MRS Communications (2019), *9*, 518–531
© Materials Research Society, 2019
doi:10.1557/mrc.2019.56

MRS Communications

Artificial Intelligence Prospective

A Bayesian framework for materials knowledge systems

Surya R. Kalidindi, Georgia Institute of Technology, Atlanta, GA, USA

Address all correspondence to Surya R. Kalidindi at **surya.kalidindi@me.gatech.edu**

(Received 7 January 2019; accepted 23 April 2019)

Abstract

This prospective offers a new Bayesian framework that could guide the systematic application of the emerging toolsets of machine learning in the efforts to address two of the central bottlenecks encountered in materials innovation: (i) the capture of core materials knowledge in reduced-order forms that allow one to rapidly explore the vast materials design spaces, and (ii) objective guidance in the selection of experiments or simulations needed to identify the governing physics in the materials phenomena of interest. The framework builds on recent advances in the low-dimensional representation of the statistics describing the material's hierarchical structure.

Introduction

The announcement of the Materials Genome Initiative (MGI)[1] as a national strategic initiative has spurred research explorations in completely new cross-disciplinary directions that are poised to cut across conventional disciplines such as materials science, manufacturing science, mechanics and design, and the emerging disciplines such as data sciences and informatics.[2–10] One of the important outcomes from these explorations is the increased recognition and appreciation of the critical role of data sciences and informatics, in conjunction with experimental and modeling sciences, in supporting the MGI vision for accelerated materials innovation. It is now becoming clear that materials data sciences and informatics can support the MGI vision in at least two ways. First, materials data sciences can offer novel toolsets for maximizing the re-usable knowledge mined from the large experimental and modeling datasets produced by materials researchers and specialists. This is particularly important because the past few decades have witnessed tremendous advances in both physics-based multiscale materials modeling[11] and multi-resolution multiscale materials characterization (e.g., Refs. 12–15), which have dramatically increased the volume and velocity of materials data. Since the advances in experimental and computational materials sciences are expected to further intensify in coming years, there is a critical need to support these advances with suitable data analytic tools. Only through the proper use of materials data analytic tools we will be able to systematically and comprehensively analyze very large datasets produced by materials researchers, mine the embedded knowledge in the previously aggregated datasets, and objectively guide future effort while increasing efficiency and productivity. Second, materials informatics can play a critical role in facilitating intimate and seamless communications between the diverse stakeholders of the materials innovation community. These communications deal with the exchange of high-value knowledge and expertise between the specialists in different disciplines contributing to the materials innovation effort. Examples would include the communication of core materials knowledge between materials researchers and design/manufacturing experts or the communication of critical insights and needs between materials experimentalists and materials modelers.[9,16,17] From the recent and ongoing explorations, it is becoming abundantly clear that the emergent toolsets in data sciences and informatics will need to be highly customized[7,17–20] and deployed in order to support the vision and goals of MGI, i.e., cut the cost and time incurred in advanced materials innovation and deployment efforts by half. If properly addressed, the introduction of modern materials data science and informatics tools can lead to a new generation of digital materials knowledge systems[8,9,16,17,21] and a complete digitization of materials innovation workflows. Such a digital transformation can in turn facilitate the design and implementation of fully automated and autonomous explorations for advanced materials discovery, development, and deployment.

Machine learning tools have proven invaluable in many applications of data sciences and informatics ranging from recommendation systems[22] to cancer detection[23] to self-driving cars.[24] They are now beginning to be applied to problems in materials science and engineering. Notable successes have been realized in automated extraction of materials data from published reports,[25,26] automated identification of salient features in images,[27,28] and the identification/prediction of material chemistries and internal structures with unusual/superior combinations of properties.[29–40] These recent applications have indeed clarified the potential benefits that could accrue from the insertion of data sciences and informatics toolsets

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

into the materials innovation workflows.[9,21] However, it should be recognized that the accuracy and utility of data-driven approaches such as machine learning are largely dictated by the amount of available data. The current practices in the material community have resulted in highly fragmented storage of the relevant data (often they reside in different file formats on different computers even in a single organization) that makes the sharing and aggregation challenging. Indeed, there are several ongoing efforts aimed at removing the barriers to data sharing and aggregation.[7,41,42] Parallel to these efforts, there is also a critical need to identify an overarching theoretical framework for the application of machine learning approaches to problems encountered in materials innovation. The adoption of such a framework will facilitate and support the curation of the data-driven models by the diverse stakeholders engaged in materials innovation efforts. In other words, the framework should allow multiple research groups to collaboratively develop, validate, and share the critically needed data-driven models. A templated model building workflow that is modular while allowing multiple options for each module would facilitate such a broad community-level collaboration.

A central tenet in the field of materials science and engineering is that the processing history controls the material's internal structure, which in turn controls the effective (macroscale) properties or performance characteristics exhibited by the material. The core materials knowledge needed to drive materials innovation is therefore most conveniently captured in the form of process–structure–property (PSP) linkages.[10,43–48] At the meso-length scales (from about 1 to about 100 μm), the relevant process variables include thermo-mechanical variables (e.g., temperature and time in a thermal treatment, reduction level in cold-rolling), the material internal structure variables aim to capture the salient details regarding the constituent (thermodynamic) phases (e.g., volume fractions, size and shape distributions of the different phases present, distributions of lattice orientations and/or grain sizes and/or grain shapes in case of crystalline constituents, grain and/or phase boundary character distributions), and the properties of interest reflect measures of various structural and functional performance characteristics of the material (e.g., elastic modulus, yield strength, failure strength, fracture toughness, and electrical/thermal conductivities). It has been recently demonstrated that the concept of PSP linkages can be extended to other material length scales such as the atomic length scales.[49] As a specific example, in the study of grain boundaries, the process variables might capture the grain boundary character defined in terms of the lattice misorientation and the grain boundary plane orientation, the material structure variables will reflect the salient features of the disordered/distorted atomic structure in the grain boundary region, and the excess grain boundary energy may serve as the property variable.

Systematic application of the PSP framework in materials innovation currently faces several hurdles. These include: (i) lack of a broadly adopted consistent mathematical framework for the rigorous statistical quantification of the material internal structure (this should be applicable to diverse material systems at multiple hierarchical length scales while being comprehensive enough to capture all of the salient details of the material internal structure; see the comprehensive review from Bostanabad et al.[50]), (ii) lack of a formal framework for capturing the uncertainty in the extracted PSP linkages (this is essential for making good decisions on where to invest the next effort in any specific material innovation effort), and (iii) lack of a formal framework for the fusion of information obtained from disparate sources (such as the different types of experiments and the different numerical simulation tools employed by material scientists and engineers to gain valuable insights into the material behavior at different length scales). The central goal of this prospective is to offer a new foundational mathematical framework that addresses the critical gaps identified above. The framework presented here combines the previously established PSP framework[10] with modern Bayesian approaches[51–55] in order to circumvent the hurdles listed above.

Statistical framework for PSP linkages

In a statistical framework, all of the variables (inputs, outputs, and model parameters) are treated as stochastic variables. For the current discussion, this requires us to treat the process, the structure, and the property variables as stochastic variables that can only be defined through suitable probability distributions (in other words, it is implicitly assumed that none of these can be precisely determined for any given material sample). Let $P \in \mathcal{R}$ denote a suitably defined single property of interest. Throughout this paper variables with multiple components (e.g., vector or matrix arrays) will be shown in bold font. Process variables can be defined generally as a set of real variables (in this treatment any categorical variables involved would be suitably mapped into binary integer variables). Let $\mathcal{P} \in \mathcal{R}^n$ denote a suitably defined n-dimensional process variable of interest. As already noted, the material structure variable aims to capture all of the salient features that might influence the property of interest. Let $\boldsymbol{\mu} \in \mathcal{M}$ denote a multi-dimensional statistical descriptor of the material internal structure, where each dimension represents a different feature (attribute). In this notation, $\mathcal{M}$ denotes the complete set of all theoretically possible values for $\boldsymbol{\mu}$. As such, $\mathcal{M}$ denotes an unimaginably large set.

The PSP linkages of interest are necessarily mediated by the physics controlling the underlying multiscale materials phenomena. In general, for a given material sample, the governing physics is not fully known a priori. Specification of the governing physics is usually accomplished in multiple forms. In its simplest form, this is accomplished through the specification of material-specific parameters in established physical laws (e.g., field equations expressed in differential or integral forms, material constitutive laws). As a simple example, the governing physics in elastic deformation of heterogeneous materials is expressed through the specification of a finite number of elastic stiffness parameters in Hooke's law (see Ref. 56).

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

In the more complex cases, the governing physics may be specified by introducing functions with unknown parametric forms (i.e., unknown model forms). In such cases, it would be convenient to introduce the unknown function using a suitable basis (e.g., a Fourier basis) which then reduces the specification of the governing physics to the specification of the weights (i.e., coefficients) of the selected basis. In a completely different but equivalent approach, the governing physics may be specified through Green's function-based influence kernels operating on material structure functions.[10,48,57,58] These kernels can also be expressed in digital forms (either through sampling or through the use of a suitable Fourier basis[59]). Let $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ denote the vector set of stochastic variables used to specify the governing physics in a given material sample in any of the mathematical forms described above. In this notation, $\boldsymbol{\Phi}$ denotes the corresponding complete set of all theoretically possible values for $\boldsymbol{\phi}$.

Let us consider the statistical formulation for a structure–property linkage. The goal for this linkage is to build a mathematical model that takes as input the distributions on the material structure and the governing physics (i.e., $p(\boldsymbol{\mu})$ and $p(\boldsymbol{\phi})$, respectively) and outputs the distribution on the property of interest (i.e., $p(P)$). Similarly, for a process–structure linkage, $p(\mathcal{P})$ and $p(\boldsymbol{\phi})$ will become the inputs, while $p(\boldsymbol{\mu})$ becomes the output. Let us now consider the information sources at our disposal for obtaining the data needed for establishing the PSP linkages in the statistical framework defined above. Broadly speaking, we have two general classes of information sources: (i) multiscale experiments and (ii) multiscale simulations. These two classes of information sources present very different avenues for exploring PSP linkages. In the experiments, one typically employs various materials characterization equipment that allow sampling of $p(P)$, $p(\boldsymbol{\mu})$, and $p(\mathcal{P})$. However, in practice, many limitations arise in efforts aimed at directly sampling these distributions; these are largely attributable to the limited capabilities of the currently available equipment (e.g., resolution limits, specialized skills needed for good sample preparation, inability to measure directly the quantities of primary interest) and available budget. Recent advances made in multiscale physics-based simulation toolsets [e.g., finite element (FE) methods,[60,61] phase-field methods,[62,63] dislocation dynamics,[64,65] and molecular statics/dynamics[66,67]] have raised the exciting possibility that they can contribute large datasets for harvesting the PSP linkages needed in materials innovation. In the context of the statistical framework introduced in this paper, the multiscale physics-based numerical simulation toolsets provide datasets that correspond to specified governing physics. As an example, in exploring structure–property linkages using physics-based simulation tools, one can hope to sample the distribution $p(P|\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}})$. In this notation, $\boldsymbol{\Sigma}$ denotes the covariance matrix (reduces to variance for a univariate distribution) for the variable set identified by the subscript. Therefore, in order to establish $p(P|\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}})$, one needs to specify the material structure and the governing physics along with their associated variances, and then execute a sufficiently large number of numerical simulations to sample the property distribution. Likewise, in exploring the process–structure linkages using multiscale simulation tools, one can hope to sample the distribution $p(\boldsymbol{\mu}|\mathcal{P}, \boldsymbol{\varphi}, \boldsymbol{\Sigma}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}})$.

The practical implementation of the statistical framework outlined above has encountered two major hurdles. The first major hurdle arises from the extremely high cost of experimentally sampling the distributions $p(P)$, $p(\boldsymbol{\mu})$, and $p(\mathcal{P})$. The cost becomes prohibitively high for any statistically meaningful sampling of these distributions, especially when one aims to include details of the material internal structure and their local mechanical/functional responses at the lower length scales. Similarly, even the sampling of distributions $p(P|\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}})$ and $p(\boldsymbol{\mu}|\mathcal{P}, \boldsymbol{\varphi}, \boldsymbol{\Sigma}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}})$ through the execution of a large number of numerical simulations employing physics-based models also incurs a high cost as these toolsets demand significant computational resources. The second major hurdle comes from the lack of a formal approach for synergistic fusion of the high-value knowledge embedded in the datasets obtained from the experiments and the simulations. In other words, a suitable framework does not exist currently that facilitates systematic and optimal learning of PSP linkages using datasets gathered from both types of information sources (i.e., experiments and simulations), while exploiting their relative merits.

In view of the challenges described above, it is argued by the author that a completely new formulation is needed to address the extraction, aggregation, and curation of the core materials knowledge needed to drive materials innovation. It is proposed that the new formulation should be centered around the problem of identifying the governing physics underlying the PSP linkages of interest. Recent work has demonstrated the capability of modern machine learning algorithms to be informed by known physics.[68] There is a clear need to extend these approaches to the challenges encountered in materials innovation. Specifically, the new formulation needed to guide materials innovation should be designed to engage optimally both the multiscale experiments and multiscale simulations as potential information sources. Furthermore, it would be highly beneficial to formulate the problem in a Bayesian framework as we expect that there will always be some uncertainty associated with both our assessment of the governing physics in a given material sample as well as the PSP linkages formulated based on the incomplete and uncertain understanding derived from the limited available data.

For the rest of this paper, we will assume that the distributions of interest are defined as normal distributions with the notation $\mathcal{N}(x|\bar{x}, \Sigma_x)$. In other words, the distributions of interest here are expressed as

$$\begin{aligned} p(P) = \mathcal{N}(P|\bar{P}, \Sigma_p); \quad p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\bar{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}); \\ p(\mathcal{P}) = \mathcal{N}(\mathcal{P}|\bar{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}); \quad p(\boldsymbol{\varphi}) = \mathcal{N}(\boldsymbol{\varphi}|\bar{\boldsymbol{\varphi}}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}}) \end{aligned} \tag{1}$$

where the bar on the top of a variable denotes its mean value. Indeed, $p(\boldsymbol{\varphi}) = \mathcal{N}(\boldsymbol{\varphi}|\bar{\boldsymbol{\varphi}}, \boldsymbol{\Sigma}_{\boldsymbol{\varphi}})$ specifies a prior distribution on

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

the governing physics, where $\boldsymbol{\Sigma_\phi}$ is expected to reflect a very large prior uncertainty in the governing physics of interest. Let $\boldsymbol{E}$ denotes a set of experimental observations and $\boldsymbol{\Sigma_E}$ the associated variance with these observations. As an example, when pursuing structure–property linkages in elastic deformation of a composite material, the set of experimental observations $\boldsymbol{E}$ would include data regarding a multitude of composite microstructures (a framework for rigorous quantification of the material internal structure will be discussed later) and the values for their associated elastic stiffness (a symmetric fourth-rank tensor), and $\boldsymbol{\phi}$ would comprise the unknown material parameters needed to define the elastic stiffness of the individual phases present in the material microstructure.

Using the notation established above, a Bayesian update of the governing physics can be expressed as

$$p(\boldsymbol{\varphi}|\boldsymbol{E}, \boldsymbol{\Sigma_E}) \propto p(\boldsymbol{E}|\boldsymbol{\varphi}, \boldsymbol{\Sigma_E})\, p(\boldsymbol{\varphi}) \tag{2}$$

where $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$ denotes the likelihood function reflecting the probability of the realization of $\boldsymbol{E}$ for specified values of $(\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$ and $p(\boldsymbol{\phi}|\boldsymbol{E}, \boldsymbol{\Sigma_E})$ denotes the (updated) posterior distribution on the governing physics. The formulation proposed in Eq. (2) offers many advantages compared with the current practices in the materials innovation tasks. First, it offers a systematic approach for discovering the governing physics in any selected multiscale materials phenomenon of interest that could potentially lead to autonomous explorations (discussed in more detail later). Second, given the high cost of multiscale experiments, it offers opportunities for optimal design of new experiments[53,55,69–71] aimed at maximizing information gain with each new experiment conducted. Third, and perhaps most importantly, it offers the best opportunity for the insertion of multiscale simulation tools in the materials innovation workflows. This is because the likelihood function $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$ evaluates probabilities that are conditional on specified governing physics; such an evaluation can only be conducted with physics-based simulation tools (it would be impossible to evaluate the likelihood function directly from experiments). Fourth, an analytical closed form solution can be established for Eq. (2) by utilizing normal distributions and maximization of the likelihood.[72–74] Alternately, the form of Eq. (2) is also amenable to the sampling of the posterior using Monte Carlo Markov chain-based approaches.[75,76]

The central impediment in the practical implementation of Eq. (2) in materials innovation tasks comes from the lack of viable approaches for the direct computation of the likelihood function $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$. As already noted, the only avenue available currently for sampling this distribution is through the use of physics-based multiscale simulations. However, this task has encountered several hurdles, which have stymied further progress. First, given the significant demands placed on computational resources by the multiscale materials models, it is currently impractical to evaluate the likelihood function directly through the execution of the physics-based toolsets. Second, the material internal structure ($\boldsymbol{\mu}$) needs to be included in $\boldsymbol{E}$ in a low-dimensional form for one to practically evaluate the likelihood function defined above. This presents a significant challenge as it needs to strike a balance between dimensionality of the representation and the ensuing accuracy.

Given all of the considerations presented above, the author proposes that the only practical way out of the current predicament is to first establish highly reliable and robust surrogate (i.e., reduced-order) models for the physics-based multiscale simulation tools, and subsequently employ the surrogate models for evaluating the likelihood function $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$. The emergent data science toolsets, including machine algorithms and codes, are ideally suited for addressing the many hurdles identified above in the practical implementation of a Bayesian framework for objectively guiding the materials innovation efforts. These opportunities are discussed next with illustrative examples where available.

Prospective applications of machine learning in materials innovation

Recent advances in the field of machine learning have opened new research avenues for addressing the major hurdles identified in the previous section. Selected exemplars of these new research directions covering a broad range of prospective applications in materials innovation are presented next.

Statistical quantification of material structure

As already mentioned, a central challenge in establishing the PSP linkages needed to drive materials innovation comes from the need to establish a low dimensional, but sufficiently accurate, statistical representation of the material internal structure, $\boldsymbol{\mu}$. Addressing this challenge requires a novel formalism for efficiently learning the salient features in the material structure. Furthermore, we desire the framework to exhibit versatility (allow generalization to all relevant structure/length scales in most material systems), extensibility (allow systematic incorporation of more detailed information as needed), high level of recoverability of the original microstructure when needed, broad applicability, and high efficacy in extracting the desired PSP linkages. Traditionally, simplistic parameters such as volume fraction (e.g., Refs. 77,78) and average particle/grain size/spacing (e.g., Refs. 79–81) have been used as statistical measures of the mesoscale material structure. It should be recognized that these simplistic measures of material structure do not meet most of the desired criteria listed above.

In order that the reader appreciates the difficulty involved in this task, an example of mesoscale three-dimensional material structure measured experimentally[82,83] in a polycrystalline metal sample is shown in Fig. 1. This particular representation of the discretized (i.e., voxelized) material internal structure involves $1670 \times 770 \times 200 \times 3 = 771.54 \times 10^6$ real numbers. This would be referred to as the high-dimensional representation of the material structure. It is impossible and unnecessary to use such a high-dimensional representation of the material structure in extracting practically useful PSP linkages. Indeed, a much smaller set of as yet unidentified statistical

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

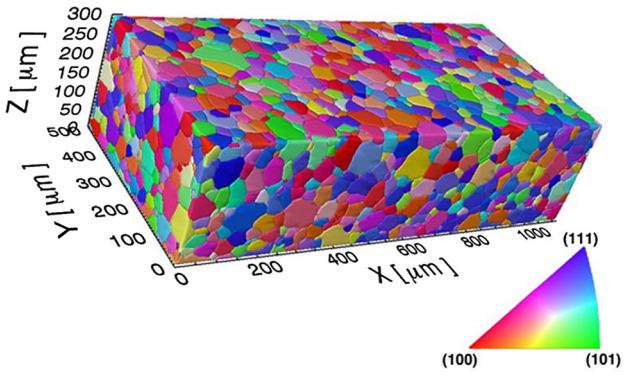


Figure 1. Mesoscale three-dimensional internal structure of beta-stabilized polycrystalline titanium containing 4300 crystals (or grains). The material structure information is represented in an array of 1670 × 770 × 200 × 3, where the first three numbers reflect the number of voxels in each of the three spatial dimensions, and the fourth reflects the number of parameters (here a set of three Euler angles) needed to specify the material state in each voxel. The color key shown in the bottom right depicts information about the Euler angles in a stereographic projection of the crystallographic orientation parallel to the *Z*-axis. Reproduced from Refs. [75,76]

descriptors of the material structure are likely to dominate the PSP linkages of interest. There is a critical need for advanced algorithms for learning these salient features. Given our interest in formulating and sharing highly versatile and extensible PSP linkages across different materials classes and application domains, it would be desirable to arrive at the low-dimensional representation of the material internal structure using unsupervised learning approaches that are not informed by the specific details of the materials processing history or the macroscale (effective) properties exhibited by the material. In other words, we should aim to formulate a consistent framework for the quantification of the material structure that can be used in formulating both the process–structure linkages as well as the structure–property linkages across multiple materials classes and application domains.

In recent years, the author's research group has advanced a framework[10] for a rigorous statistical quantification of the material internal structure that meets most of the desired criteria listed earlier. This was accomplished by bringing together the concepts of *n*-point spatial correlations and principal component analyses (PCA) for high-value low-dimensional representations of the statistics related to the material internal structure. Although the learning in this approach occurs based on the simple concept of the capture of variance in a given ensemble of material internal structures using a well-established computationally efficient algorithm (i.e., singular value decomposition), it does not negate its high utility in mining PSP linkages from experimental and simulation datasets.[35,36,38,84] It is argued by the author that the PC scores obtained in these protocols serve as highly efficient representations of $\boldsymbol{\mu}$. It has been generally observed that around 5–10 PC scores are adequate to represent the material structure with adequate fidelity.

A foundational element in the new material structure quantification framework mentioned above is the concept of a material structure function[10,48,85,86] that maps any specified combination of the material local state (defined with sufficient specificity to distinguish all of the potential local states encountered in each voxel of the material structure) and an indexed voxel to the joint probability of their co-existence. In doing so, this function takes as inputs a wide variety of descriptors of the local material state in each voxel and simply outputs a probability value that can be consistently interpreted across multiple material classes and length scales. As a specific example, for the material structure shown in Fig. 1, the relevant descriptors of the material local state would include the thermodynamic phase identifier (prescribed by a categorical variable) and the crystal lattice orientation (prescribed by a set of three continuous real numbers). If additional information at the scale of the individual voxels in this material structure becomes available (e.g., dislocation density), it can be easily appended to the local state descriptors, while the output of the expanded microstructure function maintains the same interpretation as a probability value. In this formalism, the information on grain (or phase) boundary character is automatically captured, albeit in a discretized (i.e., voxelized) manner. This is because when the phase or grain orientation changes from one voxel to the other, it implicitly captures the phase or grain boundary. Indeed, while the lattice misorientation across the boundary is captured to a higher fidelity in this formalism, the boundary plane itself is captured only to the level of accuracy allowed by the discretized representation of the material volume. However, the voxelized description of the material structure function lends itself to the use of discrete Fourier transforms in the efficient computation of the spatial correlations. This framework and the associated computational algorithms are now broadly shared through the open source code repository PyMKS.[19]

The versatility of the concept described above lies in its applicability to virtually all materials classes and all material structure/length scales. For example, in describing the material structure at the atomic length scales (could be a metal, polymer, or a ceramic), one can either use a binned local state descriptor identifying different chemical species[49] or a combination of parameters that indirectly identify the chemical species (e.g., an atomic species can be prescribed through a suitable combination of parameters such as atomic mass, atomic radius, electronegativity, etc.). In a recent application involving molecular dynamics simulated semi-crystalline polyethylene molecular structures,[87] it was found useful to aggregate blocks of atoms and label them as either crystalline or amorphous constituents based on the degree of orientation exhibited in the blocks. Spatial correlations computed on these molecular structures using the approach described above facilitated a rigorous quantification of the changes induced in the molecular structures during different stages of imposed deformation and subsequent recovery.

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

The central advantage of this new framework is that it produces an objective (unsupervised) low-dimensional representation of the material structure statistics that automatically maximizes our ability to reconstruct the statistics of the original material structure. This, of course, is a property of PCA. The reconstruction of the original material structure from the statistics, however, requires other sophisticated computational strategies.[50,88–90] Although there exist a number of other dimensionality reduction strategies (e.g., kernel PCA,[91] local linear embedding,[92] and local tangent space alignment[93]), none has been found to produce an improved low-dimensional representation of the material structure statistics for the many case studies explored thus far by our research group.

In spite of the advances summarized above, there exist several unresolved problems in the statistical quantification of the material internal structure that could benefit from further customized exploration of modern machine learning approaches. Most of the explorations using the strategies described above have been limited to two-point spatial correlations. It is now well understood that higher-order correlations (e.g., three-point spatial correlations and above) play an important role in capturing localized information[94,95] controlling failure characteristics of materials.[96,97] The techniques described above are not easily scalable to higher-order spatial correlations beyond the two-point spatial correlations. There is also the potential that we might be able to skip completely the feature engineering step in the formulation of PSP linkages using approaches such as convolution neural networks (CNNs)[98,99] or produce low-dimensional embedding for material internal structures directly using CNNs.[100]

Data-driven reduced-order PSP linkages

The task of low-dimensional representation of the material structure addressed in the previous subsection can be thought of as a pre-processing step for establishing the critically needed PSP linkages. More specifically, our interest in supporting materials innovation demands that we seek the PSP linkages with broad applicability over very large input domains. As already articulated in the previous section, the best way to insert the data from physics-based simulations into materials innovation workflows is to utilize them in evaluating the likelihood function $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$; this can only be practically pursued currently by first creating a surrogate (i.e., reduced-order) model. In other words, for each multiscale physics-based model, one would aim to establish a surrogate model mined from the results aggregated from a large number of executions of the physics-based model. In this context, the inputs to the reduced-order model would comprise all the inputs to the physics-based model (this could include the material internal structure as well as all the material parameters needed to execute the numerical simulations), and the output from the reduced-order model could be any (or all) of the outputs of interest computed by the physics-based model.

Although our ultimate goal is to evaluate the likelihood function $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$, it is much more convenient to break the problem down by first establishing a surrogate model for the deterministic predictions and then using the surrogate model for evaluating the likelihood function. As an example, in exploring structure–property linkages, our first goal will be to establish the surrogate model $\hat{P}(\boldsymbol{\mu}, \boldsymbol{\varphi})$ that closely mimics the property prediction of a physics-based model for the specified material structure and specified governing physics. In exploring process–structure models, we encounter an additional complexity because of the time variation of the material structure as a function of the process variables. In the context of data-driven models, this is best addressed as a time-series problem.[101–103] The surrogate model to be built for process–structure linkages can be expressed as $\hat{\boldsymbol{\mu}}_\tau(\mathcal{P}, \boldsymbol{\varphi}, \boldsymbol{\mu}_t, \Delta t)$, where the subscript t denotes the current time and the subscript $\tau = t + \Delta t$ denotes the time where the material structure is being predicted. In the special case of steady-state solutions that do not depend on the initial material structure, the surrogate model needed can be simplified as $\hat{\boldsymbol{\mu}}(\mathcal{P}, \boldsymbol{\varphi})$.

A large number of different options exist for the data-driven reduced-order models described above. The selection of a specific approach is largely determined by two main factors: (i) the amount of available data and (ii) a prior knowledge of the parametric model form. The simplest of the approaches is the standard regression technique based on minimization of the squares of the error, and would be the method of choice when one has a large dataset and complete knowledge of the model form. In the context of PSP linkages, a good example of the success of the standard regression technique can be found in the extraction of the localization linkages that predict the local strain field in the elastic deformation of a polycrystalline volume element of a cubic metal.[59] In this example, the model form was inspired by the best available composite theory and had over 2.5 million model parameters that were calibrated using over 1 billion training data points generated through FE simulations of candidate mesoscale polycrystalline material structures of different cubic metals. Although this involved a one-time major computational effort, the central benefit of the low-computational cost model produced in this effort is that it is applicable to virtually all polycrystalline mesoscale material structures (including all distributions of grain orientations, grain shapes, and grain sizes) of a very wide range of cubic metals (the single-crystal elastic stiffness of these materials is defined by three elastic constants often denoted as C_{11}, C_{12}, and C_{44}). Therefore, all future computations of this nature for new microstructures as well as new assignments of the cubic single-crystal stiffness parameters can be accomplished with minimal computational cost and resources. Figure 2 illustrates the predictive accuracy of the reduced-order model developed in that work.

Reduced-order model building becomes challenging when the parametric model form is unknown a priori. This is quite often the case in the formulation of PSP linkages. In such cases, one needs to consider model forms with a very large list of potential regressors and undertake either some form of feature engineering (i.e., removing the regressors with insignificant contributions or adding new ones that might exhibit

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

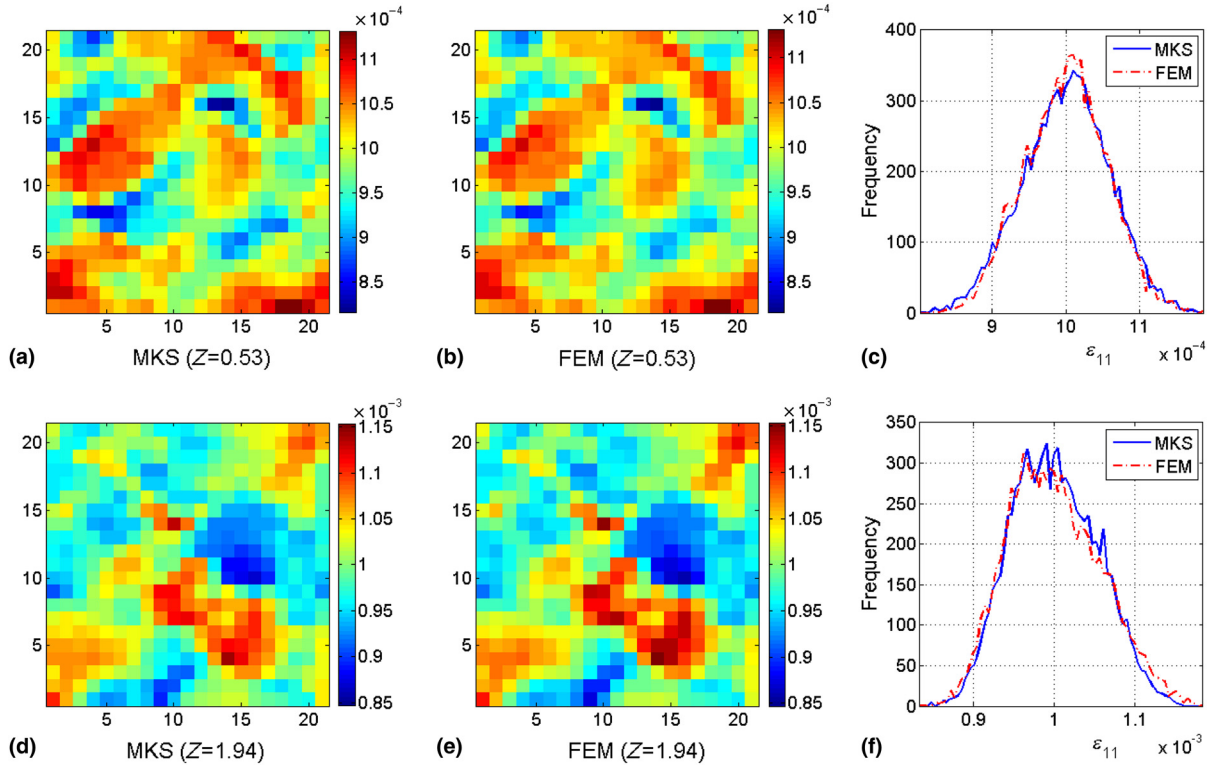


Figure 2. Illustration of the predictive accuracy of the data-driven reduced-order model (denoted as MKS) for elastic localization in mesoscale polycrystalline materials structures of cubic metals of different elastic anisotropy ratios (denoted by *Z*). Elastic strain fields in the midsections of a three-dimensional polycrystalline volume are shown. This model was produced by the standard regression technique. Reproduced from Ref. 53

significant correlations) or regularization (removing the singularity in the matrix inversion in the standard regression using techniques such as lasso[104] or ridge regression[105]) to avoid over-fits of the model. A good example of the successful application of these approaches to PSP linkages can be found in the low-computational cost surrogate models developed for the macroscale stress–strain response of two-phase composites subjected to plastic deformation.[106] In this work, the reduced-order model was built by considering a very large number of potential regressors expressed as monomials of the PC scores of the two-point spatial correlations of the mesoscale material structure (discussed in the previous subsection) and pruning this list using Pearson correlation coefficient analyses.[107] Furthermore, the overall model was designed to comprise three sub-models to allow arbitrary user-defined hardening laws for the constituent thermodynamic phases. As a result, the model is able to predict the overall stress–strain response of the composite for a large variety of material structures and user-assignments of different hardening laws to the constituent thermodynamic phases (see Fig. 3) with minimal computational cost and resources.

The example discussed above illustrates the potential benefits that could arise from the highly customized application of machine learning algorithms to formulate PSP linkages that have very broad windows of applicability. In the example presented above, the model's applicability covers the complete range of two-phase mesoscale material structures (this includes different volume fractions, size, and shape distributions) as well as a wide variety of the plastic hardening laws that could be specified by the user for each constituent phase. It is important to recognize that this level of broad applicability was accomplished by carefully injecting known physics underlying the phenomenon in various components of the overall model building effort. A simple brute-force model building effort would not have produced equivalent results.

It should be noted that the interpretability of the reduced-order models established using machine learning approaches is quite challenging. Because of their low-computational cost, one can easily probe the reduced-order models to understand the sensitivity of the different input variables in different portions of the input domain to derive some interpretability. However, the main benefit expected from the low-computational cost of these reduced-order models is that they allow rapid explorations of large design spaces (i.e., inverse solutions).

There are indeed many new avenues of research for the application of emergent machine learning approaches in the formulation of data-driven reduced-order PSP linkages. Some recent applications have highlighted the potential benefits of CNNs in formulating PSP linkages.[98,99] They circumvent

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

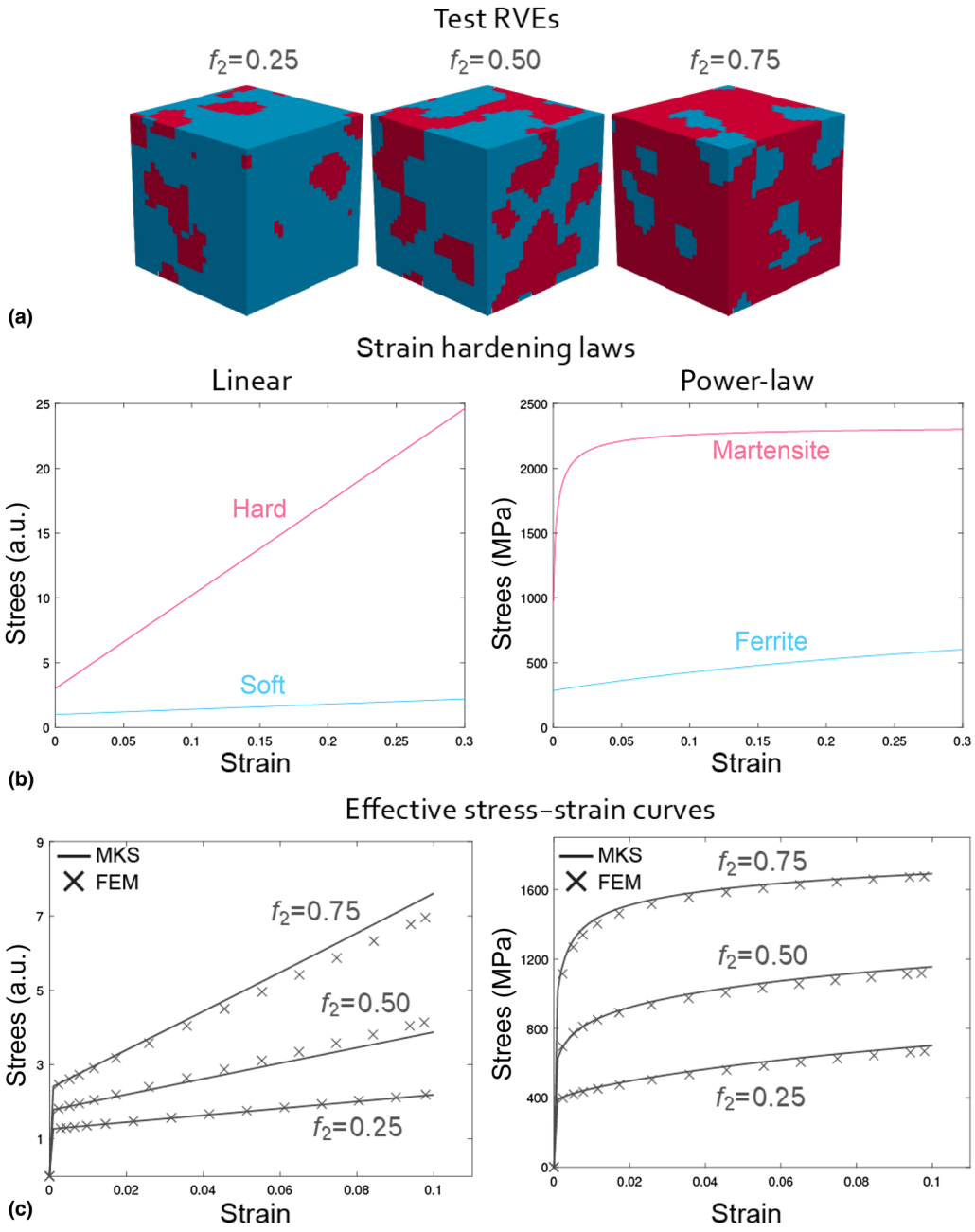


Figure 3. Illustration of the predictive capability of a low-computational cost reduced-order model developed for predicting the overall stress–strain response of a two-phase composite with arbitrary user-specified plastic hardening laws for each of the constituent thermodynamic phases. Reproduced from Ref. 96

the feature engineering step and provide easy access to a set of high-performance computational tools specifically designed for building surrogate models. Since the interpretability of the CNN models is very limited, one has to be highly vigilant in avoiding over-fits. Another area of rich opportunities is the formulation of the reduced-order process–structure linkages. As already mentioned, these are more difficult because of the time variation of the material structure. In recent work,[108] it was demonstrated that established time-series methods have potential to produce reliable and useful process–structure linkages, when combined with the concepts of the low-dimensional material structure representation discussed earlier. The examples described and reported in the literature reflect tremendous potential; there are many new research avenues to pursue before arriving at best practices for materials innovation workflows.

In addition to the machine learning approaches described above, one also needs to pay attention to the class of models based on Bayesian approaches.[51,53,109] The central advantage

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

of the Bayesian approaches is that they allow for a rigorous consideration of the uncertainty in the generated reduced-order model. It is important to note that the model error (i.e., discrepancy between the target and the prediction from the reduced-order model) is only one of the contributors to the uncertainty. Uncertainty could also arise from the inherent variance (i.e., noise) in the target value (this depends on the protocols used to generate the target value—could be either experiments or simulations) and possible over-fit of the model. Indeed, Bayesian approaches are ideally suited for situations where one desires to extract PSP linkages from relatively small datasets. Since the Bayesian approaches allow rigorous consideration of variances, they also lead more naturally to the evaluation of the likelihood function $p(\boldsymbol{E}|\boldsymbol{\phi}, \boldsymbol{\Sigma_E})$. One of the main limitations of the Bayesian approaches is that the computations generally do not scale well with either a large number of dimensions (i.e., input variables) or large datasets (i.e., large numbers of data points). That is precisely why it is important to combine the use of Bayesian approaches with the dimensionality-reduction approaches discussed earlier in order to facilitate their application to multiscale materials explorations. As for large datasets, one strategy that might work is to employ Bayesian approaches while generating and aggregating the datasets until they get sufficiently large, and then switch to approaches such as neural networks or random forests that are better at dealing with large datasets.

Among the various Bayesian approaches available for model building the most commonly used ones are Bayesian linear regression (BLR) and Gaussian process regression (GPR).[53] BLR could be the method of choice when the parametric model form is already known. One then starts with a prior distribution on the unknown model parameters, evaluates the likelihood function for the available training data, and obtains a posterior (updated) distribution on the model parameters. In many ways, this can be considered as another form of regularization and is particularly attractive for small datasets. On the other hand, GPR (also known as kriging models) may be the method of choice when building models whose parametric forms are unknown a priori. GPR essentially treats the model as a Gaussian process whose covariance is defined using specially formulated interpolating kernel functions applied on the available training data. The covariance is then used in a Bayesian framework to make model predictions for test data along with assessment of uncertainty in the predicted values. GPR is also ideally suited for building PSP linkages with small datasets. There have only been limited application of Bayesian approaches in arriving at high-value PSP linkages,[20,37] which essentially implies that there remain many different avenues for exploration in this emergent research area.

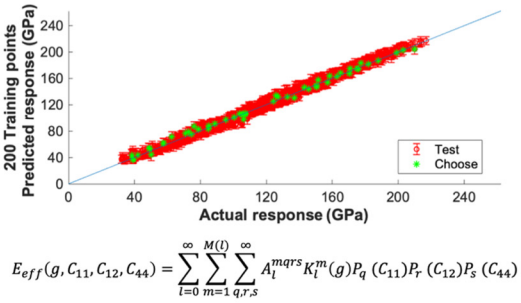
Autonomous workflows for materials innovation

The application examples described above have expounded opportunities where some preliminary investigations have already demonstrated tremendous potential. As the last example, we discuss next a prospective application area that can help bring to reality the vision of highly accelerated materials discovery, development, and deployment. This application area focuses on autonomous explorations in materials innovation. In recent work, Nikoleav et al.[110] reported successful design and implementation of highly accelerated autonomous materials innovation experiment cycles by leveraging robotics, artificial intelligence, and data sciences in order for the robot/computer to design, execute, and analyze experiments on its own. This system, called ARES (autonomous research system), demonstrated that a computer can learn and execute new experiments that result in the optimal growth of single-walled carbon nanotubes. Building on the success of ARES, it is envisioned that the scope of autonomous explorations in materials innovation tasks can be broadly expanded using the Bayesian framework presented in the previous section [see Eq. (2)].

As a first example of prospective application, consider the generation of the PSP linkages discussed in the previous subsection. In the workflows described above, it was presumed that the data needed for the extraction of the PSP linkages was already available. Often, the data generation from both experiments and physics-based simulation demands significant resources and incurs a high cost. Therefore, there is a critical need for dynamic strategies that select the new experiment (could be a physical experiment or a simulation) based on its potential for maximizing information gain in the overall effort directed at extracting the desired PSP linkages. Such dynamic strategies could help facilitate synergistic collaborations between multiple research groups. Bayesian approaches are ideally suited for such dynamic strategies because one is likely to have very small datasets as one starts to build the desired linkages. Eventually, when one aggregates a sufficiently large dataset, one might benefit from switching to other machine learning approaches.

As a specific example, consider the model building task for a structure–property linkage (the approach described here can be extended easily to a process–structure linkage). As before, let $\boldsymbol{E}$ denote a set of available data points. Employing the framework described in Eqs. (1) and (2), one would build a data-driven model, preferably using a Bayesian approach (e.g., BLR or GPR described in the previous subsection). In the context of building structure–property linkages from physics-based simulations, each data point in $\boldsymbol{E}$ would include all relevant information on a selected material structure, the imposed governing physics in the simulation, and the predicted property of interest. In such a situation, the central question that comes up is what combination of a new material structure and a new variant of the governing physics should be explored in the next physics-based simulation. Let $(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\varphi}})$ denote one such candidate combination being considered. Let $p(\tilde{P}|\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\varphi}}, \boldsymbol{E}, \boldsymbol{\Sigma_E})$ denote the predicted distribution on the property of interest using one of the Bayesian approaches (note that this is one of the main advantages of the Bayesian approaches—they predict not just the expected value but also its variance).

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press



Orientation ($\varphi_1, \Phi, \varphi_2$)	Experimental E_{eff}, GPa
339.8, 54.4, 46.1	202.9
103.7, 121.6, 49.9	199.3
232.5, 53.1, 324.0	197.7
83.2, 125.4, 30.4	195.9
3.0, 41.3, 76.4	194.0
194.7, 79.7, 317	191.1
50.0, 38.1, 250.1	190.5
114.2, 85, 173.5	173.2
170.0, 102.6, 357.9	178.3
163.6, 78.8, 168	181.2
259.9, 238.0, 145.8	189.6

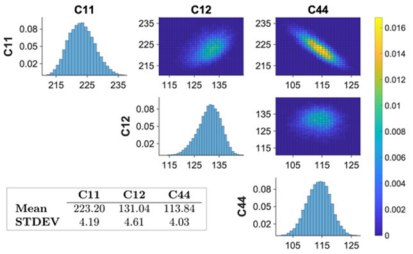


Figure 4. An illustration of the results produced by Bayesian inference to fuse the high-value information from multiscale models and multiscale measurements to estimate the parameters associated with the governing physics. In this example, the physics-based parameters to be calibrated are the single-crystal elastic constants in a polycrystalline Fe–3% Si sample. Left: Establishment of a reduced-order model predicting the effective indentation modulus as a function of user-specified values of the initial crystal orientation (where the indentation measurement might be performed) and the single-crystal elastic constants. Middle: Eleven measurements obtained from a sample.[108] Right: Posterior distributions on the unknown single-crystal elastic constants given the eleven experimental measurements.[109] The estimated distributions are in very good agreement with the expected values for these parameters.

One approach to dynamic selection of the new physics-based simulation would be to choose the specific combination of ($\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\varphi}}$) for which the variance in the predicted property is the highest as this represents the most uncertain prediction at that point in the model building effort. In other variants, one might chose the specific combination that has potential for the highest reduction in the integrated uncertainty of the predictions over the entire input domain. In other words, performing the next simulation for that specific combination is likely to yield the most valuable information for the next update of the data-driven model. One can then repeat this cycle until either the desired accuracy is attained for all the desired combinations of ($\tilde{\boldsymbol{\mu}}$, $\tilde{\boldsymbol{\varphi}}$) or until the variance shows no improvement in repeated cycles of the model building effort; the latter would imply that the best model that could be built with the available data and the selected model building approach has been built and that there is no expected benefit with further continuing the model building cycles. In the case of building structure–property models using physical experiments, one can only consider candidate material structures and the Bayesian model should be designed to provide estimates of $p(\tilde{P}|\tilde{\boldsymbol{\mu}}, \boldsymbol{E}, \boldsymbol{\Sigma_E})$. The implicit assumption in this case is that the governing physics is fixed, even though it might be unknown. Our preliminary (yet unpublished) investigations have shown that dynamic strategies based on the concepts described above can reduce the cost of the model building effort by about an order of magnitude.

The framework described above is quite versatile and is applicable to virtually all steps in the materials innovation workflow that require objective decision support. As another more sophisticated prospective application opportunity, consider the very difficult challenge of calibrating the parameters in the physics-based models to available experimental observations. This problem has been identified as one of the central bottlenecks[2,9,21,111] in the acceleration of materials innovation. This is because the physics-based multiscale models typically have a very large number of model parameters and/or alternate model form choices that need to be calibrated based on available multiscale experimental observations. In this context, it should be noted that alternate model forms can usually be expressed using a suitable basis (e.g., Fourier series) resulting in additional parameters corresponding to the weights of the basis. The data demand for systematically and rigorously exploring the calibration of physics-based models to any selected material system (defined by a combination of chemical composition and process history) is substantially large and complex. Often, one needs multimodal observations of the material phenomenon being studied using multiple experimental protocols that generally require highly specialized skills and equipment. There is also considerable uncertainty associated with the multiscale experimental observations that can be attributed to unavoidable variations in processing of the material and/or sample preparation for characterization (both structural and mechanical), resolution limits in the machines and techniques employed, fundamental limitations in currently available techniques (e.g., most microscopy techniques are only capable of characterizing the material structure on the sample surface) and operator/machine error. Therefore, it is

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

clear that one should only pursue model calibration and maturation in a suitable Bayesian framework that accounts rigorously for the uncertainty associated with each dataset.

As a conceptual illustration, we describe next how one might employ the framework presented in this paper for a specific application. Let us consider the calibration of single-crystal elastic constants (treated as the unknown physics) from a given polycrystalline sample. In recent years, spherical indentation stress–strain protocols[112–114] have been developed and demonstrated to produce highly reliable and reproducible indentation stress–strain responses in the multiple grains present in the sample. However, they do not connect directly to the values of the desired single-crystal elastic constants needed as inputs to the physics-based models. From the simulations side, while it is possible to build an FE model of the spherical indentation experiment, executing the model requires the user to guess the values of the single-crystal elastic constants. Moreover, the execution of the FE model demands significant computational resources.[115–117] As a result of these gaps and challenges, there has been a fundamental disconnect between what we can learn from the experiments and what we can learn from the numerical simulations.

In the framework presented in this paper, the problem described above would be approached in the following steps: (i) build a reduced-order model from the results aggregated by executing the FE model of spherical indentation (i.e., FE predicted values of indentation moduli) using selected values of the single-crystal elastic constants and initial grain orientations. It is important to build this reduced-order to be accurate over reasonably large subspaces of interest in the inputs (combinations of single-crystal elastic constants and initial grain orientations). This task can be accomplished using the BLR approach described earlier, as there exists prior knowledge on the parametric model form that can be used.[117] (ii) Start with a prior distribution on the unknown values of the single-crystal elastic constants and get an updated (posterior) distribution by computing the likelihood function for the available measurements of indentation moduli along with the crystal orientations where they were measured [this step needs the reduced-order model built in step (i) for efficient computation] and employing Bayesian inference techniques. (iii) Predict the distribution of indentation moduli for other grain orientations in the sample where new indentation measurements could be obtained. This is accomplished by using the posterior distribution on the unknown elastic constants together with the reduced-order model already established in step (i) from FE results. Identify the grain orientations for which the predictions exhibit the highest uncertainty and conduct new indentation measurements in those grains. (iv) Repeat the update and identify new grains for measurements. One can repeat this cycle until there is no significant improvement in the posterior distribution on the unknown single-crystal elastic constants. Figure 4 summarizes the preliminary results of one such demonstrator case study.[118,119]

Conclusions

A foundational framework has been presented for guiding objectively materials innovation efforts. This framework is based largely on Bayesian learning and is designed to be broadly applicable to the diverse range of challenges encountered in materials innovation effort. As such, it aims to capture rigorously the uncertainty involved in the extracted knowledge —a key feature to objective decision support. A number of example problems and strategies for addressing them have been discussed. It is clear that this emergent field of the application of machine learning algorithms to challenges in materials innovation offers many exciting and potentially transformative opportunities.

Acknowledgments

The author acknowledges support for this work from NIST 70NANB18H039 (Program Manager: Dr. James Warren). The author is grateful to Dave Rowenhorst (NRL), Yuksel Yabansu (GT), Marat Latypov (UCSB), and Andrew Castillo (GT) for the figures used in this paper.

References

1. Materials Genome Initiative for Global Competitiveness. National Science and Technology Council, Editor. 2011.
2. D.L. McDowell and S.R. Kalidindi: The materials innovation ecosystem: a key enabler for the Materials Genome Initiative. *MRS Bull.* **41**, 326–337 (2016).
3. M. Drosback: Materials genome initiative: advances and initiatives. *JOM* **66**, 334–335 (2014).
4. G.B. Olson and C.J. Kuehmann: Materials genomics: from CALPHAD to flight. *Scr. Mater.* **70**, 25–30 (2014).
5. J.C. Zhao: High-throughput experimental tools for the materials genome initiative. *Chin. Sci. Bull.* **59**, 1652–1661 (2014).
6. C.M. Breneman, L.C. Brinson, L.S. Schadler, B. Natarajan, M. Krein, K. Wu, L. Morkowchuk, Y. Li, H. Deng and H. Xu: Stalking the materials genome: a data-driven approach to the virtual design of nanostructured polymers. *Adv. Funct. Mater.* **23**, 5746–5752 (2013).
7. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K.A. Persson: Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
8. S. Ramakrishna, T.-Y. Zhang, W.-C. Lu, Q. Qian, J.S.C. Low, J.H.R. Yune, D.Z.L. Tan, S. Bressan, S. Sanvito and S.R. Kalidindi: Materials informatics. *J. Intell. Manuf.* 10.1007/s10845-018-1392-0 (2018).
9. S.R. Kalidindi, A.J. Medford, and D.L. McDowell: Vision for data and informatics in the future materials innovation ecosystem. *JOM* **68**, 2126–2137 (2016).
10. S.R. Kalidindi: *Hierarchical Materials Informatics* (Butterworth Heinemann, Waltham, MA, 2015).
11. P. Voorhees and G. Spanos: *Modeling Across Scales: A Roadmapping Study for Connecting Materials Models and Simulations Across Length and Time Scales*. Tech. rep. (The Minerals, Metals & Materials Society (TMS), Pittsburgh, PA, 2015).
12. E.B. Gulsoy, A.J. Shahani, J.W. Gibbs, J.L. Fife and P.W. Voorhees: Four-dimensional morphological evolution of an aluminum silicon alloy using propagation-based phase contrast X-ray tomographic microscopy. *Mater. Trans.* **55**, 161–164 (2014).
13. M.D. Uchic, M.A. Groeber, and A.D. Rollett: Automated serial sectioning methods for rapid collection of 3-D microstructure data. *JOM* **63**, 25–29 (2011).

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

14. J.F. Bingert, R.M. Suter, J. Lind, S.F. Li, R. Pokharel and C.P. Trujillo: High-energy diffraction microscopy characterization of spall damage. In Tom Proulx, Bo Song, Dan Casem and Jamie Kimberley (eds.), *Dynamic Behavior of Materials* (Springer, New York, NY, **1**, 2014), pp. 397–403.
15. U. Lienert, S.F. Li, C.M. Hefferan, J. Lind, R.M. Suter, J.V. Bernier, N.R. Barton, M.C. Brandes, M.J. Mills, M.P. Miller, B. Jakobsen and W. Pantleon: High-energy diffraction microscopy at the advanced photon source. *JOM Journal of the Minerals, Metals and Materials Society* **63**, 70–77 (2011).
16. S.R. Kalidindi, D.B. Brough, S. Li, A. Cecen, A.L. Blekh, F.Y.P. Congo and C. Campbell: Role of materials data science and informatics in accelerated materials innovation. *MRS Bull.* **41**, 596–602 (2016).
17. S.R. Kalidindi and M.D. Graef: Materials data science: current status and future outlook. *Annu. Rev. Mater. Res.* **45**, 171–193 (2015).
18. K. Rajan: Materials informatics. *Mater. Today* **8**, 38–45 (2005).
19. D.B. Brough, D. Wheeler, and S.R. Kalidindi: Materials knowledge systems in python—a data science framework for accelerated development of hierarchical materials. *Integr. Mater. Manuf. Innovation* **6**, 36–53 (2017).
20. C. Kim, A. Chandrasekaran, T.D. Huan, D. Das and R. Ramprasad: Polymer genome: a data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).
21. S.R. Kalidindi: Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. *Int. Mater. Rev.* **60**, 150–168 (2015).
22. G. Linden, B. Smith, and J. York: Amazon. com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* **7**, 76–80 (2003).
23. J.A. Cruz and D.S. Wishart: Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2**, 59–78 (2006).
24. M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L.D. Jackel, M. Monfort, U. Muller and J. Zhang: *End to end learning for self-driving cars*. arXiv preprint arXiv:1604.07316, 2016.
25. Y. Kajikawa, Y. Sugiyama, H. Mima and K. Matsushima: Causal knowledge extraction by natural language processing in material science: a case study in chemical vapor deposition. *Data Sci. J.* **5**, 108–118 (2006).
26. E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum and E. Olivetti: Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 (2017).
27. J. Nunez-Iglesias, R. Kennedy, T. Parag, J. Shi and D.B. Chklovskii: Machine learning of hierarchical clustering to segment 2D and 3D images. *PLoS One* **8**, e71715 (2013).
28. A. Chowdhury, E. Kautz, B. Yener and D. Lewis: Image driven machine learning methods for microstructure recognition. *Comput. Mater. Sci.* **123**, 176–187 (2016).
29. P. Raccuglia, K.C. Elbert, P.D.F. Adler, C. Falk, M.B. Wenny, A. Mollo, M. Zeller, S.A. Friedler, J. Schrier and A.J. Norquist: Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73 (2016).
30. B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton: Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
31. Y. Liu, T. Zhao, W. Ju and S. Shi: Materials discovery and design using machine learning. *J. Materiomics* **3**, 159–177 (2017).
32. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim: Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 54 (2017).
33. G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis and T. Lookman: Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
34. G. Pilania *et al.*: Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
35. Y.C. Yabansu, P. Steinmetz, J. Hötzer, S.R. Kalidindi and B. Nestler: Extraction of reduced-order process–structure linkages from phase-field simulations. *Acta Mater.* **124**, 182–194 (2017).
36. E. Popova, T.M. Rodgers, X. Gong, A. Cecen, J.D. Madison and S.R. Kalidindi: Process-structure linkages using a data science approach: application to simulated additive manufacturing data. *Integr. Mater. Manuf. Innovation*, **6**, 54–68 (2017).
37. A. Iskakov, Y.C. Yabansu, S. Rajagopalan, A. Kapustina and S.R. Kalidindi: Application of spherical indentation and the materials knowledge system framework to establishing microstructure-yield strength linkages from carbon steel scoops excised from high-temperature exposed components. *Acta Mater.* **144**, 758–767 (2017).
38. N.H. Paulson, M.W. Priddy, D.L. McDowell and S.R. Kalidindi: Reduced-order structure–property linkages for polycrystalline microstructures based on 2-point statistics. *Acta Mater.* **129**, 428–438 (2017).
39. M.W. Priddy, N.H. Paulson, S.R. Kalidindi and D.L. McDowell: Strategies for rapid parametric assessment of microstructure-sensitive fatigue for HCP polycrystals. *Int. J. Fatigue* **104**, 231–242 (2017).
40. H.K.D.H. Bhadeshia: Neural networks and information in materials science. *Stat. Anal. Data. Min.* **1**, 296–305 (2009).
41. A. Jain, K.A. Persson, and G. Ceder: Research update: the materials genome initiative: data sharing and the impact of collaborative ab initio databases. *APL Mater.* **4**, 053102 (2016).
42. C. Hu, C. Ouyang, J. Wu, X. Zhang and C. Zhao: NON-structured materials science data sharing based on semantic annotation. *Data Sci. J.* **8**, 52–61 (2009).
43. D.L. McDowell and G.B. Olson: Concurrent design of hierarchical materials and structures. *Sci. Model. Simul.* **15**, 207–240 (2008).
44. G.B. Olson: Pathways of *discovery designing* a new material world. *Science* **228**, 933–998 (2000).
45. G.B. Olson: Computational design of hierarchically structured materials. *Science* **277**, 1237–1242 (1997).
46. G.B. Olson: Systems design of hierarchically structured materials: advanced steels. *J. Comput. Aided Mater. Des.* **4**, 143–156 (1997).
47. D.L. McDowell, J.H. Panchal, H. -J. Choi, C.C. Seepersad, J.K. Allen and F. Mistree: *Integrated Design of Multiscale, Multifunctional Materials and Products* (Elsevier, Burlington, MA, 2009).
48. B.L. Adams, S.R. Kalidindi, and D.T. Fullwood: *Microstructure Sensitive Design for Performance Optimization* (Elsevier Science, Oxford, 2012).
49. J.A. Gomberg, A.J. Medford, and S.R. Kalidindi: Extracting knowledge from molecular mechanics simulations of grain boundaries using machine learning. *Acta Mater.* **133**(Supplement C), 100–108 (2017).
50. R. Bostanabad, Y. Zhang, X. Li, T. Kearney, L.C. Brinson, D.W. Apley, W. K. Liu and W. Chen: Computational microstructure characterization and reconstruction: review of the state-of-the-art techniques. *Prog. Mater. Sci.* **95**, 1–41 (2018).
51. P.I. Frazier and J. Wang: Bayesian optimization for materials design. In Turab Lookman, Francis J. Alexander, Krishna Rajan (eds.), *Information Science for Materials Discovery and Design* (Springer, New York, NY, 2016), pp. 45–75.
52. P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos: X-TMCMC: adaptive kriging for Bayesian inverse modeling. *Comput. Methods. Appl. Mech. Eng.* **289**, 409–428 (2015).
53. A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari and D. Rubin: *Bayesian Data Analysis*, 3rd ed. (*Chapman & Hall/CRC Texts in Statistical Science*). (Chapman and Hall/CRC, Boca Raton, FL, 2014).
54. D. Gamerman and H.F. Lopes: *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian inference* (CRC Press, New York, NY, 2006).
55. G.E. Box and G.C. Tiao: *Bayesian inference in statistical analysis* (John Wiley & Sons, 2011).
56. J.M. Whitney: *Structural Analysis of Laminated Anisotropic plates* (CRC Press, Lancaster, PA, 1987).
57. E. Kroner: Statistical modelling. In *Modelling Small Deformations of Polycrystals*, edited by J. Gittus and J. Zarka (Elsevier Science Publishers: London, 1986), pp. 229–291.
58. H. Garmestani, S. Lin, B.L. Adams and S. Ahzi: Statistical continuum theory for large plastic deformation of polycrystalline materials. *J. Mech. Phys. Solids* **49**, 589–607 (2001).
59. Y.C. Yabansu and S.R. Kalidindi: Representation and calibration of elastic localization kernels for a broad class of cubic polycrystals. *Acta Mater.* **94**, 26–35 (2015).
60. H.F. Alharbi and S.R. Kalidindi: Crystal plasticity finite element simulations using a database of discrete Fourier transforms. *Int. J. Plast.* **66**, 71–84 (2015).
61. F. Roters, P. Eisenlohr, L. Hantcherli, D.D. Tjahjanto, T.R. Bieler and D. Raabe: Overview of constitutive laws, kinematics, homogenization and

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

multiscale methods in crystal plasticity finite-element modeling: theory, experiments, applications. *Acta Mater.* **58**, 1152–1211 (2010).
62. A. Lahiri and A. Choudhury: Revisiting Jackson-Hunt calculations: unified theoretical analysis for generic multi-phase growth in a multi-component system. *Acta Mater.* **133**(Supplement C), 316–332 (2017).
63. A. Yamanaka, K. McReynolds, and P.W. Voorhees: Phase field crystal simulation of grain boundary motion, grain rotation and dislocation reactions in a BCC bicrystal. *Acta Mater.* **133**(Supplement C), 160–171 (2017).
64. A. Arsenlis and M. Tang: Simulations on the growth of dislocation density during stage 0 deformation in BCC metals. *Modelling Simul. Mater. Sci. Eng.* **11**, 251–264 (2003).
65. P. Hähner and M. Zaiser: Dislocation dynamics and work hardening of fractal dislocation cell structures. *Mater. Sci. Eng., A* **272**, 443–454 (1999).
66. S.I. Rao, D.M. Dimiduk, J.A. El-Awady, T.A. Parthasarathy, M.D. Uchic and C. Woodward: Atomistic simulations of cross-slip nucleation at screw dislocation intersections in face-centered cubic nickel. *Philos. Mag.* **89**, 3351–3369 (2009).
67. A. Leonardi and D.L. Bish: Interactions of lattice distortion fields in nano-polycrystalline materials revealed by molecular dynamics and X-ray powder diffraction. *Acta Mater.* **133**(Supplement C), 380–392 (2017).
68. L. Yang, D. Zhang, and G.E. Karniadakis: *Physics-Informed Generative Adversarial Networks for Stochastic Differential Equations*. arXiv e-prints, 2018.
69. X. Huan and Y.M. Marzouk: Simulation-based optimal Bayesian experimental design for nonlinear systems. *J. Comput. Phys.* **232**, 288–317 (2013).
70. D.J.C. MacKay: Introduction to Gaussian process. *Neural Networks and Machine Learning* 84–92 (1998).
71. C.E. Rasmussen: *Evaluation of Gaussian Processes and Other Methods for non-Linear Regression* (University of Toronto, Toronto, ON, Canada).
72. M.B. Christopher: *Pattern Recognition and Machine Learning* (Springer-Verlag, New York, 2006).
73. D.J.C. MacKay: Bayesian interpolation. *Neural Comput.* **4**, 415–447 (1992).
74. D.J.C. MacKay: *Hyperparameters: Optimize, or Integrate Out?* 1996.
75. A. Gelman: *Bayesian Data Analysis*, 2nd ed. (Chapman & Hall/CRC, Boca Raton, FL, 2004).
76. H. Haario, E. Saksman, and J. Tamminen: Componentwise adaptation for high dimensional MCMC. *Comput. Stat.* **20**, 265–273 (2005).
77. L.J Huang, L. Geng, B. Wang and LZ. Wu: Effects of volume fraction on the microstructure and tensile properties of in situ TiBw/Ti6Al4V composites with novel network microstructure. *Mater. Des.* **45**, 532–538 (2013).
78. X. Xu, S. van der Zwaag, and W. Xu: The effect of ferrite–martensite morphology on the scratch and abrasive wear behaviour of a dual phase construction steel. *Wear* **348–349**, 148–157 (2016).
79. Q. Wang, Y. Li, S. Li, R. Xiang, N. Xu and S. OuYang: Effects of critical particle size on properties and microstructure of porous purging materials. *Mater. Lett.* **197**, 48–51 (2017).
80. R. Li, R. Xin, Q. Liu, A. Chapuis, S. Liu, G. Fu and L. Zong: Effect of grain size, texture and density of precipitates on the hardness and tensile yield stress of Mg-14Gd-0.5Zr alloys. *Mater. Des.* **114**, 450–458 (2017).
81. S. Kar, T. Searles, E. Lee, G.B. Viswanathan, H.L. Fraser, J. Tiley and R. Banerjee: Modeling the tensile properties in β-processed α/β Ti alloys. *Metall. Mater. Trans. A* **37**, 559–566 (2006).
82. S.M. Qidwai, D.M. Turner, S.R. Niezgoda, A.C. Lewis, A.B. Geltmacher, D.J. Rowenhorst and S.R. Kalidindi: Estimating response of polycrystalline materials using sets of weighted statistical volume elements (WSVEs). *Acta Mater.* **60**, 5284–5299 (2012).
83. D.J. Rowenhorst, A.C. Lewis, and G. Spanos: Three-dimensional analysis of grain topology and interface curvature in a β-titanium alloy. *Acta Mater.* **58**, 5511–5519 (2010).
84. N.H. Paulson, M.W. Priddy, D.L. McDowell and S.R. Kalidindi: Data-driven reduced-order models for rank-ordering the high cycle fatigue performance of polycrystalline microstructures. *Mater. Des.* **154**, 170–183 (2018).
85. B.L. Adams, G. Xiang, and S.R. Kalidindi: Finite approximations to the second-order properties closure in single phase polycrystals. *Acta Mater.* **53**, 3563–3577 (2005).
86. D.T. Fullwood, S.R. Niezgoda, B.L. Adams and S.R. Kalidindi: Microstructure sensitive design for performance optimization. *Prog. Mater. Sci.* **55**, 477–562 (2010).
87. X. Dong *et al.*: Dependence of mechanical properties on crystal orientation of semi-crystalline polyethylene structures. *Polymer* **55**, 4248–4257 (2014).
88. D.T. Fullwood, S.R. Niezgoda, and S.R. Kalidindi: Microstructure reconstructions from 2-point statistics using phase-recovery algorithms. *Acta Mater.* **56**, 942–948 (2008).
89. D.M. Turner and S.R. Kalidindi: Statistical construction of 3-D microstructures from 2-D exemplars collected on oblique sections. *Acta Mater.* **102**, 136–148 (2016).
90. V. Sundararaghavan: Reconstruction of three-dimensional anisotropic microstructures from two-dimensional micrographs imaged on orthogonal planes. *Integr. Mater. Manuf. Innovation* **3**, 19 (2014).
91. S. Mika, B. Schölkopf, A.J. Smola, K.-R. Müller, M. Scholz and G. Rätsch: Kernel PCA and de-noising in feature spaces. In *Advances in Neural Information Processing Systems* (Massachusetts Institute of Technology, Cambridge, MA, 1999), pp. 536–542.
92. S.T. Roweis and L.K. Saul: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).
93. Z. Zhang and H. Zha: Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM J. Sci. Comput.* **26**, 313–338 (2004).
94. T. Fast and S.R. Kalidindi: Formulation and calibration of higher-order elastic localization relationships using the MKS approach. *Acta Mater.* **59**, 4595–4605 (2011).
95. D. Montes de Oca Zapiain, E. Popova, and S.R. Kalidindi: Prediction of microscale plastic strain rate fields in two-phase composites subjected to an arbitrary macroscale strain rate using the materials knowledge system framework. *Acta Mater.* **141**(Supplement C), 230–240 (2017).
96. D. Montes de Oca Zapiain, E. Popova, F. Abdeljawad, J.W. Foulk, S.R. Kalidindi and H. Lim: Reduced-order microstructure-sensitive models for damage initiation in two-phase composites. *Integr. Mater. Manuf. Innovation* **7**, 97–115 (2018).
97. N.H. Paulson, M.W. Priddy, D.L. McDowell and S.R. Kalidindi: Reduced-order microstructure-sensitive protocols to rank-order the transition fatigue resistance of polycrystalline microstructures. *Int. J. Fatigue* **119**, 1–10 (2019).
98. A. Cecen, H. Dai, Y.C. Yabansu, S.R. Kalidindi and L. Song: Material structure–property linkages using three-dimensional convolutional neural networks. *Acta Mater.* **146**, 76–84 (2018).
99. Z. Yang, Y.C. Yabansu, R. Al-Bahrani, W. Liao, A.N. Choudhary, S.R. Kalidindi and A. Agrawal: Deep learning approaches for mining structure–property linkages in high contrast composites from simulation datasets. *Comput. Mater. Sci.* **151**, 278–287 (2018).
100. N. Lubbers, T. Lookman, and K. Barros: Inferring low-dimensional microstructure representations using convolutional neural networks. *Phys. Rev. E* **96**, 052111 (2017).
101. G.E.P. Box, G.M. Jenkins, G.C. Reinsel and G.M. Ljung: *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, Hoboken, NJ, 2015).
102. P.J. Brockwell, R.A. Davis, and M.V. Calder: *Introduction to Time Series and Forecasting* (Springer, New York, NY, **2**, 2002).
103. J.D. Hamilton: *Time Series analysis* (Princeton University Press, **2**, Princeton, NJ, 1994).
104. R. Tibshirani: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)*, **58**, 267–288 (1996).
105. A.E. Hoerl and R.W. Kennard: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
106. M.I. Latypov, L.S. Toth, and S.R. Kalidindi: Materials knowledge system for nonlinear composites. *Comput. Methods. Appl. Mech. Eng.* **346**, 180–196 (2018).
107. J. Lee Rodgers and W.A. Nicewander: Thirteen ways to look at the correlation coefficient. *Am. Stat.* **42**, 59–66 (1988).

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press

108. D.B. Brough, A. Kannan, B. Haaland, D.G. Bucknall and S.R. Kalidindi: Extraction of process-structure evolution linkages from X-ray scattering measurements using dimensionality reduction and time series analysis. *Integr. Mater. Manuf. Innovation*, **6**, 147–159 (2017).
109. Q. Li, L. Gu, G. Augenbroe, C.F.J. Wu and J. Brown: *A Generic Approach to Calibrate Building Energy Models under Uncertainty Using Bayesian Inference*. In *Building Simulation Conference*. Hyderabad, India, 2015.
110. P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto and B. Maruyama: Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2**, 16031 (2016).
111. J.H. Panchal, S.R. Kalidindi, and D.L. McDowell: Key computational modeling issues in integrated computational materials engineering. *Computer-Aided Design* **45**, 4–25 (2013).
112. S. Pathak and S.R. Kalidindi: Spherical nanoindentation stress–strain curves. *Mater. Sci., Eng. R., Rep.* **91**, 1–36 (2015).
113. J.S. Weaver and S.R. Kalidindi: Mechanical characterization of Ti-6Al-4V titanium alloy at multiple length scales using spherical indentation stress–strain measurements. *Mater. Des.* **111**, 463–472 (2016).
114. A. Khosravani, L. Morsdorf, C.C. Tasan and S.R. Kalidindi: Multiresolution mechanical characterization of hierarchical materials: spherical nanoindentation on martensitic Fe-Ni-C steels. *Acta Mater.* **153**, 257–269 (2018).
115. D. Patel and S. Kalidindi: Estimating the slip resistance from spherical nanoindentation and orientation measurements in polycrystalline samples of cubic metals. *Int. J. Plast.* **92**, 19 (2017).
116. D.K. Patel and S.R. Kalidindi: Correlation of spherical nanoindentation stress–strain curves to simple compression stress–strain curves for elastic-plastic isotropic materials using finite element models. *Acta Mater.* **112**, 295–302 (2016).
117. D.K. Patel, H.F. Al-Harbi, and S.R. Kalidindi: Extracting single-crystal elastic constants from polycrystalline samples using spherical nanoindentation and orientation measurements. *Acta Mater.* **79**, 108–116 (2014).
118. S. Pathak, D. Stojakovic, and S.R. Kalidindi: Measurement of the local mechanical properties in polycrystalline samples using spherical nanoindentation and orientation imaging microscopy. *Acta Mater.* **57**, 3020–3028 (2009).
119. A. Castillo and S.R. Kalidindi: Accelerated extraction of crystal level elastic parameters via Bayesian framework. *Front. Mater.* (2019), submitted.

https://doi.org/10.1557/mrc.2019.56 Published online by Cambridge University Press