U.S. DEPARTMENT OF ENERGY

Office of ENERGY EFFICIENCY & RENEWABLE ENERGY

Advanced Manufacturing Office

Workshop on Artificial Intelligence Applied to Materials Discovery and Design

Workshop Summary Report

August 9–10, 2017

Pittsburgh, PA

Within the DOE Office of Energy Efficiency and Renewable Energy (EERE), the Advanced Manufacturing Office (AMO) partners with industry, small business, academia, and other stakeholders to identify and invest in emerging technologies with the potential to create high-quality domestic manufacturing jobs and enhance the global competitiveness of the United States.

This document was prepared for DOE/EERE's AMO as a collaborative effort by DOE AMO, Allegheny Science & Technology, and Energetics Incorporated.

Disclaimer

List of Acronyms

AI	Artificial intelligence
AM	Additive manufacturing
AMO	Advanced Manufacturing Office
API	Application programming interface
ASCR	Advanced Scientific Computing Research program of the U.S. Department of Energy
BER	Biological and Environmental Research program of the U.S. Department of Energy
BES	Basic Energy Sciences program of the U.S. Department of Energy
CALPHAD	Computer Coupling of Phase Diagrams and Thermochemistry
CESMII	Clean Energy Smart Manufacturing Innovation Institute
CIF	Crystallographic information file
CRADA	Cooperative research and development agreement
CS	Computer science
DFT	Density functional theory
DHS	U.S. Department of Homeland Security
DLMS	Direct laser metal sintering
DOD	U.S. Department of Defense
DOE	U.S. Department of Energy
DREAM.3D	Digital Representation Environment for Analyzing Microstructure in 3D
EBM	Electron beam melting
EERE	U.S. Department of Energy, Office of Energy Efficiency & Renewable Energy
FAIR	Find, access, interoperability, reuse
FOA	Funding opportunity announcement
HPC	High performance computing
GEGR	General Electric Global Research
HPC4Materials	High Performance Computing for Materials
HPC4Mfg	High Performance Computing for Manufacturing
I/UERC	Industry-University Cooperative Research Centers
ICME	Integrated computational materials engineering
IP	Intellectual property
MGI	Materials Genome Initiative
ML	Machine learning
MLMR	Materials learning for materials research
NIH	National Institutes of Health

NIP	No information provided
NIST	National Institute of Standards and Technology
NiTi	Nickel-titanium
NNMI	National Network for Manufacturing Innovation
NSF	National Science Foundation
ODE	Ordinary differential equation
PaaS	Platform as a service
PDE	Partial differential equation
PI	Principal investigator
R&D	Research and development
RD&D	Research, development and demonstration
RFI	Request for information
SEMATECH	A U.S. consortium to advance semiconductor manufacturing technologies
STEM	Science, technology, engineering, and mathematics
TRL	Technology readiness level
UQ	Uncertainty quantification
UTRC	United Technologies Research Center

Executive Summary

The Department of Energy's Advanced Manufacturing Office (AMO) held a Workshop on *Artificial Intelligence Applied to Materials Discovery and Design* in Pittsburgh, PA, in August 2017. The workshop brought together more than 100 leading scientific and technical experts to identify opportunities and technical challenges for early-stage applied research and development (R&D) that can advance the state of the art for artificial intelligence (AI) applied to materials design and discovery in energy and other related materials. The workshop covered three main topic areas: 1) Data Quantity and Quality; 2) Platforms and Infrastructure; and 3) AI for Materials Design in Specific Applications. Several themes that emerged throughout the workshop are interconnected and contribute toward the vision of developing AI-based tools for discovering and designing innovative new materials that can increase the competitiveness of U.S. manufacturing.

This workshop report summarizes the presentations, panel discussions, and breakout group discussions that took place at this event. Note that the results presented here are a snapshot of the viewpoints expressed by the experts who attended the workshop and do not necessarily reflect those of the broader materials development community.

Discussions at the workshop primarily explored issues that are common across diverse material types, such as data formatting, algorithms, models, and tools. In addition, a few industry participants shared their experiences in applying AI to alloy design for additive manufacturing and to the development of glasses and catalysts. Based on the common themes identified during the workshop, four high priority areas emerged:

Common Data Formatting and Quality: Common formatting for data was identified as a key challenge by all breakout groups. Participants suggested that any data generated comply with the FAIR principles for data management (i.e., data that is findable, accessible, interoperable, and reusable).[1] Data should encompass metadata and data uncertainty to ensure its usability. In addition to the common data format issues, challenges exist with respect to open source versus proprietary data. The government and academia should work with industry to develop new legal frameworks to make it easier for companies to share data with others.

Integrating Multi-Scale Models: AI will deliver the maximum benefit to materials discovery and design if and when multi-scale models can be integrated across multiple length and time scales. This integration includes the entire process from modeling and materials design through processing to manufacturing the advanced material. Challenges remain in defining how engineered materials will be integrated into these complex, feedstock-to-product models (e.g., dealing with material composites or compounds and groups of materials represented as systems but not as a single material). Ultimately, the goal is to use AI-based tools to quickly and efficiently design ever more complex systems involving multiple materials or hybrid materials and interfaces.

Public-Private Partnerships: Partnerships between government and the private sector have made impressive progress in materials design and are providing a continuous infusion of knowledge and data to this field. The federal government continues to fund the Materials Genome Initiative (MGI), which is developing an impressive set of tools and accompanying expertise in materials, AI, and machine learning (ML) to support high-throughput materials design. Programs at the national laboratories are critical. They offer high performance computing capabilities; user facilities for more sophisticated and expensive experimental techniques, such as light sources or neutron sources; and high-throughput tools and/or the capabilities to develop additional tools. Funding mechanisms need to enable industry and university researchers to lever these

[1] Mark D. Wilkinson, Michel Dumontier, and Barend Mons. "The FAIR Guiding Principles for scientific data management and stewardship" *Scientific Data* **3**, Article number: 160018 (2016) March 15, 2016. www.nature.com/articles/sdata201618

capabilities to make real progress in meeting the technical and scientific challenges identified during the workshop.

Cross-Discipline Education: Applying AI and ML methods to materials discovery and design is a multi-disciplinary effort requiring experts from many fields, including materials science, chemistry, mathematics, and computer science. Cross-disciplinary education and training are needed at the university and graduate levels; ideally, however, this education initiative should start earlier in grades K-12. Some public and private sector stakeholders now conduct boot camps on machine learning for materials research, but a larger, more organized effort is needed.

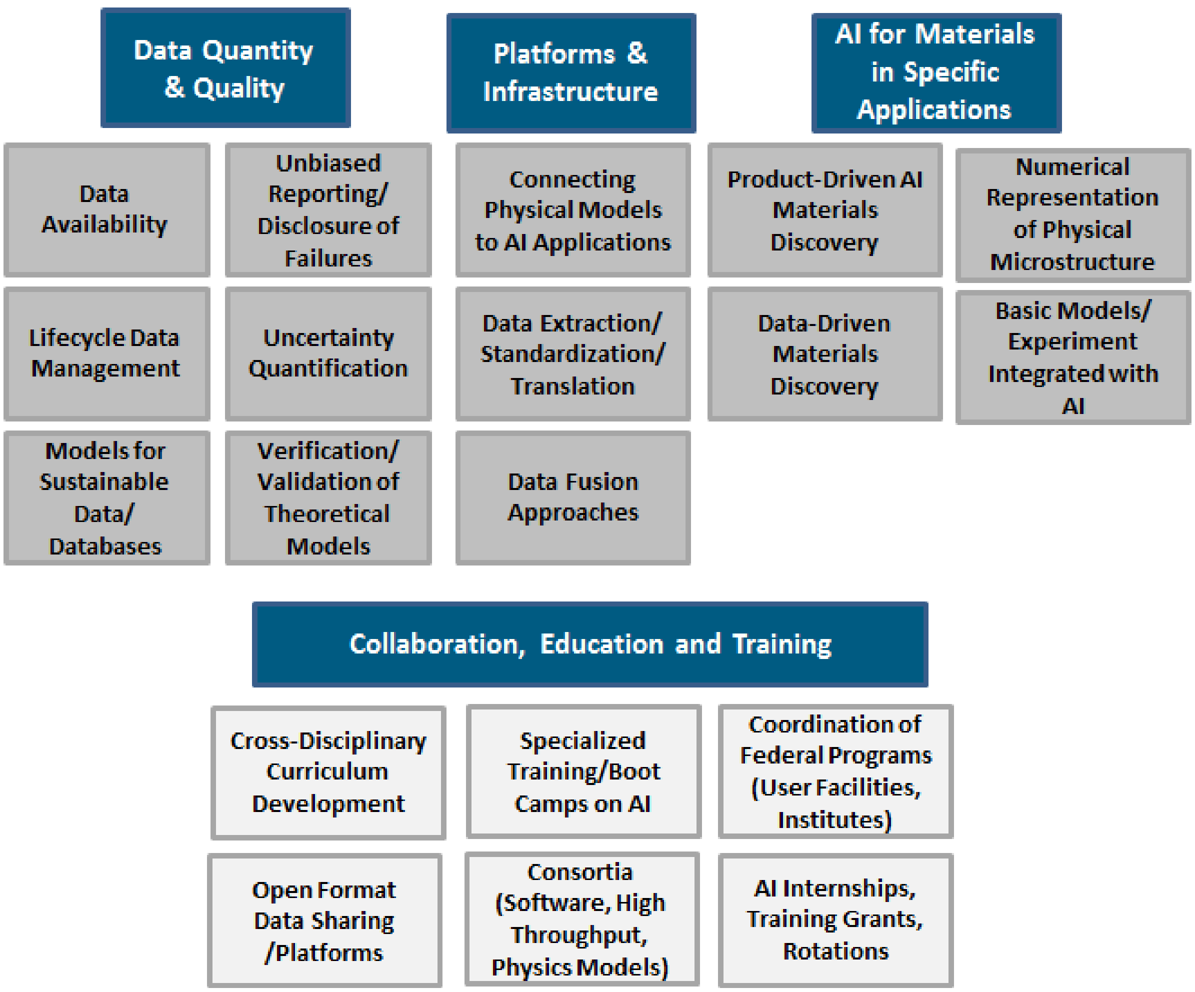

Figure E-1. Summary of Key Research and Non-Technical Priorities

Table of Contents

List of Figures

List of Tables

1 Background and Workshop Proceedings

Background

The demand for advanced materials to enable new energy technologies has exceeded the capabilities of traditional materials and chemical processes.[2] Accessibility to credible data on materials properties and behavior—and tools that can rapidly utilize this data—are critical to expedite the design and discovery of new materials which traditionally can take 10 to 20 years. The Materials Genome Initiative (MGI) for Global Competitiveness, launched in 2011, supports efforts throughout the federal government to accelerate materials discovery and design. The MGI *aims to provide the infrastructure and training that American innovators need to discover, develop, manufacture, and deploy advanced materials at least twice as fast as possible today, at a fraction of the cost.*[3,4] The initiative is enabling advances in scientific theory, modeling, simulation, high performance computing, algorithms, software, data analysis, and experimental techniques, all of which are merging to create the ability to design and engineer materials and systems more rapidly and at lower cost than traditional approaches.[5,6] To connect industry with resources in academia and the national laboratories earlier in the development process, the MGI supports broad and open access to advanced simulation tools, networks to share modeling and analysis code, access to quantitative synthesis and characterization tools, and digital data infrastructures that store a wide range of easily and reliably searchable data.[7,8]

Advances in research techniques, high performance computing, and the increasing availability of materials data is enabling significant progress in the design of innovative materials. Examples of material innovations include new, more efficient permanent magnets;[9] metals and alloys designed for additive manufacturing;[10] new electrocatalysts for energy uses;[11] and very low thermal hysteresis nickel-titanium (NiTi) based shape memory alloys.[12] Experts note progress in developing new materials using high-throughput experimental techniques and AI-based tools that can reduce materials development time by half. However, challenges still exist, and the materials design process remains relatively slow. Exponential improvements in information technology are creating new opportunities to modernize and expand materials research tools and data analytics to streamline design and discovery. For example, the ability of systems to generate and process massive amounts of data could allow for unprecedented changes in how data can be applied to intuitive, intelligent materials design.

Workshop Overview

To better understand the need for intelligent design of materials and the federal role in this area, the U.S. Department of Energy (DOE) held the *Workshop on Artificial Intelligence Applied to Materials Discovery and Design* on August 9–10, 2017. Representatives from industry, academia, the DOE national laboratories, and non-governmental organizations gathered in Pittsburgh, PA, to hear presentations by subject matter experts,

---

[2] Computational Materials Science and Chemistry: Accelerating Discovery and Innovation through Simulation-Based Engineering and Science, U.S. Department of Energy, Office of Science, July 2010, https://science.energy.gov/~/media/bes/pdf/reports/files/Computational_Materials_Science_and_Chemistry_rpt.pdf.

[3] Materials Genome Initiative for Global Competitiveness, National Science and Technology Council, June 2011, www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf

[4] Materials Genome Initiative Strategic Plan, National Science and Technology Council, Committee on Technology Subcommittee on the Materials Genome Initiative, December 2014, www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf

[5] Quadrennial Technology Review 2015, Chapter 6: Innovating Clean Energy Technologies in Advanced Manufacturing: Advanced Materials Manufacturing, U. S. Department of Energy, https://energy.gov/sites/prod/files/2016/04/f30/QTR2015-6B-Advanced-Materials-Manufacturing.pdf

[6] Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies, M.L. Green, et al., Applied Physics Reviews, 4, 2017, http://aip.scitation.org/doi/full/10.1063/1.4977487

[7] Materials Genome Initiative Strategic Plan, 2014, https://www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf.

[8] Advanced Manufacturing and new Materials, K. Bourzac, MIT Technology Review, July 2011, https://www.technologyreview.com/s/424631/advanced-manufacturing-and-new-materials/.

[9] LLNL scientists stick to a new efficient magnet, October 2017, .

[10] Design for additive manufacturing with site-specific properties in metals and alloys, S. Tammas-Williams and I. Todd, Scripta Materialia, **135**, 105-110, 2017, www.sciencedirect.com/science/article/pii/S1359646216305255

[11] Combinatorial eclectrochemistry – Processing and characterization for materials discovery, P.J. McGinn, Materials Discovery, 1, 38-53, 2015.

[12] Accelerated search for materials with targeted properties by adaptive design, D. Xue et al., Nature Communications, April 2016, .

participate in informative panel discussions with industrial experts, and participate in topical breakout sessions. Discussion topics focused on challenges and opportunities in applying AI to the discovery and design of new industrial materials.

Manufacturing remains the essential core of U.S. innovation infrastructure and is critical to economic growth and national defense. As global competition to manufacture advanced products intensifies, America's innovation ecosystem must raise its performance. Industry, academia, and government partners need to leverage existing resources, collaborate, and co-invest to nurture manufacturing innovation and accelerate commercialization.

The Advanced Manufacturing Office (AMO) within DOE's Office of Energy Efficiency & Renewable Energy (EERE) partners with private and public stakeholders to improve U.S. competitiveness, save energy, create high-quality domestic manufacturing jobs, and ensure global leadership in advanced manufacturing and clean energy technologies. AMO invests in cost-shared research and development (R&D) of innovative, next-generation manufacturing processes and production technologies that will improve efficiency and reduce emissions, industrial waste, and the life-cycle energy consumption of manufactured products. These investments help the nation harness energy efficiency in manufacturing as a competitive advantage and competitively manufacture cutting-edge clean energy technologies in the U.S. AMO is particularly interested in pre-competitive, early-stage research collaborations that might overcome some of the critical challenges associated with advanced manufacturing technology.

This workshop report summarizes the presentations, panel discussions, and breakout group discussions that took place at this event. Note that the results presented here are a snapshot of the viewpoints of the experts who attended the workshop; they do not necessarily reflect the views of the broader materials development community.

Presentations and Panel Discussions

Plenary and panel sessions featured invited experts from academia, the national laboratories, and industry. The presentations helped to assess the state of the art with respect to materials design and the specific challenges that industry faces in attempting to apply advanced AI and ML technology to industrial material design, processing, and manufacturing. Presentations delivered at the workshop are available at http://energy.gov/eere/amo/workshops.

Accelerated Search for Materials via Adaptive Learning

Turab Lookman, Physics of Condensed Matter and Complex Systems Group, Los Alamos National Laboratory

One of the biggest challenges in materials discovery and design is the very large search space of possible materials. Some experimental techniques, like the Advanced Photon Source or high-throughput techniques, generate huge amounts of data per sample (a gigabyte or more). The problem is in understanding what portion of the data is important to keep. Ultimately, the information is screened to select viable candidates for further study or for use in making predictions.

Rather than generating a lot of data and then exploiting it, Dr. Lookman and his research team are developing a more adaptive experimental design. The researchers developed a framework that uses uncertainties and maximizes the expected improvement from the best-so-far materials in an iterative feedback loop based on experiments. Dr. Lookman illustrated the framework by discussing an experiment to find NiTi-based alloys with the lowest thermal hysteresis. An initial set of 22 well-characterized training samples were synthesized in a search space of about 800,000 different alloy compositions. Nine feedback loops were performed, resulting in 36 new alloys, 14 of which had a lower thermal hysteresis value than the lowest value in the original training set of 22 samples.

The adaptive experimental design strategy uses both exploration and exploitation. In the initial stage of the experiment, researchers should focus on the samples about which they know least, because that strategy will produce the knowledge needed to gradually hone in on the search space containing the materials with the desired targeted properties. However, the most important factor is to have a well-formulated question before starting the experiment.

Analyzing Large-Scale Data to Solve Applied Problems in Materials R&D

Bryce Meredig, Co-founder and Chief Scientist, Citrine Informatics

The speed and number of new material innovations rely on wider access to materials databases. Dr. Bryce Meredig and colleagues at Citrine Informatics are taking steps to democratize materials discovery and design by creating a software platform that allows any materials scientist or chemist to access large amounts of data to create ML-based models of interest to them, ideally, in a short amount of time.

Machine learning-based tools can help scientists do more than just predict properties through modeling and optimization. The company is also using ML tools to help collect and visualize data as well as to interpret and learn from the data. Dr. Meredig provided a number of examples that show how ML-based tools have helped industry partners. In one instance, the tools helped cut R&D time for new materials in half by reducing the amount of over-testing on newly developed polymers. Fewer tests were needed on the new polymers, because the results could be supplemented with data already compiled on previously tested polymers.

Dr. Meredig discussed new methods currently being developed to accelerate materials discovery, including an active learning method called FUELS (Forest with Uncertainty Estimates for Learning Sequentially). ML guides the next experiment, based on one of five selection strategies, then folds the results back into the training data set. The goal is to find the optimal candidate after the least number of measurements.

Dr. Meredig concluded his presentation by discussing barriers to materials discovery. These barriers include restricted access to free, easy-to-use infrastructure, data and ML tools; and limited opportunities to educate physical/material scientists in machine learning, since most scientists don't learn ML tools as part of their doctorate or masters level education. If the materials community successfully addresses these barriers, then humans will be able to enlist the aid of very powerful ML tools and surpass both the purely computational materials science strategy and the purely human intuition approach.

The Materials Genome Initiative and Artificial Intelligence

A. Gilad Kusne, Materials Measurement Science Division, National Institute of Standards & Technology

The National Institute of Standards and Technology (NIST) is one of 10 member agencies actively supporting the MGI. The MGI is focused on building the infrastructure to gather the knowledge from models and data to accelerate the development of materials with targeted properties, from discovery to deployment. NIST researchers build free, helpful tools to facilitate the movement of knowledge and encourage user communities to form around materials data. Tools are available to support each of the following four NIST knowledge management principles:

Bank it: 'Smart' data ingestion tools and repository
Share it: Tools to automatically convert data into and out of standard formats
Find it: Search for resources via the materials resource registry
Check it: Use probabilistic models to assess uncertainty (e.g., confidence, or credible intervals)

NIST also supports the creation of a High-throughput Experimental Materials Collaboratory (HTEMC). The HTEMC would consist of a delocalized network of high-throughput synthesis and characterization tools and a best-in-class materials data management platform. The vision is that a researcher seeking to conduct a new high-throughput experiment could access such equipment through the collaboratory, obviating investment in expensive in-house equipment.

Recognizing the skills gap in materials design, NIST has joined with the University of Maryland, Stanford, and Johns Hopkins Universities, and Oak Ridge National Laboratory to invite materials researchers from industry, national laboratories, and academia to attend an annual three-day boot camp and workshop on machine learning for materials research. The event introduces materials researchers to machine learning theory and tools for rapid materials data analysis and encourages researchers to apply theory and tools to their own data sets.

INDUSTRY PANEL: Challenges Facing AI in Applied Materials Design

William Peter (Moderator), Director, Manufacturing Demonstration Facility (MDF), Oak Ridge National Laboratory (ORNL)**.** The MDF hosts about 700 industry customers per year. In the future, AI-based tools may be able to provide the following information to industrial customers: what materials system should be used for a particular application; what process and process parameters should be used to develop and design a desired part; and a high level understanding of expected performance of the desired part.

Joe Vinciquerra, Technology Platform Leader, Additive Materials, General Electric Global Research (GEGR). The adoption of additive manufacturing (AM) technology within GEGR is accelerating rapidly, but the biggest challenge to wider industry adoption is the slow pace of materials development for AM. The GEGR team spent 18 months on rapid material development for AM and successfully reduced the materials development cycle from about 10 years to one year; the ultimate goal is just weeks. GEGR spent a year investing in experimental methods, such as combinatorial chemistry, and ML algorithms to facilitate the design of experiments needed to expedite materials discovery and design. Obtaining the data required for AI/ML has proven to be a bigger challenge than developing the AI/ML methods themselves.

Amra Peles, Project Leader, Design for Sustainability, Pratt & Whitney. Pratt & Whitney is also investing in AM technologies, using modeling and experimentation to apply AI-based tools to help guide development. The AM process can be controlled by a number of parameters specific to the material, so materials development should ideally be done in conjunction with the materials processing. Even the design of a specific an engine part should be considered in the discovery and design process in order to get the material/part to market quickly. There is a need to also recognize the limitations of AI-based technology, for example, some predicted materials defy the laws of nature. Even when a material is successfully designed, are the processes available to rapidly scale the material?

Kishore Reddy, Staff Research Scientist/Engineer, United Technologies Research Center (UTRC). Dr. Reddy discussed the advanced analytics research that UTRC conducted as part of a service technologies initiative the company established in 2013. UTRC learned that strong interactions between business units and researchers during the development of ML-based tools enhances understanding and acceptance of those tools. There is now a strong pull for service technologies from operations, manufacturing, design, maintenance, and repair. The ultimate goal of the initiative is to be able to consider performance requirements and manufacturing constraints in the materials discovery and design process.

Adama Tandia, Research Associate, Corning Incorporated. Corning researchers began to explore ML-based tools and techniques in 2007. Researchers conducted a survey to identify the most important glass properties, then worked with the business units to develop the needed glass. This approach ensured wide acceptance of ML techniques across the company. It used to require almost two years and several million dollars to design new glass compositions. Researchers can now identify a selection of optimized glasses in a week. Based on these successes, Corning is now investigating ways to optimize processes that minimize glass defects. A key challenge is to develop a model that can accept an enormous amount of data and variables and provide a response in a fraction of a second.

Accelerated Materials Design and Discovery: An Industry-University Collaboration

Brian Storey, Program Officer, Accelerated Materials Design and Discovery, Toyota Research Institute

The mission of the Toyota Research Institute (TRI) is to improve the quality of human life by pushing the boundaries of what AI can do. Accelerated materials design and discovery is one of four R&D areas within TRI's larger AI initiative. The TRI materials program focuses on functional polymers, battery design, and fuel cell catalysts in support of Toyota's goal to develop a vehicle with zero CO_2 emissions by 2050. The challenges to meeting the 2050 goal include:

- Physics and materials needed to meet the goal
- Time scale to develop the technologies
- Ability to function successfully at an industrial/real world scale

Dr. Storey discussed the TRI approach to discovering and designing fuel cell catalysts, other than platinum, which is expensive and has performance limitations. TRI is using a range of technologies already discussed including high-throughput experiments to benchmark and measure bulk properties based on calibrated and well-controlled experiments; automated feedback based on targeted properties; and physics-based models and simulations to inform catalyst design.

Challenges exist in materials discovery and design that warrant skepticism, including a lack of clear objectives and accurate simulations. Echoing other materials experts from industry, Dr. Storey reminded the audience that a compound is not a material (an early criticism of MGI), and a material is not a system. Addressing the entire materials problem, from discovery to design, processing, and manufacturing, remains a formidable challenge. It is unclear whether AI can help solve problems that people cannot, but it may be worth trying.

Workshop Process and Breakout Sessions

The workshop brought together leading scientific and technical experts to discuss opportunities for advancing the state of the art in AI applied to materials discovery and design in energy related materials through early-stage applied research and development consistent with the missions of DOE. Breakout sessions covered three technical topics:

- **Data Quantity and Quality.** The federal government provides access to the results of federally funded research. However, the vast datasets that inform these results are not always available. Applications of AI to materials design and development require large data sets to derive any meaningful predictions of materials properties. This session explored issues related to the collection, storage, sharing, and analysis of large amounts of clean, well-organized data of the quality needed to support use of AI in advancing materials discovery and design. The session also explored the instrumentation and measurement methods needed to collect high-throughput data with the required accuracy and precision for the intended application.

- **Platforms and Infrastructure.** This session covered the challenges associated with developing and integrating the computational and design tools needed to work across materials classes and industrial sectors. These tools include searchable materials data infrastructure, algorithms, machine learning-based models, data analytics, informatics, combinatorial design methods, and other computational tools. Open architectures supporting collaborative and interactive R&D for materials design were also discussed.

- **AI for Materials Design in Specific Applications.** This session focused on the challenges associated with applying AI to the discovery and design of materials that can meet specific performance metrics for important industrial applications. Specific potential industrial applications include membrane materials, materials for extreme conditions, catalytic materials, electronic materials, and others. Case studies of success in product and process development were of interest to the group. Barriers to the widespread adoption of this approach were discussed, particularly for specific applications.

Participants in each breakout session answered questions related to the future desired state of the technologies, and challenges to achieving the desired future state. A voting process was applied to identify what individual participants perceived as the most important technical challenges to be addressed (most urgent, most significant impact if addressed). Small working groups were then organized based on the challenges, resulting in a set of proposed R&D pathways. Summaries from the breakout group discussions for the questions posed and the resulting priority R&D topics are outlined in the following chapters. Full responses from each breakout are included in Appendix A.

At the end of the workshop, all three groups were asked to discuss potential collaboration and partnership opportunities as well as related education and training requirements based on the R&D needs and pathways identified in the previous sessions. The responses from each of the breakout groups were consolidated for each set of questions posed and summarized in a separate chapter. Full responses from each group are included in Appendix B. Appendix C and D contain the meeting agenda and complete list of workshop participants.

2 Summary of Results

A. Data Quantity and Quality Breakout Session

Future Capabilities and Targets

FOCUS QUESTION 1: *What are the key capabilities, technologies, characteristics, or targets you want to see in future for AI as applied to materials discovery and design?*

During this session, participants discussed the future of AI for materials discovery as it relates to data quantity and quality requirements. The discussion centered on the following themes:

Data Format/Standardization: A lack of data standardization currently complicates the materials discovery and design process. It is often difficult to compare datasets, utilize them for analysis purposes, and combine them with additional data points due to a lack of consistency between the formatting of various data sources. Greater standardization can facilitate the use of data by researchers.

To address this issue, it was recommended that data generated for materials discovery and design embrace the FAIR Principles for data management, i.e. data should be Findable, Accessible, Interoperable, and Re-usable. Since compatibility is important to wide spread adoption of AI techniques, an open data rather than proprietary format is preferable. A standard format could be adopted by professional journals allowing R&D teams to easily find and retrieve relevant data for specific applications as well as enabling consolidation of open and proprietary data.

Data Acquisition/Quality: The materials discovery process often leverages data collected through experimentation. Experimental data can exhibit varying levels of uncertainty, depending on the conditions under which the data is measured. Many factors influence measurements, including the parameters and settings of digital equipment used during experimentation, as well as the tolerance levels of equipment utilized. Without precise knowledge of the conditions under which experiments are conducted, it is often impossible for other researchers to replicate experimental results. Reproducibility is critical to credible scientific research. Equipment should be able to provide complete information on internal parameters and settings as well as process data in a machine-readable format.

To enable reproducibility, it is important to know the uncertainty in measured data, as well as the metadata associated with the conditions under which the data was acquired. Platforms that pull data and metadata autonomously from a variety of instruments could provide this information.

To verify the experimental results of others, or to build and extend upon their work, researchers often compare their experimental results to those that have been previously published. Journal publications often show data in tables and charts, which are distributed in pdf files, or in other formats that prevent the data from being mathematically manipulated, analyzed statistically, or interpreted by software programs. Image analysis can be incorporated into analytical software to extract data from diagrams, figures, and tables. Data visualization interfaces could add multi-dimensional views (3-D) and give new perspectives to research. Advanced data analytics and visualization techniques could ultimately make published data more useful to the research community.

Technical and Scientific Challenges

FOCUS QUESTION 2: *What are the major scientific and technical challenges that limit the application of AI for materials design and discovery? What are the problems that hinder us from realizing the desired capabilities, technologies and targets identified above?*

During this session, participants discussed challenges impacting the future development of AI for materials discover, as it relates to data quantity and quality requirements. The discussion on challenges centered around the following themes:

Data Sharing: The lack of sharing of data from failed experiments was identified as a major challenge. Current research norms dictate that published results reflect successful experiments. While the sharing of successes can be helpful, withholding information concerning failed experiments can hinder research progress. Without knowledge of failed experiments, or pathways that lead to dead ends, uninformed researchers may repeat the mistakes of others. Methods that failed in the past could produce beneficial results under improved conditions. Without a complete record of past attempts, researchers are unable to fully navigate possible innovation pathways. Current norms in the academic community do not motivate/incentivize researchers to share or publish data from failed experiments; a change in this culture should be encouraged.

Another major challenge hindering data sharing is the availability of proprietary data. Data emerging from research is often proprietary, unpublished, or in a proprietary, non-standard format. Closed proprietary equipment interfaces (black box for parameters and settings) also prevent a full understanding of the conditions under which data is provided. Experimental data collected using one manufacturers equipment cannot be readily compared to data derived from a different set of instrumentation. Interoperability will require either standardized data sets, or standardized interfaces to communicate between various formats. Protocols are also needed for sharing controlled data with the research community while retaining data integrity.

Data Collection and Management: A number of challenges were identified that prevent or complicate reaching the ideal database. A major barrier is the lack of sustainable models for centralized accessible materials properties databases to support AI. Governance and business models are lacking to operate and maintain existing database solutions. Another major issue is the poor connection and integration of physical science and data science. Good methods are lacking to integrate physics models with data to provide advanced analytics. Automating experimental hardware can also increase process efficiency and enable much more data collection to take place.

Data quality is important to ensuring that materials data is usable. Without clear standards, the integrity of data cannot be guaranteed. There is currently a lack of incentives for researchers to follow standardized data collection formats and reporting. Social engineering (i.e., centralized planning) to encourage and optimize data sharing is lacking.

Data Uncertainty and Validation: The quantification of data uncertainty is important to enable comparison between data sets and understand the quality and validity of data. Uncertainty levels for model and measurements to support AI are currently unknown and experimental data is quite variable.

Tools for verification and validation of data in general are lacking. Expert review is required to verify the quality of standard datasets but can be costly. There is a currently a lack of mechanisms for non-expert review of data but this could be a viable approach. Crowd-sourcing of data labeling by non-experts is one example.

Priority R&D Pathways and Technical Approaches

Figures 2A.-1 through 2A-6 are based on worksheets completed by Data Quantity and Quality group. These are highlights of the R&D approaches needed to address the high priority technical barriers. The full R&D pathway diagrams are included in Appendix A (Figures A-1 to A-6).

Figure 2A-1 Priority R&D Pathway: Data Availability

KEY BARRIERS	SUMMARY OF APPROACH
• Data is proprietary, not published or limited (lack of measurements or models)	• Establish and develop an open-access database and collaboration platform for AI in materials development

	R&D Activities	Milestones & Targets
1-2 years	• Understand data gaps and needs • Identify relevant measurements and simulations • Identify data available in publications and open databases	Milestone: • Identify data gaps, sources of data, and needed measurements Target: • Set up first stage open database development
3-5 years	• Conduct measurements and simulation to collect data • Set up an open database as part of a collaboration	Milestones: • Needed data has been collected. • Multiple institutions can upload data and analyze it Target: • Ability for collaborating institutions to use the database
>5 years	• Fully developed integrated collaboration platform for data access and data analysis with easy user interface and maintenance plan	Milestones: • Well-developed open database and analysis framework • Long-term maintenance plan established Target: • Widespread adoption of database and framework

Figure 2A-2 Priority R&D Pathway: Lifecycle Data Management

KEY BARRIERS	SUMMARY OF APPROACH
• Inputs, data, desired outputs and metrics are varied • Analysis and workflow are non-standard and non-reported • Lack of incentive to contribute (both computer and academics)	• Track the data provenance • Develop data standards through community engagement • Create working groups to define metrics

	R&D Activities	Milestones & Targets
1-2 years	• Definition of provenance best practices and include full provenance with metadata • Identify 1-3 communities with tractable problems to standardize the desired outputs, language, and output formats • Year-long effort to engage communities' at large conferences/ working groups to develop the format	• Documentation with examples of full provenance of data for the communities identified for case study • Standard format and language for 1-3 case study examples to prototype output standardization → adopt • Creation and validation of database for test cases
3-5 years	• Expand the prototype framework to many materials communities • Incentivize data standardization and contribution with grant data management plans • Co-opt existing databases as repository	• Growth in adoption of data management practices • Validate test cases against existing best practices

>5 years	• Materials dossier • Implement interoperability between repositories • Extract from provenance best/common analysis practices and code	• Materials Application Programming Interface (API) • Drivers exist for database to talk • Building software for data in repositories; information using best practices which surfaced from provenance

Figure 2A-3 Priority R&D Pathway: Models for Sustainable Data/Database Management

KEY BARRIERS	SUMMARY OF APPROACH
• Inadequate focus on long-term support of materials data projects.	• Establish an interagency collaboration/consortium to measure the current landscape and to measure and address gaps

	R&D Activities	Milestones & Targets
1-2 years	• Conduct an inventory of existing agency efforts • Quantify gaps and redundancy in existing efforts • Begin to build a knowledge base of agency expertise to eliminate current state silo operations	• Establish an interagency group on materials data (council) • Deploy an online, open access version of knowledge-base • Establish a formal charter for a permanent interagency group/consortium (leadership/officers, by-laws, structure)
3-5 years	• Develop an economic model for sustained materials data funding • Deploy interagency demonstration projects • Refactor existing programs based on gap/ redundancy assessments • Engage academia and industry	• 1-3% of budget(s) dedicated to FAIR materials data; National Institutes of Health (NIH) model • Fund at least 5 demonstration projects (government, academia, and industry) • Establish a national, federated data services
>5 years	• Congressional mandate of materials data council • Establish new, pre-competitive consortia	• 2-5% of budget(s) dedicated to FAIR materials data (NIH model) • Industry use of demonstration project data

Figure 2A-4 Priority R&D Pathway: Unbiased Reporting and Disclosure of Failures

KEY BARRIERS	SUMMARY OF APPROACH
• Lack of unbiased reporting; negatives and failures do matter	• Establish scientific, societal, and social benefits of reporting negative results • Enable better predictions, avoid redundancy, and fill in missing knowledge gaps

	R&D Activities	Milestones & Targets
1-2 years	• Analyze professional and social factors that dis-incentivize publication of negatives • Establish case studies where knowledge of negatives mattered • Explore pathways toward dissemination of negatives outside "publication"	• Outreach and education/PR campaigns on "report negative" and "negative is not failure" • AI architecture based on deep and adversarial data • Incorporation of negatives in physical model
3-5 years	• Develop AI workflows and strategies to incorporate negatives • Develop awareness of "negative" vs. "bad" experiment • Registry of funding programs; use regulatory tools to recognize reporting "negative"	• Robust prediction of materials, devices, and technological tools for comparing special cases • Minimize stigma of "negative" • Incorporation of robust workflows that allow management of negative results; avoid "production line" trap
>5 years	• Establish integrated knowledge space supporting lateral instrumental networks of "similar" tools • Develop ways to incorporate negatives to explore latent connections in knowledge space	• Accurate and robust universal case for AI design of materials • Establish robust basis for advanced manufacturing past fourth technology revolution and scientific security of the country

Figure 2A-5 Priority R&D Pathway: Uncertainty Quantification

KEY BARRIERS	SUMMARY OF APPROACH
• Lots of input variables with unknown error scales • Merging data from different sources (experimental or theory); these have different classes of errors • Model and prediction errors that occur when building models; formal theory is lacking to describe this uncertainty	• Fund and build theory; communicate and gain confidence from the materials community • Deploy tools to help make better informed decisions

	R&D Activities	Milestones & Targets
1-2 years	• Develop theory to handle uncertainty for all classes of machine learning methods • Communicating the importance of such metrics • Better decision-making, knowing uncertainty will exist in every measurement/ prediction	• Milestone: Awareness • Standardization of error reporting • Motivate people to take up challenge • Cannot get rid of uncertainty; be able to work with it

3-5 years	• Standardizing data collection practices • Discovery of errors (benchmarking) by comparing different sources/models • Version control practices of models/experimental recipes to know when things work or not	• Milestone: Change of culture • Gain confidence in using information from different sources • Kaggle-type approach where models are scrutinized via crowd-sourcing
>5	• Build capabilities to make developments accessible to end users	• Milestone: Deployment • Get end users to adopt the technology

Figure 2A-6 Priority R&D Pathway: Validation & Verification of Theory-Based Models

KEY BARRIERS	SUMMARY OF APPROACH
• Simulation and experiment data must be compared in the same domain/space (e.g., X-ray diffraction to X-ray diffraction pattern) • Bias in simulation data/model must be quantified	• Machine learning approach to parameter tuning in physics-based models • Investigate methods for simulation model bias quantification • Curated experimental databases/standards

	R&D Activities	**Milestones & Targets**
1-2 years	• Identify available databases for validation and verification study • Identify software tools for mapping experimental and simulation data into the same domain; e.g., crystallographic information file (CIF) to X-ray diffraction • Research diffraction methods for quantifying bias	• Summary paper of validations
3-5 years	• Application of methodologies to select datasets	• Evaluation and solution of model bias in specific context
>5 years	• Build tools for general use	• Software tools • Closed loop parameter tuning in physics models

B. Platforms and Infrastructure Breakout Session

Future Capabilities and Targets

FOCUS QUESTION 1: ***What are the key capabilities, technologies, characteristics, or targets you want to see in future for AI as applied to materials discovery and design?***

The Platforms and Infrastructure Breakout Session participants had a wide-ranging discussion regarding desired future capabilities and targets for the application of AI in materials discovery and design. Outside of platforms and infrastructure, issues related to data and data management were also identified as priority areas in order for AI approaches to be widely used in the materials development area. The discussion of desired future capabilities centered on the following themes:

Data Algorithms and Models: There is a need to connect physical models and algorithms to AI applications and machine learning. Algorithms are needed that extract both physical and chemical information from materials data sets; alternatively, the capability is needed to derive physics-based models from AI. If open source codes are used, documentation of successes and failures should be encouraged and routine. Wider adoption of technology would be enabled if the ML models provide insights on how and why a material is predicted to satisfy the required targeted properties. Integrating multi-scale models would facilitate materials design from material design to processing to manufacturing. Ultimately, models could also evaluate the trade-offs in design versus cost versus supply versus performance. Algorithms for uncertainty quantification (UQ) for the model, data and results are also needed.

AI System Capabilities: In the broader context of AI, it will be important that the overall system have far-reaching ‘intelligence’ and analytical capabilities. This includes the ability to distinguish between ‘good’ versus ‘bad’ data, and to be able to establish the quantitative relationship between chemistry, microstructure, and properties. Ideally, the AI system could have the embedded capacity for peer review and uncertainty analysis of data and outcomes, and be able to identify what gaps in data need filling. An advanced functionality would be the capability to use either composition processing or properties as an input, and then get the either as the output result.

Tool Framework, Function, and Interaction: A number of capabilities were identified that would enable materials discovery and design. These include, for example, open source platforms and codes for AI and ML that work across materials disciplines; examples include R and RStudio.[13] The ability to utilize complex and multi-modal data sets is important. A standard application programming interface (API) is needed to allow system-to-system interaction and data transfer. A mobile capability identified is an ‘app’ store of post-processing tools that would allow new methods to be published and used quickly by the research community. Search functions are also needed that will help the user steer closer to what he/she is looking for without having to experiment or spend time applying multiple keywords.

Data Capture and Analytics: The ability to retrieve and fuse all relevant data from disparate sources was identified as a major goal. For the capture of data, an important issue will be having the capability to extract data from legacy documents and sources and to be able to correctly translate the information and place it in the right context. The idea is to consolidate a variety of open source and proprietary data, and allow for seamless integration between various types of structured and unstructured data. Data interoperability between sources, data curation, and automatic integration of data will be key factors. It will also be important to be able combine and fuse data coming from numerous disparate sources so that all data can be utilized by the developed algorithms, models, and tools. Data sharing is another priority area and common theme across sessions.

[13] R (https://www.rstudio.com/) is a free software environment for statistical computing and graphics. RStudio (https://www.rstudio.com/) makes R easier to use, and includes a code editor, debugging and visualization tools.

Technical challenges related to data sharing need to be resolved, and mechanisms and incentives to encourage industry to share data should be developed. Data libraries and other publication infrastructure should be developed to enhance data sharing and collaboration.

Technical and Scientific Challenges

FOCUS QUESTION 2: *What are the major scientific and technical challenges that limit the application of AI for materials design and discovery? What are the problems that hinder us from realizing the desired capabilities, technologies and targets identified above?*

Discussion of technical and scientific challenges centered on themes similar to those in future capabilities. The identified major challenges included:

Data Extraction and Fusion: A number of data challenges were discussed that are specific to developing platforms for material design. Data extraction from older legacy sources was identified as major challenge. When such legacy data is collected, it is difficult to be able to interpret, understand and categorize it in the proper context. Lack of generalization and design of experiments and historical data that evolved over time has created data sets that are not standardized and more difficult to interpret. Proprietary data issues also arise when searching across multiple databases. Combining and fusing data (so it can be readily analyzed and utilized) from disparate sources is also a major obstacle. The non-standardization of data creates a challenge for data fusion, as do proprietary data formats. Standardized data across a range of time and length scales for multi-scale models are lacking, and microstructure data needs to be characterized in a standardized way that is usable by ML models.

Models and Tools: One of the most critical challenges identified is that AI algorithms are not designed to take physical models and prior knowledge into account. Regarding outputs, AI algorithms are not designed to provide a physical interpretation of the results. The result is a need to be able to connect physical models to AI algorithms and applications. Physics-based microstructural models for additive manufacturing need to be developed prior to developing a machine learning model. Common databases are needed for process parameter driven properties as well as thermodynamics and kinetics and all data needs to be utilizable by AI/ML models. The different ‘language’ in materials science disciplines (metals, ceramics, composites) and the lack of a uniform set of synthesis and characterization techniques may hinder the creation of generalized machine learning models. When models are developed, it is important that they are not “black-box” models; material scientists need to be able to understand how the models work. The developed tools should have capabilities to better understand and interpret anomalous and outlier points in data. In the area of tools, software maintainability is a challenge. Older programs and codes may still be needed, but programming languages and platforms change over time, necessitating software updates and changes.

Software Challenges: Upgrading of software and codes that were written years ago to be compatible with and keep up with changing computer language and platforms remains a significant challenge. Computing platforms are highly dynamic and continuously evolving. Material scientists much spend time and effort maintaining and upgrading software; the issue is exacerbated when multiple types of software are interconnected. Incentives are needed to academics and other researchers to produce software that is commercially compatible and/or interoperable with available platforms.

R&D Pathways and Technical Approaches

Figures 2B-1 through 2B-3 are based on worksheets completed by the Platforms and Infrastructure group. These topics represent the challenges that were identified as those most important to address. The full R&D pathway diagrams with milestones, detailed R&D activities, outcomes, and stakeholders are included in Appendix A (Figures B-1 to B-3).

<table>
<tr><th colspan="3">Figure 2B-1 Priority R&D Pathway: Connecting Physical Models to AI Applications</th></tr>
<tr><th colspan="2">KEY BARRIERS</th><th>SUMMARY OF APPROACH</th></tr>
<tr><td colspan="2">• AI algorithms are not designed to take physical models and prior knowledge into account (input)
• AI algorithms are not designed to provide a physical interpretation of the results (output)
• Lack of physical interpretations are especially problematic for deep learning</td><td>• Multifaceted approach that combines research and development activities with opportunities to convene the community around the problem
• Use Ai to build bridge between chemical and physical models processing and structural outcomes</td></tr>
<tr><th></th><th>R&D Activities</th><th>Milestones & Targets</th></tr>
<tr><td>1-2 years</td><td>• Identify systems that are good candidates to implement physical constraints on learning input/output
• Identify gaps in physical knowledge salient to materials problems; identify areas that are well-understood with physical models
• Apply computer science interpretability techniques to materials science classification tasks (i.e., what does the computer use to classify microstructural images?)
• Engage computer scientists to define the physical constraint and physical interpretability problem space</td><td>• Sponsor a study to bring computer scientists and materials scientists together around this issue
o A national academies-style report on bridging machine learning and materials science
• Define a funding opportunity announcement (FOA) around implementing physical models in AI approaches to materials science
o Fund a set of seed projects in this area selected such that outcomes will help to guide future program development (including measures such as internal rate of return)</td></tr>
<tr><td>3-5 years</td><td>• Use AI to utilize thermodynamic and kinetic models to build a bridge between processing, structure, and properties for a real system
• Use AI to “correct” the systematic approximation errors in density functional theory (DFT) results to better match experimental data
• Identify problems with opportunities to explore the three classes of outlier data: anomalies, physical outliers, and clusters of outliers (require their own distributions)</td><td>• Successful proof-of-concept for AI supplementing physical models, including DFT and Computer Coupling of Phase Diagrams and Thermochemistry (CALPHAD)
• Develop a program around understanding outliers in AI data sets</td></tr>
<tr><td>>5 years</td><td>• Use AI to build a bridge between models and experimental results at multiple length and time scales
• Use AI to integrate data from models, simulations, and experiments, with appropriate weighing of the data in different regimes</td><td>• Execute examples of the long-term action items
o AI provides a statistical surrogate for physics-based simulations or experiments.</td></tr>
</table>

Figure 2B-2 Priority R&D Pathway: Data Extraction and Standardization/ Standardization and Translation from Traditional Data Sources

KEY BARRIERS	SUMMARY OF APPROACH
• Barrier to extracting data from legacy documents (even those directly machine readable) is on a similar scale to language translation – a very large amount of context will be needed in order to understand and categorize individual pieces of data	• Two-part solution: (1) Create a new AI tool that provides a contextual approach to analyzing legacy articles and documents to extract data (e.g., linking figures to text); and (2) establish new standards for articles to ensure that they are fully searchable and data-extractable

	R&D Activities	Milestones & Targets
1-2 years	• Articulate the barrier to legacy data extraction in terms of natural language processing with the objective of engaging the computer science (CS) community • Establish a new standard for articles and reports in materials science to define topics and tags that should be embedded. Engage the CS community, e.g., via those working on the ontology of scientific reporting • Survey other topical areas for state-of-the-art, e.g., cheminformatics, bioinformatics • Develop a schema/ontology for data exchange in materials science	**Milestones** • Demonstrate that a searchable article is feasible with demonstrations of at least two examples **Targets** • Run workshop(s) to get community input on data standards • Run workshop on digitization of legacy documents: best practices, etc. • Run workshop on natural language approach to analyzing legacy documents
3-5 years	• Implementation of data exchange standard(s), e.g., with major publishers and laboratories • Acquire a substantial data set based on legacy document analysis • Develop a publicly accessible database that supports the natural language approach to extracting data from legacy documents	**Milestones** • Demonstrate efficacy of a data schema and ontology • Demonstrate an approach to extracting useful data from legacy documents that includes development of contextual framework **Targets** • Have more than two research groups to provide a corpus of documents (including journal articles) embedded in a searchable format according to an established ontology • Implement a publicly accessible database that supports the natural language approach to extracting data from legacy documents
>5 years	• Legacy data extraction in terms of natural language processing • Development of flexible methods for scraping data from existing databases	**Milestones** • Demonstration of a data base of documents that are fully searchable in a given topical area in Materials Science • Demonstration of a natural language approach to acquiring data from legacy documents

Figure 2B-3 Priority R&D Pathway: Data Fusion		
KEY BARRIERS		SUMMARY OF APPROACH
• Fusing data from different and disparate sources is challenging • There is a lack of standards for data generation, data storage, data provenance, and metadata		• Push data format standards as early as possible in data generation life cycle • Address future data first, then legacy data • Data associated with publication and other deliverables should be subject to FAIR principles
	R&D Activities	**Milestones & Targets**
1-2 years	• Push for community-scale adoption of data formatting and metadata standards • Establish publicly accessible data repositories • Establish infrastructure for data discovery	• Prototype data standards • Prototype data repository and registry infrastructure
3-5 years	• Establish community-wide groups to identify and manage an ICME community data fusion challenge • Work toward data fusion procedures for heterogeneous (multiscale, temporal, spatial) datasets from different sources	• Define a foundational ICME data fusion challenge problem
>5 years	• N/A	• N/A

C. AI for Materials Design in Specific Application Areas Breakout Session

Future Capabilities and Targets

FOCUS QUESTION 1: *What are the key capabilities, technologies, characteristics, or targets you want to see in future for AI as applied to materials discovery and design?*

This session looked at specific challenges associated with applying AI to the discovery and design of materials that can meet the specific performance metrics for important industrial applications. Examples of industrial applications include membrane materials, materials for extreme conditions, catalytic materials, electronic materials, and others. Participants were asked to share case studies of success in product and process development using AI and to list barriers to the widespread adoption of an AI approach for their specific applications. Issues related to scientific fundamentals and data needs/management were also identified as priority areas in order for AI approaches to be widely accepted and used in manufacturing applications.

The discussion of future capabilities and targets focused on the following areas:

Data and Algorithms: Universal algorithms for AI and ML are needed that bridge the length scales of measurement across timescales (e.g., interfaces, microstructures, meta-stability, kinetics). This includes numerical representation of microstructures connecting material structure to properties with codes incorporated in AI modeling of materials. Numerical methods could be developed to represent a generic material structure that contains a set of required features. The physical structure of materials can be better related to the required mechanical properties overall size and time scales from synthesis to degradation with better scientific information. There should be a better focus on the evolution and advancement of AI as new physical principles and material structures are developed from the nano-scale to the synthesis level. New methods are needed that can integrate physical-based multi-scale models with ML. These should be transferable so they can be used for the discovery of different types of materials and applications. Ideally, techniques would enable the rapid synthesis of a range of chemistries relevant to future processing targets.

Models and Tools: The ideal AI modeling framework would include rich datasets; physics-based processing capabilities with reproducible outputs; robust and accurate predictions; be applicable to shape alloys; and include both thermal and mechanical properties prediction. Overall better experimental validation is needed for material structure and property relationships. An important capability is experimental methodologies that can characterize materials in real-time while synthesizing to enable validation of the AI method. Material structures and processes require proper representation to interoperate with AI software tools. Experiments and physical constraints should guide AI and the solutions it can provide. For processing, high-throughput metrology and ML algorithms that "train" well with small data but integrate well with an AI model require development to interpret AI-derived solutions and generate new physical models. Integration of physical modeling with appropriate AI tools and human intuition can be used to guide both material and process development.

AI System Capability: Ideally AI/ML methods would enable industry researchers to design complex systems involving multiple materials and interfaces, not just materials (e.g., battery substrates) and design materials based upon part geometry, and physical and performance requirements. Integration of smart material discovery with smart manufacturing requires a data thread from discovery to product output. The 'smart factory' idea would enable adjustment of production processes when materials with variable properties are used. There is also a strong need for rapid synthesis and characterization that is scalable to production that could be facilitated by AI. A suggested timeframe for such design-driven materials selection and qualification is from idea to product in under 3 years.

An AI machine is envisioned with characteristics that start with an application analyzer, when then determines specified properties (AI generator), selects materials candidates (AI selector with auto-test quality), and then recommends a candidate material and characteristics. Advanced AI/ML techniques would enable documentation of successes and failures when running codes, and allow for the use of low quality or minimal data sets. Other desired characteristics include rapid characterization and high throughput techniques for generation of large datasets, and the capability for in suit/ in operando materials characterization.

Technical and Scientific Challenges

FOCUS QUESTION 2: *What are the major scientific and technical challenges that limit the application of AI for materials design and discovery? What are the problems that hinder us from realizing the desired capabilities, technologies and targets identified above?*

Models and Tools: The most significant challenge identified is incorporating the fundamentals of science (e.g., thermodynamics, physics) into AI modeling of materials. This is also a common theme throughout the future desired capabilities. Models also need to be linked across multiple time and length scales and be ableto quantify uncertainty as it propagates through the component design and further up the value chain. It is challenging to be able to go back and forth between the numerical representation of the physical structure and the actual physical structure. Overall, better experimental and computational data is needed that integrates the fundamentals of science and underlying physical principles (chemistry, physics, thermodynamics, etc.). Integration of smart materials discovery is a major challenge and if addressed could lead to innovative, advanced methods that are much faster and more targeted. The lack of ability to rapidly synthesize a range of chemistries that are relevant to future processing is a major barrier where AI approaches could be quite effective.

Fundamental Data and Scientific Principles: A significant barrier exists in the limited understanding of the underlying physical principles of materials development as applied to AI. Without this understanding the development of new materials with this approach is constrained. Understanding the physical limitations of different materials and design, how to manipulate microstructure, and the lack of physical and chemical properties for a wide range of systems al contribute to this challenge.

Data Shortfalls: Current methodologies of testing and validation of materials are time-consuming and disconnected. There is a lack of non-contact and non-destructive equipment for rapid characterization of material properties during production, which could provide the process data needed to improve predictive models and quantify uncertainties. It is challenging to collect machine-accessible data that is interoperable across multiple platforms. Central databases of structural, process, experimental, and computational data are lacking. A national clearing house is needed for data/information along with an API interface for large volumes/open data and accessible to industry. Integrated learning methodologies that simultaneously process structure, property, and performance to capture variability in manufacturing and reduce risk of new materials would better meet manufacturers' needs. During production, there should be a capability to evaluate human/machine interactions and interfaces.

R&D Pathways and Technical Approaches

Figures 2C-1 to 2C-3 are based on worksheets completed by the AI in Materials Design in Specific Applications Areas group. These topics represent the challenges that were identified as those most important to address. The full R&D pathway diagrams with milestones, detailed R&D activities, outcomes, and stakeholders are included in Appendix A (Figures C-1 to C-3).

Figure 2C-1 Priority R&D Pathway: Product Functionality-Driven Accelerated Materials Discovery, Manufacturing, and Deployment by AI

KEY BARRIERS	SUMMARY OF APPROACH
• Current methodologies of testing and validation of materials are long and disconnected • Lack of tools for rapid synthesis and characterization that is scalable • Lack of equipment that is non-contact and non-destructive for rapid characterization • Lack of integrated learning methodologies that simultaneously process structure, property, and performance to reduce risk of new materials • Being able to capture variability in vendors	• Start with state of the art review; identify 5 gaps and deploy existing solutions, including in-situ sensors, modeling, and uncertainty quantification • Enable a hub for coupling materials discovery and manufacturing

	R&D Activities	Milestones & Targets
1-2 years	• Concerted effort to identify gaps in state-of-the-art • Evaluate and deploy existing tools not currently used in materials discovery • Synergistic modeling of multi-materials • Integrate in-situ sensors for real-time monitoring	• A review on state-of-the-art, gap, and needs • Clearinghouse for validated sensors and software modeling • Experimental and modeling uncertainty quantification (UQ) deployment
3-5 years	• Hub for smart materials manufacturing (public-private partnerships) • Autonomous synthesis, characterization, and validation • Strengthen characterization and property relationships • Tools to identify non-equilibrium states	• Reduce lifecycle of discovery and scale-up by 30% • Tool development (prototype demonstration) for synthesis and characterization • Tools to interrogate non-equilibrium modality
>5 years	• Materials tricorder: ability to rapidly identify materials for composition, process, and source • Cradle-to-cradle lifecycle assessment	• 50% reduction in cycle time for material discovery to in-production at lower energy costs • Materials systems performance prediction • Fully integrated system

Figure 2C-2 Priority R&D Pathway: Data Driven Materials Discovery Paradigm

KEY BARRIERS	SUMMARY OF APPROACH
• Availability of high-throughput measurement technologies, machine learning algorithms that train well with small data, and integrated AI/robotic models	• Augment/supplement serendipity and “expert intuition” in materials research with a more consistent/reproducible AI models

	R&D Activities	Milestones & Targets
1-2 years	• High-throughput precise experiment technology (e.g., model system straightforward, “easy-to-measure”) • Database management	• Sufficiently accurate AI model o No physics counterpart: Establish baseline standard o With physics counterpart: Win out in some metric • Predictive property

3-5 years	• Better algorithm to train on limited data • More interpretable model/human-in-the-loop • Expanded metadata collection; incentivizing data-sharing (centralization)	• Minimum experimentation o More "intelligent" serendipity: Reduce number of experiments versus random trial • Rediscovery o Easier-to-train better performance of combined model • Guidance
>5 years	• Robotics • Integrate AI system with other systems in the production chain	• AI automation o Combine AI model with robotics: "fully automated" material discovery o Move downstream requirements to early-stage discovery – "manufacturability"

Figure 2C-3 Priority R&D Pathway: Fundamental Models and Experimental Data Integrated with AI-based Approaches to Develop Optimal Materials & Processes

KEY BARRIERS	SUMMARY OF APPROACH
• Representing material structures and processes so they interoperate with AI software tools • Lack of central databases of structural data, process data, and experimental and computational data • Interpreting AI-derived solutions to generate new physical models	• Integrate physical modeling with appropriate AI tools and human intuition to guide material and process development • Use experiments and physical constraints to guide AI and the solutions they provide

	R&D Activities	Milestones & Targets
1-2 years	• Conduct inventory of software, tools, databases, protocols, and data representations in different domains that can be integrated with AI tools • Understand gaps in existing tools	• Domain-specific road maps to get from current software, databases, etc. to where AI tools can be used to design materials and processes
3-5 years	• Support software tools to integrate open-source AI software with open-source physical modeling code • Develop protocols to represent disparate materials and processes in databases and centralized databases that can be accessed by researchers applying AI techniques	• Adapt existing AI and physical modeling software to new protocols to grow an AI platform for designing materials and processes • Visualization tools for representing data and developing workflows/pipelines for designing materials with AI assistance • Domain-specific software for prediction of specific material properties (catalytic rates, bandgaps, melting points)
>5 years	• An ecosystem of domain-specific software tools and databases and protocols that allow researchers to use evolving AI tools to design materials and processes • Demonstrations in several domains that AI techniques can be successful in designing new and improved materials and processes	• Development of software tools, databases, communications protocols (data exchange, file formats) integrated with AI software tools that researchers can use to develop new and improved materials and processes

Figure 2C-4 Priority R&D Pathway: Universal Numerical Representations of Physical Microstructure

KEY BARRIERS	SUMMARY OF APPROACH
• Need universal numerical representation of physical microstructure derived from experimental and computational data for connectingDa material structure to properties	• Integrate current “state-of-the-art” for image processing and numerical representation into materials science • Methodically explore the existing/possible feature-space that exists in inorganic materials, and intelligently find/develop appropriate numerical methods for representation of a generic structure/ material that could contain those features.

	R&D Activities	Milestones & Targets
1-2 years	• Integrate conventional state-of-the-art image processing techniques into materials science domain • Develop framework for creating supportive datasets	• Ability to compute image-based features automatically on some subset of features • Collect and share datasets • Release public tools
3-5 years	• Use AI to explore feature-space from experimental and computational data, such as micrographic, compositional, crystallographic, atomistic, etc., data	• Identify AI selected features for universal representation • Determine method to combine multiple data sources into a single numerical representation • Collect and share datasets • Release public tools and updates
>5 years	• Develop universal numerical representation for inorganic materials compatible with current state-of-the-art for organic materials.	• Validate the reality of the selected universal features • Publish and promote use of universal numerical representation • Validate that representation can be numerically manipulated and produce meaningful results (i.e., simulate heat treatment, forge, other processing) • Release public tools and updates

D. Collaboration, Partnerships, and Education/Training

Applying AI/ML methods to materials discovery and design is a multi-disciplinary effort requiring experts across many fields, including, but not limited to, material scientists, chemists, mathematicians, and computer scientists. The field also requires advanced tools and infrastructure that is costly and difficult to maintain without the required expertise. These factors and the fact that the discipline is relatively new, arguably since the early 2000's, requires robust collaborations between the public and private sectors and identification of gaps in education and training needed to ensure that the U.S. can remain competitive in this area.

Participants were asked to answer the questions below considering the R&D needs and pathways identified. The results are a summary of the combined responses; combined raw data is available in Appendix B.

FOCUS QUESTION 3a: *Considering the R&D needs and pathways identified, what opportunities exist for collaborative efforts with the DOE labs and industry? What types of partnerships are envisioned that would be most successful in reaching targets and goals?*

Collaborative Organizations and Initiatives: As there are a number of AI-focused efforts within the Federal Government, better coordination among these groups was identified as a good step in addition to incorporating program such as the High Performance Computing for Manufacturing (HPC4Mfg) and Materials (HPC4Mtls). Coordination with the national laboratories could also occur through the user facilities supported by DOE Office of Science or through the National Network for Manufacturing Innovation Institutes. An umbrella cooperative research agreement was suggested to make collaborations with the national laboratories easier or establishing internship programs at the labs for academia and industry. More focused collaborations were encouraged through establishment of grand challenges on specific problems or 'hackathon' type programs.

Data Sharing and Accessibility: Development and consensus on an open format for data/databases is needed. The Federal Government could lead or advise a collaborative effort to provide mechanisms and mandates for a common data sharing/storage platform, especially for projects supported by the government.

Other Areas/Topics for Collaboration: Other suggestions included forming consortia around topics such as software development. Collaborations with national labs were suggested for specific tasks such as development of high throughput tools, addressing simulations at different scales, and high throughput physics simulations. Intellectual Property (IP) is recognized as an important issue that needs to be discussed such that the government and industry can be intentional about possible consequences.

FOCUS QUESTION 3b: *What education and workforce challenges need to be addressed? What skillsets or disciplines need to be further developed?*

Curriculum Development and Academic Programs: There is a general consensus that cross-disciplinary education and training is needed among material science, computer science, and engineering at the university and graduate levels. Ideally, the education should start as early as K-12 level. Undergraduates should be trained in additional skills such as coding best practices, scripting for data analysis instead of relying on ready tools, design of experiments, and anomaly removal.

Workforce Development: Specialized online training programs and bootcamps covering AI topics were identified as a way to train a seasoned workforce. Internships, training grants, and rotations could be used as alternative methods to fund workforce training. Mechanisms should be created to identify under-represented populations as a way to increase the talent pool.

Tools and Other Resources: The government can take advantage of the wealth of data it owns and communicate state-of-the-art and industry standards to the research community. Good case studies should be collected to showcase AI/ML models that can be used as training tools.

Appendix A

Appendix A.1: Data Quantity and Quality Breakout Session Results

Future Capabilities and Targets

FOCUS QUESTION 1: *What are the key capabilities, technologies, characteristics, or targets you want to see in future for AI as applied to materials discovery and design?*

Table A-1. Data Quantity and Quality – Future Capabilities/Targets

FAIR (Findable, Accessible, Interoperable, Reusable)

- Embrace "FAIR" data principles: Findable, Accessible, Interoperable, Reusable
- NIST-provided standard or template(s) for data that is intended to be shared
- Tools and methods for tracking data provenance (carrot for sharing; incentivizing positive data management practices)
- Open data format (as opposed to proprietary format)
- Code, raw data, and literature that is findable and accessible, as opposed to the publication of research results with no access to the underlying datasets

Data Visualization

- An interface that allows researchers to traverse chemical space as though flying through it (data visualization tools)
- Simple, informative user interface/visualization (field specific)

Data Quality and Completeness

- Data quality: know what uncertainty is in data; metadata about conditions of acquisition
- Basic characterization: data should be reported as cuts through parameter space; physical quality vs. temperature, magnetic field, pressure, etc.
- Robust data collection for all vertices of materials & models linking them (data and physics)
- Studies on the minimum amount of data needed to learn from and make predictions
- Datasets that are wide not deep (i.e., many input and output variables) and few observations

New Targeted Databases

- Dynamic data on properties or surrogate measurement
- Microstructure library that is searchable
- Databases for metastable materials

Data Acquisition

- How to pull data & metadata from variety of instrument platforms autonomously
- Data output tools from equipment manufacturers: equipment should be able to provide complete information on its internal parameters and settings, as well as process data, in a machine readable format
- Seamless integration of in-situ/onboard measurement into analysis
- Code base for translating between different types of instrumentation to ensure that measurements from different equipment are comparable
- Automated data analysis and metric extraction (also automate tag generation)

Table A-1. Data Quantity and Quality – Future Capabilities/Targets

• Easy capture of R&D data from various labs, equipment, processes steps • Sensor data interpretation and fusion (accurate) from different sources or locations
Standards
• Standard data sets that can be utilized for comparing algorithms • Benchmark databases that allow calculations and predictions to be compared with results from experiments • Data repositories that are compatible with each other
Other
• Data management infrastructure applicable across the entire life cycle of material development/production • High throughput characterization tools (tools designed for large datasets) • Easy to access, easy to use; “black box” tools, known to be reliable and consistent, offering automatic feature selection and automatic best model selection • Predict alternative technologies/trees/cost analysis • Tighter tolerances in processing to hit narrow chemistry/microstructure/property window • Methods that consider multiple performance properties • Make material search and manufacture easy (as easy as Googling the information) • Fund science data as a national utility (tax all science agencies); big science agencies are currently biased toward capital budgets, not data

Technical and Scientific Challenges

FOCUS QUESTION 2: What are the major scientific and technical challenges that limit the application of AI for materials design and discovery? What are the problems that hinder us from realizing the desired capabilities, technologies and targets identified in Session 1?

Workshop participants were asked to vote on the barriers and challenges they perceived as most important. The number of votes received (indicating participants' highest priorities) is shown by asterisks; the vote totals are listed in parentheses.

Table A-2. Data Quantity and Quality – Challenges and Barriers

Data Sharing & Publication

- Sharing failed data ************(12)
    - Publishing good and bad results in terms of data practices
    - Reporting "failed" results
- Proprietary data *********(9)
    - Data is not readily available: it is proprietary, unpublished, or not in the right format
    - Closed proprietary interfaces for equipment prevent researchers from fully understanding the data provided
    - Proprietary data formats
    - Proprietary or controlled data: how to share with community; integrity if censored
    - Proprietary data provision

Data Review & Validation

- Uncertainty quantification *******(7)
    - Uncertainty levels for models, measurements, etc., are currently unknown
    - Experimental data is variable
- Non-expert review **(2)
    - Standard data sets must be evaluated (labeled) by experts (high cost for expert time to accomplish this)
    - Enable crowd sourcing of labeling of data by the non-expert
- Lack of verification and validation of data and tools *(1)

Data Collection & Management

- Sustainability models **********(10)
    - Who will manage the database?
    - No business model to operate and maintain existing solutions
    - No financial support available for mundane, but necessary, work
- Physics and data integration *******(7)
    - Poor connection between physical science and data science
    - How to integrate physics models with data to go beyond data alone
- Lack of incentives for researchers to follow the data collection formats, standards, reporting, etc.; rewarding policy for contributing data; social engineering of data sharing ******(6)
- Lack of data injection pipes/code repositories **(2)
- The slow speeds associated with data collections can limit data collection efforts overall; e.g., synchrotron X-ray collection *(1)

Table A-2. Data Quantity and Quality – Challenges and Barriers

- How do you deal with missing or incomplete data?
- Efficient categories with relevant inputs to simplify data entry (if too complex, entering data becomes cumbersome)
- Legacy production data may not be as robust as data collected in the future. Can they be used together? Can legacy data really be trusted?
- Database vs. data crawlers
- Storing metadata and more, will result in more data reams; this can create a problem for handling data
- “Useful” vs. “non-useful” data: analysis and storage

Data Measurement

- Material properties depend on both composition and processing *(1)
- Descriptors: molecule and crystal are easy; everything else is not *(1)
- Discrepancies in materials synthesis, processing, characterization, testing, reporting *(1)
- Lack of real-time measurements though they have the potential to improve the efficiency of measurement processes (as opposed to time consuming ex-situ measurement) *(1)
- Automating experimental hardware can increase process efficiency and enable much more data collection to take place *(1)
- Instrument data dictates/becomes materials data

Other

- Material scientists are not computer scientists. Methods need to be intuitive for material scientists. *(1)
- Encourage collaboration between computation and experimental efforts
- More robust AI; sometimes AI itself isn’t adequately mature
- AI can only suggest candidates; discovery of materials can only be done in the lab
- Some questions are not suitable to address by AI
- Interactive visualization (human-AI interface); developing for emerging hardware (virtual/augmented/mixed reality)

Priority R&D Pathways Data Quality and Quantity: Figures A-1 through A-6

R&D Pathway: Data Availability

KEY BARRIERS:	SUMMARY OF APPROACH:
• Data is proprietary, not published or limited (lack of measurements or models)	• Establish and develop an open-access database and collaboration platform for AI in materials development

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Understand data gaps and needs • Identify relevant measurements and simulations • Identify data available in publications and open databases	**Milestone:** • Identify data gaps, sources of data, and needed measurements **Target:** • Set up first stage open database development
3-5 years	• Conduct measurements and simulation to collect data • Set up an open database as part of a collaboration	**Milestones:** • Needed data has been collected. • Multiple institutions can upload data and analyze it **Target:** • Ability for collaborating institutions to use the database
>5 years	• Fully developed integrated collaboration platform for data access and data analysis with easy user interface and maintenance plan	**Milestones:** • Well-developed open database and analysis framework • Long-term maintenance plan established **Target:** • Widespread adoption of database and framework

Stakeholders & Potential Roles

Stakeholders

Industry End Users: Provide stakeholder guidance to make the effort relevant to industry

Industry/Materials/Equipment Suppliers: Provide relevant nonproprietary data

Academia: Contribute simulation and measurement data

National Labs: Contribute data and maintain database

Government: Enable the incubation of this collaborative effort; support long-term maintenance of database and framework

Benefits/Impacts

Improved energy efficiency/footprint – High: Faster development of material for energy efficient technologies

Reduces costs – High: Substitute lower cost materials with enhanced performance

Accelerates innovation – High: Rapid development and deployment of solutions

Enhances industry competitiveness – High: Improve competitiveness through lower cost, faster development, and better performance

Faster materials/product development – High: Open data and tools are key to faster development cycle

Enhances manufacturability– Medium: The open data framework can be used for processing parameters in addition to materials data

Figure A-1. Data Availability

R&D Pathway: Lifecycle Data Management

KEY BARRIERS:	SUMMARY OF APPROACH:
• Inputs and data are varied • Desired outputs are varied • Metrics are varied • Analysis and workflow are non-standard and non-reported • Lack of incentive to contribute (computer and academics)	• Tracking the data provenance • Develop data standards through community engagement • Working groups to define metrics

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Definition of provenance best practices and include full provenance with metadata • Identify 1-3 communities with tractable problems to standardize the desired outputs, language, and output formats • Year-long effort to engage communities' at large conferences/ working groups to develop the format	• Documentation with examples of full provenance of data for the communities identified for case study • Standard format and language for 1-3 case study examples to prototype output standardization → adopt • Creation and validation of database for test cases
3-5 years	• Expand the prototype framework to many materials communities • Incentivize data standardization and contribution with grant data management plans • Co-opt existing databases as repository	• Growth in adoption of data management practices • Validate test cases against existing best practices
>5 years	• Materials dossier • Implement interoperability between repositories • Extract from provenance best/common analysis practices and code	• Materials Application Programming Interface (API) • Drivers exist for database to talk • Building software for data in repositories; information using best practices which surfaced from provenance

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Provide input and grow adoption of database usage **Industry/Materials/Equipment Suppliers:** Provide input; include relevant metadata provenance; contribute non-proprietary data (and provenance) **Academia:** Documenting provenance (including inlays, methods, etc.) input on standardization; early adopters **National Labs:** Early adopters; drive workshop activities; drive/host implementation of databases **Government:** Incentivize "good" behavior with respect to provenance; contribute to democratization of publications; support repository in early stages

Benefits/Impacts

Improved energy efficiency/footprint – Low: Potential downstream effects due to increased R&D rate
Reduces costs – High: Reduces duplication for analysis
Accelerates innovation – High: Best practices percolate

Enhances industry competitiveness – Medium: Aggregation of data for everyone
Faster materials/product development – High: Aggregation of data
Enhances manufacturability– Low: Aggregation of data

Figure A-2. Lifecycle Data Management

R&D Pathway: Models for Sustainable Data/Database Management

KEY BARRIERS:	SUMMARY OF APPROACH:
• Inadequate focus on long-term support of materials data projects.	• Establish an interagency collaboration/consortium to measure the current landscape and to measure and address gaps

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Conduct an inventory of existing agency efforts • Quantify gaps and redundancy in existing efforts • Begin to build a knowledge base of agency expertise to eliminate current state silo operations	• Establish an interagency group on materials data (council) • Deploy an online, open access version of knowledge-base • Establish a formal charter for a permanent interagency group/consortium (leadership/officers, by-laws, structure)
3-5 years	• Develop an economic model for sustained materials data funding • Deploy interagency demonstration projects • Refactor existing programs based on gap/redundancy assessments • Engage academia and industry	• 1-3% of budget(s) dedicated to FAIR materials data; National Institutes of Health (NIH) model • Fund at least 5 demonstration projects (government, academia, and industry) • Establish a national, federated data services
>5 years	• Congressional mandate of materials data council • Establish new, pre-competitive consortia	• 2-5% of budget(s) dedicated to FAIR materials data (NIH model) • Industry use of demonstration project data

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Provide input on needs, challenge problems, scientific and technical gaps of materials data **Industry/Materials/Equipment Suppliers:** Provide input on needs, challenges, and gaps **Academia:** Program participation and contribution **National Labs:** Program participation and contribution; provide unique expertise **Government:** Initiate interagency program; provide funding to support

Benefits/Impacts

Improved energy efficiency/footprint – Medium: Data centers are more efficient than individual principal investigators (PIs)

Reduces costs – High: Economy of scale

Accelerates innovation – High: Gained efficiency of coordinated effort

Enhances industry competitiveness – Medium: New sustained databases*

Faster materials/product development – Medium: New sustained databases

Enhances manufacturability– Medium: New sustained databases*

* Interrelated, as data is the foundation

Figure A-3. Models for Sustainable Data/Database Management

R&D Pathway: Unbiased Reporting – Disclosing Failures

KEY BARRIERS:	SUMMARY OF APPROACH:
• Unbiased reporting: Negatives matter!	• Establish scientific, societal, and social benefits of reporting negative results → Better prediction, avoid redundancy, and fill in missing knowledge gaps

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Analyze professional and social factors that dis-incentivize publication of negatives • Establish case studies where knowledge of negatives mattered • Explore pathways toward dissemination of negatives outside "publication"	• Outreach and education/PR campaigns on "report negative" and "negative is not failure" • AI architecture based on deep and adversarial data • Incorporation of negatives in physical model
3-5 years	• Develop AI workflows and strategies to incorporate negatives • Develop awareness of "negative" vs. "bad" experiment • Registry of funding programs; use regulatory tools to recognize reporting "negative"	• Robust prediction of materials, devices, and technological tools for comparing special cases • Minimize stigma of "negative" • Incorporation of robust workflows that allow management of negative results; avoid "production line" trap
>5 years	• Establish integrated knowledge space supporting lateral instrumental networks of "similar" tools • Develop ways to incorporate negatives to explore latent connections in knowledge space	• Accurate and robust universal case for AI design of materials • Establish robust basis for advanced manufacturing past fourth technology revolution and scientific security of the country

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Manufacturing, materials development **Industry/Materials/Equipment Suppliers:** AI companies, database companies **Academia:** Software, algorithms, out of the box ideas; changing academic culture ("publication") **National Labs:** High performance computing (HPC); leading innovation; data repositories; instrumentation networks **Government:** Government funding; leverage industrial support

Benefits/Impacts

Improved energy efficiency/footprint – High: Obviate redundancy; expand knowledge space

Reduces costs – High: Obviate redundancy; expand knowledge space; do things once

Accelerates innovation – High: Expand knowledge space

Enhances industry competitiveness – High

Faster materials/product development – High

Enhances manufacturability– Medium

Figure A-4. Unbiased Reporting and Disclosure of Failures

R&D Pathway: Uncertainty Quantification

KEY BARRIERS:	SUMMARY OF APPROACH:
• Lots of input variables with unknown error scales • Merging data from different sources (experimental or theory); they have different classes of errors • In building models, there are model and prediction errors; formal theory is lacking to describe it	• Fund and build theory, communicate, and gain confidence from the materials community • Deploy tools to help make better informed decisions

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Develop theory to handle uncertainty for all classes of machine learning methods • Communicating the importance of such metrics • Better decision-making, knowing uncertainty will exist in every measurement/ prediction	• Milestone: Awareness • Standardization of error reporting • Motivate people to take up challenge • Cannot get rid of uncertainty; be able to work with it
3-5 years	• Standardizing data collection practices • Discovery of errors (benchmarking) by comparing different sources/models • Version control practices of models/experimental recipes to know when things work or not	• Milestone: Change of culture • Gain confidence in using information from different sources • Kaggle-type approach where models are scrutinized via crowd-sourcing
>5	• Build capabilities to make developments accessible to end users	• Milestone: Deployment • Get end users to adopt the technology

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Feedback with academia/national labs on areas of improvement; implement process/product developments **Industry/Materials/Equipment Suppliers:** Develop easy-to-use platforms that incorporate uncertainty theory **Academia:** Take up the challenge and build new theory **National Labs:** Communicate results/developments **Government:** Fund fundamental theory development

Benefits/Impacts

Improved energy efficiency/footprint – Low: Depends on project focus

Reduces costs – Medium: Reduced cost of trials

Accelerates innovation – High: Better informed decisions on next steps

Enhances industry competitiveness – High: U.S. early adoption

Faster materials/product development – High: Can optimize faster

Enhances manufacturability– Medium: Depends on project focus

Figure A-5. Uncertainty Quantification

R&D Pathway: Validation and Verification of Theory-Based Models

KEY BARRIERS:	SUMMARY OF APPROACH:
• Simulation and experiment data must be compared in the same domain/space (e.g., X-ray diffraction to X-ray diffraction pattern) • Bias in simulation data/model must be quantified	• Machine learning approach to parameter tuning in physics-based models • Investigate methods for simulation model bias quantification • Curated experimental databases/standards

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Identify available databases for validation and verification study • Identify software tools for mapping experimental and simulation data into the same domain; e.g., crystallographic information file (CIF) to X-ray diffraction • Research diffraction methods for quantifying bias	• Summary paper of validations
3-5 years	• Application of methodologies to select datasets	• Evaluation and solution of model bias in specific context
>5 years	• Build tools for general use	• Software tools • Closed loop parameter tuning in physics models

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Identify most impactful simulation methods to improve **Industry/Materials/Equipment Suppliers:** Supply materials for experiment databases that result in consistent results **Academia:** Research methodologies for bias estimation **National Labs:** Build experiment database standards **Government:** Fund research

Benefits/Impacts

Improved energy efficiency/footprint – NIP

Reduces costs – NIP

Accelerates innovation – NIP

Enhances industry competitiveness – NIP

Faster materials/product development – NIP

Enhances manufacturability– NIP

Figure A-6. Validation and Verification of Theory-Based Models

Appendix A.2. Platforms and Infrastructure Breakout Session Results

Future Capabilities and Targets

FOCUS QUESTION 1: *What are the key capabilities, technologies, characteristics, or targets you want to see in future for AI as applied to materials discovery and design?*

Table B-1. Platforms and Infrastructure – Future Capabilities/Targets

Data Capture & Management
• Data interoperability to manage data and curate data. • Automatic integration of disparate data sets. • Intuitive data retrieval: “easy” ways for R&D teams to quickly harvest data relevant to specific applications • Flexible, configurable open-source data mining and public repository of scientific literature. • Integration with journals – new data is available as soon as possible. • Consolidate various open source and proprietary data. • Data fusion between experimental data and modeling data • Seamless data ingestion between structured and unstructured data • Use artificial intelligence to capture experts’ knowledge and processes • Version control of publications, with notification to anyone that used the data, requesting/suggesting those researchers rerun their analysis with the updated data • Domain-specific recommender systems to guide researchers to data collections and other resources. • Data collection through the end use stage • Image analysis to extract data from diagrams, figures and tables in literature • Traceable analysis – data is tagged with sufficient metadata for reproducibility of all processing steps. • Advanced natural language processing capability to extract data and metadata from published scientific papers • Store a methods graph as metadata to compare and improve repeatability; machine readable methods section • Evaluate consistency and reliability of publicly available data. What experiments might need to be repeated more carefully? • A format that makes journal articles/publications mineable and connects the text with the data set; published papers are metadata. • Collaboration between different communities on data standards ready for machine learning use. • One format to cite papers and other sources that is machine friendly – to facilitate text mining.

Table B-1. Platforms and Infrastructure – Future Capabilities/Targets

AI System Capability

- Have knowledge management capability whereby design tool can explain what it knows – ability to track evolution of knowledge acquired by AI system in time.
- AI should be able to distinguish "good" data versus "bad" data.
- The capability to use either composition processing or properties as an input, and get the other as the output result.
- Ability to identify duplication of efforts rapidly.
- Evaluate "trade-offs" in design vs. supply vs. cost vs. performance
- Able to establish the quantitative relationship between chemistry, microstructure, and properties.
- AI should be able to identify what gaps in data need filling
- Able to identify a correct approach to address a materials problem
- Learn new physics from outliers
- AI should do the Peer Review

Data Algorithms & Models

- AI algorithms that extract physics and chemical insights from materials data.
- New machine learning algorithms for extrapolations; algorithms that are specific to materials science
- Interpretable machine learning models; models that are not a "black box," it provides insight on why or how it predicted something.
- Build models from large data with feature interactions; many features should come together to give model a predictive capability.
- More connections between physical models and machine learning/artificial intelligence.
- Derive physics based models from AI
- Integrating multi-scale models, especially from new models in different fields
- Build data-driven models that generalize very well
- Flexible data models and formats to facilitate interoperability and reuse of data and preserve metadata
- Searchable database for machine learning models

Tool Framework, Function, and Interaction

- Standard application programming interface (API) for system-to-system interaction and data transfer.
- Function to develop metadata catalog
- AI tools for complex and multi-modal data sets
- A common AI framework backed with experimental data (multi-modal) to enable prediction at scale
- An "app store" type interface for data post-processing tools – enable easily publishing new methods
- Search capabilities that steer you closer to what you are looking for, without need to experiment with keywords, especially when you do not know what the right keywords are
- Able to have platforms that will work across materials disciplines
- Need an ecosystem that maximizes metadata (in addition to the data)
- Uncertainty quantification (UQ) for feedback to model/data help on extending design allowable spaces
- 10x-100x computing cost reduction for property prediction per point
- More open source platforms and codes for AI and ML (example: R and RStudio[14])

[14] R (https://www.rstudio.com/) is a free software environment for statistical computing and graphics. RStudio (https://www.rstudio.com/) makes R easier to use, it includes a code editor, debugging & visualization tools.

Table B-1. Platforms and Infrastructure – Future Capabilities/Targets

Data Sharing & Collaboration
• Libraries that extend learnings from known classes of materials to new classes • Publication infrastructure to share: a) data, b) code, and c) findings. • Development of libraries for R (programming language) specifically for materials research • New legal frameworks to make it easier for companies to share data without problems for universities' non-profit status • Artificial intelligence should generate proposals, and groups would offer funding for AI to solve the problem • AI networking suggestions - capability to suggest other researchers who are (or may be) working on related/relevant technical challenges
Education and Training
• Educate the next generation of engineers and scientists • Educational infrastructure in machine learning/artificial intelligence for scientists and engineers (currently education in ML/AI is just for computer scientists) • Educate materials engineers how to build statistical surrogates (metamodels) • Easily discoverable tutorials with data sets and examples of analysis.

Technical and Scientific Challenges

FOCUS QUESTION 2: *What are the major scientific and technical challenges that limit the application of AI for materials design and discovery? What are the problems that hinder us from realizing the desired capabilities, technologies and targets identified in Session 1?*

Workshop participants were asked to vote on the barriers and challenges they perceived as most important. The number of votes received (indicating participants' highest priorities) is shown by asterisks; the vote totals are listed in parentheses.

Table B-2. Platforms & Infrastructure – Challenges and Barriers

Data

- Data extraction and standardization ***************(15)
    - Clustered data, historical data from evolutionary operation, lack of generalization and design of experiments
    - Standards for data in reports and publications
    - Contextual extraction of information from journal papers, reports, etc.
    - Material science has a long history – data mining of old papers would be very helpful
    - How to search across open & proprietary databases. How proprietary data can be protected – not extracted.
- Challenges in fusion of data from different sources *****(5)
    - Lack of standards in data generated; merging data from unequal sources
    - Access to machine-readable materials data
    - Infrastructure to make good use of government-paid-for data is not present; quality verification is not present either
    - Proprietary data formats for synthesis and characterization tools (microscopy metadata, control of additive manufacturing machines); lack of data input file standards
    - Lack of flexibility in data formats; data formats need to be future-proof
- Lack of standardized data representations across range of time and length scales (multiscale models) **(2)
- Process-structure-property data acquisition technique is not well developed for big data science **(2)
- Characterization of microstructure that is ML ready - in a standardized way
- Lack of data from well-controlled materials systems
- How to describe/organize material microstructures for machine learning/AI)?

Models & Tools

- Physical models are needed that are compatible with AI **********************(22)
    - Thermodynamic and kinetics database for different material systems; materials science lacks a common database for process parameter-driven properties
    - The materials science community is concerned that machine learning results will not be placed on a physical basis. Conversely, there is value in using results to reveal unexpected trends and correlations that provoke new physical investigation.
    - Convince computer scientists that physical interpretability is essential; scientists need to understand black-box model
    - Need to develop physics-based microstructural evolution model in additive manufacturing before creating machine learning model.

Table B-2. Platforms & Infrastructure – Challenges and Barriers

- AI for materials needs tools that can identify: a) anomalous points, b) physically outlier points, and c) sets of outlier points that require their own distribution (analogy to extreme value statistics) ***(3)
  - Which outlier is the good one? Can machine learning predict? Need algorithms to predict out of the box.
  - Lack of machine learning algorithms with the following capabilities: Interoperability, inclusion of physics, learning from outliers (instead of rejecting outliers)
- Semantic classification of machine learning models **(2)
- At each scale, infrastructure for merging the multi-modal experimental data for model validation and uncertainty quantification is lacking **(2)
- Scalable platforms and algorithms; what is available currently is for smaller studies **(2)
- Can generalized machine learning models be used for specific material subjects? Metals vs. ceramics vs. composites. Different "languages" in different materials science disciplines. *(1)
- No uniform set of synthesis/characterization techniques across materials *(1)
- Need more accurate computational prediction of properties of complex materials. Examples: liquids, polymers, interfaces.
- How can self-perpetuating reinforcement be avoided? Model predicts a trend, so in exploring that trend area, more data just reinforces the trend.
- Even with established computational methods, unknown parameter values needed to adjust, tune method to a particular problem.

Software

- Software maintainability – Code written years ago might still be needed today, but language and platforms change. ****(4)
  - Now materials scientists need to spend effort maintaining software. Even worse when multiple pieces of software are linked together. Academic researchers have no incentive to create commercial-grade software, it would be a waste of their time.
- Data visualization and recommender systems *(1)
- Availability of high quality software, such as Dakota *(1)
- Academic researchers need incentives to produce good software (extensible, user-friendly, stable) *(1)
- Frameworks that can put together related bits of open source software, e.g. DREAM3D[15] *(1)
- No good materials design software has been developed yet

Fundamental Science Challenges

- Materials properties depend on process variables; AI needs to be able utilize process variable data
- Combinatorial space is enormous. How many possible configurations need data?
- Property vs. behavior; they are different

Education/Human Resources

- How to educate materials scientists in artificial intelligence/machine learning? Who teaches it? What does it supplant? What curriculum? ********(8)
- Realize our weakness/strength. Communications between domains, statisticians, and high performance computing.
- Need training on effectively communicating materials problems to computer scientists. Materials science tutorials to computer scientists?

[15] DREAM.3D stands for: Digital Representation Environment for Analyzing Microstructure in 3D. http://dream3d.bluequartz.net/

Table B-2. Platforms & Infrastructure – Challenges and Barriers

• Researchers need incentives to standardize their data/metadata
Other
• Lack of pre-packaged "challenge problems" to allow computer scientists to push development of approaches without significant collaboration **(2)

Priority R&D Pathways Platforms and Infrastructure: Figures B-1 through B-3

R&D Pathway: Connecting Physical Models to AI Applications

KEY BARRIERS:	SUMMARY OF APPROACH:
• AI algorithms are not designed to take physical models and prior knowledge into account (input) or provide a physical interpretation of results (output); both are especially problematic for deep learning	• A multifaceted approach that combines research and development activities with opportunities to convene the community around this problem.

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Identify systems that are good candidates to implement physical constraints on learning input/output • Identify gaps in physical knowledge salient to materials problems; identify areas that are well-understood with physical models • Apply computer science interpretability techniques to materials science classification tasks (i.e., what does the computer use to classify microstructural images?) • Engage computer scientists to define the physical constraint and physical interpretability problem space	• Sponsor a study to bring computer scientists and materials scientists together around this issue o A national academies-style report on bridging machine learning and materials science • Define a funding opportunity announcement (FOA) around implementing physical models in AI approaches to materials science o Fund a set of seed projects in this area selected such that outcomes will help to guide future program development (including measures such as internal rate of return)
3-5 years	• Use AI to utilize thermodynamic and kinetic models to build a bridge between processing, structure, and properties for a real system • Use AI to "correct" the systematic approximation errors in density functional theory (DFT) results to better match experimental data • Identify problems with opportunities to explore the three classes of outlier data: anomalies, physical outliers, and clusters of outliers (require their own distributions)	• Successful proof-of-concept for AI supplementing physical models, including DFT and Computer Coupling of Phase Diagrams and Thermochemistry (CALPHAD) • Develop a program around understanding outliers in AI data sets
>5 years	• Use AI to build a bridge between models and experimental results at multiple length and time scales • Use AI to integrate data from models, simulations, and experiments, with appropriate weighing of the data in different regimes	• Execute examples of the long-term action items o AI provides a statistical surrogate for physics-based simulations or experiments.

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Provide input on needs, problems, state-of-the-art, and data **Industry/Materials/Equipment Suppliers:** Think of Citrine, Thermo-Calc, Autodesk, etc. Provide data, open source tools, APIs, and the glue that connects various activities. Develop anonymous/obfuscating data sharing paradigms. Help accelerate technology transfer. **Academia:** Develop, optimize, and validate algorithms, software, data sets, etc. **National Labs:** Same as academia, and also provide data repositories and standards (NIST). **Government:** Support foundational R&D; provide links between stakeholders. Incentivize data sharing and open source software.

Benefits/Impacts

Improved energy efficiency/footprint – Medium: Potential for more efficient materials and process design
Reduces costs – High: Fewer experimental and computational cycles, and potential for improved outcomes
Accelerates innovation – High: New science

Enhances industry competitiveness – High: Reduces costs, timeframes, and outcomes
Faster materials/product development – High: Reduces timeframe
Enhances manufacturability– Medium: Potential for more efficient materials and process design; potential for improved outcomes

Figure B-1. Connecting Physical Models to AI Applications

R&D Pathway: Data Extraction and Standardization/ Standardization and Translation from Traditional Data Sources

KEY BARRIERS:	SUMMARY OF APPROACH:
• The barrier to extracting data from legacy documents (even those directly machine readable) is on a similar scale to language translation in the sense that a very large amount of context will be needed in order to understand and categorize individual pieces of data.	• Two part solution: (1) Creating a new AI tool that provides a contextual approach to analyzing legacy articles and documents to extract data (e.g., linking figures to text); and (2) establishing new standards for articles to ensure that they are fully searchable and data-extractable.

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Articulate the barrier to legacy data extraction in terms of natural language processing with the objective of engaging the computer science (CS) community • Establish a new standard for articles and reports in materials science to define topics and tags that should be embedded. Engage the CS community, e.g., via those working on the ontology of scientific reporting • Survey other topical areas for state-of-the-art, e.g., cheminformatics, bioinformatics • Develop a schema/ontology for data exchange in materials science	Milestones • Demonstrate that a searchable article is feasible with demonstrations of at least two examples Targets • Run workshop(s) to get community input on data standards • Run workshop on digitization of legacy documents: best practices, etc. • Run workshop on natural language approach to analyzing legacy documents
3-5 years	• Implementation of data exchange standard(s), e.g., with major publishers and laboratories • Acquire a substantial data set based on legacy document analysis • Develop a publicly accessible database that supports the natural language approach to extracting data from legacy documents	Milestones • Demonstrate efficacy of a data schema and ontology • Demonstrate an approach to extracting useful data from legacy documents that includes development of contextual framework Targets • Have more than two research groups to provide a corpus of documents (including journal articles) embedded in a searchable format according to an established ontology • Implement a publicly accessible database that supports the natural language approach to extracting data from legacy documents
>5 years	• Legacy data extraction in terms of natural language processing • Development of flexible methods for scraping data from existing databases	Milestones • Demonstration of a data base of documents that are fully searchable in a given topical area in Materials Science • Demonstration of a natural language approach to acquiring data from legacy documents

Stakeholders & Potential Roles

Stakeholders	**Industry End Users:** Essentially all manufacturers deal with materials. Therefore, materials informatics affects all manufacturers. These entities will define needs and priorities, and, in some instances, data (e.g., legacy documents for analysis). **Industry/Materials/Equipment Suppliers:** See above Industry End Users. **Academia:** Develop, e.g., natural language approach to analysis. In partnership with other actors, define, implement and demonstrate standards for schema and ontologies. Advocate for open access to data and publications. **National Labs:** See above for Academia. **Government:** Provide incentives, funding, support and advance policies in this area. Provide database. host workshop for near term

Benefits/Impacts

Improved energy efficiency/footprint – Medium: Impact is indirect but ultimately important; i.e., comes via impact of R&D efficiency
Reduces costs – Medium: Everyone spends a lot of time looking at old papers so this could save a lot of time for many people
Accelerates innovation – High: There is a large (enormous) amount of legacy data that is otherwise very hard to include in AI

Enhances industry competitiveness – Medium: Lack of data is a major bottleneck so this area has the potential to accelerate industry R&D (maybe High?)
Faster materials/product development – High: Materials development depends critically on data and legacy data is difficult to use unless available digitally
Enhances manufacturability– Low: No direct connection, although there is a "path to relevance"

Figure B-2. Data Extraction and Standardization/ Standardization and Translation from Traditional Data Sources

R&D Pathway: Data Fusion

KEY BARRIERS:	SUMMARY OF APPROACH:
• Fusing data from different and disparate sources is challenging • There is a lack of standards for data generation, data storage, data provenance, and metadata	• Push data format standards as early as possible in data generation life cycle • Address future data first, then legacy data • Data associated with publication and other deliverables should be subject to FAIR principles

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Push for community-scale adoption of data formatting and metadata standards • Establish publicly accessible data repositories • Establish infrastructure for data discovery	• Prototype data standards • Prototype data repository and registry infrastructure
3-5 years	• Establish community-wide groups to identify and manage an ICME community data fusion challenge • Work toward data fusion procedures for heterogeneous (multiscale, temporal, spatial) datasets from different sources	• Define a foundational ICME data fusion challenge problem
>5 years	• N/A	• N/A

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** N/A **Industry/Materials/Equipment Suppliers:** N/A **Academia:** N/A **National Labs:** N/A **Government:** NIST, DOE, Army, Air Force – promoting standards, community testing, fundamental research

Benefits/Impacts

Improved energy efficiency/footprint – Medium: Reduced repeated experiments – indirect means

Reduces costs – High: Reduced time and effort

Accelerates innovation – High: Enables deep exploration of design space problem

Enhances industry competitiveness – High

Faster materials/product development – High

Enhances manufacturability– High

Figure B-3. Data Fusion

Appendix A.3: AI for Materials Design in Specific Applications Breakout Session Results

Future Capabilities and Targets

FOCUS QUESTION 1: *What are the key capabilities, technologies, characteristics, or targets you want to see in the future for AI as applied to materials discovery and design for specific applications?*

Table C-1. Specific Applications – Future Capabilities/Targets

Data Capture & Management

- Data made available for open access; ubiquitous open data libraries and databases (no limited/tiered/restricted access)
- More complete database, especially in structures
- Unified database – common format and open access
- Image analysis software to quantify scientific images
- Rapid data analysis system for converting raw data to meaningful data
- Robust workflow system to easily extract relevant information
- Robust information capture pipeline for past/present/future data

AI System Capability

- Ability to address and make predictions about meta-stability or synthesis of inorganic materials
- High-accuracy (smart) prediction of key materials (performance) properties that helps steer R&D pathway (materials synthesis)
- Ability to impact design of complex systems involving multiple materials and interfaces (not just materials), such as batteries, etc.
- Materials process/structure/property – reduce the uncertainty of propagation of these parameters into design by providing a user-friendly tool/interface
- Capability for in-situ/in-operando materials characterization
- Rapid characterization and high throughput techniques for generation of large data sets
- Ability to physically run experiments and make some decisions, then repeat
- Capability to select dopants for new alloys based upon part geometry and physical requirements
- Effective representations, specifically for ML
- Approaches to estimate upper and lower bounds on a property based upon either physics or data mining
- Time guestimate for optimization/design based on present amount of data
- Robust predictive theories for circumstances where fundamental physics theories do not exist
- Faster design-driven materials selection and qualification during additive manufacturing to get get the timeframe from idea to product to be under 3 years
- Robust statistical descriptors for microstructures with error estimation during image analysis
- High throughput computational materials property design
- AI machine with the following characteristics: [application analyzer] → determine specified properties → [AI generator] → select material candidates → [AI selector] (with auto-test quality) → determine candidate material
- AI for optimizing powder atomization to achieve better yields of desired particle size distribution
- AI for additive processing and ability to capture complexity of multi-microstructures and/or materials (hybrid structures)
- Using AI models to create a “smart factory” – adjusts processes with varying material properties

Table C-1. Specific Applications – Future Capabilities/Targets

- Rapid assessment from literature of what is state-of-the-art before starting to optimize a material for a set of properties.
- Transfer learning for multiple materials applications
- "Remembering" previous systems for which a neural network has been trained (as applied to new systems)
- Need AI/ML to bridge the gap between coupon data and component-specific, location-specific property
- Need to close the loop between ML and experimentation

Data Algorithms & Models

- Material structure and property relationships: need better models and experimental validation of the models
- Going from micrograph to the model must be seamless; need a generic description of a polymer (a very difficult task)
- Universal AI/ML algorithms that bridge length of measurement across and timescales (interfaces, microstructures, meta-stability, kinetics)
- AI algorithms for sparse data
- Advanced AI/ML techniques requiring low-quality data sets
- Development of new methods that can integrate physical-based multiscale models with ML methods; the methods should be transferable to that they can be used for discovery of different types of materials and for different applications
- Documentation of successes and failures during running of open source codes
- Integration of models and standardization of information transfer
- An ideal AI modeling framework:
    - Input: rich dataset
    - Processing: physics-based, reproducible
    - Output: Accurate prediction, robust
    - Materials: Shape alloys
    - Property: Thermal, mechanical
- Guided verification and validation of AI models and hybrid systems (AI and physics); this is an ideal world for AI
- Physics-informed AI with uncertainty quantification for materials discovery
- AI that integrates physics models, subject matter expertise, sensitivity analysis, and uncertainty quantification
- ML algorithms for manufacturing process data and correlation to quality. Is there a library for gas turbine manufacturing? (Isn't clear if there is.)
- Most of the approaches are heading toward a data rich domain; need to develop approaches related to minimal data as well

Tool Framework, Function, and Interaction

- Easy-to-use tools and software
- User-friendly open platforms that protect data but allow materials engineers to rapidly target specific properties and find resulting materials
- Open-source codes to implement AI
- Software that integrates existing AI code with chemistry code and all open source
- Watson for manufacturing (natural language); natural language-based approaches need to be explored (mathematics and linguistics)

Table C-1. Specific Applications – Future Capabilities/Targets

- Standardized use cases and tools for AI verification and validation
- Development of experimental methodologies that can characterize in real time materials while synthesizing them to validate AI
- AI frameworks across design, process, structure, and property relationships for additive manufacturing and other manufacturing/production
- Decision science applied to autonomous experimentation; when there are many ML planners, must decide what is needed

AI Application Areas

- Additive manufacturing (AM)
    - Recommended materials based on design and operability parameters for AM
    - High-temperature materials via AM
    - Accelerated parameter optimization for new additive materials (alloys)
    - Prediction and miniaturization of residual stress from AM
    - Optimize build plates based upon material and part geometry for DLMS (direct laser metal sintering) and EBM (electron beam melting)
    - Group-based (factory) learning for process improvements in AM
- Genetic programming of atomically precise materials: 1) catalysts (such as enzymes), and 2) membranes
- Fingerprint generation for inorganic materials
- Need AI for processing recipe for process temperature and time of metals and alloys
- AI-driven adoption of casting alloys and alloys
- AI application to component risk estimation from multiple failure modes (such as temperature, mechanical, chemical, etc.)
- Use of AI for: 1) real-time analysis and ML of materials data (especially high-throughput data), and 2) autonomous synthesis of materials samples (libraries)

Education and Training

- Education/training in ML that is advanced in materials science
- Basic tools for AI training
- AI trained and used directly by materials scientists rather than the AI programmer
- Need software engineering with materials research
- Applied mathematical and numerical methods skills for materials research
- Good case studies (perhaps in book form) for materials and manufacturing

Other

- While applying AI to materials, it is important to consider “context” and “values”
- Top capabilities and targets:
    - Automated synthesis platforms
    - Better experimental and computational data
    - National clearinghouse for data
    - Interactions between human and machine interface
    - API interfaces for “big” data
    - Evolution and advancement of AI integrating code and having chemistry/physics basis

Technical and Scientific Challenges

FOCUS QUESTION 2: *What are the major scientific and technical challenges that limit the application of AI for materials design and discovery? What are the problems that hinder us from realizing the desired capabilities, technologies and targets identified in Session 1?*

Workshop participants were asked to vote on the barriers and challenges they perceived as most important. The number of votes received (indicating participants' highest priorities) is shown by asterisks; the vote totals are listed in parentheses.

Table C-2. Specific Applications – Challenges and Barriers

Data/Data Acquisition/Data Management

- Lack of systematic way of collecting data ***(3)
- Machine-accessible data that is interoperable across multiple platforms **(2)
- Characterization tools to capture lots of data quickly **(2)
- Reproducibility of published data (reporting, metadata) *(1)
- Selection/development of new sensor technologies and development of (real-time) characterization of obtained information *(1)
- In-situ physical data collection in real time – stress, phase, microstructural chemical etc. – during material processing *(1)
- Material and processing is initial value problem; AI should be coupled with sensors; fusion and qualification *(1)
- Lack of high throughput precise experiments *(1)
- Lack of experimental data *(1)
- Need more and "correct" data
- Is there a big data problem?
    - Database - Compared to Google or Facebook, sufficient databases are lacking; MGI has one database
    - Computational - How can the tools be obtained?; access is needed to the national labs
    - Experimental: high throughput; equipment; non-contact/non-destructive; auto-motor; beam-time
- Proprietary file formats; for proprietary data, who would have all the information about parameters of your project?
- Infrastructure for "data discovery" across functional areas (digital thread); infrastructure in terms of the discoverability of the data across many functional areas needs to communicate
- Deciding what data is relevant to store in a large, unified database
- Lack of sharing industry data challenges
- Inconsistencies in datasets/conditions; this affects the output of AI-based discovery
- Tools for better data: quality, quantity, and accessibility
- Notion of threshold of quantity of (imperfect) data (i.e., how big a set of data is needed to predict something meaningful?)

Models & Tools

- Fundamentals of science (e.g., thermodynamics and physics) should be incorporated into AI modeling of materials **********(10)
- Linkage of physics-based models across multiple scales with data analytic approach **********(10)
- Physical structure related to numerical representation; there needs to be a two-way street of physical structure to numerical representation and vice versa *********(9)

Table C-2. Specific Applications – Challenges and Barriers

- Integration of smart materials discovery *******(7)
- Length scales, time scales, and degradation processes *******(7)
- Develop techniques to rapidly synthesize a range of chemistries relevant to future processing ******(6)
- Lack of predictive materials models with different fidelity and efficiency ****(4)
- Bridging the gap between modeling (theoretical) data and experimental data time and complexity ****(4)
- Quantification of uncertainty propagation into design and component risk ***(3)
- Multi-scale uncertainty quantification in process-structure-property modeling chain ***(3)
- Finding the right features for learning algorithms ***(3)
- Strategies for model selection and what to use? ***(3)
- Need a way to have AI intelligently understand materials; uncertainty and an "issue of trust" ***(3)
- Applying AI to multi-physics simulations *(1)
- Physical interpretability: fidelity vs. feasibility *(1)
- Qualification/trust of quantification and interpretability of AI results *(1)
- Predictive accuracy (strongly correlated materials and realistic chemical conditions) *(1)
- Sparse directed AI/ML; how does AI operate on sparse datasets that are focused in the directions of interest? *(1)
- Data visualization tools *(1)
- Need new algorithms/methods for small data ML
- Lack of theory and working simulation
- Lack of fast testing/iteration cycle
- Metadata, time, and spatial scales needed for the models and validating models to go across the scales; scarce data lacking pedigree and metadata; models across time and space scale needed to fill gap
- "Black box" AI; need relatable and interpretable AI/ML; hypothesis generation and testing
- AI to include design validation (low population) in learning algorithms having impact on complex boundary conditions impact
- Learning methods that leverage physical knowledge/other symmetries
- No one-to-one mapping of property-structure-process when doing inverse design
- Scalable representations/features

Software

- Lack of easy-to-use-software
- Scientific challenge: accuracy of models, computer speed. Technical challenge: the way software code is written can result in code "islands" disconnected from the overall program software**(2)

Fundamental Science Challenges   Model/Tool Development and Strategies

- Is it possible to develop new materials without understanding underlying physical principles? ********* (9)
- Understanding physical limitations of different materials and designs
- How to manipulate microstructure?
- Synthesis hurdle: lack of available synthesis information
- Lack of physical/chemical properties specified for a wide range of systems
- Material-specific physics that require unique AI solutions for each
- Too many discrete R&D problems and experimental tests involved (many missing links) requiring time and money

Table C-2. Specific Applications – Challenges and Barriers

Education/Human Resources
• Insufficient training of manufacturing engineers in the benefits and capabilities of AI and ML ***(3) • Better knowledge sharing between materials scientists and computer scientists **(2) • AI-guided data generation to fill in sparseness to enhance prediction **(2) • Lack of training programs to develop technical "pentathletes" **(2) • Lack of computer infrastructure and AI-trained researchers *(1) • Lack of formal education in AI for students and researchers in materials science • Education of how to use AI: required skills and science, and knowing limitations • Development of curriculum/short course for existing materials scientists/engineers; "How do I get existing workforce knowledgeable on AI?" • Lack of coordination between academic disciplines • Many material scientists are traditionalists reluctant to take on something new. Those trained in AI/ML are creating converts
Other
• Lack of cost-sharing mechanisms, such as a technical application for cost-sharing *(1) • Lack of clear objectives for AI

R&D Pathway: Product Functionality-Driven Accelerated Materials Discovery, Manufacturing, and Deployment by AI

KEY BARRIERS:	SUMMARY OF APPROACH:
• Current methodologies of testing and validation of materials are long and disconnected • Lack of tools for rapid synthesis and characterization that is scalable • Lack of equipment that is non-contact and non-destructive for rapid characterization • Lack of integrated learning methodologies that simultaneously process structure, property, and performance to reduce risk of new materials • Being able to capture variability in vendors	• Start with state of the art review; identify 5 gaps and deploy existing solutions, including in-situ sensors, modeling, and uncertainty quantification • Enable a hub for coupling materials discovery and manufacturing

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Concerted effort to identify gaps in state-of-the-art • Evaluate and deploy existing tools not currently used in materials discovery • Synergistic modeling of multi-materials • Integrate in-situ sensors for real-time monitoring	• A review on state-of-the-art, gap, and needs • Clearinghouse for validated sensors and software modeling • Experimental and modeling uncertainty quantification (UQ) deployment
3-5 years	• Hub for smart materials manufacturing (public-private partnerships) • Autonomous synthesis, characterization, and validation • Strengthen characterization and property relationships • Tools to identify non-equilibrium states	• Reduce lifecycle of discovery and scale-up by 30% • Tool development (prototype demonstration) for synthesis and characterization • Tools to interrogate non-equilibrium modality
>5 years	• Materials tricorder: ability to rapidly identify materials for composition, process, and source • Cradle-to-cradle lifecycle assessment	• 50% reduction in cycle time for material discovery to in-production at lower energy costs • Materials systems performance prediction • Fully integrated system

Stakeholders & Potential Roles

Stakeholders	**Industry End Users:** Define goal & targets, provide feedback on tools & technology **Industry/Materials/Equipment Suppliers:** Provide tools, software, sensors (and UQ); develop technology with greater technology readiness level (TRL) **Academia:** Underlying software, sensors, and high-quality data **National Labs:** Be a hub for high performance computing, infrastructure (demo), standards & validation (professional) **Government:** Support; business model; coordinate agencies

Benefits/Impacts

Improved energy efficiency/footprint – Medium/High: System level improvement (not one step)

Reduces costs – High: Reduce life-cycle R&D

Accelerates innovation – High: Tools will be there to give knowledge

Enhances industry competitiveness – High: Rapid innovation cycle; component cost

Faster materials/product development – Medium/High: Cyclic economy possible

Enhances manufacturability– Medium/High: All of the above

Figure C-1. Product Functionality Driven Accelerated Materials Discovery, Manufacturing, and Deployment by AI

R&D Pathway: Data-Driven Materials Discovery Paradigm

KEY BARRIERS:	SUMMARY OF APPROACH:
• High-throughput measurement technologies, machine learning algorithms that train well with small data, integrated AI/robotic model	• To augment/supplement serendipity and "expert intuition" in materials research with a more consistent/reproducible AI models

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• High-throughput precise experiment technology (e.g., model system straightforward, "easy-to-measure") • Database management	• Sufficiently accurate AI model o No physics counterpart: Establish baseline standard o With physics counterpart: Win out in some metric • Predictive property
3-5 years	• Better algorithm to train on limited data • More interpretable model/human-in-the-loop • Expanded metadata collection; incentivizing data-sharing (centralization)	• Minimum experimentation o More "intelligent" serendipity: Reduce number of experiments versus random trial • Rediscovery o Easier-to-train better performance of combined model • Guidance
>5 years	• Robotics • Integrate AI system with other systems in the production chain	• AI automation o Combine AI model with robotics: "fully automated" material discovery o Move downstream requirements to early-stage discovery – "manufacturability"

Stakeholders & Potential Roles	
Stakeholders	**Industry End Users:** Provide experimental data **Industry/Materials/Equipment Suppliers:** N/A **Academia:** Basic research into algorithm, high-throughput technologies **National Labs:** Basic research into algorithm, high-throughput technologies **Government:** Provide incentives for data management/sharing

Benefits/Impacts

Improved energy efficiency/footprint – Medium: More targeted experiments, less experiments

Reduces costs – Medium: Less experiments reduce cost and "trial-and error"; rediscovery approach reduces experiments costs

Accelerates innovation – High: More targeted experiments, less experiments

Enhances industry competitiveness – High: Faster time from conception to final product

Faster materials/product development – High: More targeted experiments, less experiments

Enhances manufacturability– Medium: Integration with other systems in production chain

Figure C-2. Data-Driven Materials Discovery Paradigm

R&D Pathway: Fundamental Models and Experimental Data Integrated with AI-based Approaches to Develop Optimal Materials & Processes

KEY BARRIERS:	SUMMARY OF APPROACH:
• Representing material structures and processes to interoperate with AI software tools • Central databases of structural data, process data, and experimental and computational data • Interpreting AI-derived solutions to generate new physical models	• Integrate physical modeling with appropriate AI tools and human intuition to guide material and process development • Using experiments and physical constraints to guide AI and the solutions they provide

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Conduct inventory of software, tools, databases, protocols, and data representations in different domains that can be integrated with AI tools • Understand gaps in existing tools	• Domain-specific road maps to get from current software, databases, etc. to where AI tools can be used to design materials and processes
3-5 years	• Support software tools to integrate open-source AI software with open-source physical modeling code • Develop protocols to represent disparate materials and processes in databases and centralized databases that can be accessed by researchers applying AI techniques	• Adapt existing AI and physical modeling software to new protocols to grow an AI platform for designing materials and processes • Visualization tools for representing data and developing workflows/pipelines for designing materials with AI assistance • Domain-specific software for prediction of specific material properties (catalytic rates, bandgaps, melting points)
>5 years	• An ecosystem of domain-specific software tools and databases and protocols that allow researchers to use evolving AI tools to design materials and processes • Demonstrations in several domains that AI techniques can be successful in designing new and improved materials and processes	• Development of software tools, databases, communications protocols (data exchange, file formats) integrated with AI software tools that researchers can use to develop new and improved materials and processes
	Stakeholders & Potential Roles	
Stakeholders	**Industry End Users:** No information provided (NIP) **Industry/Materials/Equipment Suppliers:** NIP **Academia:** NIP **National Labs:** NIP **Government:** NIP	

Benefits/Impacts

Improved energy efficiency/footprint – NIP
Reduces costs – NIP
Accelerates innovation – NIP
Enhances industry competitiveness – NIP
Faster materials/product development – NIP
Enhances manufacturability – NIP

Figure C-3. Fundamental Models and Experimental Data Integrated with AI-based Approaches to Develop Optimal Materials & Processes

R&D Pathway: Physical Structure Related to Numerical Representations

KEY BARRIERS:	SUMMARY OF APPROACH:
• Need universal numerical representation of physical microstructure derived from experimental and computational data for connecting material structure to properties	• Integrate current "state-of-the-art" for image processing and numerical representation into materials science • Methodically explore the existing/possible feature-space that exists in inorganic materials, and intelligently find/develop appropriate numerical methods for representation of a generic structure/ material that could contain those features.

R&D Approach

	R&D Activities	Milestones & Targets
1-2 years	• Integrate conventional state-of-the-art image processing techniques into materials science domain • Develop framework for creating supportive datasets	• Ability to compute image-based features automatically on some subset of features • Collect and share datasets • Release public tools
3-5 years	• Use AI to explore feature-space from experimental and computational data, such as micrographic, compositional, crystallographic, atomistic, etc., data	• Identify AI selected features for universal representation • Determine method to combine multiple data sources into a single numerical representation • Collect and share datasets • Release public tools and updates
>5 years	• Develop universal numerical representation for inorganic materials compatible with current state-of-the-art for organic materials.	• Validate the reality of the selected universal features • Publish and promote use of universal numerical representation • Validate that representation can be numerically manipulated and produce meaningful results (i.e., simulate heat treatment, forge, other processing) • Release public tools and updates

	Stakeholders & Potential Roles
Stakeholders	**Industry End Users:** Input on needs (what are end goals?) and what data exists **Industry/Materials/Equipment Suppliers:** Prototyping, validation, and structural expertise **Academia:** Develop software utilizing intelligently-designed algorithms and AI framework **National Labs:** Compile software into robust and generic architecture **Government:** Computational support

Benefits/Impacts

Improved energy efficiency/footprint – Medium
Reduces costs – Medium
Accelerates innovation – High: Streamlines validation of complex computational models of material structures

Enhances industry competitiveness – High: Broadens the applicability of AI to materials; improves accuracy of fit
Faster materials/product development – Medium
Enhances manufacturability– Medium

Figure C-4. Physical Structure Related to Numerical Representations

Appendix B

Appendix B.1: Collaboration, Partnerships, and Education/Training (Data)

Responses from Data Quantity and Quality Group

FOCUS QUESTION 3a: *Considering the R&D needs and pathways identified, what opportunities exist for collaborative efforts with the DOE labs and industry? What types of partnerships are envisioned that would be most successful in reaching targets and goals?*

Table A-3a. Data Quantity and Quality – Collaboration & Partnerships

Collaborative Organizations and Initiatives

- Coordinate this work with existing High Performance Computing for Manufacturing (HPC4Mfg) program and "High Performance Computing for Materials" program (HPC4Materials)
- National Network for Manufacturing Innovation (NNMI) Institutes
- Master lists of universities (contacts) active in AI and their areas of interest; could be gathered by NIST or AMO
- Collaboration of national labs around user facilities (light sources/neutron sources) for key measurements
- Embrace & support a high throughput experiment materials laboratory
- Endorsement and active participation from professional and scientific societies
- "Hackathon" type programs, which don't require self-funded travel money
- Look at Data Carpentry organization

Data Sharing and Accessibility

- Consensus on common "language" for data and databases
- Online/open access knowledge base on material sciences: capture projects, partners, data
- Leverage efforts by DOD, NSF and NIST on FAIR data standards
- AMO should support consortium of national labs with industry advisory board to establish an open data and tools framework
- National labs could mandate users to publish all data and provide a platform for doing that; develop a new policy
- Working groups at conferences to develop data standards (provide travel support); include academics, labs, industry
- Energy-material networks; common data
- Round robin collaboration on materials data; e.g., evaluate variance in experiments

Other Areas/Topics for Collaboration

- Partnerships to create/refine/validate physical models
- Collaboration with each national lab that is active in tackling simulation at a different scales
- Industry & labs to develop high throughput tools; use Cooperative Research and Development Agreement (CRADA)
- Software developers and labs scripts; create a software consortium?
- Lateral instrument networks and knowledge spaces
- Automate materials characterization at beam lines – just send samples
- Involve and include stakeholders in technology development

FOCUS QUESTION 3b: *What education and workforce challenges need to be addressed? What skillsets or disciplines need to be further developed?*

Table A-3b. Data Quantity and Quality – Education & Training

Curriculum Development & Academic Programs

- Cross pollination by introducing materials science domain problems in computer science/statistics education
- Enhance programming/computer science education in undergraduate science programs
- Statistics for materials scientists
- Train undergraduates in scripting for data analysis instead of relying on ready tools
- Leveraging undergraduate involvement in data collection & processing
- Integrated curriculum for physics and data; causation and correlation
- Integration of informatics with experiments and "experimental" design
- Teach code structure best practices universally
- Ordinary differential equations (ODE)/partial differential equations (PDE) ←→ statistics/machine learning
- Start pipeline early; K-12 for tools/approaches to AI for materials
- Incorporate data analytics in graduate education
- Needed skills
    - Data pre-processing
    - Anomaly removal and background removal
    - Statistical analysis
    - Design of experiments

Workforce Development

- Gain basic competency in materials or AI (non-expert)
- Specialized online training programs for seasoned workforce generation
- Support Machine Learning for Materials Research (MLMR) boot camp
- Online courses for materials scientists on machine learning and vice versa – drive convergence
- AI/data science training for materials science engineers
- Train domain scientists in data analysis and machine learning
- Programming for materials scientists; AI/machine learning algorithms for materials scientists
- Multiple skills are needed (material scientists, process engineering, physics, mathematics, and computer skills); more training for specific topics; DOE support for travel to trainings
- "Hacking" skills (see software carpentry)
- Trade school training
- Under-represented populations; lots of talent left on the table

Tools and Other Resources

- User interface design; need an intuitive interface
- Human/computer interaction, e.g., worker "sanity check" of AI decision
- Automate data collection to not impede or exacerbate work flow
- Data stewardship
- AI awareness training; e.g., case studies/success stories in material discovery

Appendix B.2: Collaboration, Partnerships, and Education/Training (Platforms)

Responses from the Platforms and Infrastructure Group

FOCUS QUESTION 3a: *Considering the R&D needs and pathways identified, what opportunities exist for collaborative efforts with the DOE labs and industry? What types of partnerships are envisioned that would be most successful in reaching targets and goals?*

Table B-3a. Platforms & Infrastructure – Collaboration & Partnerships

Collaborative Organizations and Initiatives
• Create an artificial intelligence & materials science center as part of NSF Industry-University Cooperative Research Centers (I/UERC) Program • Seed funds to support AI-materials science collaboration where physical models are relatively well-established. • Establishment of challenge problems to engage community in focus efforts • Materials science & machine learning: 1) challenge problems, 2) summer schools, 3) conference/workshop
Data Formatting and Sharing
• Providing mechanisms and mandates for inter-agency sharing of data and a common data storage/sharing platform. • DOE labs could facilitate/provide infrastructure for curating publications/reports in standard formats • Government can form large teams to solve hard problems; e.g., a natural language approach to data scraping from legacy documents will need a large team – obvious candidate for lab-academia collaboration • Providing community space, intelligence support - to test-run and qualify data format fusion procedures • DOE labs could provide large-scale study of devices from industry. Information on devices provided confidentially to company. Aggregated data provided to all. Example: Batteries – DOE national labs could conduct large scale study data collection over time on batteries – performance, failure. DOE lab studies correlations. • Anonymous data publishing for industry (e.g., data related to defects or failures)
Other Areas/Topics for Collaboration
• Super computers at national labs should be used for expensive high-throughput physics simulations to build statistical surrogates. • Labs, industry, academia create a set of visible, publicized materials discovery success stories that used artificial intelligence. • Sponsor a study into machine learning and materials development; produce report like the Integrated Computational Materials Engineering (ICME) ones with gaps; roadmap best practices

FOCUS QUESTION 3b: *What education and workforce challenges need to be addressed? What skillsets or disciplines need to be further developed?*

Table B-3b. Platforms & Infrastructure – Education & Training

Workforce Development

- Establish & educate a "materials informatics profession," similar to what has emerged in biomedicine.
- Future generation of materials scientists need to be cognizant of AI and ML
- Challenge – Needing PIs/educators with knowledge in both AI and materials science.
- Short boot camp for industrial practitioners
- ML boot camp for late-career materials scientists (they have domain knowledge)
- Champion/bring Data Carpentry[16] to materials; "AI Carpentry" would bring AI skills to materials researchers unfamiliar with AI through boot-camps
- Data science education for materials scientists/engineers is a challenge, as it is done on one's free, non-existent time. Continuous education in advanced statistics, data science, programming languages is now necessary for most materials engineers, but cannot be done at work.
- Provide internship opportunities for students and more established professionals to learn AI methods
- Diversity – Create mechanisms to ensure diversity in materials informatics education & workforce
- Is there a cultural barrier to adoption of AI methods in materials science?
- Industry sharing both successes and failures of artificial intelligence to improve education

Curriculum Development & Academic Programs

- Communications skills to support collaboration across disciplines, especially materials and computer science. Joint AI-materials project-based courses with computer scientists?
- Education on application of AI for materials science & engineering undergraduates
- Develop an AI/ML curriculum for materials scientists, understanding that current faculty are not (yet) educated in this area.
- Resolve balance between specialization vs. generalization. Are materials informatics engineers needed? Or should all materials scientists be trained in informatics?

Tools and Other Resources

- "Matlab" like tool in universities & industry; e.g., Citrine
- Move past "blind" use of AI to understand what AI outputs really mean and their limitations; teach users what is going on "under the hood"
- Creation of accessible repositories connecting datasets and AI algorithms with context metadata; Github; provide richer context for a non-expert in AI with tutorials and helpful explanations
- Create a public machine learning repository for materials research for training/learning purposes.
- Simplify complex data to allow for greater cross discipline communication and collaboration. Interpret→define→access→adopt→fuse!

Education and Training (from Future Capabilities and Targets Session)

- Educate the next generation of engineers and scientists
- Educational infrastructure in machine learning/artificial intelligence for scientists and engineers (currently education in ML/AI is just for computer scientists)

[16] Data Carpentry (www.datacarpentry.org/) is non-profit organization that develops and provides data skills training to researchers. Software Carpentry (https://software-carpentry.org/) is a volunteer non-profit organization dedicated to teaching basic computing skills to researchers.

Table B-3b. Platforms & Infrastructure – Education & Training

• Educate materials engineers how to build statistical surrogates (metamodels) • Easily discoverable tutorials with data sets and examples of analysis.
Education/Human Resources (from Technical and Scientific Challenges Session)
• How are materials scientists educated in artificial intelligence/machine learning? Who teaches it? What does it supplant? What curriculum? • Realize our weakness/strength. Communications between domains, statisticians, and high performance computing. • Need training on effectively communicating materials problems to computer scientists. Materials science tutorials to computer scientists? • Researchers need incentives to standardize their data/metadata
Other (from Technical and Scientific Challenges Session)
• Lack of pre-packaged “challenge problems” to allow computer scientists to push development of approaches without significant collaboration

Appendix B.3: Collaboration, Partnerships, and Education/Training (Applications)

Responses from the Specific Applications Group

FOCUS QUESTION 3a: *Considering the R&D needs and pathways identified, what opportunities exist for collaborative efforts with the DOE labs and industry? What types of partnerships are envisioned that would be most successful in reaching targets and goals?*

Table C-3a. Specific Applications – Collaboration and Partnerships

Collaborative Organizations and Initiatives
• Multiple types of collaborations/partnerships are needed, e.g., labs &industry; tool & software; prototype demonstration; and partnership of smart manufacturing and manufacturing hubs with concerted development • Cross pollination with strong existing AI efforts in DOD, DHS, other agencies, and programs such as High Performance Computing (HPC), Advanced Scientific Computing Research (ASCR), Basic Energy Sciences (BES), and Biological and Environmental Research (BER) • Manufacturing demonstration facility-like structure that enables public-private partnerships • Computing demonstration facility • Reduce disconnect between labs and industry; leverage local lab relationships with industry (vs. program focus) • Create easy-to-implement agreements for industry to work with labs in the application of AI for materials discovery and design; perhaps have CRADAs (this has worked for others) • Expand scope and create R&D testbeds within Clean Energy Smart Manufacturing Innovation Institute (CESMII). • Follow a consortium model of the semiconductor industry (i.e., SEMATECH) • Consortium with industry, federal, and academia to inform and communicate AI for materials • If the DOE “big ideas” crosscuts still exist, this would be an excellent candidate: Office of Energy Efficiency and Renewable Energy – Office of Electricity – Office of Science • Structured AMO-organized task forces for AI applications • An umbrella CRADA or HUB (national laboratory/industry) to address a specific challenge; a co-design approach; perhaps have a one-month CRADA between national labs and industry • Application-specific FOA by DOE • Internships – rotation between universities, labs, and industry: seniors (universities); post-docs; engineers (industry) • University internships with access to lab capabilities and driven by industry goals
Data Sharing and Accessibility
• Consolidate multiple federal agency joint efforts for the collection of data (NIST, DOE, DOD) • Public/private partnerships that create mechanisms to leverage industry data (and create value) • Make data or AI submission mandatory for federally-funded R&D projects • Development of standard/open platform (platform as a service, or PaaS)
Other Areas/Topics for Collaboration
• Need for public/private partnerships to move AI materials forward with goal-oriented research focus (e.g., low hanging fruit, systems of interest) that show effectiveness • Collaboration with national labs and academic institutes on automation, visualization, machine learning, validation

Table C-3a. Specific Applications – Collaboration and Partnerships

- MGI algorithm development and implementation into industrial research
- Multi-disciplinary approach to affect all areas from federal research to prototyping advanced technology
- Software/industry relationship with DOE for availability of tools to start-ups
- DOE-sponsored mentor-protégé programs (AI challenge-specific; e.g., Cyclotron Road)
- Tech transfer collaboration: industry provides data, DOE labs and academia provide models and expertise
- Open challenges for AI-based materials study
- Improve ML/AI opportunity: all-in-one
- DOE labs/academia should run high-risk/reward research for industry
- Roadmaps of industry – specific needs in fundamental science
- Ability to loosen the U.S. manufacturing requirement on national laboratory-generated intellectual property (IP)
- Intelligent methods to navigate IP and minimize barriers; there is a fear of losing IP
- In terms of IP, think about what DOE wants and see what industry is willing to do; try to be intentional about consequences
- Webinars (Facebook, LinkedIn) to engage students and early researchers
- Use a crowd funding model to solve problems; don't wait for the government to solve problems

Education and Training (from Future Capabilities and Targets Session)

- Education/training in ML that is advanced in materials science
- Basic tools for AI training
- AI trained and used directly by materials scientists rather than the AI programmer
- Need software engineering with materials research
- Applied mathematical and numerical methods skills for materials research
- Good case studies (perhaps in book form) for materials and manufacturing

FOCUS QUESTION 3b: *What education and workforce challenges need to be addressed? What skillsets or disciplines need to be further developed?*

Table C-3b. Specific Applications – Education and Training

Curriculum Development & Academic Programs

- Encouraging universities to educate more disciplines in AI techniques; currently AI is only on computer science side
- Mandatory AI coursework for engineers
- Lack of inter-disciplines in materials and computer science is a challenge
- Develop new AI curricula applied to materials science
- Better understand cultural barriers between materials science and computer science
- Graduate student exchanges between groups
- Doctoral programs integrating interdisciplinary components of AI and work experience
- Summer schools/coding schools for undergraduate and graduate students
- Define data and model problems as lab projects in online education from kindergarten through 12th grade (K-12) and at universities with prizes
- To attract young talent, make it look “cool”

Workforce Development

- Bootcamps, workshops, etc., to teach AI to materials scientists
- Train scientists at labs in AI to be deployed to industry to learn about problems; boot camps and workshops
- Use short courses and training workshops to improve skillsets of material scientists
- Short courses in AI for technicians on the shop floor
- Advanced AI course for existing materials engineers and scientists
- Training grants for experimentalists, programmers, synthetic chemists, physicists, biologists
- Piggyback off the strong workforce development programs that the advanced manufacturing institutes are already doing; e.g., existing science, technology, engineering, and mathematics (STEM) programs
- Internships and rotations between universities, industry, and labs
- Technology transfer strategy to transition what is developed in the lab to the workforce

Other

- Sharing of datasets and training to use models; this helps to understand limitations
- Good case studies to showcase how models can work in AI sciences
- Government needs to learn and communicate state-of-the-art and industry standards to academics, material labs, and startups

Education/Human Resources (from Technical and Scientific Challenges Session)

- Insufficient training of manufacturing engineers in the benefits and capabilities of AI and ML
- Better knowledge sharing between materials scientists and computer scientists
- AI-guided data generation to fill in sparseness to enhance prediction
- Lack of training programs to develop technical “pentathletes”
- Lack of computer infrastructure and AI-trained researchers
- Lack of formal education in AI for students and researchers in materials science
- Education of how to use AI: required skills and science and knowing limitations

Table C-3b. Specific Applications – Education and Training

• Development of curriculum/short course for existing materials scientists/engineers; "How do I get existing workforce knowledgeable on AI?" • Lack of coordination between academic disciplines • Many material scientists are traditionalists reluctant to take on something new. Those trained in AI/ML are creating converts

Appendix C. Agenda

Workshop on Artificial Intelligence Applied to Materials Discovery and Design

August 9 – 10, 2017, Pittsburgh Marriott City Center, Pittsburgh, PA

<table>
<tr><th colspan="4">Day 1 (Wednesday, August 9)</th></tr>
<tr><td>7:00 – 8:15 am</td><td colspan="3">Registration and Continental Breakfast</td></tr>
<tr><td>8:15 – 8:30 am</td><td colspan="3">Opening Remarks and Workshop Objectives
Brian Valentine, AMO Technology Manager</td></tr>
<tr><td>8:30 – 9:10 am</td><td colspan="3">Accelerated Search for Materials via Adaptive Learning
Turab Lookman, Physics of Condensed Matter and Complex Systems Group, Los Alamos National Laboratory</td></tr>
<tr><td>9:10 – 9:50 am</td><td colspan="3">Analyzing Large-Scale Data to Solve Applied Problems in R&D
Bryce Meredig, Co-founder and Chief Scientist, Citrine Informatics</td></tr>
<tr><td>9:50 – 10:30 am</td><td colspan="3">The Materials Genome Initiative and Artificial Intelligence
A. Gilad Kusne, Materials Measurement Science Division, National Institute of Standards & Technology</td></tr>
<tr><td>10:30 – 10:45 am</td><td colspan="3">Break</td></tr>
<tr><td>10:45 – 11:55 am</td><td colspan="3">Panel on Challenges Facing AI in Applied Materials Design
William Peter (Moderator), Director, Manufacturing Demonstration Facility, Oak Ridge National Laboratory
Amra Peles, Project Lead, Pratt & Whitney
Kishore Reddy, Research Engineer, United Technologies Research Center
Adama Tandia, Research Associate, Corning Incorporated
Joseph Vinciquerra, Technology Platform Leader, Additive Materials, GE Global Research</td></tr>
<tr><td>11:55 am – 12:00 pm</td><td colspan="3">Breakout Session Groups and Instructions</td></tr>
<tr><td></td><td>Data Quantity and Quality
Salons 2-3
Collection, storage, sharing and analysis of large amounts of clean, well-organized, quality data. Includes instrumentation needed for collection with the required accuracy and precision.</td><td>Platforms and Infrastructure
Salon 4
Challenges in developing and integrating computational and design tools working across materials classes and sectors; searchable data infrastructure, ML models, algorithms, etc.</td><td>AI for Materials Design in Specific Applications
Salon 5
Specific challenges with applying AI to materials for important industrial applications, i.e., battery materials, materials for extreme conditions, catalytic materials, etc.</td></tr>
</table>

12:00 – 1:00 pm	Lunch
1:00 – 2:30 pm	**Breakout Session 1: Future Desired Capabilities/Targets for AI in Materials Design**
2:30 – 2:45 pm	Break
2:45 – 4:15 pm	**Breakout Session 2: Technical and Scientific Challenges**
4:20 – 4:40 pm	**Day 1 Report Outs**
4:40 – 5:00 pm	**Department of Energy Perspectives on Artificial Intelligence in Materials** *Dr. Mark Johnson*, Director, DOE Advanced Manufacturing Office
Day 2 (Thursday, August 10)	
7:30 – 8:30 am	Registration and Continental Breakfast
8:30 – 8:45 am	**Day 1 Recap** *Brian Valentine*, AMO Technology Manager
8:45 – 9:15 am	**Keynote Presentation: Accelerated Materials Design and Discovery: An Industry-University Collaboration** *Brian Storey*, Program Officer, Accelerated Materials Design and Discovery Program, Toyota Research Institute
9:15 – 11:00 am	**Breakout Session 3: R&D Pathways and Technical Approaches**
11:00 – 11:15 am	Break
11:15 am – 12:15 pm	**Breakout Session 4: Collaboration, Partnerships, and Education/Training**
12:15 – 1:15 pm	Lunch
1:15 – 1:45 pm	**Day 2 Report Out**
1:45 – 2:00 pm	**Closing Session**
2:00 pm	**Adjourn**

Appendix D. Workshop Participants

Name	Organization
Gibson Asuquo	DOE Advanced Manufacturing Office
Adrian Barbu	Florida State University
Mostafa Bedewy	University of Pittsburgh
Jack Beuth	Carnegie Mellon University
Debangsu Bhattacharyya	West Virginia University
Tyler Borchers	Arconic Technology Center
Isaac Chan	DOE Advanced Manufacturing Office
Yu-Chin Chan	Northwestern University
Cristian Ciobanu	Colorado School of Mines
Fred Crowson	Energetics Incorporated
Jun Cui	Ames Laboratory, Iowa State University
Ismaila Dabo	Penn State University
Brian DeCost	Carnegie Mellon University
Stephan V. Dekwiatkowski	Fine Optics Inc.
Sanket Deshmukh	Virginia Tech
Ram Devanathan	Pacific Northwest National Laboratory
Alden Dima	National Institute of Standards and Technology
Wayne Dudding	DOE/National Energy Technology Laboratory
Christopher Evans	Savvior
Zak Fang	University of Utah
Chris Fish	McAllister & Quinn
Thomas Frewen	United Technologies Research Center
Kevin Gallagher	PPG Industries, Inc.
Jacquelynn Garofano	United Technologies Research Center
Michael Garris	National Institute of Standards and Technology
Robert Gemmer	DOE Advanced Manufacturing Office
Garrett Goh	Pacific Northwest National Laboratory
Tara Gonzalez	DOE Advanced Manufacturing Office

Name	Organization
Alison Gotkin	United Technologies Research Center
Martin Green	National Institute of Standards and Technology
T. Yong Han	Lawrence Livermore National Laboratory
Mark Hartney	SLAC National Accelerator Laboratory
Kai Ming Ho	Iowa State University
Elizabeth Holm	Carnegie Mellon University
Huilong Hou	University of Maryland, College Park
Diana Hun	Oak Ridge National Laboratory
Alojz Kajinic	Arconic Technology Center
Sergei Kalinin	Oak Ridge National Laboratory
Matthew Krug	Arconic Technology Center
A. Gilad Kusne	National Institute of Standards and Technology
Paul Leu	University of Pittsburgh
Turab Lookman	Los Alamos National Laboratory
Steven Lopez	Harvard University
Tommi Makila	Energetics Incorporated
Sudipto Mandal	Carnegie Mellon University
Paul Mason	Thermo-Calc Software Inc
Bryce Meredig	Citrine Informatics
Lincoln Miara	Samsung Electronics
Theresa Miller	Energetics Incorporated
George Muntean	Pacific Northwest National Laboratory
Manh Cuong Nguyen	Ames Laboratory, U.S. DOE
Rahmi Ozisik	Rensselaer Polytechnic Institute
Amra Peles	Pratt & Whitney
William Peter	Oak Ridge National Laboratory
Laura Pullum	Oak Ridge National Laboratory
Liang Qi	University of Michigan
Anthony Rollett	Carnegie Mellon University

Name	Organization
Benjamin Sanchez	Harvard University
Karthikeyan Saravanan	University of Pittsburgh
Christian Schafmeister	Temple University
Dennis Sheberla	Harvard University
Yunfeng Shi	Rensselaer Polytechnic Institute
Stephen Sikirica	DOE Advanced Manufacturing Office
Aarti Singh	Carnegie Mellon University
Jan Steckel	DOE/National Energy Technology Laboratory
Brian Storey	Toyota Research Institute
Ramesh Subramanian	Siemens Energy
Bobby Sumpter	Center for Nanophase Materials Sciences
Adama Tandia	Corning Research & Development Corporation
Emmanuel Taylor	Energetics Incorporated
Michael Taylor	University of Pittsburgh
Laura Thornton	Edison Welding Institute
Albert To	University of Pittsburgh
Richard Todaro	Allegheny Science & Technology
Brian Valentine	DOE Advanced Manufacturing Office
Vinod Veedu	Oceanit
Joseph Vinciquerra	GE Global Research
Hongliang Xin	Virginia Tech
Wei Xiong	University of Pittsburgh
David Yaron	Carnegie Mellon University
Phillip Yu	PPG Industries, Inc.

U.S. DEPARTMENT OF ENERGY

Office of ENERGY EFFICIENCY & RENEWABLE ENERGY

For more information, visit:
energy.gov/eere/amo

DOE/EE-1757 • March 2018