NISTIR 7898

Building the Materials Innovation Infrastructure: Data and Standards

James A. Warren
Ronald F. Boisvert

http://dx.doi.org/10.6028/NIST.IR.7898

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 7898

Building the Materials Innovation Infrastructure: Data and Standards

James A. Warren
Materials Science and Engineering Division
Material Measurement Laboratory

Ronald F. Boisvert
Applied and Computational Mathematics Division
Information Technology Laboratory

http://dx.doi.org/10.6028/NIST.IR.7898

November 2012

U.S. Department of Commerce
Rebecca Blank, Acting Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director

Building the Materials Innovation Infrastructure: Data and Standards

A Materials Genome Initiative Workshop

Summary Report

May 2012 Workshop

NIST

National Institute of Standards and Technology

Workshop Report:
Building the Materials Innovation Infrastructure: Data and Standards

A Materials Genome Initiative Workshop

Summary Report for Workshop held

May 14-15, 2012

Washington, DC

Prepared for the National Institute of Standards and Technology

November 2012

Acknowledgments

This report summarizes the results of the Workshop "*A Materials Genome Initiative Workshop – Building the Materials Innovation Infrastructure: Data and Standards*" held May14-15, 2012, in Washington, DC. The workshop was a collaborative effort of the National Institute of Standards and Technology (NIST) and a representative Steering Committee (members shown below). Thanks are extended to the Steering Committee co-chairs and members for their extensive efforts in guiding this workshop and successfully reaching the outcomes. Thanks also go to Energetics Incorporated for their assistance in facilitating the workshop and preparing this report.

Steering Committee Co-chairs
James A. Warren, NIST
Ronald Boisvert, NIST (co-chair)

Steering Committee Members
John Allison, University of Michigan
Cynthia Friend, Stanford Linear Accelerator Laboratory
David Furrer, Pratt & Whitney
Kristin Persson, Lawrence Berkeley National Laboratory

Cover photo credits (from left to right)

NIST – L. Bendersky, 2011.
Microsoft Online – MH 900178448, 2012.
NIST – B. Young, 2004.
Microsoft Online – MH 900402686, 2012.
National Renewable Energy Laboratory – D. Schroeder 2012, Research Support Facility.

(this page intentionally left blank)

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

1.1 THE MATERIALS GENOME INITIATIVE

The Materials Genome Initiative (MGI) is a multi-agency multi-stakeholder effort to develop the infrastructure needed to enable the materials science community to design, develop, manufacture, and deploy advanced materials at least twice as fast as possible today, at a fraction of the cost. This new Materials Innovation Infrastructure (MII) will leverage advances in materials modeling, computing, and communications to accelerate advanced materials design and deployment in the United States across many fields. More widespread implementation of advanced materials will contribute to new products with enhanced functionality, and potentially enhance U.S. global competitiveness.

Major advances in theory and modeling have led to a remarkable opportunity for the use of computational simulation in predicting the behavior of material systems. However, such computational tools are not in widespread use today due to limitations in their capabilities, a lack of expertise needed to employ them, and a general lack of confidence in accepting conclusions that are not empirically based. Similarly, advances in networked communications have led to remarkable opportunities for the sharing of technical information, such as materials property data. This too has had limited use due to the lack of suitable data repositories, standards, and incentives for sharing.

A recent report emphasizes the growing importance of manipulating, mining, managing, analyzing, and sharing scientific data.[1] The issues outlined in this report are highly relevant to the MII, which will coordinate large amounts of diverse scientific information related to materials.

The MGI is addressing some of these issues by developing a MII that includes: (1) accurate models of materials performance validated using experimental data, (2) open-platform frameworks to ease the development and interoperation of simulation codes, (3) software that is modular and user-friendly with applicability to broad user communities, and (4) data repositories built on community standards and outfitted with modern search, retrieval, and analysis tools. Achieving these objectives will help leverage existing Federal investments in computational capabilities and data management, and provide an integrated approach to materials science and engineering.

[1] Atkins, D.; Baker, S.; Dietterich, T.; Feldman, S.; Hey, T.; Lyon, L. *National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization Final Report*; National Science Foundation: March, 2011, http://www.nsf.gov/od/oci/taskforces/.

1.2 WORKSHOP OVERVIEW

The *Materials Genome Initiative Workshop – Building the Materials Innovation Infrastructure: Data and Standards* was held on May 14-15, 2012 in Washington, DC to help define the cross-cutting and domain-specific data challenges facing the creation of the MII. The workshop was hosted by the National Institute of Standards and Technology (NIST) as part of its broad-based efforts to develop new integrated computational, experimental, and data informatics tools. Attendees at the workshop included over125 experts from industry, national laboratories, government agencies, and academia with interdisciplinary interests ranging from the physical and material sciences and engineering to mathematicians and computer scientists.

1.2.1 Workshop Process

Participants were asked to identify challenges, gaps, and opportunities for the development of the MII in the following areas:

- *Data Representation and Interoperability*. This includes an assessment of the types of data that should be collected and the most efficient digital representation. Also of interest is the identification of needed metadata (i.e., additional information that describes properties of the data itself and is necessary to make the data usable). Standards may play a particularly useful role here. If the community can agree on specific data formats and communication protocols, then the retrieval, exchange, and reuse of digital data is greatly enhanced.
- *Data Management*. This includes techniques and tools needed for the organization and maintenance of digital data repositories. Some of the main issues include incentives for making contributions to repositories, methods for access and use that respect intellectual property rights, and long-term sustainability of data archives.
- *Data Quality*. In general, data is of high quality if its origins are known, the methods by which it was obtained are fully documented, and its uncertainty has been carefully quantified. Data quality can be facilitated if uniform formats for documenting data provenance are available, and if reliable techniques and tools exist for uncertainty quantification of results produced from both experiment and simulation.
- *Data Usability*. Data is usable if users can find what they need, can easily extract it from available repositories, and have access to techniques and tools to perform the necessary follow-up analyses.

The workshop generated input based on two separate breakout topics: 1) length scale of the phenomena under study, and 2) a set of representative technical application areas (TAAs).

1.2.1.1 Data Challenges at Length Scales

The design and assessment of materials requires modeling and experimentation at all length and time scales. Unfortunately, no single model can feasibly include all such details and their effects. As a result, computational models typically focus on a particular length scale, i.e., atomic, nano and molecular, micro, and macro. The modeling techniques and data needs for each of these scales are potentially quite different. Additional data challenges are created when there is a need to use separate models operating on different length scales to obtain a combined result. The breakout sessions in this area focused on the data and data infrastructure issues relevant to different length scales, considering data derived from both computational and experimental techniques. A key metric for all length scale discussions was how much the time required for materials design could be reduced if challenges were addressed.

1.2.1.2 Data Challenges for Technical Application Areas

New materials are typically developed with a particular application in mind. The ease of development depends on whether the particular data and tools needed for that application are available and sufficiently capable. Thus it is critical to consider data issues in the context of specific technical application areas (TAAs). TAA breakouts included in the workshop are shown in Figure 1.1; these also provided additional context for linking the issues identified for length scales and associated time scales.

Figure 1.1. Workshop Technical Application Areas

- Electrochemical Storage
- High Temperature Alloys
- Catalysis
- Lightweight Structural Materials

The rationale for selecting these TAA included: (1) the existence of an established community; (2) coverage of a broad range of materials systems and problems; and (3) representation of important topics aligned with goals of the participating federal agencies. The TAAs selected are only a few of those to be considered within the MGI; they represent a starting point for how to integrate application-oriented viewpoints into the materials science and engineering data challenges. A key metric for all TAA discussions was how much the time required for materials design (specific to the TAA) could be reduced if data challenges were addressed.

1.3 ORGANIZATION OF THE REPORT

This report summarizes the concepts generated at the workshop and is organized around the length scale and technical application area topics described previously. Each length scale chapter includes an overview of the topic and a discussion of the major short-term and long-term challenges. For each TAA, an overview of the topic and data challenges is provided, as well as more detailed descriptions of the higher priority challenges (i.e., related research and development (R&D) activities, milestones, and potential benefits).

It should be noted that the ideas presented here are a reflection of the workshop attendees and not necessarily the entire community of interest. However, considerable effort was made to ensure that participants represented all aspects of the field. Since results for each topic area were generated by independent focus groups, the opinions presented for each may differ.

2 Length Scale Challenges

The challenges identified for the various length scales considered are described in the following sections. For all length scales a voting process was employed to enable categorization of the challenges into high, medium, and low priority, and into the timeframes corresponding to the greatest potential impact. Short term impact was defined as less than five years; long term impact as five or more years. Some challenges were noted to have both short- and long-term impacts; these are discussed separately in each section. It is acknowledged that all of the challenges identified are impediments to the material innovation and design cycle. The categorization scheme was employed to provide added perspective on the relative importance and timing for impacts if addressed.

2.1 Macro Length Scales

The macro length scale is the vantage point where objects, actions, and time can be clearly measured and observed by the human eye without significant additional equipment. At this scale, the properties and behavior of various material systems are, in a sense, a global average of the properties at individual points. Significant amounts of data can be captured for analysis and prediction of materials performance and ultimately tied to service life.

2.1.1 Overview of Macro Length Data Challenges

The macro length scale data challenges that can potentially slow the materials design cycle are listed in Table 2.1 and further described in the following sections.

2.1.2 Top Short-Term Macro Length Challenges

Unknown metadata requirements: While it is generally agreed that the standard inclusion of metadata with primary data would greatly aid in the usability, representation, and interoperability of data, the exact metadata required to realize these benefits are not known. Additional studies are needed to create a list(s) of necessary metadata fields that would benefit all collected data, both experimental and computational.

Data transfer through the supply chain: In addition to information sharing protocols to protect intellectual property, steps need to be taken to ensure that data moving from fundamental science to product development be properly handled to avoid real or perceived intellectual property constraints.

Translation to different formats: In data collection systems, on-going configuration changes, such as different hardware and/or operating systems, will likely necessitate translation of information into a format that can be used by receiving systems – either now or in the future. Furthermore, basic translation methods between the identified formats would facilitate data interoperability and complement any implemented data standardization. Fundamental requirements for software packages would also facilitate the accessibility and interoperability of new and legacy computing architectures.

Table 2.1 Macro Length Scale Challenges
(♦ = one vote for potential short-term impact/<5 years)
(● = one vote for potential long-term impact/5 years and beyond)

Data Representation and Interoperability	
High Priority	• Inadequate understanding of metadata required for primary data ♦♦♦♦♦♦♦♦♦♦ (10) • Ensuring data life (maintaining and archiving) ●●●●●●●●●(9) • Identifying/determining data requirements ♦♦♦♦●●●●(4,4) • Definition of data or metadata for particular applications with standards for software to facilitate linkage ♦♦♦♦♦♦●(6,1)
Medium Priority	• Cost of maintaining database updates ●●●●●(5) • Extending macro models to capture new phenomenon where parameters do not reflect the physics (e.g., deformation resistance for thinning in crystal plasticity) ♦♦●●(2,2) • Knowing which entities are working on relevant topics (e.g., Alloy Phase Diagram International Commission) ♦●●(1,2)
Lower Priority	• Lack of common ontologies or vocabularies (and software systems that can use them) to describe concepts, properties, objects, and relations in the material science domain ♦♦ (2) • Generation of grade-specific rather than just generic data (e.g., for plastics) • Establishing model material systems which represent classes of mechanisms and features
Data Management	
High Priority	• Protecting intellectual property of data and R&D results ♦♦♦♦♦♦♦♦●●●●●●●(8,7) • Lack of a national database for materials data; open source and accessible for use by various eco-systems ♦♦♦♦♦♦●●●●●●●●●●●● (6,12) • Transferring data across supply chains (i.e., intellectual property issues) ♦♦♦♦♦♦●●●(6,3)
Data Quality	
High Priority	• Lack of data quality index that includes the quality of specific properties such as modulus, low-cycle fatigue life, and best method specimen finish ●●●●●●●(7)
Medium Priority	• Performing validation studies on commercial materials ♦♦●●●(2,3) • Generation of data with perspective on future uses ♦♦♦●(3,1) • Capturing full multi-axial response of material behavior with all associated conditions (e.g., micro structure, boundary conditions, and processing history, finish) ♦♦●●(2,2) • Validating modeling and experimental data and creating protocols for data collection and interpretation ●●●(3) • Detailing computational or experimental design with calibration against an accepted standard to assure data quality ♦♦♦(3) • Useful standards for experimental data (e.g., collection methods, interpretation) ♦♦●(2,1)
Lower Priority	• Identifying the key parameters that link one model type (e.g., thermo) to another (e.g., mechanical and strain) ♦♦ (2) • Obtaining path dependent information that impacts structure (e.g., chemistry and cooling rate)♦●(1,1) • Gaining broad buy-in to minimum or sufficient provenance so that data is acceptable (i.e., a standard that "proves" data quality) ♦ (1)

Table 2.1 Macro Length Scale Challenges (♦ = one vote for potential short-term impact/<5 years) (● = one vote for potential long-term impact/5 years and beyond)	
	• Determining the role of surface characterization and bulk measurements (e.g., deterministic vs. probabilistic tools) ●(1) • Generating experimental data (cross method correlation, repeatability)
Data Usability	
High Priority	• Standardizing interpretation of export control requirements for materials data and models ♦♦♦♦♦♦●●●●●●●(6,7) • Standardizing formats and metadata requirements for reporting and databasing test data ♦●●●●●●●●●●(1,10) • Creating a centralized repository with a standard that researchers can apply ♦●●●●●●●●●(1,9) • Translation of data to different formats to interface with different software applications ♦♦♦♦♦♦(6)
Medium Priority	• Creating tools for merging and analyzing multimodal, 3-D data (e.g., multiple types of information, chemical properties, and structures) per pixel ♦●●●●(1,4) • Building automated upload and transfer tools for standardized data exchange formats (e.g., Free Data Capture) ♦♦♦●(3,1)
Lower Priority	• Lack of clear intellectual property definitions and definition determination ♦(1) • Lack of data tools to support fusion of different types of data, validation of data accuracy/consistency, and combining of multi-scale data types • Attaching metadata to data during collection from the instrument

2.1.3 Top Long-Term Macro Length Challenges

National database: One or more central repositories are needed to warehouse data from all entities that generate materials data. These central repositories would be openly accessible and contain standardized information uploaded by researchers and others in the field. The standard formats and metadata requirements would depend on the sectors served by and contributing to the database. These standards would allow any interested party to act as a repository, enabling a federation of the database structure. The GenBank at National Institutes of Health is one example of this data warehouse concept.[2] The Materials Project, a collaborative effort between the Massachusetts Institute of Technology (MIT) and Lawrence Berkeley National Laboratory (LBNL), is a data warehouse (albeit at the atomic scale) focused on materials relevant to the Materials Genome.[3]

Ensuring data life: Standards for maintaining and archiving collected data needs to be established in conjunction with a national data warehouse. Protocols are needed to ensure that data can be retrieved on-demand with high fidelity and immediate utility with all necessary translation or conversion methods already in place.

---

[2] Ostell, J. M., Integrated Access to Heterogeneous Data from NCBI. *IEEE Engineering in Medicine and Biology* **1995,** *14* (6), 730-736, http://dx.doi.org/10.1109/51.473267.
[3] https://www.materialsproject.org

Data quality index: Many issues can affect data quality, including data consolidation from various systems, external data that is not easily integrated, duplicate data, and various measurement methods, processes, and equipment. Data inaccuracies can lead to large productivity losses and disappointing project returns. With the large increase in data generation, data quality and early detection of unreliable data is extremely valuable.

2.1.4 Macro Length Challenges with both Short- and Long-Term Impacts

Intellectual property of data and R&D results: Intellectual property protection ranks highly in both short- and long-term impacts due in part to the large financial implications for product and model development. Intellectual property generated by an organization is a collection of unique items that provide knowledge and economic benefit to the organization, and is a source of competitive advantage of differentiation. When creating the MII, the intellectual property of data must be protected not only to safeguard the interests of individual organizations but to promote trust among the users.

Short-term goals could include identifying intellectual property concerns of the community and developing protocols for information sharing. Once a framework is in place to address intellectual property concerns, maintenance and monitoring of the framework would be necessary over the long term to balance pre-competitive data sharing with proprietary or competitive information.

Standardized export control requirements: Export control regulations are federal regulations designed to limit the exchange of selected commodities, services, software, technical data, and other information to foreign countries or persons, either in the U.S. or abroad. Items of concern involve dual use technologies (i.e., technology which can be used for both peaceful and military aims). An example is rocket technology which was developed to carry humans into space but also provided the knowledge for development of intercontinental ballistic missiles.

Most industrial countries have export controls on designated dual-use technologies without the permission of the government, and penalties for non-compliance can be substantial. The application of these regulations when exchanging material data and material models is complicated. Transfer restrictions depend on the technology specifics and the destination country. Standardized interpretation of these regulations could facilitate long-term data-sharing and collaboration related to materials data and models. As regulations change over time it will also be important to incorporate updates to the standard interpretation, disseminate periodic information to stakeholders, and determine ownership of the standard interpretation over the long-term.

2.2 MICRO LENGTH SCALE

The micro length scale focuses on the microstructural features that are relevant to many of the manufactured specialty materials. Some of the phenomena encompassed at the micro length scale include diffusion, phase changes and boundaries (e.g., phase diagrams), and phase field modeling. Materials microstructure data can be generated from experimental observations or computer simulations.

2.2.1 Overview of Micro Length Data Challenges

The micro length scale data challenges that can potentially slow down the materials design cycle are listed in Table 2.2 and further described in the following sections.

Table 2.2 Micro Length Scale Challenges
(♦ = one vote for potential short-term impact/< 5 years)
(● = one vote for potential long-term impact/5 years and beyond)

Data Representation and Interoperability	
High Priority	• Generation of data compatible with commercial codes and usable by non-material scientists ♦♦●●●●●●●●●●●●● (1,13) • Choosing an adequate schema (e.g., organization and fields) that is appropriate for precompetitive research, evolvable, and flexible enough to handle future modifications ♦♦♦♦♦♦♦♦♦♦♦♦♦♦ (14) • Lack of key information: full pedigree, test and volumetric data for chemistry, phase, orientation, defects at appropriate scale to represent property changes, universal identifiers or identification of materials (e.g., important microstructural characteristics and defects) ♦♦♦●●●●●● (3,6) • Developing intuitive application programming interfaces (APIs) that allow users to scale, link-up, and integrate different types of data to and from models ♦●●●●●●● (1,7)
Medium Priority	• Standardization of data formats (at least within subcategories) to provide consistency for research done in relative isolation ♦●●●●●(1,5) • Replacing poorly-defined data with definitions for curation and enabling use of the data in some form (e.g., interaction of tools) ♦♦♦●●(3,2) • Lack of methods for determining similarities of microstructures and standard validation and verification tests ♦●●● (1,3)
Lower Priority	• Handling sensitive property data (e.g., regulatory) ♦♦♦ (3) • Developing robust and self-consistent phenomenological theory ♦●● (1,2) • Lack of an anonymous, proprietary data match-making tool ♦♦ (2) • Establishing requirements for code that can analyze data generated by unrelated instruments or simulations ♦ (1)
Data Management	
High Priority	• High cost and time required to archive and manage data (incentives are lacking) ♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦●● (15,2) • Lack of data scientists ♦♦♦♦♦♦♦♦♦●●●●● (9,5) • Lack of means to reference or cite a dataset to maintain "ownership" during data sharing ♦♦♦♦♦♦♦♦♦♦●●●● (10,4) • Lack of sufficient data storage capabilities: ●●●●●●●●●●●● (12) – Automated data storage protocols that add metadata to facilitate both machine and human-centric search and find strategies

Table 2.2 Micro Length Scale Challenges (♦ = one vote for potential short-term impact/< 5 years) (● = one vote for potential long-term impact/5 years and beyond)	
	– Federated data storage networks[4] that enable data access – Data storage retrieval and transformation that is compliant with International Traffic in Arms Regulations • Lack of standards, centralized storage and maintenance of experimental and computed data, including data storage solutions available to the public ♦♦♦♦♦♦♦♦♦♦♦●●●●●●● (11,7)
Medium Priority	• Unknown formats and quality of structured and unstructured data ♦♦♦● (3,1)
Lower Priority	• Decreasing the cost of computing ♦♦♦ (3) • Comparing data management based on economic security, national security and defense, and common global needs ♦● (1,1) • Rising cost of unused data; determining when to delete data
Data Quality	
High Priority	• Developing uncertainty quantification tools for modeling and simulation to automatically calculate prediction uncertainty, taking into account the origin and accuracy of input data and computational methods ●●●●●●●●●●●●●●●●●●● (19) • Obtaining key information to support quality, including pedigree of data, amount and quantity, verification and replication ♦♦♦●●●●●● (3,6) • Standardizing work-flows to enhance collaboration opportunities ♦♦●●●●● (2,5)
Medium Priority	• Gaining greater involvement from instrument manufacturers; most original data has relevant metadata ♦♦♦♦ (4)
Lower Priority	• Establishing standards for accessing and improving data quality and preserving legacy data ●● (2)
Data Usability	
High Priority	• Lack of tools to easily find existing data; text mining issues arise with legacy data ♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦♦●●●●● (25,5) • Lack of tools for retrieving data (similarity metrics), evaluating information, or assessing knowledge gain associated with datasets, and lack of standards for integrating these methods ♦♦♦♦♦♦●●●●●●●●● (6,9) • Scaling tools to terabyte data sets ♦●●●●●●●● (1,8)

[4] Federated storage refers to storage resources that are governed by a common management system with rules for data storage, management, and migration throughout the storage network. Storage capacity would be managed by controllers or appliances controlling multiple arrays.

2.2.2 Top Short-Term Micro Length Challenges

Common data schema: A number of challenges with short term impact relate to choosing and developing a common schema (e.g., organization and fields). The schema should be evolvable and flexible so it can handle future adaptations. A general schema appropriate for use at the precompetitive research level could help promote data sharing during the early stages of materials design and development. A significant challenge for data usability is the lack of tools to easily locate existing data, especially legacy data where text mining issues often arise. An example of useful common schema is the Chemical Markup Language (CML) Schema (http://www.xml-cml.org/schema/). CML and the Polymer Markup Language extensions for material information have been described in a recent article.[5] Other relevant markup languages include MatML (http://www.matml.org/), ThermoML (http://trc.nist.gov/ThermoML.html), and UnitsML (http://unitsml.nist.gov/).

Data archiving: Archiving data poses a challenge for data management because it requires both time and financial investments. Companies and organizations with limited funds may not perceive archiving as a priority. Research entities in particular may not be able to archive data due to the time and resources required. Incentives may be needed to help promote and enable more data archiving. An example of this in the public sector is the National Science Foundation's Extreme Science and Engineering Discovery Environment.

Data ownership: Many organizations maintain ownership of data to retain competitive or other advantages. While this is a reasonable organizational priority, it can significantly restrict the level and types of open sharing that will be needed to create the MII. Before sharing can be more widely enabled, mechanisms are needed to protect company or organizational data when necessary while minimizing restrictive ownership protocols.

2.2.3 Top Long-Term Micro Length Challenges

Data compatibility: Ensuring that data is both compatible with commercial software and usable by non-materials scientists is a primary challenge for enabling widespread use of datasets. With the creation of a standard set of basic codes, dataset could be made broadly compatible, improving accessibility to virtually any individual or organization.

Handling of massive data sets: Datasets are rapidly growing in complexity and scale. The future challenge will be to ensure that information access and usability tools can handle much larger datasets, including those up to terabyte levels.

---

[5] Adams, N.; Winter, J.; Murray-Rust, P.; Rzepa, H. S., Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language. *Journal of Chemical Information and Modeling* **2008,** *48* (11), 2118-2128, http://dx.doi.org/10.1021/ci8002123.

Error and uncertainty quantification: The quantification of error and uncertainty is a significant challenge with all data sets. Error quantification is tied to data analysis, particularly at similar time, length, and dimensional material complexity scales.

Data storage: A major challenge for data management over the longer term involves the development of automated data storage protocols and federated data storage networks. It would be beneficial to develop standardized work flow processes as well as data storage protocols that automatically add metadata to facilitate both machine and human centric search and find strategies.

2.2.4 Micro Length Challenges with both Short- and Long-Term Impacts

Data scientists: Data scientists employ techniques and theories from many fields, including math, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products. The lack of access to skilled data scientists is a contributing factor in many of today's shortcomings in data management and exploitation. The shortfall is expected to increase as the amount and complexity of data generated continues to grow at an exponential pace.

Data retrieval and integration tools: Data retrieval and mining methods and associated analyses and algorithms (e.g., similarity metrics) are lacking today and are needed to improve data usability now and in the future. Tools to evaluate the information or knowledge gain associated within datasets (i.e., understanding and measure of data utility) are also limited. Standards are lacking for integrating or linking disparate retrieval methods, which will impede the linking of datasets from diverse sources. All of these issues will become more critical as the amount of data collected increases.

Central data storage: A publically availability centralized storage and maintenance system for both experimental and computational data is a priority challenge for both the short and long term. Data standards that allow for consistent management of day are a key component. However, the cost and time to develop these systems will be significant and as a result they will be difficult to develop and maintain without national or international participation.

2.3 Nano and Molecular Length Scale

The nano and molecular length scale acts as a bridge between the micro and atomic length scales and is increasingly important to materials due to the current push towards nanotechnology and nanoscience. The molecular level was included in this length scale in order to encompass phenomena largely at the atomic scale (with even more complexity) and growing interest in new properties that emerge at the nanoscale.

2.3.1 Overview of Nano and Molecular Length Data Challenges

The nano and molecular length scale data challenges that can potentially slow the materials design cycle are listed in Table 2.3 and described below.

Table 2.3 Nano and Molecular Length Scale Challenges
(♦ = one vote for potential short-term impact/<5 years)
(● = one vote for potential long-term impact/5 years and beyond)

Data Representation and Interoperability	
High Priority	• Creating a taxonomy to systematically map out materials spaces down to the nanoscale ♦●●●●●●●●●●●●●●● (1,15) • Developing stringent standards for what is considered "data" ♦♦♦♦●●●●●●●●(4,8) • Limited ability to move from data to credible information and decision making ♦♦♦●●●● (3,4) • Informing the community about web standards for data representation ♦♦♦♦♦♦ (6)
Medium Priority	• Creating representations of particle distributions in nanocomposites (beyond atomic force microscopy pictures) ♦♦♦♦ (4) • Transitioning between models (lack of a single model representation) ♦●●● (1,3) • Lack of pre-competitive data for model development ♦♦♦♦ (4) • Consistently using proper definitions (e.g., "a property that emerges under certain conditions relative to a material that exists at those conditions") ♦♦♦ (3) • Overcoming data model dependency at the nanoscale when models are not currently known (i.e., complexity at a scale larger than nanoscale may lead to emergent behavior) ♦●● (1,2)
Lower Priority	• Lack of the following: ♦ (1) − Potential functions to model all species of interest through bond breakage − Standards for reporting validation of potential functions − True multi-scale simulation capability − Nanoscale data for validation • Continuously improving timescale data • Relating the product to the materials property data
Data Management	
High Priority	• Lack of incentives to disseminate data ♦♦♦♦♦♦♦♦♦♦● (10,1) • Constructing a metadata interface between disparate databases ♦♦♦♦♦♦♦♦♦♦ (10) • Establishing a flexible framework to allow new forms of data and storage ♦♦♦♦♦♦♦♦●● (8,2)
Medium Priority	• Accessing proprietary experimental data and using it with other databases and simulation data ♦●●●● (1,4) • Ensuring data flexibility in the MII ●●●● (4)
Data Quality	
High Priority	• Establishing/improving the provenance of experimental data and models ♦♦●●●●●●●● (2,8) • Avoiding a strict focus on specific accuracy targets in simulations ♦●●●●●●● (1,7) • Systematically comparing and validating density function theories against experiments and higher-order theories ●●●●●● (6)

Table 2.3 Nano and Molecular Length Scale Challenges (♦ = one vote for potential short-term impact/<5 years) (● = one vote for potential long-term impact/5 years and beyond)	
Medium Priority	• Life prediction of materials without physical testing ♦♦●●● (2,3) • Lack of methods to ensure proper attribution and create a culture of sharing ♦♦♦♦♦ (4) • Providing a sufficient pedigree of data to support product warranties ♦●●●●(1,4) • Properly defining standard reference data and archival data ♦♦●●(2,2)
Lower Priority	• Evaluating the quality of data ♦(1) • Providing access to experimental data in open literature with sufficient logic or condition information ♦ (1) • Determining the qualitative and quantitative accuracy of error bars
Data Usability	
High Priority	• Representing real material data in process-structure and structure-property relationships ♦♦♦♦♦♦♦♦●(8,1) • Creating a system to balance the needs of the individual and community while sharing and distributing data ●●●●●●●● (8)
Medium Priority	• Limited network of data and models (with uncertainties) to make multi-physics predictions; lack of tools for sensitivity analysis and decision making ♦♦●●(2,2) • Determining if data is reproducible ♦♦♦● (3,1) • Determining appropriate amount of metadata and meaning behind data collected ♦●● (1,2)
Lower Priority	• Linking data with reality via model verification and/or repeatable control of materials manufacturing equipment ♦● (1,1) • Developing data ontology • Complementing data with tools to build virtual models

2.3.2 Top Short-Term Nano and Molecular Length Challenges

Data dissemination and sharing: Sharing data poses a challenge as organizations are concerned about protection of scientific work and intellectual property. More widespread, open data dissemination will be necessary to build a successful MII, but incentives are currently lacking to encourage this higher level of data sharing. Mechanisms or incentives to encourage data dissemination could potentially be provided via authorship credit, acknowledgements, a reward system, or other incentives. An example of this concept is the World Wide Web Consortium, a community-driven site that serves as a comprehensive and authoritative source for web developer documentation. A recent study illustrates what motivates scientists to share data.[6]

Metadata interfaces between databases: Another significant challenge for data management is developing and constructing a standard metadata interface between disparate databases. Databases vary in form, amount, type, quality, and properties, and methods are needed to access the metadata when utilizing or linking different databases. This would allow data to be easily shared even through different software systems.

---

[6] Tenopir, C.; Allard, S.; Douglass, K.; Aydinoglu, A. U.; Wu, L.; Read, E.; Manoff, M.; Frame, M., Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE* **2011,** *6* (6), e21101, http://dx.doi.org/10.1371/journal.pone.0021101.

Representing real materials data in process-structure/structure-property relationships: Process-structure and structure-property relationships are essential to the understanding and development of materials. Today challenges exist in representing real nano and molecular materials data so that it can be effectively used in process-structure and structure-property relationships. While application of real materials data to these relationships is a significant challenge, if achieved it would provide greater capability to identify the most suitable materials for structural applications while avoiding overdesign.

Flexible data frameworks: Establishing a flexible data management framework poses a significant challenge. An extensible framework must allow for the incorporation of new forms of data and storage for emerging models while accommodating current data and storage protocols. A flexible framework must also be able to integrate new data into existing and emerging models to facilitate useful data sharing.

2.3.3 Top Long-Term Nano and Molecular Length Challenges

Taxonomy for mapping materials spaces: A significant challenge with long-term impact for data representation is the lack of taxonomy or coding scheme to systematically map out materials spaces down to the nanoscale. A mapping of materials spaces to corresponding applications (including situation-specific data) would provide valuable information that could be shared with the materials community to facilitate and speed materials design.

Stringent standards for data: The current lack of standards for what qualifies as 'data' makes it difficult to integrate and share different types of data between different applications and disciplines. More stringent data definition standards are needed to enable data identification and facilitate sharing and transfer and allow researchers to better utilize external sources of data in their models.

Provenance of experimental data and models: Establishing the provenance of experimental data and models continues to be a challenge. Provenance establishes the history, transformation, and ownership of the data, model, or other object. As data is increasingly shared, users will need to be assured that experimental data or models from various sources is of high quality by understanding its origin and measurement conditions. Provenance is also a key element in using data to provide credible analysis and information that effectively supports decision-making.

Balancing needs of individuals/communities to share/distribute data: A system is currently lacking that can balance the needs of individuals and the larger community when sharing and distributing data. The need of researchers and scientists to "own" certain results and data should be balanced with the broader interests of the community. In this context, scientists would be encouraged to view themselves as part of a collective network working within the larger community.

Avoiding strict focus on specific accuracy targets in simulation: The degree of accuracy required for simulation is situational (i.e., for some applications 80% "right" might be sufficient). A total focus on accuracy can lead to a lack of progress, as the simulation/model can always be improved. Perfect accuracy is impossible, so the target accuracy should be application specific.

2.3.4 Nano and Molecular Length Challenges with both Short- and Long-Term Impacts

Of the challenges identified in this venue, most were selected as having either a strongly short-term or long-term impact. Some less pressing challenges, such as those below, could have comparable short- and long-term components.

- Determining the appropriate amount of metadata or meaning of the data collected
- Properly defining standard reference data and archival data
- Predicting the life of materials without physical testing
- Limited ability to move from data (with qualifying data provenance) to information and decision making

2.4 Atomic Length Scale

The atomic length scale encompasses several modeling techniques that involve direct simulation of atoms, as well as experimental techniques that probe this length scale. Some of the methods that the atomic length scale encompasses include Density Functional Theory (DFT) and Quantum Monte Carlo (QMC) analysis.

2.4.1 Overview of Atomic Length Data Challenges

The atomic length scale data challenges that can potentially slow the materials design cycle are listed in Table 2.4 and described below.

Table 2.4 Atomic Length Scale Challenges (♦ = one vote for potential short-term impact/< 5 years) (● = one vote for potential long-term impact/5 years and beyond)	
Data Representation and Interoperability	
High Priority	• Lack of standards for experimental data representation, including format, experimental conditions, level of information, and analysis tools; large number of formats are used for reporting data ♦♦♦♦♦♦♦●●●●●●●●●●● (7,11) • Dealing with a variety of rapidly evolving methodologies and codes, shifting of best practices, and changes in default settings over time ●●●●●(5)
Medium Priority	• Lack of taxonomies (bottom-up or top-down) resulting in limited ontologies and protocols for database structure and property definitions (e.g., common metadata definitions) ♦♦●● (2,2) • Improving the use of large amounts of simulation data, particularly for soft matter (liquids, colloids, polymers, foams, gels, etc.) ♦♦● (2,1)

Table 2.4 Atomic Length Scale Challenges (♦ = one vote for potential short-term impact/< 5 years) (● = one vote for potential long-term impact/5 years and beyond)	
Lower Priority	• Complete disclosure of equilibration procedures prior to production runs of the simulation ♦♦(2) • Lack of methods to encode (and form queries) for complex molecular structures ♦(1) • Developing materials relevant to working conditions; experimental testing and theoretical modeling often use conditions that are not sufficiently realistic ♦(1) • Limited publicly available data – interoperable databases are needed to enable data mining to fill gaps
Data Management	
High Priority	• Inadequate allocation of resources for data management beyond the immediate scope of the current project ♦♦♦♦♦♦●●●●●● (6,7) • Lack of local and global data storage, storage tools, and data collection solutions (i.e., flexible access, permissions, scalable infrastructure) ♦♦♦♦♦♦(6)
Medium Priority	• Lack of tools enabling online data storage online ♦♦♦♦♦●●(5,2) • Inadequate database longevity and maintenance ♦♦♦♦(4) • Limitations due to the lack of computer science and information technology expertise among materials scientists ♦♦●● (2,2)
Data Quality	
High Priority	• Lack of defined validation error bars for computational data ♦♦♦♦♦♦♦♦♦●●●●●● (9,6) • Lack of experiment round-robin tests designed to specifically validate computed properties ♦♦♦♦♦●●●●(5,4) • Inability to adequately compare experimental and computational data (i.e., error analysis, validation) ♦♦♦♦♦♦♦(7) • Insufficient incentives for data sharing ♦●●●●●● (1,6) • Inability to audit calculations, results, and conclusions so they can be reproduced and understood ♦♦♦♦♦(5) • Limited integration of data generation and analysis workflow with storage of results ♦♦●●● (2,3)
Medium Priority	• Limited means to measure reliability of results and data, including error bars, random test and audit of data, and understanding of methods ♦♦♦(3) • Unacceptable propagation of errors and uncertainty quantification using higher-order approximation methods; systematic uncertainty quantification for large datasets ♦●●(1,2)

Lower Priority	• Lack of agreed-upon experimental datasets to use as standards to validate theoretical methods ●●(2)
	• Lack of benchmarks to support data quality:
	  – Poorly defined benchmark problems for quantitatively validating computations at extreme points of the design space ♦♦ (2)
	  – Insufficient experimental benchmarks for testing energy accuracy of computations for surface species and nanoparticles between 1 and 6nanometers in diameter ●(1)
	  – Absence of relevant benchmarks when new methods are introduced ♦(1)
	• Lack of trust in data ●(1)
	• Inadequate method development for accurate description of materials and chemical activities or properties ●(1)
	• Lack of methods to improve the energy accuracy of fast methods such as DFT ●(1) Lack of methods to systematically cancel known systematic errors in data (e.g., electronic structure) to make it more useful ●(1)

Table 2.4 Atomic Length Scale Challenges (♦ = one vote for potential short-term impact/< 5 years) (● = one vote for potential long-term impact/5 years and beyond)	
Data Usability	
High Priority	• Lack of a single searchable database providing properties linked to composition, microstructure, and scale(s) of the experiment of computation ●●●●●●●●●(9) • Developing open source interfaces to support many data standards; lack of common interface for searching across databases ♦♦♦♦♦♦♦♦♦♦(10) • Lack of complete workflow documentation, including all information necessary for reproducing calculations and experiments, for all properties and codes (i.e., providing inputs, parameters) ●●●●●●●●●(9)
Medium Priority	• Lacking data quality evaluation "scoring" by later users and scoring accountability (note the name of the scorer) ♦♦♦(3)
Lower Priority	• Limited data storage and technologies that integrate with data analysis for more scalable data mining ●●(2) • Inability to use existing popular, well-developed, mass-scale data solutions (e.g., Flickr, Google) ●●(2) • Tools for high-throughput experimental measurement of energy stability of nanomaterials, complex materials (bulk), and surface species ♦●(1,1) • Information tools available at the database site to prevent data transfer issues ●(1) • Lack of accurate materials imaging tools (e.g., measurement and visualization of concentration gradients and flow) ♦(1) • Maintaining compatibility of software (platform and software evolution) ●(1)

2.4.2 Top Short-Term Atomic Length Challenges

Computational data validation: Research in general would benefit from error analyses and improved data validation efforts. Data quality is hampered by a lack of reportable, defined uncertainty metrics to enable more rigorous validation of computational data. Uncertainty metrics would facilitate more direct assessment of the accuracy and precision of experimental and computational data. In addition to these metrics, methods for measuring the reliability of results and data are generally limited, including random test and audit of data and understanding of the methods used to produce data. Benchmarks are also lacking to validate computational data particularly at the extreme points of the design space.

Inadequately detailed research papers in some cases have highlighted the need for detailed validation and verification of results. In addition, metadata could be useful for validation if it could be attached to electronic structure data to reflect technical failure or success of a particular method.

Comparability of data: Another challenge impacting data quality is the inability to adequately compare experimental and computational data. It is important to understand the limits of using computational and experimental data and the potential pitfalls of treating them in a similar way. The lack of complete workflow documentation and the high cost of data access compound the problem and highlight the need for publicly available, transparent data.

Database interfaces: Database interfaces remain a challenge for data retrieval and usability. Web interfaces that are often used to enable databases can be perceived as an impediment to research. There is growing interest in allowing APIs to access databases in an automated manner rather than via web interfaces. The "pymatgen" (http://pypi.python.org/pypi/pymatgen) Python library, for example, programmatically accesses the MIT/LBNL Materials Project data (https://www.materialsproject.org).

It is also difficult to search across multiple large datasets without a means to effectively integrate databases. The development and broader use of open source interfaces that support a range of data standards will be important to overcoming data retrieval and usability concerns.

2.4.3 Top Long-Term Atomic Length Challenges

Data compatibility and usability: The lack of standards for representing experimental data poses a significant challenge. Data is often produced in proprietary formats, which can render further analysis by the community at large difficult or impossible. There are also large numbers of formats for reporting data (i.e., input files used for different software packages). These challenges demonstrate the need for improved methods of data representation that provide information on the relevance of the data for a particular application. Information sharing efforts and the exchange and uploading of data would benefit significantly from data standardization.

Disparate databases and datasets: One significant challenge is the absence of a single searchable database that provides properties linked to composition, microstructure, and scale of the experiment or computation. The lack of integration of multiple datasets and property sets in a single search environment make searches for specific structures particularly difficult. Tools are needed to automatically extract data from the expanding materials science literature; this will require the ability to search for useful data in large datasets (i.e., database interfaces).

Workflow documentation and automated data retrieval: A scarcity of complete workflow documentation creates a considerable challenge for data retrieval and usability by research entities. Failure to disclose necessary inputs, parameters, and other critical information behind the data or model makes it difficult to reproduce or improve upon the work of other researchers.

Locating relevant data in the open literature is also becoming increasingly difficult as the amount of computational reporting continues to rapidly grow. Finding data can be so costly and time-consuming that it is not feasible given the resources available for the research. An automated method for extracting data, hypotheses, and other key information from the literature would help to address this growing challenge.

2.4.4 Atomic Length Challenges with both Short- and Long-Term Impacts

Lack of standardization: One of the challenges identified as having both short- and long-term impact potential is the lack of standardized formats to describe experimental conditions, analysis tools, and other practices that would enable comparison of existing and new research in a meaningful way. This challenge highlights the need for a standard format or translators to standardize experimental and computational data both now and in the future.

Data validation and error bars: This challenge can have both short- and long-term impacts on data quality. Evaluating the accuracy and precision of experimental and computational data is difficult given the lack of uniform standards and procedures for reporting the margins of error. As noted, this impacts the ability to assess the accuracy and precision of the data and to compare datasets from similar experiments. This is an on-going challenge that will require sustained attention to ensure information fidelity.

Data management resources: Allocation of resources for data management beyond current project scope is often inadequate and impacts the longevity and utility of data, both in the short and long term. In general, the tendency for data to be filed and forgotten limits the use of valuable data from past research. Both the lack of standardized formats for storing large simulation datasets and the absence of an accessible repository for storing computational data demonstrate the need for proactive allocation of resources to manage data so that it can remain useful for the long term.

Linked data is an approach that could be relevant for efficient data management. Linked data provides a publishing environment where not just documents but data are accessible on the Web, extending the Web into a global data space with open standards (e.g., Web of Data). A number of resources are available described the concept of linked data, which is gaining interest in the scientific community. [7,8]

---

[7] Heath, T.; Bizer, C., *Linked Data: Evolving the Web into a Global Data Space*. 1st ed.; Morgan & Claypool: 2011; Vol. 1,http://dx.doi.org/10.2200/S00334ED1V01Y201102WBE001.

[8] Berners-Lee, T., Tim Berners-Lee on the next Web. In *TED Talks*, TED Conferences, LLC: 2009, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html.

3 TECHNICAL APPLICATION AREAS

3.1 ELECTROCHEMICAL STORAGE

Electrochemical energy storage is an application area of growing interest as the adoption of alternative sources of energy increases. Sources such as solar and wind are intermittent, and storage capabilities are needed to enable energy capture and on-demand use. Developing electrochemical storage is also important to alternative transportation technologies such as hybrid and electric vehicles. Today's technologies for energy storage are inadequate to support the innovative, next generation grid and transportation capabilities envisioned for the future. New materials have the potential to increase the efficiency of devices for energy harvesting, storage, and conversion and create viable solutions for storing alternative energy.

3.1.1 Overview of Data Challenges for Electrochemical Storage

The data challenges related to materials design for electrochemical storage are listed in Table 3.1 and further described in the following sections.

Table 3.1 Electrochemical Storage Challenges
(● = one vote)

Frameworks, Methodologies, and Tools	
High Priority	• Developing and standardizing a simulation and analysis toolbox ●●●●● (5) • Creating a central data repository and software structure ●●●● (4)
Medium Priority	• Building a database with proper indices for searching for solid state materials and properties (e.g., similar to the American Chemical Society's SciFinder research discovery tool for organic compounds) ●● (2)
Lower Priority	• Building open source development environments for the creation of digital frameworks (with data consistency checking and model linking) that are reusable, robust, re-composable, and easily used and accessed by non-materials scientists and electrochemists (e.g., the North Carolina State framework); the environment would have the ability to rapidly create workflows for non-experts ● (1) • Designing experimental tests for materials benchmarking ●(1) • Establishing an architecture of databases and systems that can drive layers of information and data by emulating computer science methods ● (1) • Developing a common nomenclature and vocabulary to communicate chemical reaction data across disciplines (e.g., fuels cells to batteries) ● (1) • Extending the MGI framework to enable portability, scalability, and adaptability ●(1) • Developing methods to analyze and integrate data from experiments and models (e.g., an open source interface or a set of basic/common formats) • Developing a concrete use case for an energy storage problem to demonstrate the benefits of integrating shared distributed data and metadata

Table 3.1 Electrochemical Storage Challenges (● = one vote)	
Data Capture and Evaluation	
High Priority	• Defining and generating data on materials and systems (safety, lifetime, reliability, and cost) ●●●●● (5)
Medium Priority	• Evaluating correct structure property data ●●● (3) • Defining the top twenty queries that the MGI should address ●●● (3) • Capturing data from unpublished failures ●●● (3) • Establishing ongoing efforts to digitize existing phase diagrams from old databases and publications ●●● (3) • Researching fundamental data including thermodynamics data and reaction rates (e.g., dissolution rate or the transport properties specific to energy storage) ●● (2) • Creating a system where suppliers provide materials properties with products ●● (2)
Lower Priority	• Identifying the most costly and time-consuming priorities in the field of electrochemical storage ● (1) • Gathering all known experimental data and defining figures of merit • Investigating commonalities across various types of databases
Cross-cutting	
High Priority	• Communication and interaction of researchers at different stages of development (e.g., atomistic to end-of-line) to focus research on relevant materials properties ●●●●●(5)
Medium Priority	• Generating relevant, high-impact data using a distinguishing methodology (e.g., electrolyte transport properties) ●●(2) • Researching the flow of data and models as a means of establishing provenance ●● (2) • Investigating data correlations across applications and end uses (e.g., bad data can yield good data science) ●● (2)
Lower Priority	• Dividing experimental data modeling needs by technology segment (i.e., grid, transportation, or end life) ●(1) • Enhancing communication and information dissemination across disciplines by creating a new material ontology

3.1.2 Top Challenges for Electrochemical Storage

Based on the results of the voting process as shown in Table 3.1, a set of the top data challenges impacting the design of new materials for electrochemical storage were identified. If addressed these challenges have the potential to speed up the materials design cycle. Top challenges are summarized below and described in more detail in Figures 3.1.1 through 3.1.4.

Define and Generate Data on Materials and Systems: Data on both materials and the systems into which these materials are inserted must be defined and generated to incorporate considerations of safety, longevity, reliability, and cost into design. This will require identification of key metrics and targets for these essential parameters. Models for life prediction that are enabled by appropriate information and allow researchers to access the data should be developed. Lab scale-up to metric tons (at tolerable impurity levels) and full cell and systems modeling are needed as well. This would decrease the time for actual testing of real batteries from years to weeks or days via modeling. Other benefits include identification of key materials for modeling, being able to determine the feasibility of scaling up material to production, and knowing how a cell will act in a completed battery pack. Figure 3.1.1 provides additional details about this challenge.

Standardization of a Simulation and Analysis Toolbox: One of the main data challenges for electrochemical storage is the development and standardization of a simulation and analysis toolbox via nomenclature standards. Specific tasks include: (1) building community standards for experimental and computational analysis for batteries similar to ASTM (American Society for Testing and Materials) International standards; (2) building a system of metadata tags; and (3) creating a standards methods database that can be revised as needed over time. With a simulation and analysis toolbox researchers will spend less time reproducing others' results, standards will become R&D proxies for pre-competitive collaboration, and overall materials genome discussions will become more systematic. Figure 3.1.2 provides additional details about this challenge.

Central Repository and Software Infrastructure: The creation of a central data repository and software infrastructure incorporating information related to electrochemical storage materials could provide numerous benefits, but remains a challenge for electrochemical storage. Such a repository would make collaboration and data sharing much easier. A leader or manager for the central data repository must be identified along with an institution where it can reside. A business model should be developed to demonstrate how the repository could be sustained, including incentives to maintain the infrastructure. Champions (respected entities/individuals within the community) of the repository would help to ensure continued support. Successful creation of a staffed central repository and software infrastructure could potentially allow cross-disciplinary access to data, enable users to operate analysis tools using shared data, and tie together data from different sources. Figure 3.1.3 provides additional details about this challenge.

Communication across Different Stages of Development: To better focus research on relevant materials properties, extensive communication must be established between researchers working at different stages of development, from atomistic to deployment to in-service. Activities include events that bring together academia and industry to identify needs, identification of figures of merit for calculated properties, R&D to evolve figures of merits and to inform researchers of what data to collect, and the development of electrochemical materials descriptors (rather than "properties"). Communication across development stages will lead to a better understanding of how and why materials perform. Predictive models and the time-to-market of materials discoveries will also improve with more open lines of communications. Figure 3.1.4 provides additional details about this challenge.

Figure 3.1.1 Define and Generate Data on Materials and Systems

There is a current need to define and generate data on both materials and systems to incorporate considerations of safety, longevity, reliability, and cost into design. Key performance targets must be identified to overcome this challenge.

Short-term Win

Accelerated discovery and commercialization of new materials and systems

Long-term Advance

Commercialization of new materials via cycle life models and better scale-up methods

Short-term Activities

- Identify key performance targets such as safety, life, cost, toxicity, and reliability
- Develop predictive models for cycle life (with appropriate data and data access for researchers)
- Perform scale-up from lab-scale to metric tons including tolerances for impurity levels and cost
- Model full cell and systems

Long-term Activities

- Verify experimental vs. modeling cycle life
- Research experimental data for modeling inputs
- Develop fundamental mechanisms of failure

Short-term Milestones

- Determine impurity levels (in parts per million or parts per billion)
- Develop predictive tools for measuring cycle life
- Develop integrated modeling from battery to full system

Long-term Milestones

- Establish definitions of failure mechanisms

Short-term Outcomes

- <u>TAA Benefits</u>
  - Common platform for all battery models and systems
- <u>Contribution to MII</u>
  - Multi-scale modeling capability
  - Collaboration across disciplines (e.g., chemistry, physics, engineering, and others)

Long-term Outcomes

- <u>TAA Benefits</u>
  - Common platform for all battery models and systems
- <u>Contribution to MII</u>
  - Valuable data to support future materials work

Figure 3.1.2 Standardization of a Simulation and Analysis Toolbox

A simulation and analysis toolbox and nomenclature standards would be beneficial to designing materials for electrochemical storage. These will reduce the time needed to reproduce work and provide a more systematic

Short-term Win

Less time spent reproducing results

Long-term Advance

Acceleration of electrochemical storage field through data reuse

Short-term Activities

- Develop a set of standard experimental techniques as a guideline for investigating new materials for battery components and proxies
- Develop standard techniques for computational research including code, optimization paths, configurational choices, and other simulation requirements

Long-term Activities

- Develop a community website to house and document common standards and deliver best practices

Short-term Milestones

- Draft toolbox standard compliant with standards organizations and material science battery/energy storage community
- Establish micro-formats to enable search of standards
- Complete a series of workshops to evolve and extend the draft standard

Long-term Milestones

- Track rate of growth and use of the standards in materials science

Short-term Outcomes

- <u>TAA Benefits</u>
  - Community adoption of standard nomenclature in papers
- <u>Contribution to MII</u>
  - Decreased time required for researchers to reproduce experiments and simulations from other researchers

Long-term Outcomes

- <u>TAA Benefits</u>
  - Useful legacy data for future materials design
- <u>Contribution to MII</u>
  - Increased confidence of the research community in reusing data and conclusions as a result of an established toolbox standard

Figure 3.1.3 Central Repository and Software Infrastructure

A central data repository needs to be created and will require leadership, selection of a facility, development of software infrastructure, and a plan for sustained maintenance and growth over time.

Short-term Win

Connecting data at length scales

Long-term Advance

Acceleration of the design process and improved data validation

Short-term Activities

- Identify respected "champions" to ensure sustained support
- Develop a funding/business model
- Choose an institution to house the repository
- Build an incentive infrastructure

Long-term Activities

- Facilitate the development of better batteries
- Incorporate automatic links to publications
- Change the materials science and engineering culture for how data is shared and stored
- Improve communication across existing boundaries

Short-term Milestones

- Establish a group of community members authorized to upload data
- Form a governance board
- Attract more than 10,000 active users and/or contributors

Long-term Milestones

- Gain active participation from the battery community
- Obtain long-term commitment from an institution

Short-term Outcomes

- <u>TAA Benefits</u>
  - Access to data across community disciplines
  - Ability to operate specialized analysis tools on data provided by others
  - Association of data from different sources
- <u>Contribution to MII</u>
  - Integration of new tool s by improving and evolving the interface to data

Long-term Outcomes

- <u>TAA Benefits</u>
  - Sustained source of usable legacy data
- <u>Contribution to MII</u>
  - Integration of new tools by improving and evolving the interface to data
  - Reduction of the design process length for new materials

Figure 3.1.4 Communication across Different Stages of Development

Communication needs to be established between researchers working at different stages of development (e.g., atomistic to deployment to in-service) to focus on relevant materials properties. This will increase understanding of materials, improve models, and shorten the time to market for new materials.

Short-term Win

Identification of key priorities, parameters, and issues of greatest impact to industry

Long-term Advance

More effective navigation of the materials genome across domains

Short-term Activities

- Conduct workshops between academia and industry to identify pre-competitive needs
- Identify figures of merit for calculated properties
- Perform R&D to evolve figures of merit and inform researchers of data collection needs

Long-term Activities

- Develop electrochemical materials descriptors rather than “properties"
- Advance data mining techniques
- Develop new analytics for navigating materials applications

Short-term Milestones

- Develop a clear roadmap of energy storage materials properties goals with shortcuts
- Collect and disseminate industry-academia findings with shortcuts

Long-term Milestones

- Accomplish a 'Fundamental Engineering Problem ' on atomistic-to-design modeling chain of analysis
- Publish descriptor database and data analytics

Short-term Outcomes

- TAA Benefits
  - Better understanding of why and how materials perform
  - Better predictivity of models
  - Improved time to market for designed materials.
- Contribution to MII
  - Broad perspectives on materials data needs and development

Long-term Outcomes

- TAA Benefits
  - Shortened time to design new classes of electrochemical materials
  - Expanded design limits of a particular material
- Contribution to MII
  - Creation of “technology-agnostic” descriptor database
  - Formation of an open, shared toolkit for data analytics allowing rapid validation of data and models

3.2 HIGH-TEMPERATURE ALLOYS

High-temperature alloys are used in technologies where the material must withstand extreme temperatures, such as in jet engines. These alloys must be able to exhibit creep resistance, high strength at extreme temperatures, and high fatigue life, as well as degradation resistance at high temperatures.

3.2.1 Overview of Data Challenges for High-Temperature Alloys

The data challenges related to materials design for high temperature alloys are listed in Table 3.2 and further described in the following sections.

Table 3.2 High-Temperature Alloy Challenges
(● = one vote)

Data Quality/Pedigree	
Medium Priority	• Referencing data for phase-based properties to build multicomponent "databases" (e.g., CALPHAD) ●●● (3) • Understanding statistical confidence but not knowledge ●●● (3)
Lower Priority	• Flawed or missing fundamental data for many systems • Determine what data history should be required • Determine the volume of data required to achieve a needed level of certainty • Develop data quality assessment metrics that are quantitative • Reduce amount of time required for the review of independent data sets and consensus of results by experts
Data Retrieval/Usability	
High Priority	• Lack of a definition for high-temperature materials metadata that enables easier data retrieval ; lack of defined universal identifiers for materials so data can be located in an established database ●●●●●●●●●●●●●●●●(16)
Medium Priority	• Visualization of generated information and data ●●●(3) • Determining the representative volume element for each material or property critical for high-temperature systems ●●●(3)
Lower Priority	• Determine the grain size and orientation before initial use and at inspection intervals (non-destructive evaluation) ●●(2) • Identify the key microstructural attributes that are most influential to properties ●(1) • Develop data mining and analysis tools for large data sets
Material Applications	
High Priority	• Formulation of foundational engineering problems to provide context for data challenges ●●●●●●●●●●(10) • Application data needs for materials and quantification of properties ●●●●●●●●●(9)
Lower Priority	• Linking models between different length scales ●●(2)

Table 3.2 High-Temperature Alloy Challenges (● = one vote)	
Data Sharing	
Lower Priority	• Lack of requirements and a coherent community for obtaining and sharing data throughout the high-temperature materials supply chain (e.g., by developing a common certified data format, data sharing definitions) ●●(2) • Lack of incentives for sharing data throughout supply chain that outweigh the immediate costs, risks, and inertia ●(1) • Data encapsulation (for mobility)
Materials Degradation	
High Priority	• Develop a thorough understanding of the inherently dynamic nature of high-temperature materials in service ●●●●●●(6) • Developing a thorough understanding of the damage and aging phenomena for new high-temperature materials to increase their adoption ●●●●●(5)
Medium Priority	• Combining experimental and computational exploration of ternary and higher component systems ●●●(3) • Obtaining high-temperature corrosion data (especially localized corrosion) for multiple environments ●●(2)
Lower Priority	• Understanding and predicting microstructure and thus mechanical property degradation under a complex environment at high temperatures ●(1) • Enabling science-based processing (e.g., microstructure evolution models, defect formation)
Data Capture	
High Priority	• Automatically capturing data and metadata when generated, and subsequent transfer and sharing with other stages in the work flow ●●●●●●●●●●●●●●(14) • Long test times and high costs for testing ●●●●●●●●●●●(11)
Medium Priority	• Determining where/how to store share large amounts of test data (e.g.10 gigabytes of 3-D structural information) that is generated daily on high-temperature materials ●●●●(4)
Lower Priority	• Eliminating duplications of work and pooling data ●(1) • Developing rapid throughput characterization protocols • Developing standards for data collection (e.g., test and model methods) • Characterization and validation of multi-axial data for both deformation and life

3.2.2 Top Challenges for High-Temperature Alloys

Based on the results of the voting process as shown in Table 3.2, a set of the top data challenges impacting the design of high temperature alloys were identified. If addressed these challenges have the potential to speed up the materials design cycle. Top challenges are summarized below and described in more detail in Figures 3.2.1 through 3.2.5.

Data Retrieval for High-Temperatures Materials: For high-temperature materials, definition of metadata needs to be established so that data searches retrieve all relevant information, but does not retrieve extraneous/unnecessary information .Universal identifiers for materials should be determined when the database is built so data can be easily located. Pilot programs could be used to help demonstrate data retrieval methods. Also, to ensure that all data and metadata has been captured and can be retrieved, a database schema for high-temperature alloys needs to be built, tested, and implemented. Achieving these objectives would facilitate validation of models and lead to reduced testing requirements and costs. Figure 3.2.1 provides additional details about this challenge.

Designs for Fatigue Optimization: Models and associated data that provide predictive capabilities for high-temperature materials properties require further development. This involves design for thermal mechanical fatigue (TMF) optimization and component design for location specific high-temperature fatigue optimization. Related activities include: (1) development of TMF resistant materials; (2) tools to link designer requirements and materials capabilities; and (3) the development of materials data and associated material models to accelerate development of advanced superalloys for high-temperature applications. The detailed design requirements linking model-based alloy design efforts must also be captured. Benefits resulting from overcoming this challenge include reduced cost of energy and reduced cycle time for the design of new materials. Figure 3.2.2 provides additional details about this challenge

High-quality Fundamental Phase Data: Inexact fundamental multi-component phase data prevents the use of higher mechanistic process-structure and structure property models. Fundamental phase-based property models should be defined and pure element data with uncertainties should be established. Also, an infrastructure should be established for assessments and to determine error propagation for high-order systems. The result will be more accurate, high-order process-structure and structure-property models as well as the ability to design materials faster and more affordably. Figure 3.2.3 provides additional details about this challenge.

Characterization of Time- and Environment-Dependent Degradation Phenomena: Challenges for the high-temperature degradation of materials include characterizing structural and surface degradation, extrapolating short-term tests to life predictions, and the inherent complexity and high costs of simulations. Protocols should be utilized for in-service monitoring and inspection data as feedback for developing models and assessing components. Communication protocols that protect intellectual property and provide relevant data should be established to encourage data sharing. Benefits include improved communication between end-users and the supply chain, validated predictive models, and improved life assessments. Figure 3.2.4 provides additional details about this challenge.

Data Capture for High-Temperature Alloys: Current methods of materials characterization are expensive and time consuming; such issues are exacerbated when characterizing extreme temperature phenomena. To overcome this challenge a user-friendly, fully automated approach is needed for capturing and storing data and metadata. An enabling model and mechanism validation would also dramatically reduce physical testing requirements. A significant reduction in time to complete a full materials characterization could be achieved if this challenge is overcome. Figure 3.2.5 provides additional details about this challenge.

Figure 3.2.1 Data Retrieval for High-Temperature Alloys

Data retrieval methods need to be enhanced with improved definitions for high-temperature materials metadata. Universal identifiers for materials should be designed in the early stages of database development.

Short-term Win

Accelerated search for high-temperature data

Long-term Advance

Verified model for fast data mining and information retrieval

Short-term Activities

- Lay foundation for universal identifiers
- Build consistent descriptors
- Build a framework for data retrieval
- Develop a pilot test database

Long-term Activities

- Build database scheme
- Develop data mining tools
- Explore enhanced visualization

Short-term Milestones

- Identify representative team members from both academia and industry
- Develop a project charter
- Develop a database using developed definitions and identifiers
- Gather feedback from the user community

Long-term Milestones

- Implement and test database schema
- Implement and test database data mining tools
- Develop interface tools for visualization

Short-term Outcomes

- TAA Benefits
  - Required data or metadata is gathered
  - Culture is changed through involvement of others
- Contribution to MII
  - Realization of proof of concept with the prototype

Long-term Outcomes

- TAA Benefits
  - Enabling of data mining
  - Identification of new correlations
- Contribution to MII
  - Ability to build better models
  - Reduced need for testing and reduced costs
  - Cut in half development time

Figure 3.2.2 Designs for Fatigue Optimization

In the future, materials should be designed with thermal mechanical fatigue (TMF) optimization and specific high-temperature fatigue optimization. Models can aid in the development of fatigue property databases for materials used in high temperature applications.

Short-term Win

Realization of energy savings from efficient materials designs

Short-term Activities

- Establish material models
- Collect data and validate models
- Apply models to TMF and overall property performance predictions

Short-term Milestones

- Create databases for fatigue data
- Develop models and generate data
- Validate model or current methods
- Apply these milestones to the prediction of new alloys and designs

Short-term Outcomes

- TAA Benefits
  - Reduced energy costs
  - Reduced use of materials
  - Optimized development cycle
- Contribution to MII
  - Creation of models for future use
  - Integration of material models with design

Long-term Advance

Development of infrastructure to support a future energy efficiency increases

Long-term Activities

- Develop materials that include usable models in addition to data sheets

Long-term Milestones

- Apply methods to other alloy systems and component applications

Long-term Outcomes

- TAA Benefits
  - Rapid design changes
  - Reduced energy costs
  - Improvements to automotive and aeropace industries
- Contribution to MII
  - Development of integrated computational science and engineering skill sets within industry

Figure 3.2.3 High-Quality Fundamental Phase Data

Generation of fundamental multi-component phase data will enable the use of higher mechanistic process structure and structure property models. These models will relate material process, structure, and properties to reduce new material design time.

Short-term Win

Accurate reference and phase data for pure elements

Long-term Advance

Reliable fundamental data for multi-component phases

Short-term Activities

- Begin defining models for fundamental phase properties
- Define priority elements (e.g., nickel, titanium, and cobalt)
- Establish pure element data with uncertainties

Long-term Activities

- Define error propagation for higher-order system assessments
- Establish infrastructure for user to contribute assessments and associated data
- Develop incentivities to encourage others to contribute assessments
- Define metrics to rate quality of an assessment

Short-term Milestones

- Define models for fundamental phase properties
- Define priority elements (e.g., nickel, titanium, and cobalt)
- Establish pure element data with uncertainties , initially with priority elements then continuing with all the elements

Long-term Milestones

- Define and distribute error propagation methods
- Delineate metrics for quality assessment
- Develop infrastructure for assessment development
- Populate the assessment development infrastructure

Short-term Outcomes

- TAA Benefits
  - Improved accuracy of higher-level models
  - Improved higher-order assessments
- Contribution to MII
  - More accurate pure element data, which is important for the entire MII
  - Defined fundamental phase-based property models resulting in improved efficiency

Long-term Outcomes

- TAA Benefits:
  - More accurate high-temperature models
  - Faster design and qualification of new materials
- Contribution to MII
  - More accurate multi-component databases
  - More efficient development of multi-component databases
  - Faster and cheaper materials design

Figure 3.2.4 Characterization of Time- and Environment-Dependent Degradation Phenomena

The objective is to increase the understanding of high-temperature degradation through computational models that simulate material degradation and through experimentation. This can be achieved by characterizing structural and surface degradation, extrapolating short-term tests for life predictions, and reducing inherent complexity and the high costs of simulating operations.

Short-term Win

Improve communication between end-users across the supply chain

Long-term Advance

Utilize in-service monitoring and inspection data as feedback for model development and component assessment

Short-term Activities

- Establish communication protocols that protect intellectual property and provide relevant data

Long-term Activities

- Prioritize damage mechanisms
- Improve in-situ non-destructive evaluation and integrate these measurements into materials and design models
- Increase research into environment degradation

Short-term Milestones

- Define the business case for data sharing
- Define a forum for data sharing
- Create a mechanism for data sharing

Long-term Milestones

- List prioritized mechanisms
- Assess state-of-the-art and complete gap analyses
- Validate model predictions with field data

Short-term Outcomes

- TAA Benefits
  - Greater access to data that supports predictive models
  - Better next generation materials
- Contribution to MII
  - Common communication protocols
  - Improved communication between end-users and supply chain

Long-term Outcomes

- TAA Benefits
  - Validated predictive models
  - Improved life assessments
  - Improved materials development
- Contribution to MII
  - More relevant application data

Figure 3.2.5 Data Capture for High-Temperature Alloys

The time and costs of high-temperature material characterization need to be reduced. These reductions can be achieved through activities such as automated data capture and computational models.

Short-term Win

Development of an automated approach to storing data and metadata

Long-term Advance

Streamlined and robust data collection for high-temperature alloys

Short-term Activities

- Develop a user-friendly automated approach for storing data and metadata
- Define model-driven protocols for physical testing
- Identify model materials that represent relevant classes of mechanisms

Long-term Activities

- Enable invisible, zero-cost data capture
- Enable model validation with dramatically reduced physical testing
- Develop accelerated testing techniques

Short-term Milestones

- Identify a "standard" material characterization project to demonstrate feasibility of data and metadata capture, storage and retrieval
- Develop an accelerated testing protocol for a finite, standard set of material properties and accompanying physics-based models

Long-term Milestones

- Set up company materials information system to automatically incorporate data from materials test vendor
- Achieve a 50% reduction in physical tests required for new materials qualification
- Reduce by 50% the time to measure all relevant properties for physical and virtual testing techniques

Short-term Outcomes

- TAA Benefits
  - Reduced cost of data capture
  - Reduced time for full characterization
- Contribution to MII
  - Transferrable protocols to other materials and application areas

Long-term Outcomes

- TAA Benefits
  - Commercialization of new materials in the timeframe of design process
- Contribution to MII
  - Faster development of other families of materials, components, and assemblies via protocols

3.3 CATALYSIS

Catalysis is the change in the rate of a chemical reaction resulting from the use of a catalytic agent (catalyst). Catalysts enable reactions that might otherwise be blocked or slowed by a kinetic barrier by lowering or altering the requirements for the reaction to take place. Catalysts also enable reactions without being consumed in the process. Important sub-classes are electrocatalysts and photocatalysts, which enable electrochemical and photochemical conversions (e.g., as in fuel cells and devices for sun-driven fuels production). Catalysts already play a large role in energy and environmental technologies; advances in catalytic materials could enable important breakthroughs and lead to important new opportunities.

3.3.1 Overview of Data Challenges for Catalysis

Today challenges in data quality, data management, data representation and data usability are hindering progress in the field of catalysis and development of catalytic processes, which could provide significant social benefits. The following catalytic processes were identified as those that would benefit the most from an improved data infrastructure. Note that this list is not intended to be comprehensive; additional feedback from the catalysis community will be needed to ensure the most appropriate processes and challenges are being considered.

1. Selective (catalytic) oxidation of methane to methanol (potential for production of inexpensive liquid fuels from abundant resources)
2. Biomass conversion to both chemicals and liquid fuels
3. Greenhouse gas conversion
4. Water splitting and oxygen reduction
5. C-C_3 feedstock conversion (e.g., shale gas)

The data challenges related to catalytic materials design and development are listed in Table 3.3 and further described in the following sections.

Table 3.3 Catalysis Challenges (● = one vote)	
Data Representation	
High Priority	• Inconsistent, incomplete structure/activity data/correlations ●●●●●●●●●●●(11) • Inadequate representation of low symmetry surfaces with adsorbates (ex: amorphous) in a searchable format ●●●●●●●●(8)
Medium Priority	• Lack of a defined list of critical physical descriptors ●●●●●●(6)
Lower Priority	• Lack of machine learning tools to derive descriptors • Description and prediction of microscopic forms (e.g., for oxide catalysts) • Scaling issues (e.g., scalability) for data and relationships
Data Quality/Pedigree	
Medium Priority	• Obtaining experimental data on energy of selected systems chosen as well-defined benchmarks for extensions via theory ●●●●●●(6) • New, affordable, and accurate predictive computational methods for designing catalysts to benchmark faster models ●●●●●●(6)

Table 3.3 Catalysis Challenges (● = one vote)	
Lower Priority	• Inadequate assessment tools for reliability of experimental and computational data ●●(2) • Lack of tools for evaluating similarity (dissimilarity) ●(1)
Catalyst Characterization	
High Priority	• Lack of descriptions and models of complex materials (e.g., support-catalyst interactions); mesoscale processes in particular are poorly modeled/understood ●●●●●●●●●●●(11) • Define Industrial focus point around which to apply databases/tools/chemical classes ●●●●●●●●(8) • Incomplete documentation of catalyst experimental characterization; standards are lacking to enable reproducibility ●●●●●●●●(8)
Medium Priority	• Lack of available data; raw data (computational and experimental) is not routinely provided yet is needed to reproduce results and enable new analyses ●●●●(4)
Lower Priority	• Bridging the gaps from computations to surface science to real catalysis ●●●(3) • Long development times; limited understanding of time requirements and stages of the discovery process ●●(2) • Development of theoretical experimental tools that bridge different length and time scales • Reduce the number of steps involved in modeling kinetics
Data Sharing	
High Priority	• Lack of experimental methods for high throughput screening ●●●●●●●●●(9)
Medium Priority	• Inability to incorporate experiments and calculations into a common database (different data structures) ●●●●●(5)
Lower Priority	• Lack of a mechanism or tool to readily interface models with a data repository ●(1)

3.3.2 Top Challenges for Catalysis

Based on the results of the voting process as shown in Table 3.3, a set of the top data challenges impacting the design of new materials for catalysis were identified. If addressed these challenges have the potential to speed up the materials design cycle. Top challenges are summarized below and described in more detail in Figures 3.3.1 through 3.3.5.

Relationships between Fundamental Elements of Data: There is a need to better understand the fundamental elements of catalysis data and the relationships within data. Improved metadata formats are also needed to enable more advanced searches. The development of a comprehensive list of such metadata (loosely described here as a "descriptor") is an important first step to better understanding the data. Addressing these challenges would lead to the development of common terminology and data standards for a master list of characteristics. Another benefit is improved capabilities for robust design methods of new catalyst materials s, which would enable discovery and accelerate the development of new materials. Figure 3.3.1 provides additional details about this challenge.

High Accuracy Methods for Modeling Catalysis and Reactions: This challenge is related to the limitations of commonly used modeling methods that include DFT and QMC, among others. While QMC has improved accuracy over DFT methods, it is limited in scope, utility, and speed. Improved accuracy of modeling overall would allow for more extensive verification and validation of existing methods, yielding potentially revolutionary improvements in accuracy in the prediction of difficult applications. High accuracy methods would also enhance capabilities for benchmarking and validation of models. Figure 3.3.2 provides additional details about this challenge.

Mesoscale Catalyst Modeling and Mesoscale Structure-Activity Relationships: The increasing complexity of catalyst systems highlights the need for more efficient, reliable mesoscale modeling tools that are important to catalysis design. The large magnitude and configuration space associated with this challenge requires both new tools and methods for materials design. Models are needed that incorporate structural dynamics and electronic structures; these can be used as a basis for addressing the durability of catalysts in a fundamental way, in addition to other industry-relevant problems. Developing mesoscale models and tools could accelerate development of potentially new catalysis systems and would enhance the available knowledge base to enable faster development of materials. Figure 3.3.3 provides additional details about this challenge.

High Throughput Approaches to Integrated Experimental- and Computational-Driven Materials Design: It is challenging to integrate experimental and computational data in a way that will drive materials design. For example, coupling fast computing with physical models and fundamental property data would better enable development of materials to meet energy and environmental needs by accelerating the process of catalyst discovery. High throughput approaches would increase data sharing and availability and enable the materials science community to issue data along with publications. These steps would allow for rapid development of new catalysts. Figure 3.3.4 provides additional details about this challenge.

Integrated Predictive Tools for Heterogeneous Catalytic Reaction Mechanisms: The materials science community is challenged by the lack of integrated predictive tools for heterogeneous catalytic reaction mechanisms. Catalytic reactions are a series of bond-breaking and formation steps, the overall rate of which is determined by the catalytic material and reaction conditions. For this reason, tools are needed to determine the mechanisms governing the critical catalytic transformations and support key predictive experimental and theoretical methods - enabling more rapid prototyping of heterogeneous catalytic processes. Figure 3.3.5 provides additional details about this challenge.

Figure 3.3.1 Relationships between Fundamental Elements of Data

Identification of the fundamental elements of data (i.e., master list of descriptors) and the relationships between elements is key to accelerating materials design and process development. Metadata, such as chemistry and

Short-term Win

Improved screening of new catalyst candidates

Long-term Advance

Quantified structure-property relationships to accelerate design and process development

Short-term Activities

- Articulate a master list of descriptors that includes: structure, reactivity, stability, size, shape, and binding energy
- Document the history of catalysts use to identify shortcomings
- Define chemical "neighborhood," class, or space descriptors and informatically describe them for screening

Long-term Activities

- Research correlations among catalyst types and descriptors and quantify structure-property relationships
- Model experiments on a chemical "neighborhood" for a demonstration classification
- Cross-validate tests of chemical "neighborhoods"

Short-term Milestones

- Identify a list of descriptors agreed on by the community
- Establish data representation standards
- Agree on definitions

Long-term Milestones

- Identify gaps in property data
- Systematically evaluate composition and conditions to address gaps in the most important descriptors

Short-term Outcomes

- TAA Benefits
  - Common terminology
  - Data standards and ontology for master list of characteristics
- Contribution to MII
  - Foundation for predictive model development, gap analysis for experiments, and benchmark calculations

Long-term Outcomes

- TAA Benefits
  - Robust prediction of new catalyst candidates
  - Discovery of new materials with accelerated development
- Contribution to MII
  - Framework for data and insight lead to expansion of knowledge base
  - Fundamental classes and understanding can be generalized for other applications and design

Figure 3.3.2 Higher Accuracy Methods for Modeling Catalysis and Reactions

The capabilities, accuracy, scope, utility, and speed of DFT and QMC methods related to catalysis development need to be improved and expanded.

Short-term Win

Increased usability of existing codes (e.g., Quantum Monte Carlo Package (QMCPACK))

Long-term Advance

Improved scope (e.g., solid-state coupled cluster) and speed (e.g., better algorithms for all)

Short-term Activities

- Increase amount of user-friendly QMC codes
- Create more accessible GW/Bethe-Salpeter equation (BSE) codes
- Improve pseudopotentials
- Improve functional dispersion corrections

Long-term Activities

- Increase periodic, solid-state, many-body coupled cluster methods
- Improve scaling in system size (e.g., local methods)
- Hybridize codes and methods
- Investigate trade-off between accuracy and speed with algorithms and theoretical development

Short-term Milestones

- Adoption by users (outside of developers) of QMC GW/BSE
- Create high-level reference data for developing functional potentials, pseudopotentials, and web database with documentation

Long-term Milestones

- Apply coupled cluster methods to solids to possibly show proof-of-principle
- Create a periodic 100 atom system (crystal) with coupled cluster with single and double excitation
- Converge diffusion Monte Carlo on 100 atoms (solid/periodic) with less than one week computing time
- Develop carbon monoxide and water on platinum(III) to within 10 to 20 millielectron volt with any method

Short-term Outcomes

- <u>TAA Benefits</u>
  - Increased use of better methods within the community
  - More extensive verification and validation in existing DFT/pseudopotentials
- <u>Contribution to MII</u>
  - More powerful methods generally applied to materials problems

Long-term Outcomes

- <u>TAA Benefits</u>
  - Revolutionary accuracy in prediction of very hard applications
  - Benchmarking and validation of models
- <u>Contribution to MII</u>
  - Faster, more accurate materials development

Figure 3.3.3 Mesoscale Catalyst Modeling and Mesoscale Structure-Activity Relationships

Mesoscale modeling tools can be improved by including structural dynamics as well as electronic structure to enable high fidelity simulation for catalysis and materials discovery.

Short-term Win

New mesoscale models and experimental tools with temporal and spatial resolution

Long-term Advance

Established knowledge base and tool set for accelerating mesoscale material design

Short-term Activities

- Identify model systems for mesoscale model development and experimental calibration
- Establish working groups to bridge atomistic and microscale modeling for new methodologies
- Implement experimental tools (e.g., transmission electron microscope , X-ray) for inoperando mesoscale measurements

Long-term Activities

- Define error propagation for higher-order system
- Establish infrastructure for users
- Contribute assessments and define metrics for quality assessment

Short-term Milestones

- Illustrate model system with identified data constructs for future system implementation

Long-term Milestones

- Create accessible knowledge base for industrially relevant catalytic processes

Short-term Outcomes

- <u>TAA Benefits</u>
  - Creation of a group of models to be used as a basis for tackling industry-relevant problems
  - Ability to address durability in a fundamental way
- <u>Contribution to MII</u>
  - Capacity to bridge more established scales (e.g., macro and atomistic) with a new level of modeling

Long-term Outcomes

- <u>TAA Benefits</u>:
  - Formation of new material systems for catalysis
  - Broadened available knowledge base for faster development
- <u>Contribution to MII</u>
  - Continued ability and infrastructure to bridge scales

Figure 3.3.4 High Throughput Approaches to Integrated Experimental- and Computational-Driven Materials Design

Effective integration of experimental and computational data can accelerate catalyst discovery by taking advantage of fast computing, physical models, and fundamental property data.

Short-term Win

Rapid identification of potential new catalytic materials

Long-term Advance

Expanded and improved quality of fundamental data infrastructure needed to amplify and accelerate the successful discovery of new catalytic materials

Short-term Activities

- Identify "descriptors" for key reactions
- Correlate existing materials composition and structure with descriptor values
- Validate choice of descriptor
- Carry out a computational search for new materials which meet criteria for a promising catalyst target

Long-term Activities

- Obtain experimental data on selected benchmark systems which enable vast extension via computational theory
- Develop methods for high throughput generation of experimental data
- Improve methods for finding predictors
- Develop faster computational methods that have higher accuracy

Short-term Milestones

- Descriptors for key reactions
- Demonstrated computational search for new materials based on selected criteria

Long-term Milestones

- Validated high throughput methods for experimental data generation
- Demonstration of higher accuracy using selected benchmarks

Short-term Outcomes

- TAA Benefits
  - Increased data sharing and availability
  - Increased data disclosure with publications
- Contribution to MII
  - Integrated system containing validated data

Long-term Outcomes

- TAA Benefits
  - Rapid development of new catalysts
- Contribution to MII
  - Sustained system for integration of validated data

Figure 3.3.5 Integrated Predictive Tools for Heterogeneous Catalytic Reaction Mechanisms

The mechanisms governing critical catalytic transformations need to be established. The identification of key predictive experimental and theoretical tools will be essential to this effort.

Short-term Win

Establishment of simple qualitative descriptors from known reactions that permit rapid prediction of candidate catalysts

Long-term Advance

Increased comprehension of reaction mechanisms, including those with complex aspects

Short-term Activities

- Establish set of qualitative descriptors from known reactions that permit rapid prediction of candidate catalysts for a specific class of reactions and related new classes of reactions
- Analyze and validate existing data to determine descriptors and create informatics tools for data validation and analysis
- Extend the application of the design concept to more complex reactions
- Refine the descriptors for increased accuracy

Long-term Activities

- Provide new tools and approaches for decreasing development and prototyping time frame
- Expand design tools to be more comprehensive and include complexity of materials and reaction conditions
- Secure stable resources for computer-based projects
- Engage industry

Short-term Milestones

- Demonstrate proof of principle that descriptors in a model system lead to improved catalytic performance for a specific reaction

Long-term Milestones

- Sustained development of validated predictive tools for reaction mechanisms

Short-term Outcomes

- <u>TAA Benefits</u>
  - Rapid prototyping of heterogeneous catalytic processes
- <u>Contribution to MII</u>
  - Data infrastructure for reaction mechanisms

Long-term Outcomes

- <u>TAA Benefits</u>
  - Determination of mechanisms for catalytic reactions using more complex materials, including multi-component, non-crystalline and porous frameworks
  - Greater understanding of reaction mechanisms and the evolution of catalytic materials under a range of conditions, including complex environments
- <u>Contribution to MII</u>
  - Data infrastructure for reaction mechanisms

3.4 Lightweight Structural Materials

Lightweight structural materials enable industrial, commercial, and transportation systems to push performance limits and often times exceed customer expectations while matching or improving energy efficiency, safety, durability, utility, and/or environmental goals. For example, novel alloys based on magnesium and on titanium aluminides are lightweight and have promising property profiles that could make them suitable for certain applications, such as the power trains of automobiles or structural components made from cast and wrought alloys.

3.4.1 Overview of Data Challenges for Lightweight Structural Materials

The data challenges related to design and development of lightweight structural materials are listed in Table 3.3 and further described in the following sections.

Table 3.4 Lightweight Structural Material Challenges
(● = one vote)

Specific to Lightweight Structural Materials	
High Priority	• Establishing a data repository(s) for aluminum, titanium, and magnesium ●●●●●●●●●●●● (12) • Developing models of structures or materials comprising multiple materials, such as fiber plus matrix in a composite; two metals plus weld; or single crystal plus precipitate in a polycrystalline structure ●●●●●●●●●● (10) • Developing phase and interface properties, homogenization theories/models ●●●●●● (6) • Modeling joints ●●●●●● (6)
Medium Priority	• Lack of good data on elastic coefficients ●●●(3)
Lower Priority	• Understanding how constituent material properties relate to a specific (laminate) orientation (i.e., in a composite structure) ●●(2) • Establishing preferred orientation for texture ●●(2) • Modeling processes to convert material to usable goods
Generic to Lightweight Structural Materials	
High Priority	• Developing standards for defining data quality "data readiness level" ●●●●●●●●●(9)
Medium Priority	• Drive data cost to zero; create, manage, explore ●●●●●(5) • Creating data "social networks" ●●●●●(5) – Data and models: what's available and who's willing to share? • Establishing an expert rating system for data and data reproducibility ●●●●(4) • Establishing ownership and maintenance of the MGI database ●●●●(4) • Lack of taxonomy, e.g., a rules-based classification system for items across disciplines, length scales, experiments, and analyses ●●●(3)
Lower Priority	• High cost behind data – unique or expensive data not available to potential users ●●(2) • Filtering data (i.e., one kind of information);combining targeted data across domains ●●(2) • Creating a large data network that is transparent, self-organizing, and flexible ●(1) • Understanding model readiness level ●(1)

3.4.2 Top Challenges for Lightweight Structural Materials

Based on the results of the voting process as shown in Table 3.4, a set of the top data challenges impacting the design of new lightweight structural materials were identified. If addressed these challenges have the potential to speed up the materials design cycle. Top challenges are summarized below and described in more detail in Figures 3.3.1 through 3.3.5.

Optimized Joining Methods: Multi-scale models are needed to determine optimized joining methods for lightweight materials. These models would predict material performance in areas such as corrosion susceptibility and rates, static strengths, fatigue, and dynamic strength. Initial multi-scale models development could begin with selected materials and joining processes. New and continuously updated validation data will also be needed to ensure optimal model outputs. Development of good multi-scale models for joining will improve the ability to design durable hybrid structures with disparate materials bonding profiles. Figure 3.4.1 provides additional details about this challenge.

Models to Connect Phase Properties: To create structures of different sizes and shapes, engineers need to know bulk properties, which depend strongly on phase properties, topology, and interfaces. Today, inadequate models exist that can connect the bulk with phase properties. Development of these models will enable predictive properties capabilities that could be utilized by both experts and non-experts involved in materials design and development. Figure 3.4.2 provides additional details about this challenge.

Data Repositories for Aluminum, Titanium, and Magnesium: Aluminum, titanium, and magnesium are predicted to be the engineering workhorses for many structural applications in the future due to their high strength to weight properties. These material systems comprise a multitude of alloy types and configurations. Currently, a robust data repository for each alloy type does not exist, so non-validated data and information is sometimes used for model development and potentially structure design. An engineering performance data repository would provide a source of reliable data to facilitate high confidence part designs. Figure 3.4.3 provides additional details about this challenge.

Large Reduction in Data Costs: Materials data should be available to the materials community of interest as well as the broader scientific and industrial community, and should be increased by orders of magnitude to create a high fidelity MII. Creating, collecting, and storing potentially immense amounts of data is expensive. It is a significant challenge exists to drive these data costs down, but it will be essential to creating a robust MII. Figure 3.4.4 provides additional details about this challenge.

Standardized Definitions for Maturing Levels of Computational Models, Experimental Data, and Simulation: A standard set of data readiness levels should be developed to characterize data and models and to provide an assessment of the fidelity of collected information which can be used in various decision-making circumstances. These metrics, similar to the technology readiness levels used by the U.S. Department of Defense (DOD) and the National Aeronautics and Space Administration, would characterize the maturity of data or models. This could be factored into a structure's design. Figure 3.4.5 provides additional details about this challenge.

Multi-Material Models that include Texture and Structure: Today's materials modeling techniques that take into consideration texture with structure are inadequate or non-existent. This capability is extremely difficult to achieve and why it is absent or limited. However, elucidating computational methods to achieve such a linkage could open up new design space for lightweight materials and other material systems. Figure 3.4.6 provides additional details about this challenge.

Figure 3.4.1 Optimized Joining Methods

To optimize joining methods, the relevant properties (e.g., static, fatigue, dynamic strength, and corrosion resistance) of joined materials must be identified. The development of multi-scale models will enable prediction of bonding possibilities, bonding strength, and optimized joining methods between new materials. The data must be continuously validated and updated for on-going model optimization.

Short-term Win

Use of data and models to design joints for parts in a variety of industries

Long-term Advance

Creation of multiscale models to design and predict the joining of disparate materials

Short-term Activities

- Identify materials and joining processes likely to lead to significant cost and weight savings
- Understand conventional state-of-the-art approach to problem by others
- Define objectives and new framework
- Ascertain technical gaps
- Determine national return on investment

Long-term Activities

- Create a repository of joint performance data that includes variability of in bound material and process
- Standardize techniques for joint testing
- Validate models at wide range of length scale
- Verify model across a suitable range of materials, processes, and scales

Short-term Milestones

- Identify specific (representative) joint geometries and materials
- Define state-of-the-art for joint optimization
- Characterize new framework (high-level)

Long-term Milestones

- Create a repository of joint performance under relevant service conditions
- Standardize techniques for joint testing after being placed in relevant service conditions
- Develop an expert system for predicting joint variability

Short-term Outcomes

- <u>Benefit to TAA</u>
  - Better models for existing joint and material configurations
  - Faster implementation of lightweight materials
- <u>Contribution to MII</u>
  - Development of multi-scale-aware framework for improving existing major industry sector
  - Connection of material and design to end performance

Long-term Outcomes

- **<u>Benefit to TAA</u>**
  - Faster more dexterous use of broader range of materials
  - Improved reliability through better corrosion performance
  - Reduction in weight and cost by improved confidence in joint strength data
- **<u>Contribution to MII</u>**
  - Resolution of key issues over a broad range of scales in joint design and durabilty (e.g., bonding is at molecular level, corrosion at microstructure level, fatigue at all scales, static strength at macro but driven by micro)

Figure 3.4.2 Models to Connect Phase Properties

Models that link bulk properties with phase, topology, and interface properties can aid in quantitative prediction of component performance.

Short-term Win

Identification of important factors connecting phase properties with interface properties

Long-term Advance

Quantitative prediction of component performance to accelerate design process based on a complete understanding of phase and interface properties

Short-term Activities

- Generate phase properties
- Encourage awareness and understanding that orientation is comparable in importance to phase distributions in determining the properties of magnesium oxide, titanium, and aluminum-lithium
- Establish twin interface (e.g., magnesium, titanium, titanium-aluminum, precipitation, hardening, fracture)

Long-term Activities

- Enable design engineers to quantitatively predict component performance and properties using microstructure variables and phase attributes
- Facilitate quantitative prediction of aggregate properties based upon phase, interface, etc.
- Use interface properties (e.g., energies, diffusivities) in design as a degree of freedom

Short-term Milestones

- Identify specific coarse-graining (homogenization) approaches most promising for MGI
- Identify missing properties
- Identify strategies to obtain properties at phases, not normally obtainable as single crystals (e.g., Guinier and Preston zones)

Long-term Milestones

- Obtain community acceptance of orientation as comperable in importance phase distributions for magnesium oxide, titanium, and aluminum-lithium
- Interface mechanism in computational modeling (e.g., Finite Element Analysis) and not just cohesive zone
- Develop methods (e.g., models, experts) to quantify interface properties over huge orientation space

Short-term Outcomes

- Benefit to TAA
  - Move from wizard, experience, and empirical to quantitative and predictive description of impact of phase, interface, and topology properties on bulk properties
- Contribution to MII
  - Identification of important quantities for phase and interface properties

Long-term Outcomes

- Benefit to TAA
  - Faster, cheaper designs of lightweight structures by including phase, interface, and topology as design variables
- Contribution to MII
  - Identification of methods for incorporating phase and interface properties into higher scale models (integration techniques and models)

Figure 3.4.3 Data Repositories for Aluminum, Titanium, and Magnesium

A performance data repository needs to be created for the workhorse engineering materials (i.e., aluminum, titanium, and magnesium). This will improve capabilities for alloy model development.

Short-term Win

Definition of scope, funding, needs, and logistics of data repository

Long-term Advance

Establishment of repository (i.e., models with validated data) as primary data collection for efforts in light metals

Short-term Activities

- Define data to be included
- Secure resources for development of repositories
- Enable Internet access across multiple databases
- Identify where the data should reside

Long-term Activities

- High fidelity models of materials (e.g., chemistry, micro, and properties) that are depenendent on composition, temperature, pressure
- Acquire fundamental “data” for all binary and ternary systems in aluminum, titanium, and magnesium
- Validate models with descriptions as inputs and experiments

Short-term Milestones

- Identify and explore current sites (e.g., ASM International's Medical Materials Database)
- Demonstrate single point search
- Establish a flexible protocol/schema that builds on existing work
- Establish a community of interest via incentives and other mechanisms

Long-term Milestones

- Demonstrate model utility for industrial and scientific sector problems
- Establish repository as primary data collection for efforts in light metals

Short-term Outcomes

- <u>Benefit to TAA</u>
  - Save money by avoiding duplication
  - Tighter coupling between “academia” and industry
  - Potential performance/cost benefit to U.S. industry
- <u>Contribution to MII</u>
  - Success of pilot project exemplar demonstrated for other areas

Long-term Outcomes

- <u>Benefit to TAA</u>
  - Pervasive use of light metals
  - Establishment of community repository
  - Perpetual, self-sustained system
- <u>Contribution to MII</u>
  - None identified

Figure 3.4.4 Large Reduction in Data Costs

Data costs should be driven toward zero while collecting significantly more technical data. Less costly tools for data generation will be needed to accomplish this. Social networks and data repositories could also provide greater access of data and tools to communities of interest.

Short-term Win

Creation of more usable and relevant data for users to analyze

Short-term Activities

- Map major data types used in lightweight structural materials
- Estimate costs for creating, managing, and exploring the data types
- Issue challenges to reduce by at least 10x the cost to create data, for lightweight structural materials for example
- Create social network and data repository to share data and foster collaboration

Short-term Milestones

- Write report of data types and classes used in lightweight structural materials field
- Segment data generators and obtain cost models
- Identify challenges with prizes to change data cost structure
- Launch social network with data repository and measure data flow and user participation

Short-term Outcomes

- Benefit to TAA
  - More data for analysis
  - Engagement of nontraditional experts and disciplines
- Contribution to MII
  - Generalized method for data creation at greatly reduced cost
  - Formation of a socially networked environment for data creation and knowledge discovery

Long-term Advance

Enabling a more in-depth material understanding while lowering the cost to store data

Long-term Activities

- Create new simulation and predictive models for data production at low cost
- Miniaturize existing physical data creation techniques

Long-term Milestones

- Release simulation and predictive models and tools for general use
- Create new automated equipment to produce high quality data for years
- Launch social network for lightweight structural materials data sharing, challenges, and partnering

Long-term Outcomes

- Benefit to TAA
  - More well documented material properties for design and manufacturing applications
- Contribution to MII
  - Establishment of general methods and environment to accelerate discovery of novel materials, new material applications, and knowledge network formation

Figure 3.4.5 Standardized Definitions for Maturing Levels of Computational Models, Experimental Data, and Simulation

The establishment of data and model readiness levels (DMRL), similar to the technology readiness levels used by DOD and NASA, would aid in characterization of data and model readiness for decision-making.

Short-term Win

Identification of a DMRL for use in lightweight materials design

Short-term Activities

- Identify and gather a community of interest
- Define strawman DMRLs
- Present draft to committee
- Iterate drafts with community and committee

Short-term Milestones

- Identify community
- Hold kick-off meeting
- Draft DMRLs

Short-term Outcomes

- Benefit to TAA
  - Facilitates common understanding of model and data readiness for lightweight materials
- Contribution to MII
  - Application of same metrics throughout MII
  - Facilitation of data and model usage for innovation

Long-term Advance

Identifies the areas of lightweight materials data and models that require greater confidence

Long-term Activities

- Publicize existence of these metrics
- Promote usage of DMRLs in relevant funding and sponsorship opportunities

Long-term Milestones

- Gain widespread adoption of DMRLs throughout industry and government

Long-term Outcomes

- Benefit to TAA
  - Same as short-term outcomes
- Contribution to MII
  - Same as short-term outcomes

Figure 3.4.6 Multi-material Models that Include Texture and Structure

Beneficial improvements to materials models include relating structure performance to texture, grain morphology, constituent particles, and other structural data to reveal non-obvious performance factors (e.g., anisotropy).

Short-term Win

Identification of relationships between materials' interfaces and structural performance

Long-term Advance

Creation of models for predicting composite structures based on constituent materials properties

Short-term Activities

- Perform a literature survey to determine state of the art relationships between interfacial structure and structural performance
- Establish detailed testing and analysis for state of the art
- Determine domains of applicability
- Expand domains to materials such as composites, metals, and polycrystal line structures

Long-term Activities

- Determine important vehicle properties
- Determine constituent properties relevant to vehicle performance
- Establish a constituent link to the aggregate properties
- Develop required data sets at all relevant length scales to inform materials decisions

Short-term Milestones

- Create state of the art report
- Generate intermediate characterization of interface influence on tensile strength for one material
- Use new information to design interface
- Perform interface tests
- Establish data format for modeling codes.

Long-term Milestones

- Determine final vehicle performance criteria
- Establish final needed constituent propertes
- Develop, refine, and validate constituent models
- Disseminate to community

Short-term Outcomes

- <u>Benefit to TAA</u>
  - Database on interface influence
  - New best practice interface design
- <u>Contribution to MII</u>
  - Specific contributions to MII database
  - Template for user practice for data input to MII

Long-term Outcomes

- <u>Benefit to TAA</u>
  - Reduced cost of development of new materials
- <u>Contribution to MII</u>
  - Generation of data for database
  - Creation of constitutive model

4 Cross-cutting Challenges

A number of cross-cutting themes can be found throughout this report. In many cases these challenges impact the entire community involved in materials design. Some of the leading themes are summarized below.

Strong leadership for community efforts: Strong community leadership will be needed to foster the level of data generation, analysis, and sharing needed to support a successful MII. The materials community includes many disciplines and applications and these groups of interest can be somewhat isolated. As a result, there is limited leadership to support or champion efforts that will benefit the community at large. A cultural change towards a data-sharing philosophy will require leadership to build community-wide support and understanding of the value proposition. This will continue to be a challenge but is vital to the success of the future MII.

Data sharing: Incentives and structures are needed to encourage data sharing. Mechanisms are currently lacking to balance the needs of organizations and the larger community for sharing and distributing data. Both public and private organizations can require ownership of certain results and data for scientific or business reasons. Mechanisms are needed to balance those needs with the broader interests of the community while protecting ownership of discoveries, intellectual property, and competitive differentiators. A reward system for shared datasets, coupled with a structure that protects data at some level or credits data, could provide incentives for data dissemination. Digital Object Identifiers (DOI®) could be important idea for sharing data. These create a framework for identification, managing intellectual content and metadata, linking users with content sources, and enabling automated management of media.[8,9]

Computational validation: Validation of computational models was cited across all length scales as well as applications as a high priority. Computational validation tools require a long time to develop and are a general impediment to the success of the MII. Systematic and proper evaluation of the data acquisition process (and data) for model validation is another key challenge. Of particular interest is the ability to measure and prove that validated computational tools can greatly speed the design of new materials; be effectively applied to 'real' products; and reduce time and cost of testing materials.

Central data repository: The need for centralized and accessible data is a common challenge cited by all groups and noted as a high priority. Such repositories are currently limited and lack

[9] http://www.doi.org/

standard formats which makes it more difficult for data to be used by the community at large. Data management in general can be costly, requiring resources not just for generation but for collection, archiving, and maintenance. Some of the general requirements for a central data repository include:

- Standardized formats
- Definition of what data should be stored and maintained
- Data standards for uploading data
- Extensible framework for new data types that emerge
- Usable by non-materials scientists
- Means for sustained data life and maintenance

A HUB-based infrastructure serving a large number of users is one approach. In this case users could have access to a standard repository as well as opportunities and support for interacting with data. Similar successful efforts (e.g., nanoHUB) could be explored to gain insights on how to build and effectively operate a HUB-based repository.

Large data sets: The lack of methods for proper storage, transmission, and analysis of extremely larger datasets was noted as a cross-cutting challenge for a number of areas. The amount of data generated and accessible is growing exponentially and this trend is expected to continue. New methods will be needed to effectively manage and extract useful information from these massive datasets.

Data interfaces and interoperability: Effectively using available data is challenged by a lack of open source interfaces and interoperability between databases. Resolving these issues is a high priority for all domains. The lack of metadata interfaces between databases is a major issue, along with the ability to combine experimental and computational data and computations from different sources. Translating to different formats to interface with different software applications is currently problematic. Flexible, common data schemas and interfaces are needed to resolve some of these challenges. Other requirements for data compatibility include integration of multiple property datasets in a single searchable environment and compatibility with commercial codes.

Data standardization: Stringent standards are needed for both computational and experimental data are considered a high priority, including data curation and a data quality index to describe the quality of specific properties based on testing conditions. Standard data formats and metadata requirements for reporting and databasing of raw test data are also a priority; these will greatly facilitate data sharing and understanding of data quality and context.

APPENDIX A: PARTICIPANT LIST

Plenary Speakers
Axel Szalay, Johns Hopkins University
Gerhard Klimeck, Purdue University
Ian Foster, Argonne National Lab; University Chicago
Mladen Vouk, North Carolina State University
Workshop and Report Preparation
Energetics Incorporated

Sean Agnew	University of Virginia
John Allison	University of Michigan
Clare Allocca	National Institute of Standards and Technology
Steven Arnold	National Aeronautics and Space Administration
Michael Ashby	University of Cambridge
Mark Asta	University of California, Berkeley
Rick Barto	Lockheed Martin Advanced Technology Laboratories
Laura Bartolo	Kent State University
Chandler Becker	National Institute of Standards and Technology
Amber Boehnlein	Stanford Linear Accelerator Laboratory
Ronald Boisvert,	National Institute of Standards and Technology
Donald Boyce	Cornell University
Mary Brady	National Institute of Standards and Technology
Lisa Brasche	Iowa State University
Curt Breneman	Rensselaer Polytechnic Institute
Charles Campbell	University of Washington, Department of Chemistry
Carelyn Campbell	National Institute of Standards and Technology
Tof Carim	U.S. Department of Energy, Office of Science and Technology Policy
Gerbrand Ceder	Massachusetts Institute of Technology
Anne Chaka	Pacific Northwest National Laboratory
Long-Qing Chen	Penn State University
John Christensen	Robert Bosch LLC
Stephen Christensen	Boeing Research & Technology
Julie Christodoulou	Office of Naval Research
James Cotton	Boeing
Paul Crooker	Electric Power Research Institute
Thomas Devereaux	Stanford Linear Accelerator Laboratory
Andrew Dienstfrey	National Institute of Standards and Technology
Michael Fahrmann	Haynes International, Inc.
Barry Farmer	Air Force Research Laboratory

Tim Finin	University of Maryland, Baltimore County
Ian Foster	University of Chicago; Argonne National Laboratory
Stephen Freiman	Freiman Consulting
Cynthia Friend	Stanford Linear Accelerator Laboratory
Martin Fritts	National Institute of Standards and Technology
Marc Fry	Granta Design
David Furrer	Pratt & Whitney
Edwin Garcia	Purdue University
Bruce Garrett	Pacific Northwest National Laboratory
Edward Glaessgen	National Aeronautics and Space Administration, Langley Research Center
Sharon Glotzer	University of Michigan
Carlos Gonzalez	National Institute of Standards and Technology
Jeffrey Greeley	Argonne National Laboratory
Dan Gunter	Lawrence Berkeley National Laboratory
Francois Gygi	University of California, Davis
Jeff Hammond	Argonne National Laboratory
Britt Hedman	Stanford Linear Accelerator Laboratory
Scott Henry	ASM International
Daryl Hess	National Science Foundation
Michael Hill	University of California, Davis
Hanchen Huang	University of Connecticut
Kerry Hughes	The Dow Chemical Company
Warren Hunt	The Minerals, Metals & Materials Society
Patrick Hurley	Johnson Controls
Russell Irving	General Electric Company
Greg Jackson	University of Maryland
Duane Johnson	Ames Laboratory
William Joost	U.S. Department of Energy
Surya Kalidindi	Drexel University
Ursula Kattner	National Institute of Standards and Technology
Steven Kaye	Wildcat Discovery Technologies
Robert Kee	Colorado School of Mines
John Kieffer	University of Michigan
John Kitchin	Carnegie Mellon University
Gerhard Klimeck	Purdue University
William Knowlton	Boise State University
Boris Kozinsky	Robert Bosch LLC
Paul Krajewski	General Motors Company
Charles Kuehmann	QuesTek Innovations LLC
Sergey Levchenko	Fritz Haber Institute of Max Planck Society
Alexis Lewis	U.S. Naval Research Laboratory
Mei Li	Ford Motor Company
Kenny Lipkowitz	Office of Naval Research
Ping Liu	Brookhaven National Laboratory

Zi-Kui Liu	The Pennsylvania State University
Miron Livny	Wisconsin Institutes for Discovery
Benjamin Mann	Ayasdi, Inc
Radenka Maric	University of Connecticut
David Marques	Elsevier
Emmanuelle Marquis	University of Michigan
Will Marsden	Granta Design
David Martinsen	American Chemical Society
Paul Mason	Thermo-Calc Software Inc
Matthew Miller	Cornell University
Bhubaneswar Mishra	Courant Institute of Mathamatical Sciences at New York University
Amy Moll	Boise State University
Charles Moosbrugger	ASM International
Dmitri Novikov	United Technologies Research Center
Vidvuds Ozolins	University of California, Los Angeles
Clare Paul	Air Force Research Laboratory
Kristin Persson	Lawrence Berkeley National Laboratory
Tresa Pollock	University of California, Santa Barbara
Mohamed Rahmane	General Electric Company, Global Research
Krishna Rajan	Iowa State University
Jud Ready	Exponent
John Rodgers	Innovative Materials Technologies Inc.
Gregory Rohrer	Carnegie Mellon University
Robert Rudd	Lawrence Livermore National Laboratory
Joseph Salvo	General Electric Company, Global Research
Nanda Santhanam	Autodesk
Paul Saxe	Materials Design, Inc.
Sadas Shankar	Intel Corporation
Jeff Simmons	Air Force Research Laboratory
David Skinner	Lawrence Berkeley National Laboratory
Lewis Sloter	US Department of Defense
George Spanos	The Minerals, Metals & Materials Society
Ram Sriram	National Institute of Standards and Technology
Ranjan Srivastava	University of Connecticut
Alejandro Strachan	Purdue University
Veera Sundararaghavan	University of Michigan
Alexander Szalay	Johns Hopkins University
Sam Thamboo	General Electric Company
Ward Thomas	Sentient Corporation
Katsuyo Thornton	University of Michigan
Sally Tinkle	Nano Science and Technology Institute, National Nanotechnology Coordination Office
Dallas Trinkle	University of Illinois, Urbana-Champaign
Anton Van der Ven	University of Michigan

Krystyn Van Vliet	Massachusetts Institute of Technology
Gregory Voth	The University of Chicago
Mladen Vouk	North Carolina State University
Charles Ward	Air Force Research Laboratory
James Warren	National Institute of Standards and Technology
Charles Welch	U.S. Army, Engineer Research and Development Center
Terry Wong	Pratt & Whitney Rocketdyne
Ye Xu	Oak Ridge National Laboratory
Matthew Zaluzec	Ford Motor Company
Ji-Cheng Zhao	The Ohio State University
Wayne Ziegler	Army Research Lab
Frank Zok	University of California, Santa Barbara
Alex Zunger	University of Colorado, Boulder

APPENDIX B: ACRONYMS

API	application programming interface
BSE	Bethe-Salpeter equation
CALPHAD	calculation of phase diagrams
DFT	Density Functional Theory
DMRL	data and model readiness levels
DOD	U.S. Department of Defense
DOE	U.S. Department of Energy
DOI	Digital Object Identifiers®
LBNL	Lawrence Berkeley National Laboratory
MGI	Materials Genome Initiative
MII	Materials Innovation Infrastructure
MIT	Massachusetts Institute of Technology
NIST	National Institute of Standards and Technology
QMC	Quantum Monte Carlo
R&D	research and development
TAA	technical application area
TMF	thermal mechanical fatigue