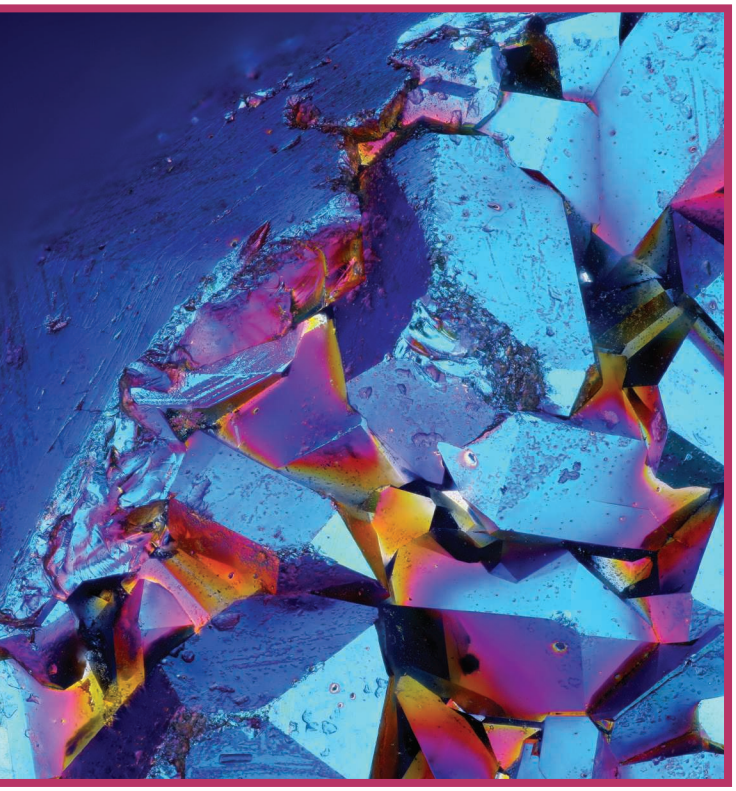
Chandler A. Becker and James A. Warren

Progress in Materials Data Availability and Application

A review



©SHUTTERSTOCK.COM/TOMATITO

Digital Object Identifier 10.1109/MSP.2021.3121563
Date of current version: 28 December 2021

Materials researchers are generating data in ever-increasing volumes, and these data are, in turn, fueling the discovery and design of new materials. These changes are enabled by advances in data capture via laboratory information management systems (LIMS), high-throughput experiment and computation, improved data collection capabilities, automated experiment selection and execution, and new analytical methods based on artificial intelligence (AI) and machine learning (ML) approaches. Data infrastructure is being developed with increasing numbers of repositories as well as data-sharing and standardization efforts.

Incorporating materials theory and knowledge is an exciting and complex area where expertise in data analytics and signal processing can make an impact in real-world applications. These combined techniques in data science and the materials domain provide many opportunities to apply sophisticated data processing and analysis approaches to materials data and its attendant design and discovery problems.

Introduction and motivation

Materials are all around us. Most human-made materials are "developed" from more basic raw materials and are a critical part of everything from aircraft and buildings to medical devices and consumer electronics. While most materials development has been done empirically (for millennia) due to the complexity of materials and their processing behaviors, there are strong motivations to incorporate newer approaches that better integrate computation, data science, and theory in the form of physical and chemical laws. Integrating these strategies is one approach to designing new materials faster, more cheaply, and with greater confidence, which is a key focus of the Materials Genome Initiative (MGI) [1]. The development of new materials is crucial to addressing a wealth of global challenges in health care, the environment and sustainable growth, energy production, and communication.

Materials research and development (MR&D) is inherently very broad in scope, heterogeneous, and cross-disciplinary. This is both an opportunity and a challenge. Numerous types of materials exist (e.g., metals, ceramics, plastics, and

semiconductors) with very different associated properties, such as strength, viscosity, elasticity, electrical resistance, and so on. The type of material data and relevant models, with associated physical underpinnings, are important considerations for use in materials design. Additionally, materials physics and chemistry have been extensively addressed by theory, so these insights can be applied to gain understanding where data may be more sparse. This feedback between experimental data and modeling also creates opportunities to contribute to the theories with new data being acquired. As part of this process, there is a need/role for expertise in signal and other sophisticated data processing approaches.

As an example to make these issues more transparent, we consider in some detail the additive manufacturing (AM) process, which is related to 3D printing. During AM, parts are often built by adding heated layers of material, generally metallic or polymeric, which then cool and solidify. A recent review of these approaches and applications can be found in [2]. While AM has recently been applied to various industrial problems of importance, including aerospace and biomedical, there are still considerable challenges to be addressed as it becomes more widely used. These include manufacturing process reliability, data acquisition, material and part characterization, validation, certification for use, and tracking of additional testing or characterization throughout the life of the material [2], [3].

In one of the common approaches to AM, laser powder bed fusion, a laser is used to raster over a freshly deposited thin layer of metal powder, locally melting and then solidifying the powder to create a new level of solid material in a part [2]. In many ways, the part can be considered a large number of welded particles. After many passes of depositing powder and scanning the laser, an entire part is created. Once it is made, there may be additional processing, such as heating, to improve the performance by making the composition, or distribution of elements, more consistent. These heat treatments can also improve the spatial distribution and sizes of pores or voids that occur when particles join and leave spaces between them. Heating can also relieve stresses that result from the materials not having time to relax fully before solidifying completely.

These steps are important since processing and structure are key to the material and part properties, which determine whether a part can be used in a particular application. Properties of interest include (among others) the so-called "melt pool" dimensions, i.e., the region where the metal has been liquified by the laser and then resolidified; local structure of the materials on the scale of micrometers; and final part geometries on the order of centimeters.

One example can be seen in Figure 1 (from [4]), which shows the local structure for one of the materials studied (IN625, a nickel-based superalloy). Nickel-base superalloys have an internal microscopic structure (microstructure) in which cuboidal precipitates of one composition are surrounded by channels of the same material with a different composition. This type of microstructure helps provide strength to the material; thus, understanding both how these structures arise and what determines their properties is essential to predicting and controlling the behavior of the final part.

Even if the same manufacturing procedures are used, there can be large variability within parts, including some that break before being completed. Models can help researchers understand why parts fail or how to optimize the manufacturing processes used, part geometries, or materials employed but only if the models are sufficiently accurate. For example, models could be used to examine processing and property relationships (e.g., how does the speed of the laser relate to the strength of the part?). Additionally, there is a need to define what certifications are necessary to approve a part for use in a critical application and how to track individual samples and parts, with associated data and metadata, throughout their lifecycles.

Generating accurate models and benchmarking them generally requires substantial amounts of experimental data. The

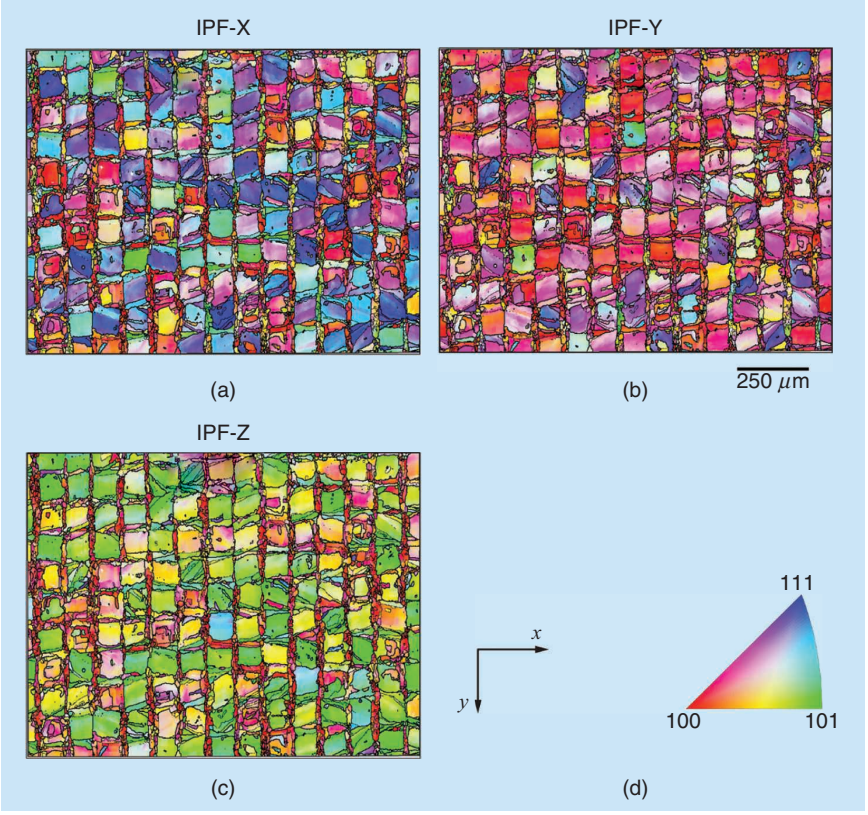


FIGURE 1. The local crystal orientations in a nickel-base superalloy used in AM as determined by electron backscatter diffraction. MR&D provides many interesting and challenging data analytics problems. IPF: inverse pole figure (X, Y, and Z referring to coordinates used for the AM build). (Source: [4], which contains additional descriptions of characterization methods and other metadata; used with permission.)

Additive Manufacturing Benchmark Test Series (AM-Bench) effort [3] is working to address these issues via highly characterized and instrumented part builds and the release of curated data (https://ambench.nist.gov and https://www.nist.gov/AMBench) in support of model validation. Samples are built with specific geometries and then characterized in detail. Characterization methods include microscopy, X-ray analysis, thermography, and others [3], [4]. This requires signals from multiple sources and instruments plus the ability to process and collate multiple sets of data that are not necessarily well aligned in time or space.

Some of these issues are discussed as related to Challenge 4 of the Air Force Research Laboratory (AFRL) AM Modeling Challenge, which addressed 3D microstructure reconstructions based on high-energy diffraction microscopy [5]. That work also highlights the importance and challenges of automation (serial sectioning) and image analysis in these systems as well as decisions about when to use various experimental techniques (destructive and nondestructive) and data fusion approaches. It also discusses challenges related to data acquisition, storage, and processing.

By making the data available in a curated form and supporting modeling challenges that make use of the data, the AM-Bench and AFRL efforts demonstrate the types of impact that the distribution of well-characterized data can have as well as where there may be opportunities for new types of analysis to further materials understanding. If models, such as those used in AM, are generated and validated, this can reduce the number of experiments required to design or optimize a new material or part by selecting those to perform that give the most information. This could be done with significant human involvement or, increasingly, by using automated approaches.

Also, models (physical and/or numerical) can be used as surrogates for experiments if they are sufficiently trusted. Finite element analysis is a good example of this at the continuum scale for designing parts—materials properties and equations are employed instead of having to create a new experiment for each small change in a design. This results in faster and more reliable materials development and use. However, the accuracies of the models are dependent on factors including the representation of the equations governing materials properties and the underlying data used to fit the models.

As an example of the use of AM-Bench data, Zhu et al. [6] developed a physics-informed neural network model that integrates momentum, mass, and energy conservation laws as part of the learning process and subsequently examined temperature profiles and the fluid dynamics of melt pools to better predict solidification behavior. ML in AM more generally is reviewed in [7] and [8] with a particular focus on the geometric design, process parameters, and in situ detection of anomalies that could adversely affect performance.

Data for materials discovery and design

The AM example described highlighted one major area of data in materials R&D, which is focused on understanding, modeling, and utilizing models and data to develop and refine materials manufacturing processes. Another major area of application is in materials discovery and design. This topic has been reviewed in more detail elsewhere, e.g., [1] and [9]–[13], and interested readers can find considerably more discussion than is covered here.

Various materials data resources are available from experimental and computational sources, and there are both specialized and general repositories to support data archiving and distribution, with examples reviewed in references such as [9] and [11]. Computational databases and repositories include the results of simulations and theory, modeling, and AI/ML. Repositories based on quantum mechanical simulations are particularly common and increasingly being used to filter very large materials parameter spaces (e.g., composition ranges) to find optimal candidates for experimental study. Many of these are based on high-throughput methods that can be used to generate large numbers of data points or records using scripts and automated approaches. The majority of them have application programming interfaces (APIs) to facilitate integration with other data sources.

Experimental sources come from a wide variety of techniques, such as scattering, microscopy, and many different characterization methods. These approaches can produce massive amounts of data, which are increasingly available to downstream researchers via databases and repositories. The available data sets and repositories have been used with varying degrees of success in materials discovery and design problems, as discussed in [1] and [10]. In particular, there are far more computational (e.g., quantum mechanical calculations) than experimental data available due to the relative difficulty of acquiring, annotating, and sharing them.

Experimentally, high-throughput methods have been growing in use as a way of increasing the volume of data collected [1], [10], [14]. Combinatorial methods, for example, can generate data on a wide range of systems in a relatively quick manner by allowing a range of material compositions to be simultaneously created and then tested. These approaches have varying degrees of human involvement, as Bayesian optimization and similar approaches can be used to guide the selection of experiments by people or robots.

Autonomous workflows are considerably expanding the amount of data collected and made available in structured forms appropriate for deeper analysis [15]–[18]. In particular, they are increasingly being adopted to optimize experimental designs and reduce the need for researchers and technicians to be involved in all steps of the design, execution, and analysis of a given effort. One example of the combination of autonomous experimentation and Bayesian optimization is in the discovery of a new material for use in photonic switching devices [15]. However, human-in-the-loop approaches allow people to provide decision-making capabilities and provide critical checks, which can give better results overall [10].

One significant challenge is in acquiring and managing scientific data, particularly when they are produced in large volumes and quickly (high velocity). This includes the basic infrastructure necessary to capture the raw or minimally processed data as well as file formats (including some that are closed and proprietary). The volume of data may be too large for storage, and, thus, preprocessing may be required. Once data have been acquired, they need to be organized and labeled with metadata that can be later used for analysis and sharing. These activities can require

considerable time and effort on the part of researchers. Backups and long-term archiving are also considerations, both technically and from a cost perspective. Some of the issues related to transmission electron microscopy in particular are discussed by Spurgeon et al. [19].

Scientific data ecosystems are emerging to more effectively address these challenges [9], [20]–[22] with supporting infrastructure. For example, LIMS efforts, such as NexusLIMS [23], support automated data and metadata capture for electron microscopy. Metadata associated with microscopy sessions are logged and made available along with the images themselves. Repositories, such as the Materials Data Facility [22] and similar, are also accessible for data distribution and archiving, and they house data of many different types. These kinds of repositories also provide a place for data supporting publications that augment the supplementary material.

Advanced computing approaches are also supporting the growth in materials data. Butler et al. [12], for example, include various listings of software packages and tools available for data analytics and ML. Exascale computing is on the horizon and will provide powerful new computing and analysis capabilities. Efforts such as SciServer are also enabling collaborative data access and analysis [24], which allows experts to interact and work together in new and exciting ways.

Increasingly, the availability of high-quality materials data is also enabling improved materials design approaches. Materials design is the creation of new or improved materials with a particular set of target properties [1], [14]. While there can be overlap with materials discovery efforts, design is really more about optimization to solve particular problems. New and optimized materials can then be integrated into new products. Data and, increasingly, computation underpin these approaches. Materials design is currently an area of great interest, including the application of AI/ML approaches. These techniques within materials science are reviewed in a number of papers, including [1], [9]–[13], and [25]. One consideration in materials research is that data collection is often difficult and expensive, so there is a particular need for methods that can work with smaller data sets [26].

Available data facilitate better analysis, including verification and validation as well as uncertainty quantification (UQ), a growing part of MR&D. This topic is treated in depth in [27], which includes discussions of epistemic and aleatoric uncertainties, that is, uncertainties stemming from imperfect knowledge and random fluctuations, respectively, as well as methods for propagating uncertainties (including the Monte Carlo analysis). In the context of ML, model error assessment is discussed in, e.g., [10]. UQ considerations are especially important when linking data and models across length and timescales, a critical component of materials design where behaviors at the macroscale (e.g., failure) can be controlled by atomic-level defects.

In addition to formal UQ, data being considered for reuse must also be assessed for physical relevance and correctness, which can be challenging in a multidisciplinary effort. For example, spurious results/predictions may be misleading when incorporated into another effort that then relies on those "wrong" results. Extrapolating from areas of reliable application is also a risk [28]. Thus, it is necessary to account for the underlying physics where possible and develop minimal standardized benchmarks for materials systems [9]. Also important is to consider how representative the data are for the system. Do they represent an edge case or something that is broadly true? The significance of these questions highlights the need for metadata to document what the data represent, how they were acquired and processed, and information to facilitate appropriate reuse.

Trends and outlook

Many of the trends related to data in MR&D are similar to other fields. These include AI/ML and the automation of computation as well as experimentation with attendant larger volumes of data [12]. Given the tremendous impact such methods can have on all of MR&D, these trends represent especially fertile ground for interdisciplinary collaboration.

Data (databases, repositories, and so on) and software are now being distributed by many projects, which increases the need for data fusion approaches. With increasing availability, there are also important considerations around data reuse. One that is especially important for MR&D is the requirement that disparate research communities must understand what data generated outside their own community represent and how to work with them. Domains often use similar words to describe different concepts, and vice versa.

This is true even within the field of MR&D, and it applies even more broadly between MR&D and other domains. As a specific example, "registry" could mean alignment (as in data sets being aligned in space) or a system for federated data cataloging and access (such as the National Institute of Standards and Technology Materials Resource Registry). Experts in different domains, even those working on the same project, may default to their own preferred use of the term and need to work together to develop common understandings. When data are used across disciplines, it becomes crucial to employ metadata to document the context and any conversions or transformations that have been applied.

The application of findable, accessible, interoperable, and reusable (FAIR) concepts [29] into systems and data design can help with this, especially via common formatting, metadata, and documentation. The Open Databases Integration for Materials Design (OPTIMADE) consortium [30] is one example of this in which common APIs are being developed to facilitate the exchange and reuse of density functional theory calculations. Acquiring data and metadata from machines in formats that can be readily used and reused supports FAIR, and well-documented APIs are critical for developing automated approaches.

While FAIR is an important approach, there are also considerations that might prevent data from being made available (e.g., proprietary or access-control concerns), which complicate efforts to manage them. Recently, the field of MR&D has been paying greater attention to all of these and developing efforts to address them. Groups, such as the Research Data Alliance (RDA), Materials RDA (MaRDA), professional societies, and similar associations, are fostering communities to establish norms and standards, including guidelines, software, and so on, for data and metadata

management. This will make it easier to utilize the data via advanced analysis techniques.

To effectively solve the challenge of a data infrastructure that integrates materials domain expertise, signal processing methodologies may be applied along with tools from other disciplines. AI/ML is being used to both accelerate existing efforts and develop insights, though the physical relevance of the results is critical in the application to MR&D. To support these goals, new platforms are being developed for curation and storage, computation, and access. Numerous approaches include working together to optimize data acquisition, analysis, and use as well as identifying commonalities in streams of data from a numerical viewpoint. However, with the growth in data volume and automation, the opportunities are greater than ever.

Authors

Chandler A. Becker (chandler.becker@nist.gov) received her Ph.D. degree in materials science and engineering from Northwestern University. She is a materials researcher in the Office of Data and Informatics, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA. Her research interests include developing the data infrastructure needed to enable reliable materials data interoperability and reuse, particularly tools and approaches for robust analysis and computational reproducibility.

James A. Warren (james.warren@nist.gov) is the director of the Materials Genome Program at the National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, 20899, USA. He cofounded and directed the NIST Center for Theoretical and Computational Materials Science and served as the leader of the Thermodynamics and Kinetics Group. In 2011, he was part of the ad hoc committee appointed by the Office of Science and Technology Policy that crafted the founding white paper launching the Materials Genome Initiative (MGI). Since then, he has been coordinating interagency efforts to achieve the goals laid out in the MGI. His research interests include developing models of materials phenomena and their associated solution techniques.

References

[1] C. Suh, C. Fare, J. A. Warren, and E. O. Pyzer-Knapp, "Evolving the materials genome: How machine learning is fueling the next generation of materials discovery," *Annu. Rev. Mater. Res.*, vol. 50, no. 1, pp. 1–25, 2020, doi: 10.1146/annurev-matsci-082019-105100.

[2] T. D. Ngo, A. Kashani, G. Imbalzano, K. T. Q. Nguyen, and D. Hui, "Additive manufacturing (3D printing): A review of materials, methods, applications and challenges," *Composites B, Eng.*, vol. 143, pp. 172–196, Jun. 2018, doi: 10.1016/j.compositesb.2018.02.012.

[3] L. Levine et al., "Outcomes and conclusions from the 2018 AM-bench measurements, challenge problems, modeling submissions, and conference," *Integr. Mater. Manuf. Innov.*, vol. 9, no. 1, pp. 1–15, Mar. 2020, doi: 10.1007/s40192-019-00164-1.

[4] M. R. Stoudt, M. E. Williams, L. E. Levine, A. Creuziger, S. A. Young, J. C. Heigel, B. M. Lane, and T. Q. Phan, "Location-specific microstructure characterization within IN625 Additive manufacturing benchmark test artifacts," *Integr. Mater. Manuf. Innov.*, vol. 9, no. 1, pp. 54–69, Mar. 2020, doi: 10.1007/s40192-020-00172-6.

[5] M. G. Chapman et al., "AFRL additive manufacturing modeling series: Challenge 4, 3D reconstruction of an IN625 high-energy diffraction microscopy sample using multi-modal serial sectioning," *Integr. Mater. Manuf. Innov.*, vol. 10, no. 2, pp. 129–141, Jun. 2021, doi: 10.1007/s40192-021-00212-9.

[6] Q. Zhu, Z. Liu, and J. Yan, "Machine learning for metal additive manufacturing: Predicting temperature and melt pool fluid dynamics using physics-informed neural networks," *Comput. Mech.*, vol. 67, no. 2, pp. 619–635, Feb. 2021, doi: 10.1007/s00466-020-01952-9.

[7] Z. Jin, Z. Zhang, K. Demir, and G. X. Gu, "Machine learning for advanced additive manufacturing," *Matter*, vol. 3, no. 5, pp. 1541–1556, Nov. 2020, doi: 10.1016/j.matt.2020.08.023.

[8] C. Wang, X. P. Tan, S. B. Tor, and C. S. Lim, "Machine learning in additive manufacturing: State-of-the-art and perspectives," *Additive Manuf.*, vol. 36, p. 101,538, Dec. 2020, doi: 10.1016/j.addma.2020.101538.

[9] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, "Data-driven materials science: Status, challenges, and perspectives," *Adv. Sci.*, vol. 6, no. 21, p. 1,900,808, Nov. 2019, doi: 10.1002/advs.201900808.

[10] D. Morgan and R. Jacobs, "Opportunities and challenges for machine learning in materials science," *Annu. Rev. Mater. Res.*, vol. 50, no. 1, pp. 71–103, Jul. 2020, doi: 10.1146/annurev-matsci-070218-010015.

[11] R. Batra, L. Song, and R. Ramprasad, "Emerging materials intelligence ecosystems propelled by machine learning," *Nature Rev. Mater.*, vol. 6, no. 8, pp. 655–678, Aug. 2021, doi: 10.1038/s41578-020-00255-y.

[12] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, Jul. 2018, doi: 10.1038/s41586-018-0337-2.

[13] S. M. Moosavi, K. M. Jablonka, and B. Smit, "The role of machine learning in the understanding and design of materials," *J. Am. Chem. Soc.*, vol. 142, no. 48, pp. 20,273–20,287, Dec. 2020, doi: 10.1021/jacs.0c09105.

[14] R. Arróyave and D. L. McDowell, "Systems approaches to materials design: Past, present, and future," *Annu. Rev. Mater. Res.*, vol. 49, no. 1, pp. 103–126, 2019, doi: 10.1146/annurev-matsci-070218-125955.

[15] A. G. Kusne et al., "On-the-fly closed-loop materials discovery via Bayesian active learning," *Nature Commun.*, vol. 11, no. 1, p. 5966, Dec. 2020, doi: 10.1038/s41467-020-19597-w.

[16] H. S. Stein and J. M. Gregoire, "Progress and prospects for accelerating materials science with automated and autonomous workflows," *Chem. Sci.*, vol. 10, no. 42, pp. 9640–9649, 2019, doi: 10.1039/C9SC03766G.

[17] N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, and G. Ceder, "Toward autonomous design and synthesis of novel inorganic materials," *Mater. Horiz.*, vol. 8, no. 8, pp. 2169–2198, 2021, doi: 10.1039/D1MH00495F.

[18] G. Pilania, "Machine learning in materials science: From explainable predictions to autonomous design," *Comput. Mater. Sci.*, vol. 193, no. 8, p. 110,360, Jun. 2021, doi: 10.1016/j.commatsci.2021.110360.

[19] S. R. Spurgeon et al., "Towards data-driven next-generation transmission electron microscopy," *Nature Mater.*, vol. 20, no. 3, pp. 274–279, Oct. 2020, doi: 10.1038/s41563-020-00833-z.

[20] N. Brandt, L. Griem, C. Herrmann, E. Schoof, G. Tosato, Y. Zhao, P. Zschumme, and M. Selzer, "Kadi4Mat: A research data infrastructure for materials science," *Data Sci. J.*, vol. 20, no. 1, p. 8, Feb. 2021, doi: 10.5334/dsj-2021-008.

[21] A. Dima et al., "Informatics infrastructure for the materials genome initiative," *JOM*, vol. 68, no. 8, pp. 2053–2064, Aug. 2016, doi: 10.1007/s11837-016-2000-4.

[22] B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, and I. Foster, "A data ecosystem to support machine learning in materials science," *MRS Commun.*, vol. 9, no. 4, pp. 1125–1133, 2019, doi: 10.1557/mrc.2019.118.

[23] J. Taillon, T. F. Bina, R. L. Plante, M. W. Newrock, G. R. Greene, and J. W. Lau, "NexusLIMS: A laboratory information management system for shared-use electron microscopy facilities," *Microsc. Microanal.*, vol. 27, no. 3, pp. 511–527, 2021, doi: 10.1017/S1431927621000222.

[24] M. Taghizadeh-Popp et al., "SciServer: A science platform for astronomy and beyond," *Astron. Comput.*, vol. 33, p. 100,412, Oct. 2020, doi: 10.1016/j.ascom.2020.100412.

[25] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Comput. Mater.*, vol. 5, no. 1, p. 83, Dec. 2019, doi: 10.1038/s41524-019-0221-0.

[26] D. E. P. Vanpoucke, O. S. J. van Knippenberg, K. Hermans, K. V. Bernaerts, and S. Mehrkanoon, "Small data materials design with machine learning: When the average model knows best," *J. Appl. Phys.*, vol. 128, no. 5, p. 054901, Aug. 2020, doi: 10.1063/5.0012285.

[27] Y. Wang and D. L. McDowell, "Uncertainty quantification in materials modeling," in *Uncertainty Quantification in Multiscale Materials Modeling*, New York, NY, USA: Elsevier, 2020, pp. 1–40.

[28] M. K. Horton, S. Dwaraknath, and K. A. Persson, "Promises and perils of computational materials databases," *Nature Comput. Sci.*, vol. 1, no. 1, pp. 3–5, Jan. 2021, doi: 10.1038/s43588-020-00016-5.

[29] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, p. 160,018, 2016, doi: 10.1038/sdata.2016.18.

[30] C. W. Andersen et al., "OPTIMADE, an API for exchanging materials data," *Scientific Data*, vol. 8, no. 1, p. 217, Mar. 2021, doi: 10.1038/s41597-021-00974-z. SP