a more complete description of diffraction from dynamical diffraction theory.

Summary of Phase Identification and Quantitative Analysis — To summarize, Figure 4 shows the ways in which assumptions about the sample and measurement could be relaxed to make the analysis more difficult. This, therefore, also shows the features of a database that would be needed to address those complexities. For example, to build and test an algorithm that could handle phase identification regardless of the instrument geometry, a database is needed that has Bragg-Brentano, Debye-Scherrer, and GIXRD measurements. If this algorithm also needs to be able to account for strain effects, then examples of strain effects are needed at each geometry. Similar logic can be applied to additional complexities, adding new dimensions to consider, and building toward a general database needed for a general algorithm. With all of these routes to support complexity, workshop participants indicated that it is important to incorporate known physics.

3.4.2. Broad AI for Materials Science and Engineering

Rapid innovation in algorithm and model development, broadly speaking, presents a significant opportunity for industry to introduce targeted autonomous solutions for repetitive problems that typically require human expertise and experience. These technologies can be developed and adopted most efficiently if industry and other end-users prioritizes which tasks ML methods will have the largest impact on their productivity. These can include methods for combining multimodal data streams, automated experiment planning, metadata assessment, and automated selection of analysis tools. Though the workshop focused on X-ray methods, workshop participants generated several cards for algorithm and model development that apply to materials research more broadly. In this section we present these issues through examples specific to XRD. The cards within this category are shown in Tables 16-18.

Table 16. Goal note cards for Broad AI for Materials Science and Engineering

Votes	Content
5	Data transformation and preconditions scheme to improve ML interpretation
4	Multiple technique analysis results coordinated for consistent result
3	Predicting metadata (e.g., wavelength) based on a raw X-ray diffraction pattern alone
3	Minimize human interaction (robustness -> trust) easy to modify (code-free)
3	Automatic identification of next analysis. Example: microscopy, single crystal XRD.
3	How to deal with time series data and its analysis (as environment etc. changes or ages in certain environment)
2	How to measure robot wear and tear given data quality erosion without human oversight
2	Composition Mapping (3D)
1	More ML tools & data related to manufacturing
1	3D Grain Mapping: HEDM
0	Data storage / pipeline should allow for re-interpretation
0	Super Edisonian Combinatorial approach vs targeted discovery. I.e., many samples vs as few samples as possible

Table 17. Challenge note cards for Broad AI for Materials Science and Engineering

Votes	Content
5	How to maximize current generation methods?
5	What quality standards should apply to data used to train ML algorithms
4	Privacy preserving learning with regard to database models
4	Measurement condition determination: Time vs Quality
4	Transferability for feature extraction
3	Combining data from XRD, XRF, and electron microscopy to determine the structure (crystalline)
3	How do we harness many computers to analyze large numbers of samples simultaneously? What about QA/QC?
3	AI isn't compared to expert
2	Good methods for domain adaptation to go from synthetic data to realistic data
2	AI adoption: – issues with variable conflation / convolution limit statistical power of AI models (or any statistical models) without express design in which case, use Design of Experiments.
2	The samples to be analyzed should be suggested by a trained machine (system)
2	Reversible data-model matching
2	Set up network and workflow for AI to act on – but not necessarily tied directly to hardware – When to ship your samples for TEM?
2	Feedback on Sample Prep or Measurement Quality
2	How to make trained models instrument/sample agnostic
2	Productive workflows can include human-in-the-loop: Use AI to automatically generate solutions, human expert vetoes problematic suggestions
1	In the future the XRD pattern could be related to many materials properties using AI
1	How to teach analytical machines to overcome sample prep issues consistently
1	Automation tools need to make it easy to program custom models
1	How do we link XRD lab together to form a singular material processing mind for selecting and analyzing specific materials?
1	Physics-informed data samples, contextualized data / sampling
1	Data mapping between different characterization techniques, technical barrier
1	Data Fusion: – alignment (time, space) – conceptual – relevant fusion depends on desired response: in f(x,y) x & y change
0	Flowing XRD data into the decision model then to the synthesis tool
0	Make ML assessment tools more readily available
0	Match making between type of data and methods
0	How do we postprocess highly displaced data en-mass to bring the re-test rate down
0	Implementing algorithms with ELNs
0	Data is collected as fast as possible. Not great for building libraries.

Table 18. Potential Solution for Broad AI for Materials Science and Engineering

Votes	Content
7	Generative AI for data search and queries
7	Exploratory querying/match making how to help users finding/uniting questions
7	AI or ML to identify poor sample prep or other artifacts
6	Quantitative data quality assessments
4	Knowledge Graph of what questions can be answered given particular measurements
4	Privacy-preserving + federated learning
3	LLMs for prompt engineering to match methods to data
2	Protocols for splitting train/split similar materials – split by sample group or by space group or element – split by fidelity, make sure lab scale diffraction in test set
2	Data collection protocols for small samples: – multiple measurements – vibrating samples
2	Transferability assessment with clever similarity analysis (multimodal similarity)
1	More Hybrid Metrology + Analysis
1	Develop measurements for success / confidence in machine refinements to determine pass / fail thresholds for machine learning that is relevant to materials questions.
0	Pushing bar on ML work to move field forward
0	Better decision points on what techniques needed (e.g., lab vs synchrotron vs neutron)

Harnessing the AI Revolution — Workshop participants discussed the application of data-driven methodologies to many aspects of materials science and engineering. High-ranking potential solutions suggest that some combination of generative AI, large language models (LLM), and knowledge graphs (KG) has the potential to improve search capabilities, including based on similarity (e.g., matchmaking). A number of the cards focused on general practices in using data-driven methodologies, such as data cleaning, addressing missing metadata, addressing unbalanced datasets, and how to split datasets for training and testing. Language modeling and knowledge graph technology represent largely unexplored opportunities to develop high-quality interconnected datasets of raw data, metadata defining samples and experimental protocols, and high-quality text emphasizing the expert subjective evaluation of the analysis.

Managing Risk and Maximizing Returns — Some workshop participants expressed some hesitancy towards data-driven methodologies for a number of reasons. they indicated that it is difficult to determine how to optimally adopt data-driven approaches to maximize a return on investment. Furthermore, they highlighted the desire to maximize the effectiveness and impact of current data-driven capabilities. These cards suggest a general disconnect between industry and academia. Academia in the materials community is in constant pursuit of novel approaches in data-driven methodologies, whereas the materials industry needs to be strategic in adopting certain approaches that will provide a lower-risk, stable return on investment over industrially-relevant periods of time. These two extremes could potentially be accommodated by a software development model that has different release tracks where the slower release track tends to prioritize stability and the faster release track tends to prioritize novelty.

Model Evaluation Criteria — To enable robust progress in ML-driven analysis algorithms, the community must come to a consensus on useful model evaluation criteria for each of the different analysis tasks. In order to evaluate the models, the particular tasks must be robustly defined in mathematical terms. For example, in the problem of identifying phases present in an XRD pattern, the task could be posed as a retrieval problem – matching the phases to a known ground truth. In this case, the relevant metrics might include precision or recall. In the problem of phase mapping across some composition space in the absence of a predetermined ground truth, model evaluation criteria could be some measures of when the phase map has converged to a stable solution. Another category of problems are the structure (Rietveld) refinement problems. Here, there are the conventional evaluation criteria, such as the R-factors, reconstruction errors, or log-likelihoods of the pattern from the model against the measured pattern. In a ML context, additional evaluation criteria might include the accuracy of the refined parameters with respect to some known ground truth, or alternatively use the reconstruction errors and similar. Practitioners of Rietveld refinement often discuss how well the peaks from the model match the peaks observed in the measured pattern – which is a qualitative metric that is often undervalued by global quantitative measures. Another model evaluation is related to UQ, especially in the case where several related XRD patterns are refined together (with the model describing how they are related). The signal from one pattern will affect the refinement of the structure from the other. Metrics are needed to understand the strength of this information sharing. In addition to instrument calibration, the models themselves are calibrated for specific tasks, and metrics are needed for how well the models extrapolate beyond the data they were trained against – especially relevant for continuous monitoring and for data drift as new measurements are evaluated. Further, there is a need for criteria that allow for the selection of appropriate models. Metrics such as the Bayes Information Criterion provide some measure of whether the data justifies the complexity of the models. However, models could also be evaluated in terms of whether the physicochemical constraints they impose are appropriate (e.g. a model with a priori limits on what materials could or could not be present).

Detecting Measurement Quality — Workshop participants identified a need for criteria for automatically assessing the quality of materials measurements. This is a more challenging task than evaluation metrics for well-defined prediction and fitting tasks described above because many aspects of "measurement quality" are subjective. How can we know when we have a "good" measurement? Perhaps the most obvious metric is the signal-to-noise ratio, which can be estimated from Poissonian counting statistics. While this constitutes a quantitative measure of the quality, the question then becomes: what signal-to-noise ratio is good enough? The answer to that may be task-dependent. In the case of XRD-based phase ID tasks, the signal-to-noise ratio required to confidently exclude the possibility of minority phases depends on the phases being considered. If the phases considered produce very different diffraction patterns, then only a moderate signal-to-noise ratio and resolution are required to determine the presence of each phase. However, if the phases produce patterns with significant peak overlap or if the primary differences are subtle (such as weak superlattice reflections in the case of ordered vs disordered crystals), then a much higher signal-to-noise ratio and resolution and lower background would be required to distinguish them. Similar arguments could be applied to other quality metrics such as the angular ($\mathbf{q}$-space) range and resolution of the data (including effects from the beam divergence and chromaticity, detection optics, the detector, and the goniometer). These can be important metrics of quality when considering tasks like detecting slight distortions in high symmetry phases.

There are other more difficult to quantify metrics of measurement quality. How much signal is coming from the sample holder? If a sample being measured in a reflection geometry is only weakly diffracting then a significant portion of the signal reaching the detector may be from the sample holder. Is this the optimal instrument configuration for the task at hand? The choice of diffraction slit width will affect the intensity of the signal, the resolution, and how much of the sample and/or sample holder is irradiated. Were there any issues with the sample preparation? Powders with large crystalites will not provide good statistics over all orientations. Samples mounted above or below the optimal sample plane will affect the accuracy of the angles ascribed to the peaks. What is the expected background signal from the instrument, and how likely are amorphous phases in the sample? Many types of instrumental background, such as air scattering, have broad features over a wide angular range, which can be difficult to distinguish from the broad diffraction signal from amorphous materials. Furthermore, some samples can change or degrade under the irradiation used to measure them. Determining and optimizing the measurement quality in terms of the risk of sample degradation can be critical for some samples. Some of these considerations might be addressed by dedicated datasets as discussed in AI-Ready Data. Metrics for data quality are needed that consider the context of the task at hand. For example, robust and well-defined methods for assessing these aspects of quality could have a substantial impact by enabling automated optimization of instrument configuration and sample preparation.

Quality of the Analysis — The level of trust given to a particular analysis should be informed by quantitative metrics for evaluating measurement and model quality. How the form of the model – the assumptions that go into it – along with the quality of the data will determine what statements one can confidently make from a measurement. The analysis will have considered certain possibilities, but ignored others. For example, a supervised phase detection system (or conventional search/match) will not be able to detect and identify structures not in its library of structures, let alone, heretofore, unreported structures. Many ML models (and conventional analysis methods) are typically not designed to indicate when the analysis is inconclusive. They always provide a conclusion within the bounds of the model design. As a concrete example, consider a deep neural net trained on XRD patterns to predict the phase fraction from a library of structures. Given this design, such an ML model will consider every XRD pattern as being composed of structures from its training set library of structures. In other words, this ML model could never predict that the XRD pattern is from a structure that is not in its library, instead predicting that the XRD pattern is from some combination or mutation of existing structures. Practitioners must decide if those bounds allow for conclusive results.

It is important to emphasize interpretability and explainability in model design. Interpretable ML models can have a physics-informed model design such that the bounds of the model, the intermediate results, parameters, and/or predictions can be understood in terms of the context of the problem. Explainability in ML models refers to inspecting any model to understand the criteria used for the predictions, regardless of the intrinsic level of interpretability of the structure of the model. For example, attribution methods can be used to identify the influential training instances that the predictions are highly sensitive to. It is well known that ML models often extrapolate poorly, but it is not always straightforward to determine when an ML model is extrapolating, particularly in high-dimensional datasets. Many ML models construct a latent representation of the inputs. The location of a new data point in this latent representation space relative to the training points can be used to determine if the ML model is attempting to extrapolate. Understanding the model form, constructing interpretable models, and inspecting those models to explain the predictions, are all tools that a practitioner could use to judge when to trust the results of the analysis.

Search Tools — Workshop participants highlighted the need for improved multifaceted search capabilities, (e.g., data, software, algorithms, etc.). In addition to topics discussed in FAIR for Data, Models, Tools, and Instruments, workshop participants highlighted the opportunity to leverage new technologies for the use case of search. Specifically, they indicated a potential opportunity in harnessing knowledge graphs, large language models, and generative AI more broadly as potential approaches to improving search capabilities. Applications of these technologies to XRD-related search problems have been broadly underexplored in part due to a lack of standardized evaluation datasets and criteria.

Model Transferability — Participants noted the importance of developing criteria and capabilities for determining the degree of applicability and transferability of ML systems to new datasets and tasks not directly linked to the data used for training. While these are longstanding research topics in fields like computer vision, there have been relatively few specific efforts to quantify extrapolation quality and transferability for X-ray diffraction data. What is needed in the diffraction domain is more robust criteria for deciding when pre-trained ML models are applicable, when transfer learning is likely to give reasonable results, and when it is more appropriate to train new models from scratch. Such criteria will likely need to account for some of the domain knowledge particular to the field of diffraction. These topics are also closely linked to uncertainty quantification and model calibration, see Section 3.4.3.

Privacy-Preserving and Federated Learning — Workshop participants highlighted the need to develop mechanisms that enable ML/AI researchers to develop models for proprietary data while protecting that data from unauthorized disclosure. Specific examples in the XRD space are when one database holder wants to combine data with another database holder to train a model, but one or both parties do not want to share their databases with each other. This is an active area of research and development within the cryptography and computer science community. Furthermore, it is likely that materials and manufacturing stakeholders and use cases are underrepresented in current efforts. Increasing the participation of materials and manufacturing stakeholders in efforts surrounding privacy-preserving technologies and federated learning could become increasingly important given the unique complexity of materials and manufacturing data.

Fusion of Multiple Data Streams — There are important opportunities for combining multiple information sources to improve the quality and robustness of XRD analysis. A well-established example is the use of composition-based filtering in search-match procedures to narrow down structure matches based on the known chemistry of the sample. High throughput XRD systems can integrate X-ray fluorescence measurements to bring the raw information underlying this analysis directly into ML-driven analysis pipelines. This highlights the opportunity to develop and disseminate experimental protocols and modeling strategies for information fusion.

3.4.3. Uncertainty Quantification

As ML methods are intrinsically statistical in nature, workshop participants noted that UQ is of critical importance if data-driven methodologies are to be implemented in robust and reliable production applications. Strengthening UQ evaluation requires accurate evaluation of standard statistical quantities like mean average error (MAE), as well as the determination of the confidence of each individual ML prediction (error bounds), as it is individual predictions that are used in industrial applications. Obviously, no amount of model UQ can reveal or mitigate bad quality in the training data, as systematic errors in the generation of the data set will propagate through a machine learning model. Careful curation of the training and testing dataset is, therefore, a crucial step in developing trustworthy ML models, and there is a need for developing an accuracy measure to communicate the quality of such a dataset.

Other important sources of uncertainty were identified in the interpretation of the various fittings as well as in the ML models themselves, whose quantitative predictions should be evaluated. Sources of uncertainties associated with implementing a trained ML model include too localized or limited training data, ML training errors (ex. overfitting), sub-optimal model architecture or parameter learning, and out-of-distribution samples. This reflects the uncertainty in the prediction with respect to the model design, which is not captured by the uncertainty estimates output from any one model. While many techniques are available for estimating the UQ on regression predictions[54], none were suggested as a solution, possibly because the workshop participants were not experts in the ML field. It is well known that ML models do not extrapolate well. However, due to the complex design of many algorithms (e.g., deep neural nets) and the high dimensionality of the data they operate on, it is often not obvious when the ML model is, in fact, extrapolating. Robust failure detection was also identified as a critical need associated with verification and validation, as well as developing repeatable data analysis. The cards within this category are shown in Tables 19-21.

Table 19. Goal note cards for Uncertainty Quantification

Votes	Content
4	Uncertainty quantification
4	Two categories of uncertainty quantification: – Raw measurement noise – Fitting interpretation
3	Easy verification of results
1	Accuracy & HTP for diverse lab -> Needs to be understood & used by non-experts
1	Repeatable analysis: Respecting promises: inputs / outputs and semantics

Table 20. Challenge note cards for Uncertainty Quantification

Votes	Content
3	Quantify detection limits
2	Uncertainty in the x-axis poorly defined (e.g., composition)
2	Database accuracy should be quantified
2	High variability -> more open AI / automation -> less tailored AI -> more errors
1	XRD respective of data/sample bias. Fairness of AI/ML models.
1	Uncertainty quantification related to analysis quantification (regression)
1	Types of UQ: – robust failure detection – reasonable accurate error bounds

Table 21. Potential Solution for Uncertainty Quantification

Votes	Content
4	Quality standards for input data: – invest in development of UQ (uncertainty quantification) and error analysis for ML – use results to specify quality standards

3.5. Community Engagement

Workshop participants authored a number of cards centered around community engagement. One major theme was workforce development, with an emphasis on cross-disciplinary education and considering both the existing workforce and next-generation workforce. Other cards discussed culture change, intellectual property, and economic considerations. Data-driven methodologies are ultimately implemented as software. Workshop participants highlighted a number of issues related to the sustainability of both open-source and proprietary software. Finally, workshop participants highlighted the need for new and ongoing convening and coordination efforts.

3.5.1. Workforce Development

Workforce development was identified as an important need where depth of knowledge in both experimental methodology and data-driven methodologies is critical for success.

It is important to have a depth of understanding of the physical processes and geometries of the diffraction experiments to enable correct interpretation of measurement results. However, it is becoming increasingly important to have additional depth of knowledge in the machine learning algorithms to understand their implications and biases for use as analysis tools. Having a workforce trained in both disciplines is needed for the full realization of machine learning and autonomous workflows of diffraction experiments. Even non-expert practitioners need to be educated in framing their problems as ML tasks and in identifying and deploying existing solutions. The cards within this category are shown in Tables 22-24.

Table 22. Goal note cards for Workforce Development

Votes	Content
8	AI vs. existing methods? Are they known and /or already exploited?
6	Cross-disciplinary education: Statistics, experimental design, materials, physics, economics
6	Robust education for existing staff
6	Should not be a black box. Even with automation, expertise is needed.
3	XRD is not one thing – it is very versatile. The data modality depends heavily on the task.
3	Need to empower bench scientists to run their own analyses
1	Workforce training to use AI / ML tools
1	What is the expected level of expertise of people working in the lab?
0	Make XRD lessons searchable

Table 23. Challenge note cards for Workforce Development

Votes	Content
8	Need for new community members to learn what pitfalls to look out for – these are in the literature but hard to find without knowing already
7	Few materials experts in general. Even fewer with knowledge of AI / ML / automation
7	Huge lack of knowledge XRPD + AI
4	Education for sample prep
4	It's very context dependent for what other measurements to do and when.
4	ML researchers can focus too narrowly on specific established tasks, making overly restrictive assumptions.
4	Determining the problems that really need ML
3	Unknown Unknowns Problem; what questions to ask
3	Technical Barriers: Overlap – or lack – in necessary expertise, e.g., XRD analysis & infrastructure architecture
3	ML research in XRD may not be familiar with successful methods that are already established
2	Need a different way of thinking from scientists / technicians
0	Training the resource creators on data storage (including metadata)

Table 24. Potential Solution for Workforce Development

Votes	Content
8	Education! Users / practitioners need to know: – what factors matter – why they are important – where to learn more to be effective with next-generation tools (and current ones.
6	Limited personnel with necessary expertise: – work on career pipeline – collaborate routinely with industry
3	Stronger online educational materials aimed at teaching computer science students materials fundamentals – incentives / amplification for blog posts?
3	Practical co-training for student of computer science and materials to teach them to be effective cross-disciplinary collaborators. Capstone programs?
3	Educate operators of instruments on widely used software
1	Remove administrative barriers to train materials experts in coding
1	To understand what is AI, it is recommended to start one application
1	Best Practice Guide
0	Need for crystallographic coursework outside of the academic department – summer school?

Cultivating a Multidisciplinary Workforce — Workshop participants clearly articulate a desire for a future workforce that has expertise in both their core chosen discipline (e.g., materials engineering, mechanical engineering) and core topics related to ML and AI. While an increasing number of universities are evolving their degree programs to support this combination of expertise, this will not benefit the existing workforce. Additionally, they indicated a need to develop programs and resources specifically targeting the existing workforce. Topics related to core knowledge and workforce development were discussed in workshops organized by The Minerals, Metals & Materials Society (TMS)[55, 56].

Learning from Existing Resources — Workshop participants also spoke on key content of educational materials and resources. First and foremost, they noted that AI/ML should not be treated as a black box. Secondly, they indicated a need to elucidate known pitfalls. Known pitfalls are in the literature, but they are difficult for inexperienced practitioners to find. If known pitfalls are not widely recognized, new adopters may repeat known mistakes, waste their valuable time, and ultimately become frustrated.

3.5.2. Software Development and Sustainability

Workshop participants highlighted a number of issues related to the development and sustainability of software. Much of the algorithm development is occurring in an academic research environment and the pathways to production software are plagued with many challenges. Research software is highly decentralized and often poorly documented, making it challenging for industrial users to discover and reuse. Further, industrial developers have little incentive to assume the responsibility of maintenance and support for open-source software they might otherwise incorporate into or enable integration with their products. The cards within this category are shown in Tables 25-27.

Table 25. Goal note cards for Software Development and Sustainability

Votes	Content
6	How to make automated refinement coding resistant to software changes and updates or changing industry standards / updates

Table 26. Challenge note cards for Software Development and Sustainability

Votes	Content
7	Need to build new technologies on existing programs rather than write more new programs
6	Too many programming languages
4	Lack of Software Testing in Peer Review
2	How to overcome coding challenges with new software adoption? (needing to constantly rewrite)
2	Need simple UI to train or add new information to an AI model
2	The evolution of computer hardware should be taken into account when an app (AI) is developed.
1	Hard to commercialize research code (licenses, quality)
0	How to code entries in open-source software to classify entries to be used in quantitative data consistently
0	Much of open-source is written for a particular application and is not carried through to a reusable library or tool. Incentives + funding needed.
0	The evolution of ML framework changes rapidly. That means that any software will be old rapidly too.
0	Software support for broken user code not sustainable with current business model. Difficult to clearly set expectations for supported features.

Table 27. Potential Solution for Software Development and Sustainability

Votes	Content
10	Sustainability Models for Open-Source
3	Need for ongoing support mechanisms for software, not just initial development of open-source deliverable.
3	More clear support for software projects LAMMPS, Bluesky , etc. instead of funding them initially through science based proposals
3	Build on existing software solutions! – Widely adopted standards – Curated list of software – Fostering collaboration – Educate funding agencies on steady software support – Career advancement for software devs
3	Simplify Use: Incentivize use of code, which encourages reuse for new tech and encourages user-focused design
0	Remove administrative preferences for software types which are not compatible or not relevant to materials problems
0	The frameworks at the present are enough to start an AI application

Workshop participants noted that the current state of the machine learning software for XRD applications consists of numerous open-source code packages that might be implemented in hundreds of example code notebooks. The open-source nature of these code bases does have the strong advantage of being transparent in the analysis steps, imparting traceability and reproducibility of results, and allowing for open review and correction of the code. However, the decentralized state of open-source code bases for XRD analysis does raise concerns for industrial applications, namely, the discovery, stability, and responsibility of these code bases. We note that a recent strategic planning workshop hosted by the MGI Foundation[57] placed a specific emphasis on open-source software[58].

Sustaining Scientific Code — Workshop participants noted that open-source code bases are decentralized and disparately distributed. It can be difficult to discover the relevant algorithms with the latest updates, corrections, or improvements. One group may develop the initial algorithm and host it on a code repository, and an entirely separate group may modify that code and then host it on their own repository. Overall, there is a lack of a sustainability model for most open-source software within this community. Many tools are developed by graduate students or postdocs and not necessarily maintained in the long-term. They noted the need for a business model that is capable of elevating select algorithms developed in academia into robust and sustainable software packages. They noted that instrument vendors may have difficulty traversing the current state of this software landscape, thus hindering their ability to implement it with vendor's products.

Managing Dependencies — Workshop participants also discussed concerns about the sustainability of software dependencies. Any particular open-source code base for XRD analysis may be built on a foundation of other open-source dependencies. These dependencies can be many layers deep. If one of the open-source dependencies implements a change to a feature, that update may have unintended consequences as the change propagates through the chain of packages that depend on it. As such it can be daunting for industrial applications to rely on such open-source packages without assurances of the stability of the foundation. Modern version tracking and code containers can go some way to improve this situation.

Containerization — Code containers allow for the exact version state of all of the dependencies to be not only saved and archived but also allow for the distribution and implementation of that code environment. These code containers, therefore, enable reproducibility and longer-term stability of open-source algorithms. However, code containers do not guarantee the long-term support of the code base. For example, one dependency may implement a major change that requires a significant effort to re-write the packages that rely on it. While code containers archive the state of the code environment, freezing it in a state that will function long-term, they do not address the effort required to stay up-to-date as improvements or corrections are made or future compatibility. The same could also be said of proprietary code development, it is perhaps more pronounced in the open-source community.

Liability Concerns — Workshop participants expressed concerns about who is responsible for the code packages. Industrial instrument vendors have a liability to their customers to ensure that any analysis software they provide produces accurate results. With the advent and increasing popularity of machine learning algorithms, and deep learning algorithms in particular, there are few guarantees of the performance. It can sometimes be obfuscated as to how a particular deep learning model was trained – what data was used for training, for testing, and what training protocol was used. Even if the training data is provided, it can be a non-trivial task just to determine if a new point is within the training distribution and, therefore, the model is interpolating or outside that distribution where the model must extrapolate. Deep learning models are notoriously poor at extrapolating and can provide spurious predictions outside of the training set distribution. Furthermore, obtaining meaningful uncertainties on the predictions of deep learning models is an area of active research that has not seen wide adoption in the materials community. Instrument vendors are, therefore, justifiably hesitant to implement such open-source deep learning algorithms if they cannot assure their end-users that the code will not confidently make incorrect predictions.

Pairing this concern over liability with the industrial incentive to maintain a competitive advantage makes instrument vendors reluctant to provide external programmatic control (e.g., APIs) of their instruments. Enabling external programmatic control might open new liabilities (or even the perception of liabilities) were anything to go awry downstream of that programmatic control. Additionally, if software that interacts with their API changes in a way that makes API communications fail, they could be liable to deal with such failures, even if just to maintain good customer relations. Furthermore, instrument vendors have historically sought a competitive advantage not only from the hardware of the instrument but also through in-house on-board proprietary analysis software. That incentive is at odds with end-users seeking to autonomize their laboratories – end users that may want to implement their own analysis algorithms, trained on their own databases, to solve their particular research problems.

Paths Forward for Scientific Software — There are several pathways that could be followed to address each of these concerns. For example, the community might focus on developing a suite of tests to score the performance of new algorithms and models under different use conditions. For exmaple, an algorithm that seeks to predict the space group of a material from a powder XRD pattern might be tested on materials with space groups that were excluded from the training set. This would aid in the quantification of the accuracy and robustness of the predictions by the models. Scored algorithms could then be hosted on a curated repository. To ensure robustness, this repository should also host not only the training and test data but also archive and communicate the exact training protocols that were used. When also paired with code containers this repository would also enable the reproducibility and longevity of the algorithms. Such curation and scoring might encourage the popularity of the models, which in turn might encourage their long-term support. Developing such a repository would require the input and consensus of the XRD community and could, therefore, be the output of a working group of a consortium in the field.

3.5.3. Culture Change, Intellectual Property, and Economic Considerations

The economic and intellectual property landscape of the materials science and engineering community makes the development of a sustainable ecosystem of interoperable solutions challenging. The use of intellectual property rights associated with software and data formats generated by commercial instrument vendors makes implementation of these tools and data into autonomous pipelines challenging. Conversely, end-users are reluctant to share their proprietary data to improve analysis models. To overcome these barriers workshop participants highlighted the need for a clear value proposition for instrument vendors and other service providers to move towards more interoperable and open models. Workshop participants highlighted a divide between industry and academia in the incentives and constraints around adopting new methodologies and the publication of data. The cards within this category are shown in Tables 28-30.

Table 28. Goal note cards for Culture Change, Intellectual Property, and Economic Considerations

Votes	Content
4	What criteria should factor into the decisions on automation level? Need a quantitative framework for investment.
4	Work within government regulations
3	Identify decision-making points -> are there points in a process that one wouldn't want automated?
2	Utilization of data for plant costs combining results & characteristics of final product. Cost / benefit modeling with upfront
1	Cost benefit of AI/ML vs current state
1	ML Systems can have high upfront data costs so good communication between technical staff and decision-makers is needed
0	What level of utilization would make it worth investing in-house or using a user facility?
0	Improve Compliance: – Enabling social change to encourage adoption – Simplifying use – easier to comply than not

Table 29. Challenge note cards for Culture Change, Intellectual Property, and Economic Considerations

Votes	Content
4	Data plumbing and DevOps is an expensive distraction
4	Modes of collaboration to bring AI/ML to sensitive data
4	IT Security is a concern for industry data
4	Better incentive structure need for service-oriented infrastructure development
3	No one is funded to make good data
3	Regulatory + SOP inertia. Need a strategy to validate, verify, and accept new methods if they are superior.
2	Human experts are a challenge to AI decision until the learn how to cooperate
2	Barrier motivation for adoption, proper usage, and maintenance of new software and systems
2	Incremental Adoption: Technology evolves more quickly than the applications they are meant to address. Mapping/porting legacy data, analysis, is impossible
2	Synchrotron-Benchtop Divide, Academia-Industry Divide. Industry does not have synchrotron access.
2	Autonomous labs are expensive. How to get economy of scale and lower cost?
2	Sustainable growth in self-driving laboratories
1	Lack of Customer Demand
1	Standard, e.g. ASTM, prohibit new methods being adopted
1	How do we optimize the automation we have to compete with AI
0	How to maintain consistent and comparable data with new software / given customer needs / demands (/ methods)
0	Internal SOPs in industry make effort to adopt new methods prohibitive
0	Non-technical: ID administration and regulatory restrictions
0	Customer-focused development model is needed to advance data and operation in a way that does not affect stability of day-to-day scientific operations
0	Federal funding for both labs, university, and industry should include useful, service-oriented work in the support of science. At present, we have too many large innovative projects with low impact.
0	Need to work around restrictions on the use of programs, apps, languages that store and use proprietary or confidential data
0	How long should a measurement be for it to be worthwhile to automate?
0	High risk for tool vendors: Accuracy and Safety
0	Risk Management for vendors and for end-users
0	Liability for AI analysis results
0	Safety Risks: Radiation and Mechanical Movement
0	Inflated Expectations for AI
0	Less than 1 % of customers need a flexible programming interface. More demand is needed to make OEM investment viable
0	Balance between: Open-Source Software may not correct for certain artifacts and Proprietary Software may obfuscate raw data
0	Large Facility (e.g. EPICS, Bluesky) vs. Small Lab (e.g., Bespoke API)

Table 30. Potential Solution for Culture Change, Intellectual Property, and Economic Considerations

Votes	Content
9	IP (intellectual property) rights are a major blocker for industry collaboration with universities.
6	Encourage funding for translational research collaborating with industry and academia
5	Academia lacks ideas about business basics, and a sense of what topics are of general interest.
5	User Facility Model for Autonomous Systems
3	Identify the customer: Who is going to use the results and what are they trying to accomplish?
3	Make Computational Materials Science a commercially viable entity – what is the value proposition?
2	Sell software integration with machine automation
0	Basic infrastructure work is really important and needs dedicated funding and support. Not all funding should be focused on step changes.
0	Volume H and other resources exist – make them more approachable and advocate for practitioners to use it
0	Understand Needs in the Smart Lab

Value Proposition — Workshop participants indicated that industry stakeholders must see a clear value proposition and path to a return on investment if data-driven and autonomous methods are to be adopted. This requirement makes adoption quite challenging as multiple actors in the supply chain contribute to the development and optimization of a new material. When a company is developing a new material, they often rely on: (1) laboratory instruments produced by another company, (2) data management systems produced by another company, and (3) AI/ML platforms produced by another company. This means the adoption of autonomous or on-the-fly data-driven methodologies requires cooperation and standards across many actors within the supply chain.

Academic-Industry Divide — Workshop participants also noted a number of issues causing a divide between academia (where novel data-driven methodologies are often developed) and industry where they will ultimately be used to solve real-world problems. Industry and academia are driven by different incentive structures, which can inhibit technology transfer. The software aspects of this challenge are discussed in the next section. There is also a divide in hardware where academia and national labs have the funds to build and assemble customized laboratory equipment. The research and development involved in creating autonomous laboratories in academia and national laboratories is not necessarily lowering the costs and barriers to entry for industry to do the same. Thus, industry would be faced with the same startup costs. Workshop participants highlighted a desire for economies of scale and lessons learned in hardware development to be broadly shared.

Intellectual Property Issues — Some of these issues can be addressed if industry collaborates directly with academia, but workshop participants noted that intellectual property (IP) concerns are a major issue when setting up collaborations with academic institutions. Academia's mission to generate public knowledge and publish in scholarly journals is at odds with industry's mission to have a competitive advantage. Some participants noted that funding for transitional research can be helpful while others suggested a user facility model. The user facility model could also help academic and industrial researchers clearly separate the concerns of public contributions (e.g., open-source software and equipment designs) from proprietary problems (e.g., materials discovery for specific companies), where open-source platforms can also be used for both proprietary and open science. Workshop participants also noted regulatory compliance and IT security requirements, which may be less cumbersome in academia.

3.5.4. Convening and Coordination

Workshop participants highlighted the need for new convening and coordinating efforts, such as working groups and consortia. Expanded participation in working groups will be an important mechanism to make progress on specific opportunities identified during the workshop. The cards within this category are shown in Tables 31-33.

Table 31. Goal note cards for Convening and Coordination

Votes	Content
4	Wholistic adoption / integration
3	Data Consumption: Who needs the results?
0	Incentive mechanism engagement of community/data owner to share data

Table 32. Challenge note cards for Convening and Coordination

Votes	Content
3	Involvement of instrument vendors in discussion on programs and standards
3	Standardization of interfaces requires cooperation between instrument manufactures and users
3	Need community engagement around existing standards (CIF) to extend it for new functionality
3	Communication across fields (HTE, instrumentation, industry) – educational content – conference symposia
3	How to build multidisciplinary teams, need AI, HPC, UX, Domain (science) experts
2	End-users are not connected to industry needs or artifacts
2	Need to introduce industry, in particular, manufacturers who cater to industry, to programs, apps, and languages that will work with their software.
1	Too much fragmentation and parallel development – it is not possible to keep up with all the different frameworks and tools
1	Data and automation infrastructure development needs to be organized as a community level effort to avoid unsustainable duplication of work. Needs to integrate existing technology (Bluesky)
0	Every customer having their own particular solution is the worst possible situation for instrument vendors

Table 33. Potential Solution for Convening and Coordination

Votes	Content
10	Consortia for databases + models
3	Increase collaboration between manufacturers and developers
3	Form dedicated working groups for priority problems
1	Look at the OME & OPTIMADE example of vendor participation
1	Hackathons for dark data
1	Hackathons & Coordination for Software Projects
0	Larger workshop with more practitioners
0	New Working Groups for Data/Control Interface Standards

Consortia — Workshop participants expressed considerable interest in forming a consortia. While details were not discussed, there are many successful examples of public-private cooperation. On a smaller scale, examples such as the NIST Flow Cytometry Standards Consortium show how consortia can be focused around a measurement technique. On a larger scale, the numerous Manufacturing USA institutes have tackled a number of important problems. Several cards in various categories mentioned funding and perhaps some funding will be required to overcome barriers that have plagued the community for some time.

Gaining Momentum — Workshop participants enthusiastically supported this workshop, which highlighted the critical role of instrument vendors. Where many previous workshops overlooked this stakeholder, this workshop placed them centerstage. Workshop participants suggested that NIST and ICDD continue and expand communication and collaboration between instrument vendors and AI developers.

New Working Groups — Workshop participants highlighted the need to establish new working groups. This was highlighted in general terms and specifically to address FAIR for data, models, and tools. In the domain of experimental data, examples such as the Open Microscopy Environment (OME) data model and the NeXus Data Format for neutron, X-ray, and muon science are successful examples.

4. Summary and Outlook

This report presents the findings of the Autonomous Methodologies for Accelerating X-ray Measurements Workshop, which was hosted on 17-18 October 2023 at ICDD headquarters in Newtown Square, Pennsylvania. The facilitated sessions generated 322 cards, which consisted of 83 goals, 164 challenges, and 75 possible solutions.

All 322 cards were listed in Tables 4-33, which were interpreted through the context of: (1) data and metadata, (2) physical infrastructure, (3) algorithm and model development, and (4) community engagement. The most up-voted cards listed in Tables 1-3 demonstrated a clear consensus on the most important challenges and opportunities to address in the community. In order to address these, stakeholders within industry, government, and academia should establish new partnerships and devise actionable plans going forward. Our interpretation of the findings of this workshop suggests that action plans should be developed for the following future desired outcomes within the community:

1. **FAIRness of XRD and other Experimental Data** — The community achieves wide adoption of a consensus-based specification for data and metadata. This consensus-based specification provides a rich plurality of metadata fields to enable wide reuse of data and metadata. This consensus-based specification enjoys wide adoption across instrument vendors and software providers.

2. **Programmatic Control and Administration of XRD and other Experimental Equipment** — There is wide adoption of a consensus-based specification for protocols and semantics for the programmatic control and administration of experimental instruments. This consensus-based specification provides an extendable mechanism for instruments to enumerate and describe the available operations and parameters for controlling and administering the instrument. This consensus-based specification enjoys wide adoption across instrument vendors and software providers.

3. **Comprehensive Datasets for Phase Identification and Quantitative Analysis** — Available data for training data-driven methodologies covers far more real-world use cases. For example, labeled data would cover situations of diverse measurement quality, sample preparation quality, and diverse chemistry and processing conditions.

4. **Robust Tools for Phase Identification, Quantitative Analysis, and Autonomous Laboratories** — Operators of autonomous laboratories and other laboratory equipment have the ability to leverage a plethora of robust and well-documented tools for autonomous laboratories, phase identification, quantitative analysis, and other laboratory activities. Open-source and for-profit providers coexist in delivering and supporting high-quality software.

5. **Vibrant Marketplace for Autonomous Laboratory Equipment and Services** — It is possible to affordably procure a wide variety of equipment to build an autonomous

laboratory. There are a number of industry-lead standards that enable diverse components from different vendors to easily plug and play.

6. **Workforce Equipped for the Autonomous Laboratory** — Next-generation and current-generation workforce have a plethora of mechanisms to acquire skills and expertise in leveraging autonomous labs and data-driven methodologies in the laboratory.

We hope this report stimulates action within the community. We recommend that the community organize future workshops to develop consensus-based tactical action plans for the desired states described above. Many of these issues are complex and longstanding and cannot be solved without broad perspectives and participation from all applicable sectors of industry, government, and academia.

References

[1] *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security* (National Academies Press). https://doi.org/10.17226/12199. Available at http://www.nap.edu/catalog/12199

[2] Materials genome initiative. Available at https://www.mgi.gov/.

[3] Morgan D, Jacobs R Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research* 50(1):71–103. https://doi.org/10.1146/annurev-matsci-070218-010015. Available at https://www.annualreviews.org/doi/10.1146/annurev-matsci-070218-010015

[4] Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocarsly JD, Bran AM, Bringuier S, Brinson LC, Choudhary K, Circi D, Cox S, De Jong WA, Evans ML, Gastellu N, Genzling J, Gil MV, Gupta AK, Hong Z, Imran A, Kruschwitz S, Labarre A, Lála J, Liu T, Ma S, Majumdar S, Merz GW, Moitessier N, Moubarak E, Mouriño B, Pelkie B, Pieler M, Ramos MC, Ranković B, Rodriques SG, Sanders JN, Schwaller P, Schwarting M, Shi J, Smit B, Smith BE, Van Herck J, Völker C, Ward L, Warren S, Weiser B, Zhang S, Zhang X, Zia GA, Scourtas A, Schmidt KJ, Foster I, White AD, Blaiszik B 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery* 2(5):1233–1250. https://doi.org/10.1039/D3DD00113J. Available at http://xlink.rsc.org/?DOI=D3DD00113J

[5] Boiko DA, MacKnight R, Kline B, Gomes G Autonomous chemical research with large language models. *Nature* 624(7992):570–578. https://doi.org/10.1038/s41586-023-06792-0. Available at https://www.nature.com/articles/s41586-023-06792-0

[6] Nikolaev P, Hooper D, Webber F, Rao R, Decker K, Krein M, Poleski J, Barto R, Maruyama B Autonomy in materials research: a case study in carbon nanotube growth. *npj Computational Materials* 2(1):16031. https://doi.org/10.1038/npjcompumats.2016.31. Available at https://www.nature.com/articles/npjcompumats201631

[7] Rahmanian F, Flowers J, Guevarra D, Richter M, Fichtner M, Gregoire J, Stein HS Enabling modular autonomous feedback-loops in materials science through hierarchical experimental laboratory automation and orchestration. https://doi.org/10.26434/chemrxiv-2021-kr87t. Available at https://chemrxiv.org/engage/chemrxiv/article-details/6165f62235b406a76a11cbd5

[8] Sim M, Ghazi Vakili M, Strieth-Kalthoff F, Hao H, Hickman R, Miret S, Pablo-García S, Aspuru-Guzik A ChemOS 2.0: an orchestration architecture for chemical self-driving laboratories. https://doi.org/10.26434/chemrxiv-2023-v2khf. Available at https://chemrxiv.org/engage/chemrxiv/article-details/64cbe80adfabaf06ffa61204

[9] Data and AI-driven materials science group. Available at https://www.nist.gov/mml/mmsd/data-and-ai-driven-materials-science-group.

[10] DeCost BL, Lei B, Francis T, Holm EA High throughput quantitative metallography for complex microstructures using deep learning: A case study in ultrahigh carbon steel. *Microscopy and Microanalysis* 25(1):21–29. https://doi.org/10.1017/S1431927618015635. Available at https://academic.oup.com/mam/article/25/1/21/6887488

[11] Green ML, Choi CL, Hattrick-Simpers JR, Joshi AM, Takeuchi I, Barron SC, Campo E, Chiang T, Empedocles S, Gregoire JM, Kusne AG, Martin J, Mehta A, Persson K, Trautt Z, Van Duren J, Zakutayev A Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews* 4(1):011105. https://doi.org/10.1063/1.4977487. Available at https://pubs.aip.org/apr/article/4/1/011105/123789/Fulfilling-the-promise-of-the-materials-genome

[12] Hattrick-Simpers JR, Gregoire JM, Kusne AG Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge. *APL Materials* 4(5):053211. https://doi.org/10.1063/1.4950995. Available at https://pubs.aip.org/apm/article/4/5/053211/121536/Perspective-Composition-structure-property-mapping

[13] Hattrick-Simpers JR, Zakutayev A, Barron SC, Trautt ZT, Nguyen N, Choudhary K, DeCost B, Phillips C, Kusne AG, Yi F, Mehta A, Takeuchi I, Perkins JD, Green ML An interlaboratory study of zn–sn–ti–o thin films using high-throughput experimental methods. *ACS Combinatorial Science* 21(5):350–361. https://doi.org/10.1021/acscombsci.8b00158. Available at https://pubs.acs.org/doi/10.1021/acscombsci.8b00158

[14] Joress H, DeCost BL, Sarker S, Braun TM, Jilani S, Smith R, Ward L, Laws KJ, Mehta A, Hattrick-Simpers JR A high-throughput structural and electrochemical study of metallic glass formation in ni–ti–al. *ACS Combinatorial Science* 22(7):330–338. https://doi.org/10.1021/acscombsci.9b00215. Available at https://pubs.acs.org/doi/10.1021/acscombsci.9b00215

[15] Joress H, Green ML, Takeuchi I, Hattrick-Simpers JR Applications of high throughput (combinatorial) methodologies to electronic, magnetic, structural, and energy-related materials. *Encyclopedia of Materials: Metals and Alloys* (Elsevier), pp 353–371. https://doi.org/10.1016/B978-0-12-819726-4.00146-0. Available at https://linkinghub.elsevier.com/retrieve/pii/B9780128197264001460

[16] Kusne AG, Keller D, Anderson A, Zaban A, Takeuchi I High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* 26(44):444002. https://doi.org/10.1088/0957-4484/26/44/444002. Available at https://iopscience.iop.org/article/10.1088/0957-4484/26/44/444002

[17] Weaver JS, Pintar AL, Beauchamp C, Joress H, Moon KW, Phan TQ Demonstration of a laser powder bed fusion combinatorial sample for high-throughput microstructure and indentation characterization. *Materials & Design* 209:109969. https://doi.org/10.1016/j.matdes.2021.109969. Available at https://linkinghub.elsevier.com/retrieve/pii/S0264127521005232

[18] Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, Park CW, Choudhary A, Agrawal A, Billinge SJL, Holm E, Ong SP, Wolverton C Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* 8(1):59. https://doi.org/10.1038/s41524-022-00734-6. Available at https://www.nature.com/articles/s41524-022-00734-6

[19] Hattrick-Simpers JR, DeCost B, Kusne AG, Joress H, Wong-Ng W, Kaiser DL, Zakutayev A, Phillips C, Sun S, Thapa J, Yu H, Takeuchi I, Buonassisi T An open combinato-

rial diffraction dataset including consensus human and machine learning labels with quantified uncertainty for training new machine learning models. *Integrating Materials and Manufacturing Innovation* 10(2):311–318. https://doi.org/10.1007/s40192-021-00213-8. Available at https://link.springer.com/10.1007/s40192-021-00213-8

[20] Kusne AG, Gao T, Mehta A, Ke L, Nguyen MC, Ho KM, Antropov V, Wang CZ, Kramer MJ, Long C, Takeuchi I On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Scientific Reports* 4(1):6367. https://doi.org/10.1038/srep06367. Available at https://www.nature.com/articles/srep06367

[21] Jha D, Kusne AG, Al-Bahrani R, Nguyen N, Liao Wk, Choudhary A, Agrawal A Peak area detection network for directly learning phase regions from raw x-ray diffraction patterns. *2019 International Joint Conference on Neural Networks (IJCNN)* (IEEE), pp 1–8. https://doi.org/10.1109/IJCNN.2019.8852096. Available at https://ieeexplore.ieee.org/document/8852096/

[22] Lolla S, Liang H, Kusne AG, Takeuchi I, Ratcliff W A semi-supervised deep-learning approach for automatic crystal structure classification. *Journal of Applied Crystallography* 55(4):882–889. https://doi.org/10.1107/S1600576722006069. Available at https://scripts.iucr.org/cgi-bin/paper?S1600576722006069

[23] Choudhary K, DeCost B, Tavazza F Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Physical Review Materials* 2(8):083801. https://doi.org/10.1103/PhysRevMaterials.2.083801. Available at https://link.aps.org/doi/10.1103/PhysRevMaterials.2.083801

[24] Stanev V, Oses C, Kusne AG, Rodriguez E, Paglione J, Curtarolo S, Takeuchi I Machine learning modeling of superconducting critical temperature. *npj Computational Materials* 4(1):29. https://doi.org/10.1038/s41524-018-0085-8. Available at https://www.nature.com/articles/s41524-018-0085-8

[25] Wang A, Liang H, McDannald A, Takeuchi I, Kusne AG Benchmarking active learning strategies for materials optimization and discovery. *Oxford Open Materials Science* 2(1):itac006. https://doi.org/10.1093/oxfmat/itac006. Available at https://academic.oup.com/ooms/article/doi/10.1093/oxfmat/itac006/6637521

[26] Audus DJ, McDannald A, DeCost B Leveraging theory for enhanced machine learning. *ACS Macro Letters* 11(9):1117–1122. https://doi.org/10.1021/acsmacrolett.2c00369. Available at https://pubs.acs.org/doi/10.1021/acsmacrolett.2c00369

[27] Kusne AG, McDannald A, DeCost B, Oses C, Toher C, Curtarolo S, Mehta A, Takeuchi I Physics in the machine: Integrating physical knowledge in autonomous phase-mapping. *Frontiers in Physics* 10:815863. https://doi.org/10.3389/fphy.2022.815863. Available at https://www.frontiersin.org/articles/10.3389/fphy.2022.815863/full

[28] DeCost B, Joress H, Sarker S, Mehta A, Hattrick-Simpers J Towards automated design of corrosion resistant alloy coatings with an autonomous scanning droplet cell. *JOM* 74(8):2941–2950. https://doi.org/10.1007/s11837-022-05367-0. Available at https://link.springer.com/10.1007/s11837-022-05367-0

[29] Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, Sarker S, Oses C, Toher C, Curtarolo S, Davydov AV, Agarwal R, Bendersky LA, Li M, Mehta A, Takeuchi I On-the-fly closed-loop materials discovery via bayesian active learning. *Nature Communications* 11(1):5966. https://doi.org/10.1038/s41467-020-19597-w. Available at https://www.nature.com/articles/s41467-020-19597-w

[30] Tavazza F, DeCost B, Choudhary K Uncertainty prediction for machine learning models of material properties. *ACS Omega* 6(48):32431–32440. https://doi.org/10.1021/acsomega.1c03752. Available at https://pubs.acs.org/doi/10.1021/acsomega.1c03752

[31] Dima A, Bhaskarla S, Becker C, Brady M, Campbell C, Dessauw P, Hanisch R, Kattner U, Kroenlein K, Newrock M, Peskin A, Plante R, Li SY, Rigodiat PF, Amaral GS, Trautt Z, Schmitt X, Warren J, Youssef S Informatics infrastructure for the materials genome initiative. *JOM* 68(8):2053–2064. https://doi.org/10.1007/s11837-016-2000-4. Available at http://link.springer.com/10.1007/s11837-016-2000-4

[32] Greene G, Ragland J, Trautt Z, Lau J, Plante R, Taillon J, Creuziger A, Becker C, Bennett J, Blonder N, Borsuk L, Campbell C, Friss A, Hale L, Halter M, Hanisch R, Hardin G, Levine L, Maragh S, Miller S, Muzny C, Newrock M, Perkins J, Plant A, Ravel B, Ross D, Scott JH, Szakal C, Tona A, Vallone P A roadmap for LIMS at NIST material measurement laboratory. https://doi.org/10.6028/NIST.TN.2216. Available at https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2216.pdf

[33] Kaiser DL, Hanisch RJ, Warren JA, Trautt Z Materials data : a landscape analysis and potential roadmap for the NIST material measurement laboratory. https://doi.org/10.6028/NIST.IR.8364. Available at https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8364.pdf

[34] Riccardi D, Trautt Z, Bazyleva A, Paulechka E, Diky V, Magee JW, Kazakov AF, Townsend SA, Muzny CD Towards improved FAIRness of the ThermoML archive. *Journal of Computational Chemistry* 43(12):879–887. https://doi.org/10.1002/jcc.26842. Available at https://onlinelibrary.wiley.com/doi/10.1002/jcc.26842

[35] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, ’T Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Van Der Lei J, Van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3(1):160018. https://doi.org/10.1038/sdata.2016.18. Available at https://www.nature.com/articles/sdata201618

[36] FAIR for research software (FAIR4rs) working group. Available at https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg/.

[37] FAIR principles for research hardware interest group. Available at https://www.rd-alliance.org/groups/fair-principles-research-hardware/.

[38] FAIR for machine learning (FAIR4ml) interest group. Available at https://www.rd-alliance.org/groups/fair-machine-learning-fair4ml-ig/.

[39] Hall SR, Allen FH, Brown ID The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A Foundations of Crystallography* 47(6):655–685. https://doi.org/10.1107/S010876739101067X. Available at https://scripts.iucr.org/cgi-bin/paper?S010876739101067X

[40] Toby BH Classification and use of powder diffraction data. *International Tables for Crystallography*, eds Hall SR, McMahon B (International Union of Crystallography), Vol. G, pp 117–130. https://doi.org/10.1107/97809553602060000735. Series Title: International Tables for Crystallography Available at https://xrpp.iucr.org/cgi-bin/itr?url_ver=Z39.88-2003&rft_dat=what%3Dchapter%26volid%3DGa%26chnumo%3D3o3%26chvers%3Dv0001

[41] Könnecke M, Akeroyd FA, Bernstein HJ, Brewster AS, Campbell SI, Clausen B, Cottrell S, Hoffmann JU, Jemian PR, Männicke D, Osborn R, Peterson PF, Richter T, Suzuki J, Watts B, Wintersberger E, Wuttke J The NeXus data format. *Journal of Applied Crystallography* 48(1):301–305. https://doi.org/10.1107/S1600576714027575. Available at https://scripts.iucr.org/cgi-bin/paper?S1600576714027575

[42] Andersen CW, Armiento R, Blokhin E, Conduit GJ, Dwaraknath S, Evans ML, Fekete A, Gopakumar A, Grazulis S, Merkys A, Mohamed F, Oses C, Pizzi G, Rignanese GM, Scheidgen M, Talirz L, Toher C, Winston D, Aversa R, Choudhary K, Colinet P, Curtarolo S, Di Stefano D, Draxl C, Er S, Esters M, Fornari M, Giantomassi M, Govoni M, Hautier G, Hegde V, Horton MK, Huck P, Huhs G, Hummelshoj J, Kariryaa A, Kozinsky B, Kumbhar S, Liu M, Marzari N, Morris AJ, Mostofi AA, Persson KA, Petretto G, Purcell T, Ricci F, Rose F, Scheffler M, Speckhard D, Uhrin M, Vaitkus A, Villars P, Waroquiers D, Wolverton C, Wu M, Yang X (2021) OPTIMADE, an API for exchanging materials data. *Scientific Data* 8(1). https://doi.org/10.1038/s41597-021-00974-z. Available at http://dx.doi.org/10.1038/s41597-021-00974-z

[43] Awesome materials informatics. Available at https://github.com/tilde-lab/awesome-materials-informatics.

[44] The Minerals, Metals & Materials Society Building a materials data infrastructure: Opening new pathways to discovery and innovation in science and engineering. https://doi.org/10.7449/mdistudy_1. Available at http://www.tms.org/mdistudy

[45] Bahim C, Casorrán-Amilburu C, Dekkers M, Herczog E, Loozen N, Repanas K, Russell K, Stall S The FAIR data maturity model: An approach to harmonise FAIR assessments. *Data Science Journal* 19:41. https://doi.org/10.5334/dsj-2020-041. Available at http://datascience.codata.org/articles/10.5334/dsj-2020-041/

[46] Research Data Alliance FAIR Data Maturity Model Working Group FAIR data maturity model: specification and guidelines https://doi.org/10.15497/RDA00050. Publisher: [object Object] Version Number: 1 Available at https://zenodo.org/record/3909563#.YGRNnq8za70

[47] Workshop on artificial intelligence applied to materials discovery and design. Available at https://www.energy.gov/sites/prod/files/2018/03/f49/AI%20Applied%20t

o%20Materials%20Discovery%20and%20Design_Workshop%20Summary%20Report.pdf.

[48] SiLA rapid integration. Available at https://sila-standard.com/.

[49] MQTT: The standard for IoT messaging. Available at https://mqtt.org/.

[50] Elias JR, Chard R, Libera JA, Foster I, Chaudhuri S The manufacturing data and machine learning platform: Enabling real-time monitoring and control of scientific experiments via IoT https://doi.org/10.48550/ARXIV.2005.13669. Publisher: [object Object] Version Number: 1 Available at https://arxiv.org/abs/2005.13669

[51] Apache kafka. Available at https://kafka.apache.org/.

[52] Eminizer M, Tabrisky S, Sharifzadeh A, DiMarco C, Diamond JM, Ramesh KT, Hufnagel TC, McQueen TM, Elbert D OpenMSIStream: A python package for facilitating-integration of streaming data in diverse laboratory environments. *Journal of Open Source Software* 8(83):4896. https://doi.org/10.21105/joss.04896. Available at https://joss.theoj.org/papers/10.21105/joss.04896

[53] Joress H, Trautt Z, McDannald A, DeCost B, Kusne AG, Tavazza F (2024) Driving U.S. innovation in materials and manufacturing using AI and autonomous labs. https://doi.org/https://doi.org/10.6028/NIST.SP.1320. Available at https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=958246

[54] Psaros AF, Meng X, Zou Z, Guo L, Karniadakis GE Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics* 477:111902. https://doi.org/10.1016/j.jcp.2022.111902. Available at https://linkinghub.elsevier.com/retrieve/pii/S0021999122009652

[55] The Minerals, Metals & Materials Society Creating the next-generation materials genome initiative workforce. https://doi.org/10.7449/mgiworkforce_1. Available at https://www.tms.org/mgiworkforce

[56] The Minerals, Metals & Materials Society Core knowledge and skills for effective use of advanced computation and data in materials and manufacturing. https://doi.org/10.7449/coreknowledge_1. Available at https://www.tms.org/coreknowledge/10.7449/coreknowledge_1

[57] Materials genome foundation. Available at https://www.materialsgenomefoundation.org/.

[58] Materials genome foundation: Strategic planning workshops. Available at https://github.com/materialsgenomefoundation/workshops-2023/.

Appendix A. Acronyms

AE — Autonomous Experimentation

AI — Artificial Intelligence

API — Application Programming Interface.

CEC — Cation Exchange Capacity

CIF — Crystallographic Information File

DOE — Department of Energy

ELN — Electronic Laboratory Notebook

ETL — Extract Transform Load

FAIR — Findable, Accessible, Interoperable, and Reusable

FTNIR — Fourier Transform Near-Infrared Spectroscopy

GIXRD — Grazing Incidence X-ray Diffraction

HEDM — High Energy X-ray Diffraction Microscopy

HPC — High Performance Computing

HTE — High Throughput Experimentation

HTP — High Throughput

ICDD — International Centre for Diffraction Data

ICME — Integrated Computational Materials Engineering

ID — Identification

IP — Intellectual Property

IT — Information Technology

IUCR — International Union of Crystallography

JSON — JavaScript Object Notation

JSON-LD — JavaScript Object Notation for Linked Data

KG — Knowledge Graph

LAMMPS — Large-scale Atomic/Molecular Massively Parallel Simulator

LLM — Large Language Model

MAE — Mean Absolute Error

MDI — Materials Data Infrastructure

MGI — Materials Genome Initiative

ML — Machine Learning

NIR — Near Infrared spectroscopy

NIST — National Institute of Standards and Technology

NMR — Nuclear Magnetic Resonance

OME — Open Microscopy Environment

OPTIMADE — Open Databases Integration for Materials Design

PONCKS — Partial Or No Known Crystal Structure

PXRD — Powder X-ray Diffraction

QA — Quality Assurance

QC — Quality Control

QPA — Quantitative Phase Analysis

REST — REpresentational State Transfer.

RIR — Reference Intensity Ratio

SOP — Standard Operating Procedure

SRM — Standard Reference Material

TEM — Transmission Electron Microscopy

UQ — Uncertainty Quantification

UX — User Experience

XCT — X-ray Computed Tomography

XML — Extensible Markup Language

XRD — X-ray Diffraction

XRR — X-ray Reflectivity

XRF — X-ray Fluorescence

XRPD — X-ray Powder Diffraction

Appendix B. Data Availability

Photographs of all easel boards and extracted note card data is available for download:

Trautt, Zachary T., McDannald, Austin S., DeCost, Brian L., Joress, Howie L. (2024), Workshop Data on Autonomous Methodologies for Accelerating X-ray Measurements, National Institute of Standards and Technology, https://doi.org/10.18434/mds2-3498