Digital Twins for Materials

Surya R. Kalidindi[1*], Michael Buzzy[1], Brad L. Boyce[2], Remi Dingreville[2]

[1]Georgia Institute of Technology, Atlanta, GA, USA
[2] Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, NM, USA

*** Correspondence:**
Surya R. Kalidindi
surya.kalidindi@me.gatech.edu

Keywords: Digital Twins, Materials Knowledge Systems, Artificial Intelligence, Computational Materials Science, **Machine Learning.**

Abstract

Digital twins are emerging as powerful tools for supporting innovation as well as optimizing the in-service performance of a broad range of complex physical machines, devices, and components. A digital twin is generally designed to provide accurate in-silico representation of the form (i.e., appearance) and the functional response of a specified (unique) physical twin. This paper offers a new perspective on how the emerging concept of digital twins could be applied to accelerate materials innovation efforts. Specifically, it is argued that the material itself can be considered as a highly complex multiscale physical system whose form (i.e., details of the material structure over a hierarchy of material length) and function (i.e., response to external stimuli typically characterized through suitably defined material properties) can be captured suitably in a digital twin. Accordingly, the digital twin can represent the evolution of structure, process, and performance of the material over time, with regard to both process history and in-service environment. This paper establishes the foundational concepts and frameworks needed to formulate and continuously update both the form and function of the digital twin of a selected material physical twin. The form of the proposed material digital twin can be captured effectively using the broadly applicable framework of n-point spatial correlations, while its function at the different length scales can be captured using homogenization and localization process-structure-property (PSP) surrogate models calibrated to collections of available experimental and physics-based simulation data.

1 Introduction

Recent forward-looking roadmaps (Gil and Selman, 2019; Jenks et al., 2012) have identified the development of a fully digital framework that fuses human-subject matter expertise, process and performance modeling, experimental in-situ diagnostics, and data science algorithms as one of the most important areas to transform manufacturing and surveillance of components throughput their life cycle. Indeed, the digitization of product lifecycle management (PLM) has led to the emergence and deployment of digital threads (Niederer et al., 2021; Kapteyn et al., 2021; Zeb et al., 2021) in a broad spectrum of manufacturing industries. These digital threads collect, curate, and archive all of the data/information generated from all stages of the product life cycle: conceptualization, design, prototype, manufacturing, operation, and retirement (Singh and Willcox, 2018; Margaria and Schieweck, 2019). Digital threads open multiple new avenues for fostering innovation and improving the in-service performance of a wide range of products. A necessary feature of the digital threads is that they encompass both the in-silico activities (e.g., model-based or virtual engineering) and the

physical activities (e.g., measurements made during the different stages of manufacturing, testing, and operation of the product) conducted in the PLM. An important outcome from the deployment of digital threads is that they have opened new opportunities for the creation and use of in-silico analogues to the physical product. The recent advances in digital and sensor technologies (Ullo and Sinha, 2020; Mei et al., 2019) enable the in-silico objects to co-exist along with their physical counterparts. In addition to mimicking the physical products, the in-silico analogues offer unprecedented potential for consistent change management, allowing the optimization of intentional or unintentional product evolution over time. Therefore, within this context, a digital twin can be defined as a high-fidelity in-silico representation closely mirroring the form (i.e., appearance) and the functional response of a specified (unique) physical twin. Digital twins have thus far been used in the manufacturing and performance evaluation of complex engineered physical systems (e.g., turbine engines) (Zaccaria et al., 2018; Raj and Surianarayanan, 2020; Tao et al., 2018; Lim et al., 2021; Xie et al., 2021) and/or their components, where the focus has been largely on capturing accurately the macroscale geometry and the component-level performance metrics. Current digital twins do not address adequately the capture and archival of the materials data, which typically deals with physical phenomena occurring at the lower material length scales (typically ranging from the atomic to the macroscale). This disconnect is not surprising given the siloed nature of current materials research and product design/manufacturing communities. However, it is abundantly clear that a successful extension of digital twins to include the materials data/information in a comprehensive manner can allow for a holistic co-design of material, manufacturing process, and product in fully integrated innovation cycles, possibly resulting in dramatic improvements in the overall part performance.

Materials, in their own right, represent highly complex multiscale and multi-physics systems. Their production and in-service responses are controlled by a wide range of phenomena occurring at length scales ranging from the atomic to the macroscale and an equally broad range of associated time scales. Figure 1 depicts schematically the hierarchical nature of materials systems with examples of a wide variety of physical phenomena that occur at the nano- and micro-scales. Clearly, the materials phenomena occurring at the lower material length scales play important roles in controlling the macro- and component-scale performances of the part. In the current research paradigm, the considerations at the component/part scale and the material scale are studied in a mostly de-coupled manner by different groups of researchers. The former are the domain of mechanical designers and manufacturing specialists, while the latter are addressed by materials science and engineering specialists. More specifically, the field of materials science and engineering focuses on understanding how the different processing histories (e.g., thermo-mechanical processing) influence the material structure (includes information on the many aspects of order and disorder seen at different length scales cf. Figure 1) and their associated properties. However, understanding and quantifying accurately the underlying process-structure-property (PSP) relationships (McDowell and LeSar, 2016; Kalidindi, 2015) at the different material length and time scales is quite arduous. This is mainly because the diverse physical phenomena occurring at these scales are necessarily related and co-dependent with one another. Therefore, adopting a systems approach that manages the complex trade-offs between potentially conflicting multifunctional requirements at the different length scales spanning across the complete range of material and product scales would yield significant benefits.

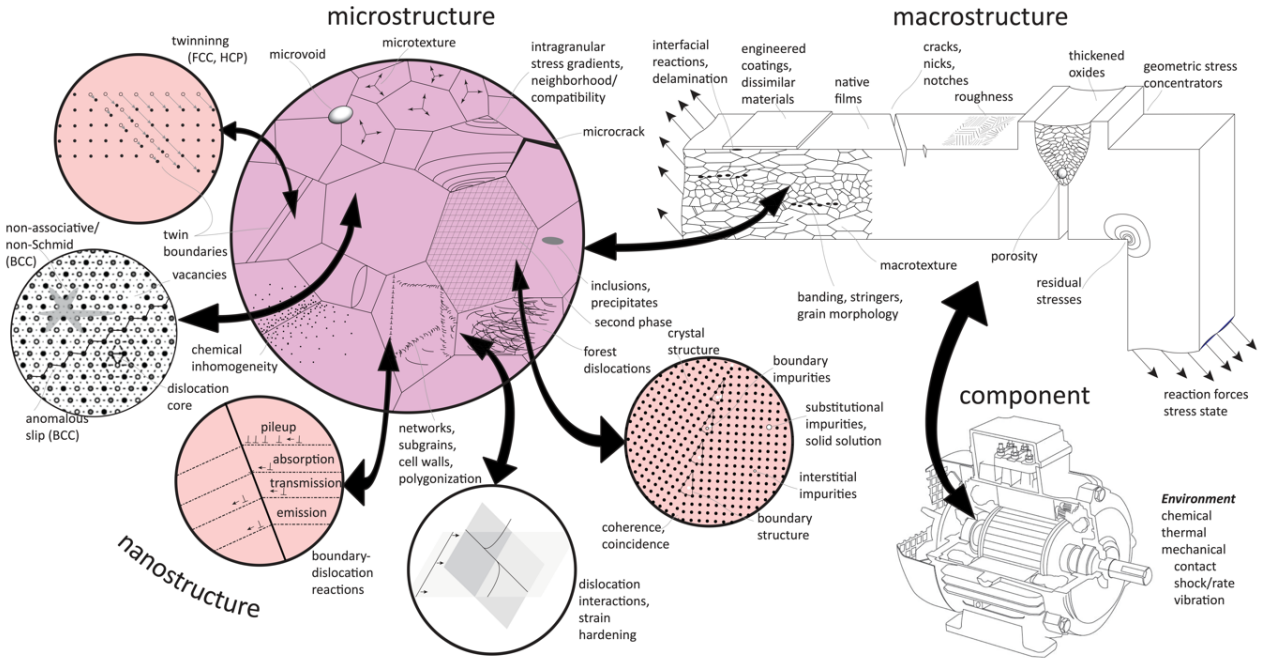

Figure 1. A schematic depiction of the multiscale and multi-physics nature of material systems and their relationship with the component performance. A comprehensive understanding of material performance requires a complete hierarchical representation of structural/chemical features, the relationship between those features and material properties, and the mechanisms that drive their evolution either through processing or service history. All arrows represent scale bridging, i.e., upscaling via homogenization and downscaling via localization.

However, this task faces many hurdles. The most significant of these hurdles comes from the fact that the relevant data, even for a selected single material system, is necessarily generated by distributed teams of researchers with the requisite expertise. For example, on the experimental front, materials data comes from a wide range of imaging modalities (e.g., optical microscopy, scanning and transmission electron microscopy, various diffraction and spectroscopic techniques, X-ray tomography, atomic force microscopy) (Polonsky and Pandey, 2021; Belianinov et al., 2018) and property evaluations (e.g., mechanical testing in different modes and at different spatial resolutions, thermal conductivity, diffusivities) (Khosravani et al., 2020: Khosravani et al., 2021). On the modelling front, the data comes from an equally disparate set of sources that aim to faithfully simulate specific selected sub-phenomena at different material length scales (e.g., density functional theory computations, molecular dynamics, discrete dislocation dynamics, kinetic Monte-Carlo simulations, cellular automata, phase-field simulations, finite element models) (Horstemeyer, 2009; Matouš et al., 2017; Panchal et al., 2013). Although each individual dataset often provides a partial insight, only a systems approach can provide the comprehensive holistic view needed to objectively drive materials innovation in an accelerated manner; this is indeed the goal of many national and

international materials research initiatives (e.g., ICME (Allison et al., 2006), MGI (de Pablo et al., 2019; National Science and Technology Council, 2011)).

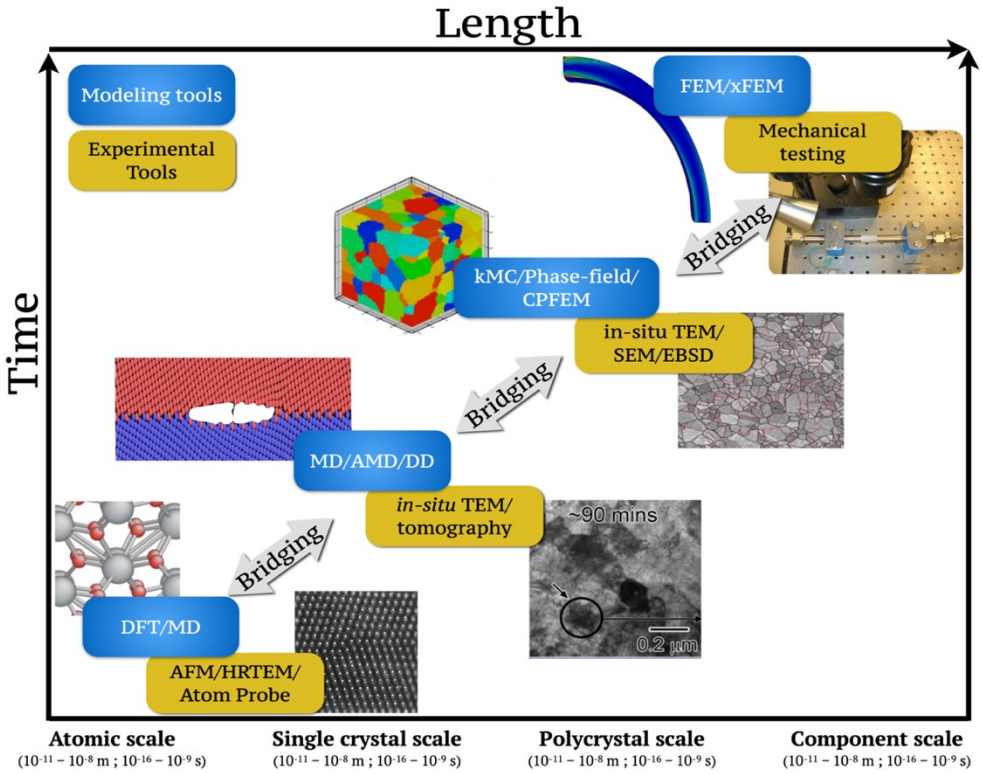

Figure 2. Modeling and experimental tools typically used to obtain relevant materials data at different length and time scales. Example of modeling tools used include Density Functional Theory (DFT), Molecular Dynamics (MD), Accelerated MD (AMD), Dislocation Dynamics (DD), kinetic Monte Carlo (kMC), Crystal Plasticity Finite Element Modeling (CPFEM), FEM, and extended FEM (xFEM). Examples of experimental tools used include Atomic Force Microscopy (AFM), High Resolution Transmission Microscopy (HRTEM), *in situ* TEM, tomography, Scanning Electron Microscopy (SEM), Electron Backscattered Diffraction (EBSD), and mechanical testing.

Figure 2 illustrates the large variety of data sources involved in formulating a systems approach to understanding and optimizing materials for desired combinations of macroscale (effective) properties. As already noted, the datasets collected from any one data source (refers to either a single experimental protocol or a single physics-based simulation tool) often provides incomplete and uncertain insights into the physics controlling the materials phenomena of interest. At a high level, it should be recognized that physics-based simulations are designed to provide predictions of the material response to imposed (thermo-mechanical) environments for user-specified physics. On the other hand, experiments are generally designed to provide observations of material response to specific imposed environments, for as yet unknown (or uncertain) materials physics. Clearly, all individual datasets (from any individual data source) should be treated as being incomplete and/or uncertain. However, if the insights from the datasets collected from the different data sources can be effectively fused in a consistent framework, it is likely to produce much more comprehensive and valuable insights. Currently, there does not yet exist an overarching mathematical framework for such data fusion. The development and utilization of such a framework is likely to open new avenues for major time and effort savings in materials-product co-design and innovation efforts by optimally guiding the effort investment (i.e., objectively identifying the next best steps based on a rigorous statistical analyses of all previously aggregated data).

As already noted, the perspectives presented above build on multiple national and international initiatives. Specifically, ICME (Allison et al., 2006), and MGI (de Pablo et al., 2019)

have articulated the need for increased use of computational tools and data sciences (including artificial intelligence/machine learning toolsets (AI/ML)) to accelerate the rate of materials discovery, development, and deployment in advanced technologies. Indeed, much progress has been made in organizing and disseminating materials data (TMS, 2017), and physics-based simulation toolsets (TMS, 2015). There has also been a strong injection of data sciences and AI/ML into materials research, especially in aspects related to data ingestion (e.g., experimental laboratory automation) (Kalidindi et al., 2019), curation (e.g., ontologies) (Morgado et al. 2020; Voigt and Kalidindi, 2021), feature engineering (Xiang et al., 2021; Kalidindi, 2020), and automated generation of surrogate models (Generale and Kalidindi, 2021; Marshal and Kalidindi, 2021). These recent advances in materials research have set the stage for the extension and application of the emerging concept of digital twins described earlier to include the multiscale details of the material. This paper establishes a roadmap for the pursuit of this goal, i.e., the extension of digital twins to include materials data over a hierarchy of length scales. Specifically, we identify the key foundational elements that currently exist and outline the gaps that need to be overcome for success in this endeavor.

2 Main Elements of Digital Twins for Materials Systems

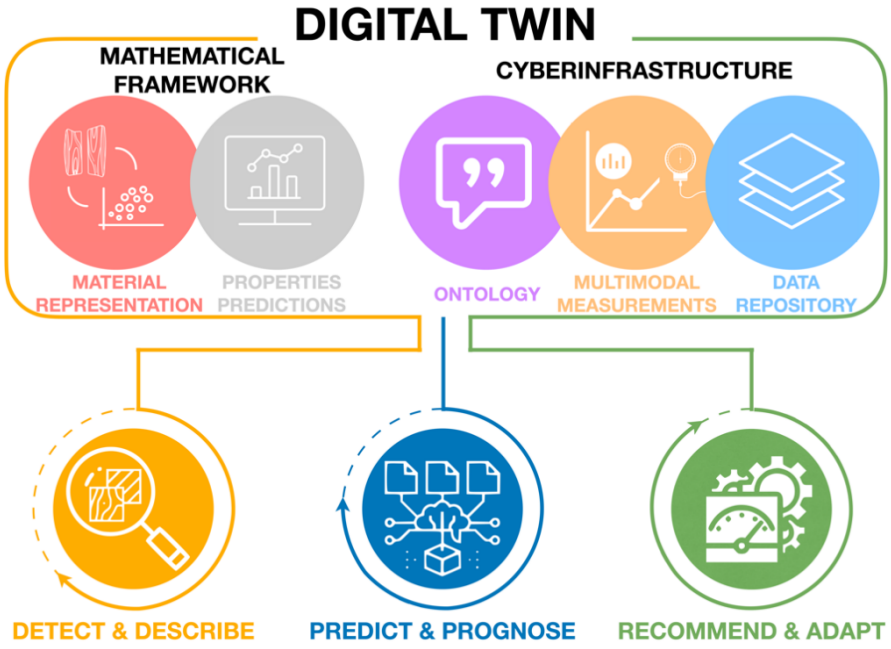

Figure 3. The main components of the proposed roadmap for building digital twins for material systems.

2.1 Physical Twin of a Material System

Digital twins of macroscale engineered components and machines typically aim to represent a uniquely identified single physical twin. For example, a digital twin might target a specific turbine engine in service on an airplane. However, in building digital twins for a material system, it becomes intractable to consider each individual material sample as the physical twin. This is not only because of the large number of distinct material samples that can be produced for a nominally specified chemical composition and processing history, but also due to the fact that non-destructive characterization techniques are not yet mature for evaluating both the three-dimensional structure of the material as well as its properties of interest. Furthermore, even with the use of destructive techniques for materials characterization, one can only hope to establish distributions that adequately quantify the material structure and properties in a stochastic framework (i.e., accounting for the

significant uncertainty associated with these quantities for any given material sample). Given these considerations, it is readily apparent that the digital twins for materials systems can only be established in a stochastic framework at the present time. In other words, we propose here that digital twins of materials systems should aim to produce multiple instantiations (as many as needed) sampled from the distributions of the possible material structure and their associated properties (with both structure and properties defined over a hierarchy of material length scales). Therefore, in our proposed framework, we will associate the digital twins of the material system to the nominal chemical composition and processing/service history that created the material samples. In doing so, we will implicitly define the material by the controllable details (each of which is identified with aleatoric uncertainty) of the generative process used to create the material samples (i.e., instantiations of the physical twin). This, we believe, will result in a much more pragmatic approach to building digital twins for material systems that will have high value for the design and in-service prognosis of engineered components and devices.

2.2 Mathematical Framework for Digital Twins of Material Systems

The mathematical framework underpinning the digital twins for material systems should address two main needs: (i) the statistical quantification of the material structure over a hierarchy of material length scales[1] and its suitable representation in practically useful low-dimensional forms, and (ii) the reliable prediction of the material properties of interest given information about the material structure and the processing/service history. These tasks indeed correspond to defining the form and the function of the digital twins for material systems. As already noted, both these tasks need to be addressed in a stochastic framework that rigorously tracks the uncertainty associated with all of the available data and propagates it into the predictions of the material properties.

2.2.1 Material Structure Representation and Quantification

The term *material structure* is used here to describe the spatial arrangement of structural and chemical heterogeneities, which constitute a material at a specified instant of time and govern its properties at that instant of time. For a given chemical composition, thermodynamics predicts an equilibrium crystallographic phase (or a multiphase mixture), and at finite temperature, an equilibrium vacancy concentration. Yet materials are rarely in their thermodynamic ground state. Essentially, an overwhelming subset of the material structural features represent metastable or unstable defects created throughout the process history. Conventionally, material structure defects have been classified based on their dimensionality as planar grain boundaries, linear dislocations, and point-wise atomic impurities; these are but the simplest examples of a myriad of complex microstructural features (see Figure 1). The material structure is not usually static but evolves when stimulated by exposure to energy (thermal, mechanical, chemical, etc.). Through state-of-the-art processing, the most perfect undoped, isotopically pure silicon single crystals have been produced to purity levels of >99.9999%. On the other hand, the most sophisticated structural alloys benefit from their complex, multiscale arrangement of the lower length scale structural features, reminiscent of the hierarchical nature of biological systems. Hence, the challenge for a digital twin of a material system is to represent the necessary complexity of the inherently high-dimensional material structure features with sufficient fidelity to capture the relevant subset that controls the material response of interest. Complicating matters, no single experimental technique is capable of comprehensively digitizing the material's complete internal structure.

[1]In PSP linkages, one associates a material structure to an instant of time. The structure is then assumed to be responsible completely for the properties exhibited by the sample. In any imposed process, the structure is assumed to evolve with time. When the structure evolves, its associated properties are also expected to evolve.

A digital twin of a material system should be able to instantiate a representative volume of the material with sufficient statistical sampling of all the relevant lower length-scale structural features and their spatial arrangements. Given the roughly eight orders of magnitude in length scales (from ~Å to ~cms) involved, it should become clear that such instantiations cannot be deterministic or unique. Therefore, what is required here is the ability to produce multiple instantiations that reflect as accurately as possible the inherent stochasticity of the material structure for a given nominal composition and process history. Laplace conjectured that by knowing every atom, its position and momentum, we could anticipate the behavior of the material (marquis de Laplace, 1814). While this statement reflects accurately the expected causal relationship between the material structure and its associated properties, it reflects a practically impossible pursuit. Therefore, we take the viewpoint that the digital twin of a material system is intended to be a minimally sufficient reduced-order representation of Laplace's "demon". A tractable digital twin of a material system should therefore utilize a versatile (broadly applicable to all material classes and length scales) low-dimensional representation of the material structure that would allow efficient learning of the functional response of the material system. Based on the earlier discussion, it is also clear that the low-dimensional representation of the material structure can only reflect suitably defined *statistical measures* at different material length scales; henceforth, such salient statistical measures of the material structure will be referred as *features*. Because of our interest in instantiating the material structure in our digital twins, it is important that the selected feature set should produce realistic, sufficiently accurate, instantiations of the material structure that can be subsequently correlated with its functional response. This is not a trivial requirement. For example, most of the conventionally used statistical measures of the material structure, such as the overall alloy composition, phase volume fractions, and the averaged grain sizes are woefully inadequate for producing the required instantiations of the multiscale material structure for our digital twins. More advanced approaches involving a richer set of microstructure statistics (e.g., orientation and mis-orientation distributions, grain aspect ratio distributions) have led to concepts such as statistically equivalent representative volume elements (Ghosh and Groeber, 2020; McDowell et. al., 2011). Some of these concepts have also been implemented in open-source codes such as DREAM.3D (Ghosh and Groeber, 2020; Groeber and Jackson, 2014).

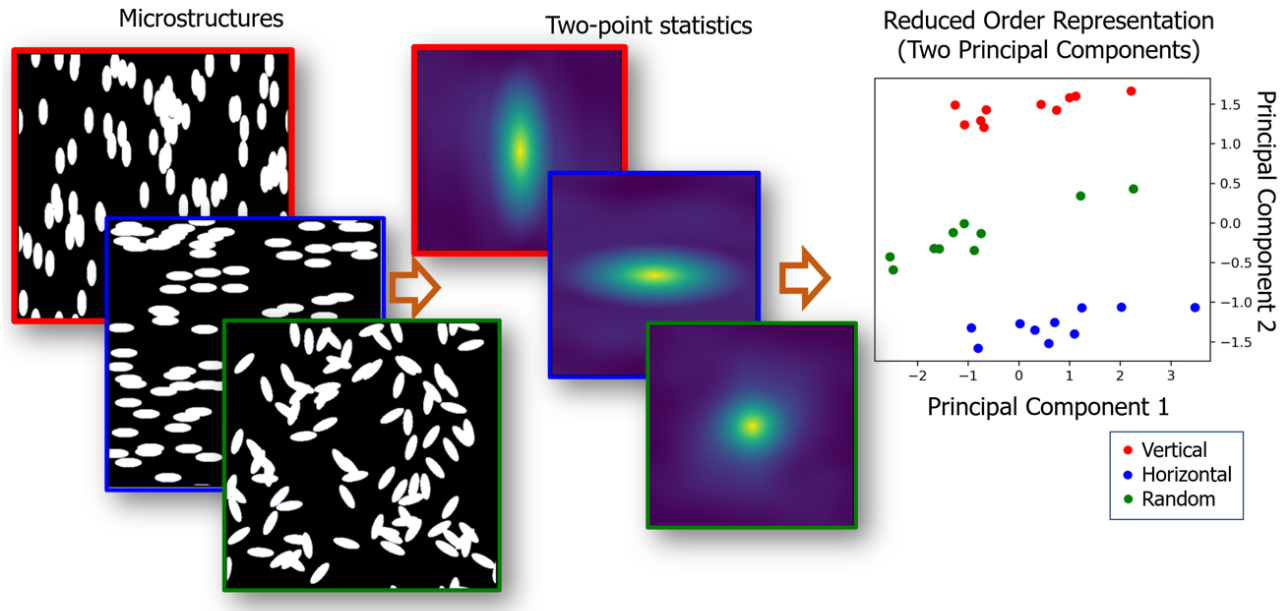

Figure 4. The MKS workflow for feature engineering of material structure. In this example, we start with microstructures belonging to three distinct classes (corresponding to vertical, horizontal, or random ellipses), with

one example of each class shown on the left. Their corresponding 2-point features are shown in the middle and reflect a large number of statistics (including volume fractions, size and shape distributions) for each microstructure. The low-dimensional representations of the microstructure statistics are shown on the right, in the subspace of the first two PC scores. The clusters in the PC plots successfully classify the microstructures in the three classes. The intra-class variance between microstructures within each class can also be quantified from the PC representations.

A comprehensive and systematic framework available today that is capable of providing the requisite feature engineering capabilities for the material structure is the framework of n-point spatial correlations (Adams et al., 2021; Torquato, 2002; Fullwood et al., 2010; Niezgoda et al., 2013; Niezgoda et al., 2011; Torquato and Stell, 1982). In recent work, Kalidindi and co-workers (e.g., Kalidindi, 2015) have developed and demonstrated an efficient and broadly applicable computational framework and toolsets for addressing this task. Broadly referred as Materials Knowledge Systems (MKS), this framework takes advantage of the computational efficiency of voxelated representations and Fast Fourier Transform (FFT) algorithms to implement the theoretical framework of n-point spatial correlations. The feasibility and benefits of this approach have been demonstrated on a wide variety of material classes and material structures at different length scales (from the atomic (Gomberg et al., 2017; Kaundinya, 2021) to dislocation length scales (Robertson and Kalidindi, 2021a) to microscale (Generale and Kalidindi, 2021)).

At its core, MKS defines and utilizes a material structure function (Kalidindi, 2015) that maps a selected combination of spatial position $\boldsymbol{x} \in \Omega$ (the physical volume of the material domain) and a local material state $\boldsymbol{h} \in$ H (includes information such as phase identifiers, chemical compositions, lattice orientations, defect densities) to suitable measures (e.g., density) that reflect the intensity of $\boldsymbol{h}$ at $\boldsymbol{x}$. Mathematically, one can express this function as $m(\boldsymbol{h}, \boldsymbol{x})$. Implicit in this definition is the expectation that H needs to be identified suitably to capture the complete set of material states of interest at the different material structure length scales. Features of the material structure can then be defined as expectations of suitably scaled moments of $m(\boldsymbol{h}, \boldsymbol{x})$. For example, the expected value of $m(\boldsymbol{h}, \boldsymbol{x})$ over Ω can provide a set of 1-point features that can be interpreted as the volume fractions of $\boldsymbol{h}$ in Ω (Kalidindi, 2015). Similarly, the expected value of $m(\boldsymbol{h}, \boldsymbol{x})m(\boldsymbol{h}', \boldsymbol{x} + \boldsymbol{r})$ over Ω can produce a set of 2-point features that can be interpreted as the joint probability of realizing $\boldsymbol{h}$ at $\boldsymbol{x}$ and $\boldsymbol{h}'$ at $\boldsymbol{x} + \boldsymbol{r}$, where $\boldsymbol{r}$ denotes a specified vector separating the two spatial points randomly selected from Ω. Although, one can define higher-order features (e.g., 3-point features), one often finds a sufficiently large number of features in the 2-point feature set, as it includes all permutations of $(\boldsymbol{h}, \boldsymbol{h}')$ over a very large domain of $\boldsymbol{r}$ (this domain includes all distinct set of all vectors of interest that can be placed in Ω). The adequacy of the set of 2-point features in capturing the salient features of the material structure (including the set of features identified in conventional practices in materials science and engineering) has been established for a broad range of material classes (Generale and Kalidindi, 2021; Latypov et al., 2019) as well as the different structure length scales (Kaundinya, 2021; Robertson and Kalidindi, 2021a; Fullwood et al, 2010) encountered in them.

The MKS framework described in Figure 4 produces a very large number of features, even when using only the 2-point feature set. For extracting a low-dimensional feature set, one needs to employ a suitable dimensionality reduction technique. Of the various options for this task, principal component analysis (PCA) has been found to be particularly attractive. First, it allows for an unsupervised learning of the salient low-dimensional features based on maximization of captured variance. Therefore, it identifies a consistent set of features that can be used across multiple PSP surrogate models, allowing for full interoperability among collections of such models. In other words, since the salient features are identified without the knowledge of the specific targets (i.e., outputs) of the surrogate model, they can be used for different targets (for example, in the predictions of very different properties of a given material system). Second, the PCA basis can be inspected and interpreted to a limited extent, allowing for the low-dimensional features to be associated with some

(limited) physical meaning. Third, since PCA essentially involves a rotational transform of the original space, it preserves distances between datapoints in the original space. Finally, the orthogonal decomposition involved in the PCA allows for practically useful reconstructions of the full feature list, i.e., a reconstruction of the high-dimensional feature list from the low-dimensional feature list. Of course, these reconstructions are approximate because of PC truncation. However, since the PC representations are maximized to capture variance, it is possible to make sure that the approximation introduced by the truncation is within acceptable tolerance. The PC scores obtained from the application of PCA on an ensemble of 2-point feature sets (one set corresponds to one material structure) serve as a highly effective low-dimensional feature set for the material structure in our digital twins. There exist a multitude of other options for dimensionality reduction of the feature space, such as isomap or kernel PCA. However, the nonlinear embeddings employed in these techniques can introduce distortions into the latent space that negate the benefits of PCA identified above (Hu et al., 2022).

As stated earlier in Section 2.1, the physical twin is not defined as a single instantiation of a material structure, but rather as the outcome of a stochastic generative process that yields instantiations that we then observe. The MKS framework described above provides a mathematically compact representation using computationally efficient tools. However, many tools (e.g., phase-field simulations, micromechanical finite element models) only take specific instantiations of the material structure as inputs. Therefore, successful creation of digital twins for materials requires the ability to move between statistical representations of material structure and their three-dimensional physical instantiations at low computational cost. While the computation of 2-point spatial correlations from instantiations is relatively easy (Cecen et al., 2016), the inverse computation is not trivial. Very recently, it has been shown that the three-dimensional material structures can be instantiated from their 2-point feature sets with minimal computational cost (Robertson and Kalidindi, 2021b). As a result of the many advantages described above, the MKS framework along with its open-source code repository PyMKS (Brough et al. 2017) offers a powerful, currently available, toolset for addressing the challenges of building digital twins of materials systems.

It is also noted that there are a number of other options based on neural networks that allow one to combine feature engineering and property prediction into a single-step framework. These approaches offer attractive avenues when one is interested in a limited number of properties as targets. If one insists on de-coupling the form and function of the digital twins (as we have argued here), then it is imperative to pursue feature engineering of the material structure independently from establishing property predictors (discussed in the next section). In this context, it should be recognized that the autoencoder-decoder networks (Herr et al., 2019) offer an interesting option. These networks do address the unsupervised feature engineering of the material structure. Therefore, the features identified from such networks can then be input into other neural networks for property predictions. This idea represents an open research avenue that merits further exploration.

2.2.2 Predictions of Material Properties

Reliable prediction of the effective properties of a given material sample is a challenging task. At a high level, the main options are to either measure experimentally the properties of interest or to leverage known physics (often delivered in physics-based simulation packages) to estimate their values. Both approaches face hurdles when one desires to produce a multiscale, digital twin for materials. On the experimental front, the effort and cost involved in measuring all of the properties of interest along with the related information (e.g., anisotropy, variances) over the multiple material length scales of interest are often prohibitive. On the modelling front, there is substantial uncertainty in the model forms and/or parameter values used in the physics-based models. It is therefore clear that neither approach by itself is optimal in getting us the requisite information. In this regard, the

recent emergence and successful application of materials data analytics tools has opened up new avenues for addressing these gaps.

Recently (Kalidindi, 2015; Kalidindi, 2020), it has been argued that process-structure-property (PSP) linkages can be defined over different material-structure length scales to capture the core knowledge needed to study multiscale material responses. It is argued here that the same PSP linkages can be utilized to predict the functional response of the material digital twin. This is because the PSP linkages can be used to update both the changes in the multiscale material structure due to the imposed service conditions (using suitably defined process-structure evolution linkages) but also their associated properties (using structure-property linkages). The required PSP linkages need to be formulated using available data that might often be disparate, incomplete and/or uncertain. Most importantly, the framework for predicting the function of the material digital twin should allow easy (and possibly frequent) updates as new data becomes available. It is also likely that one needs to chain together multiple PSP models in order to make the predictions of the function of the material digital twin.

A Bayesian framework has the potential to address scale-bridging with uncertain physics. The proposed Bayesian framework will be described next using the structure-property (SP) linkages as an example. However, they will be formulated such that they can also be easily applied to capturing process-structure linkages (PS). Typically, SP linkages are formulated to take structure variables as inputs and predict property values as output. The mapping implied in these linkages can be expressed as $P = \mathcal{F}(\boldsymbol{\mu})$, where P is a property variable and $\boldsymbol{\mu}$ denotes a vector of structure features (e.g., the PC scores of the 2-point feature set described in Section 2.2.1). Both P and $\boldsymbol{\mu}$ should be treated as stochastic variables. This naïve definition makes the governing physics implicit in the formulation of $\mathcal{F}$. It would be much more desirable for SP linkages to explicitly treat the governing physics as additional input variables to the mapping, i.e., to refine the desired mapping as $P = \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\varphi})$,where $\boldsymbol{\varphi}$ denotes the governing physics. In practice, $\boldsymbol{\varphi}$ would represent a vector of parameters defining the physical mechanisms controlling the response of the material physical twin (e.g., parameters used in constitutive modeling of the material response). This refinement is advantageous in two ways. Firstly, it allows one to treat $\boldsymbol{\varphi}$ as a stochastic vector variable, which often exhibits a significant amount of uncertainty for a selected material physical twin. Secondly, it allows for the uncertain physics to be passed between linkages. This is particularly useful for the multiscale phenomenon that occur in material systems, as the uncertain physics learned in one length scale can still be utilized at another length scale. An example of this scale-bridging is depicted schematically in Figure 5. The first linkage estimates the indentation yield strength (effective property) of a single grain given the grain's orientation (structure variable) and critically resolved shear strengths (physics variables). The second linkage estimates the bulk yield strength (property) given the two-point statistics of the grain orientations (structure variables) and the same critically resolved shear strengths (physics variables). Since the physics variables in these two linkages are the same, the uncertain knowledge of the physics variables extracted in the grain-scale data (could come from experiments and/or simulations) can be upscaled and utilized in making predictions of the effective properties at the polycrystal scale.

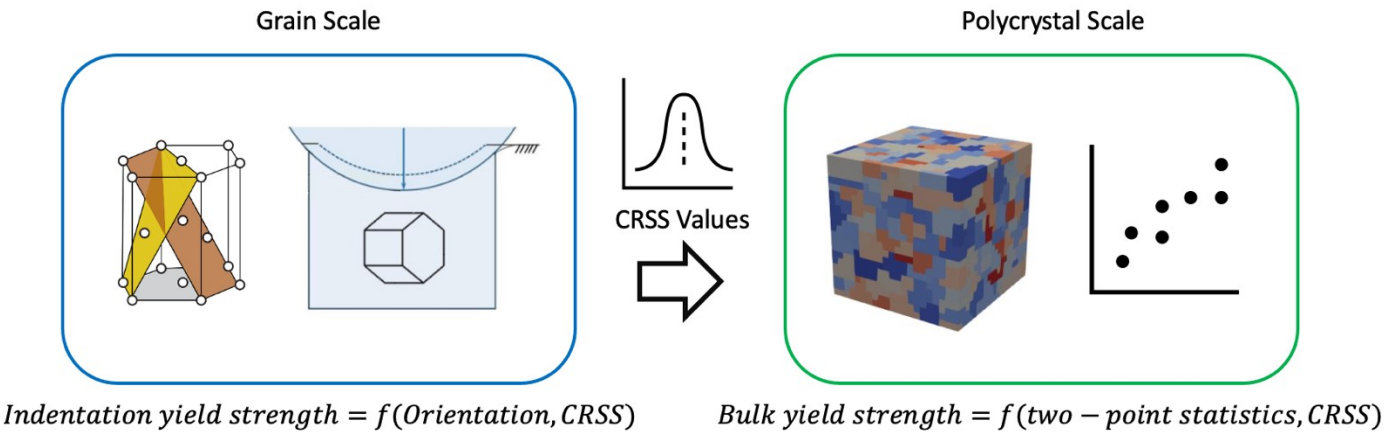

Figure 5. Schematic illustration of the scale-bridging between the response of an individual grain and the response of a polycrystalline aggregate. At the grain scale, the structure-property linkage is formulated to take grain orientation (structure variable) and critical resolved shear strength(s) (CRSS; physics variables) as input and predict the overall property of interest (e.g., indentation yield strength of grains of different orientations). This linkage is used with both experimental and modeling datasets to extract a posterior on the CRSS for a given material system (see Figure 6 for more details). At the polycrystal scale, the structure-property linkage is formulated to take the 2-point spatial correlations of orientation (structure variables) and CRSS (physics variables) to make a prediction of the bulk (effective) yield strength of the polycrystal (c.f., Paulson et al., 2017).

In establishing the material physics parameters, one has to exploit all of the available data, collected from disparate sources (e.g., experiments and physics-based simulations). Machine learning of $\boldsymbol{\varphi}$ for a selected material physical twin can be accomplished using a Bayesian update strategy expressed as:

$$p(\boldsymbol{\varphi}|\boldsymbol{E}) \propto p(\boldsymbol{E}|\boldsymbol{\varphi})\, p(\boldsymbol{\varphi}) \quad (1)$$

where $\boldsymbol{E}$ denotes the set of available experimental observations, $p(\boldsymbol{\varphi})$ is the prior (reflecting our best initial guess), $p(\boldsymbol{E}|\boldsymbol{\varphi})$ is the likelihood of realizing the observations in $\boldsymbol{E}$, and $p(\boldsymbol{\varphi}|\boldsymbol{E})$ denotes the updated posterior on $\boldsymbol{\varphi}$. Although Eq. (1) looks very simple, its practical usage for learning the controlling physics in multiscale material phenomena has been hindered by several factors. First, only the physics-based simulation tools that faithfully mimic the experiments performed to obtain $\boldsymbol{E}$ can allow for the computation of the likelihood term in Eq. (1). This is because only these tools allow arbitrary specification of the governing physics $\boldsymbol{\varphi}$. However, a brute-force application of physics-based tools for computing the likelihood is prohibitively expensive because of the extremely large number of simulations one needs to perform to accomplish this task since it entails performing simulations covering a large domain of likely governing physics for all of the experimental observations in $\boldsymbol{E}$. Second, the proportionality in Eq. (1) implies that one needs to develop and implement a suitable strategy for establishing the proportionality factor. Recent work (Castillo et al., 2021) has demonstrated that it is possible to train AI/ML models on simulation results produced by physics-based models, which can then be used to compute the likelihood term in Eq. (1). Furthermore, they would also allow for the implementation of Markov-Chain Metropolis-Hastings (MCMH) approaches for sampling the posterior in Eq. (1) without explicitly computing the proportionality factor. It is important to note that the posterior estimate of $\boldsymbol{\varphi}$ is not restricted to come from any single source of data. As an example, let us consider the situation where the data becomes available from different test modes (these could be indentation tests and micro-pillar tests for grain-scale mechanical measurements). In such situations, we need to establish different surrogate models for each test mode. Let $P_1 = \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\varphi})$ and $P_2 = \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\varphi})$ represent such surrogate models. Since the underlying microstructure and physics variables have the exact same meanings in both models, one can use both models with their respective experimental datasets for sampling a consistent posterior

for $\boldsymbol{\varphi}$. Once the posterior is established, one can establish the desired SP linkage in a stochastic framework by marginalizing as:

$$P(\boldsymbol{\mu}|\boldsymbol{E}) = \int \mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\boldsymbol{E}) d\boldsymbol{\varphi} \quad (2)$$

As noted above, the practical implementation of Eqs. (1) and (2) needs the establishment of suitable AI/ML surrogates. These usually take the form $\mathcal{F}(\boldsymbol{\mu}, \boldsymbol{\varphi})$, and can be accomplished using Gaussian Process Regression (GPR). The central advantage of the proposed strategy here is that the formulation of the needed AI/ML models is generally a one-time effort. In other words, when these are properly designed to cover large input domains in the space of the controlling physical parameters and the space of relevant material structures, they only need to be performed once (examples can be seen in prior work (Castillo et al. 2021; Castillo and Kalidindi, 2020)). This feature allows for a relatively low-computational cost update of the surrogate model as new experimental data becomes available. It is also possible to suggest new experiments that maximize the potential for improving the accuracy of the predictions (i.e., reducing the prediction uncertainty). This is most efficiently accomplished using established concepts of information gain such as the posterior predictive variance (Castillo and Kalidindi, 2019; Castillo and Kalidindi, 2020; Castillo et al., 2020), expected improvement (Ghoreishi et al., 2019; Solomou et al., 2018; Takhtaganov and Müller, 2018; Talapatra et al., 2018), and expected information gain (Pandita et al., 2019).

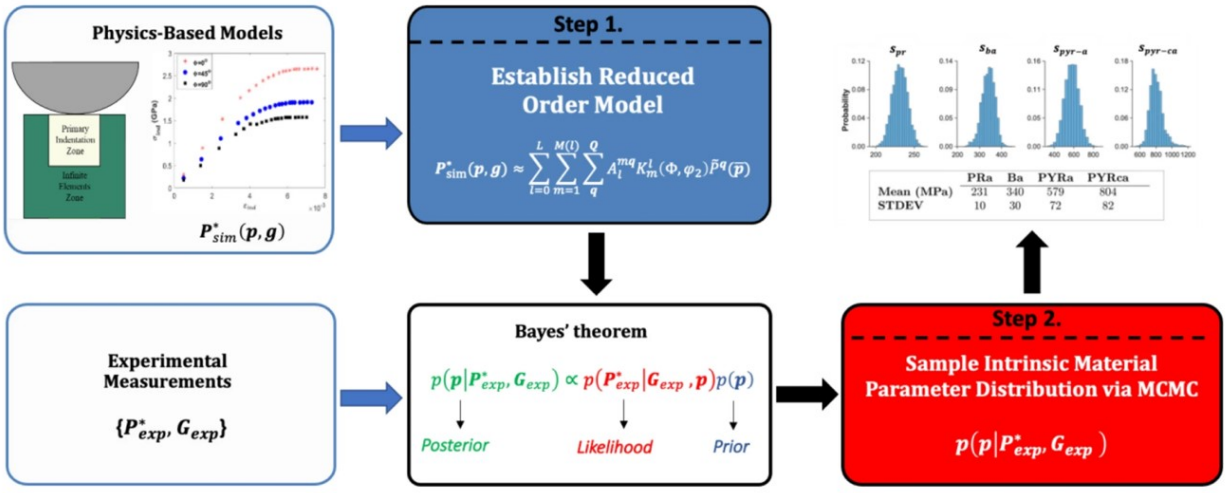

Figure 6. An example application of the Bayesian update strategy for the fusion of experimental and simulation datasets from indentation of a-Ti grains in a polycrystalline sample (Castillo et al. 2021).

An example application of the proposed Bayesian approach methodology is depicted in Figure 6, taken from the work of Castillo et al. (2021). In this example, the information from spherical indentation measurements on individual grains in a polycrystalline sample and the corresponding simulations using crystal plasticity finite element models are combined to establish distributions on the unknown values of the critical resolved shear strengths of four different families of potentially active slip systems in a selected Ti alloy. The approach described in this study resulted in at least one order of magnitude savings in both the overall cost and effort expended, when compared to the conventional approaches that employed small-scale testing to obtain the same information.

2.3 Cyberinfrastructure for digital twins of materials

Cyberinfrastructure supports the acquisition, storage, management, and fusion of data within a collaborative, but distributed, research environment. The creation of a robust cyberinfrastructure is critical to the realization of a digital twin, as digital twins exist at the confluence of multiple disparate data streams (e.g., simulation data, experimental data, real time sensor data). These data streams

present challenges in managing both the variety and volume of data ingested, as well as any associated metadata needed to ensure high utility of the data for future use. Challenges in the variety of data come from the *multimodal* nature of materials data*,* meaning that the data in question stems from a variety of data sources (e.g., different imaging or analysis modalities). For example, materials data can take many forms: scalar parameters (e.g., diffraction line profile), time series (e.g., fatigue response), and spatially resolved (2-D and 3-D) image data (e.g., SEM image, tomography scan), and each modality is accompanied by its own unique forms of metadata that describe pre-process, in-process, and post-process information. Challenges in the volume of data stem from advancements in acquisition resolution and high-throughput experimental capabilities (hyperspectral imaging, x-ray computed tomography, etc.). For example, it is now commonplace to collect a large ensemble of images with high spatial resolution at a high frame rate using a variety of microscopes (e.g., optical, scanning electron, transmission electron), producing gigabytes-to-terabytes of observations of a single material (Dingreville et al. 2016). Similarly, expanded computational resources and multiscale modeling capabilities can also generate large amounts of data related to a material's response to variety of environments (de Oca Zapiain et al., 2021). The main challenge lies in collecting and curating this large collection of heterogeneous data into the high-value *information* needed for the creation of a digital twin.

2.3.1 Data Sources

Material structure measurements capture the *state* of the material before, during, and after evolution, and material property measurements quantify various *characteristics* of evolution (e.g., resistance to evolution, evolution rates). The constellation of methods used to measure material structure and properties is extensive, and here we only mention two general trends. First, the digital data stream is becoming more entrenched in the instruments used to measure material properties. Just a generation ago, material structures were documented on film and quantification was performed by manual measurements; lab instruments utilized strip-chart recorders that created an analog graphical representation of the data. Now, not only have data streams become digitized, but increasingly, the data collection instruments are networked and remotely accessible. Yet significant concerns remain regarding the cyber vulnerability of both the data and the instrument, and institutional regulations regarding interconnectivity are highly disparate. Second, with the continuing advances in measurement sensors, data transfer, and data storage, the data streams are becoming increasingly dense, requiring thoughtful strategies for intelligent data reduction. Additionally, unconventional datasets, collected with alternative low-cost methods are proving to have utility. Previous trends in measurement science have focused on increases in precision and accuracy of data. Now, the focus is shifting to affordable high-density data streams that can provide similar or complementary information content to the existing suite of ultra-precise measurements.

The external stimuli (e.g., thermo-chemo-mechanical loading) driving material structure evolution need to be tracked through the use of suitable sensors. Sensors generally transduce various forms of energy (Table 1) into electrical signals that can be transformed into digital data. The transduction can also involve intermediate forms of energy, e.g., magnetic or optical. All forms of sensing have limits in resolution, range, accuracy, and precision. The fidelity of the digitized resolution of the external stimulus captured by the sensor is limited by the accuracy of the correlation of the electrical signal to the intensity of the imposed stimulus, and the bit-depth of the stored information. The fidelity of an environmental measurement can also be limited by the temporal and spatial resolution of the sensor. Sensor arrays allow for spatial mapping of a field (e.g., temperature field on a sample surface) of interest, with the spatial resolution limited by the spacing between individual sensors in the array. Alternately, one can acquire such information using a single sensor and rapidly scanning a region of interest; this strategy will lead to some degree of temporal disregistry between individual measurements.

Table 1. Example of energy forms that drive changes in material state and the transducers employed to observe the corresponding exposure history.

Stimuli	Application Examples	Sensor Examples
Mechanical	Vibration, Shock, Sound/Phonon, Stress, Strain	Strain gauges, piezoelectric, magnetostrictive, eddy current, accelerometer, capacitive.
Electrical	Current, magnetic fields	Voltage sensors, current sensors, resistance sensors, power sensors hall-effect sensors, giant magnetoresistance sensors, fluxgate sensor
Radiant Energy	Gamma, X-ray, UV, Infrared, Visible light, Microwave, Radio waves,	Photoresistors (LDR), photodiodes, phototransistors, charged-coupled devices, gamma ray detectors, microwave sensors, CMOS detector
Thermal	Convective, conductive, latent	Thermocouples, RTDs, Thermistors, infrared, semiconductor sensors
Chemical	Gases, liquids, solids, ions, isotopes, etc.	Hygrometer, gas sensor, pH sensor,
Nuclear	Neutron, Beta, Alpha, Proton	Gas-filled proportional detectors, ionization chambers, Geiger-Mueller tubes, scintillators, solid-state detectors
Gravitational	weight	See mechanical sensors

2.3.2 Data Management (Ontology, Data software platforms)

The high volume and high variety of materials data quickly outpaces rudimentary data organization techniques typically used by humans (project specific folder structures, ad hoc organization or note taking). We therefore require more sophisticated data management tools to manage the storage and organization of the materials data relevant to the digital twin. In their most basic forms data management tools act as *simple data repositories*, centralized locations where data is held and made accessible to others. However, simple data repositories do not necessarily provide a systematic scheme for the organization of the data or metadata therein. Digital twins require the establishment of standards and protocols to catalogue, vet, compare, and use data reliably and credibly in automated (and possibly autonomous) protocols (Kalidindi and De Graef, 2015; Sorkin et al., 2020). Consequently, data management solutions for digital twins should aim to at least meet FAIR data principles: Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al., 2016). FAIR data should have: (1) globally assigned, rich, searchable metadata with a unique persistent identifier and clear provenance; (2) standardized communication protocols for data storage and retrieval; (3) consistent, widely utilized, non-proprietary standards employed for data formatting. Data repositories generally only meet the most basic aspects of FAIR – namely accessibility. Materials databases progress further towards FAIR principles by providing greater searchability. Databases allow users to construct and carry out complex queries to search for information, and therefore improve searchability. However, their searchability is generally limited to tabular data. Furthermore, databases are also generally limited in their interoperability and reusability. In particular, they are not well suited for the materials data needed for digital twins as there is no natural way to describe the relational connections between disparate materials data (e.g., temporal variations along process paths, nested composition relationships, multimodal data describing single sample).

In order to truly realize FAIR data principles for materials data, we need to adopt emerging software tools in ontologies and linked data. *Ontologies* for data management are an open-world framework where we construct a standardized language to connect and describe objects. There currently exists many standardized languages used to describe ontologies such as OWL (McGuinness and Harmelen, 2004), RDF (Lassila and Swick, 1998), or JSONLD (Sporny et al., 2014). These languages all describe data in subject-predicate-object triples where we link the subject and the object through some rule (the predicate). One way to capture this information is through the formation of *knowledge graph* consisting of nodes (subjects, objects) and edges (predicates). Knowledge graphs allow for easily understood visual depictions of metadata, and for the application of emergent graph-based AI toolsets for the automated identification of new connections between aggregated elements of a complex heterogeneous dataset.

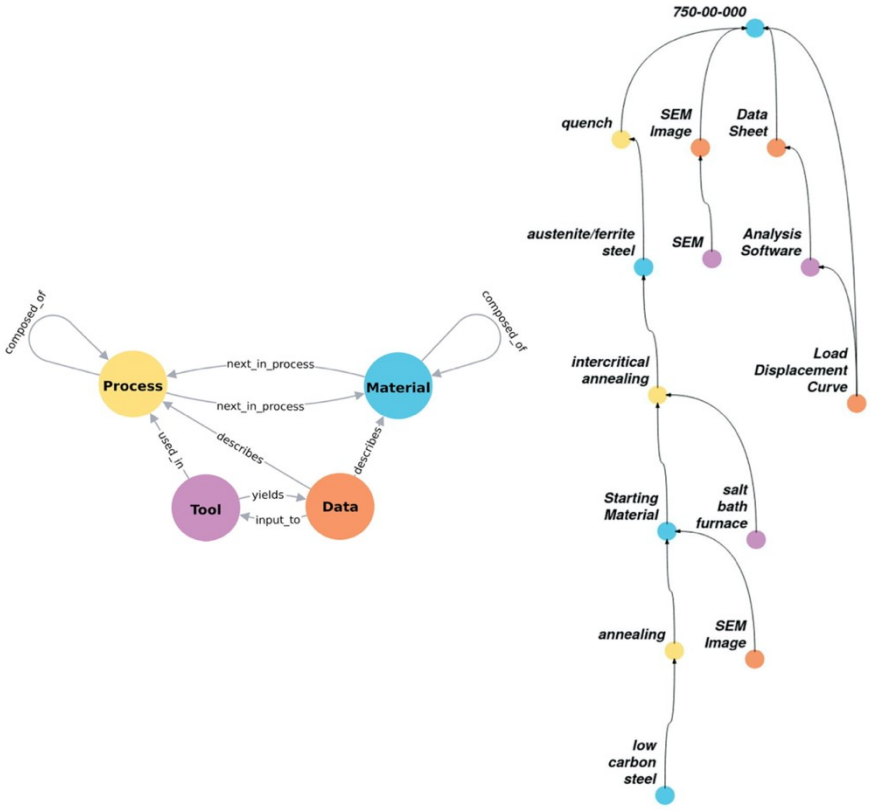

Figure 7. A) The constitutive elements of a recently proposed materials ontology (Voigt et al. 2021). The four main elements are Material (blue), Process (yellow), Tool (purple), and Data (orange) are shown in different colors along with the allowed connections between them. B) An example of a knowledge graph constructed using the ontology.

A recently proposed materials ontology (Voigt and Kalidindi, 2021) shown in Figure 7A can prove valuable in our effort to collect and curate the data needed for a materials digital twin. This ontology consists of four primary classes of entities (denoted by circles) that can serve as subjects or objects: Process, Material, Tool, and Data. A total of nine predicates (denoted by arrows) have been defined to link these objects. Process nodes hold information about process parameters, tool nodes describe the settings and characteristics of machines, and data nodes hold the payloads of interest (images, tabular data, etc). A material node describes the state of the material along a nominal process. Therefore, every time an action is taken on a material, we produce a new material node. This allows us to easily associate data with a point along a process path. As an example, a given steel (Material) produced after a specified thermo-mechanical processing route (Process) can be studied in

a microscope (Tool); the results of the study are captured in a file (Data). Figure 7B depicts an example knowledge graph for a steel. The process begins with a generic low carbon steel node (seen at the bottom of the knowledge graph). It then undergoes a standard annealing step to get a uniform starting material, and proceeds through a specialized intercritical annealing and quench steps to its final form (labelled as 750-00-000 in the knowledge graph). Along the processing route shown, we are able to connect the various data/metadata collected. For example, it is seen that both the starting material and the final material have associated SEM images. The final material also has a datasheet generated using a known software package (defined by a Tool node) which took a known load-displacement curve (defined by a Data node) as input. Ontologies allow us to systematically capture interconnected materials data and allow for the context of a dataset to be robustly described and communicated, thus enhancing the reusability of the data.

2.3.3 AI Tools

There currently exist several software packages than can be used to support the mathematical framework proposed in section 2.2. For structure quantification PyMKS (Brough et al.) offers computationally efficient tools for the feature engineering of material internal structures. PyMKS supports various data transformations needed to capture information on a wide range of material local states encountered in different material classes at different material structure length scales. PyMKS utilizes Dask, a distributed framework for developing python applications, to facilitate computations involving large datasets on supercomputers and large clusters (Rocklin, 2015). Subsequent to feature engineering, surrogate model building can be accomplished via a wide variety of popular python packages; examples include Statsmodels (Seabold and Perktold, 2010) for basic statistical models, SKLearn (Pedregosa et al., 2011) for machine learning tools, PyTorch (Paszke et al. 2019) and TensorFlow (Abadi et al., 2016) for neural networks/deep learning tools.

AI tools support digital twins beyond the needs of the mathematical framework alone. AI based segmentation strategies have gained traction, and Bayesian CNNs have recently been used to characterize the segmentation uncertainty in materials images (LaBonte et al., 2019). AI tools have also been effective in fusing multimodal materials data. Multi-input NNs have proven effective in combining data from multiple sources and different data types. For example, numeric and categorical data, assessed via multi-layer perceptron algorithms can be directly combined with image-based convolutional NNs (Azim and Aggarwal, 2014). While data streams are typically experimental, it can sometimes be beneficial to integrate high-fidelity simulation data from traditional high-performance computing approaches (e.g., atomistic modeling, phase-field, finite element) to augment “missing” experimental data or to represent functional dependencies/sensitivities that were not exposed in the experimental datasets. For instance, well-established experimental methods such as diffraction measurements are being implemented into computational models as a complement of the interpretation of experimental results (Coleman et al., 2014; Kunka et al., 2021). Alternatively, researchers have recently used generative machine learning algorithms such as generative adversarial network (GAN) to generate large materials and process libraries (Banko et al., 2020).

3 Applications

The ability to use a digital twin to provide an accurate picture of the corresponding physical twin at any given point in time is expected to significantly improve the guidance to subject-matter experts towards rational (and optimized) material/process improvements. Additionally, predictions of component performance can drive upstream changes in design or manufacturing process. To date, the development of detection and prognosis-driven planning strategies has largely focused on tuning individual process parameters such as temperature or materials composition for example, despite the urge to devise efficient strategies for the selection of multiple interdependent variables to

substantially accelerate and improve scientific discovery. Digital twins open up new opportunities to enable such strategies and accelerate autonomous experimental design and exploration. Autonomous experiments are emerging in materials research leading to the acceleration of materials design and discovery (Nikolaev et al., 2016; Gongora et al., 2020; Hase et al., 2018; Pendleton et al., 2019; Correa-Baena et al., 2018; Häse et al., 2019). The idea is to integrate automation with some form of machine learning or artificial intelligence framework to accelerate experimentation or to guide and discover the next set of experiments. Most of the work to date is dedicated to materials discovery, *i.e.*, autonomously predict and synthesize materials with targeted properties. For instance, Nikolaev et al. (2016) presented a closed-loop iterative method that automatically analyzes experimental results from carbon nanotubes grown from chemical vapor deposition to design or alter the next set of growth experiments to best reach a designated design target growth. Expanding autonomous loops to encompass more complex workflows will require the integration of the digital twin elements described in Section 2 with the automation of expert decisions. One interesting direction is to use the digital twins as a tool to autonomously test hypothesis during an experimental design. In this case, the practitioner would simply state the Process, Material, Tool, and Data and have the automation process decide whether the hypothesis is supported or refuted in order to decide on the potential next set of experiments. In this context, each automated trial would be guided by the knowledge collected and curated by the digital twin.

One particular application domain of interest for digital twins is the material/process exploration in additive manufacturing, with origins in rapid prototyping. There are extensive model-based simulations of the additive manufacturing process, ranging from powder packing through the entire laser-matter interaction and solidification process that can be taken as input into the Bayesian update strategy described in 2.2.2. The range of physical considerations in this process are daunting. In addition to these process models, there are complementary and similarly extensive set of structure-property models. Currently, a comprehensive digital representation of the entire spectrum of governing equations is beyond the state-of-the-art. A digital twin composed of many surrogate models utilizing the Bayesian update strategy could be formulated to optimize the parameters of these models for use in material design as well as process optimization.

4 Conclusions

Digital twins of the components in devices have enabled the in-service monitoring, prognosis, and design of complex systems. This work proposes both the conceptual framework and the cyberinfrastructure required to extend the concept of digital twins to the material level. Digital twins for materials provide a statistical in-silico materials representation of both structure and performance. The proposed framework consists of a materials representation based on n-point spatial correlations and PCA, a performance prediction framework centered around a two-step Bayesian framework, and a cyberinfrastructure that leverages new material ontologies for the management of multimodal materials data. Together, these foundational elements offer new opportunities for the extension of current digital twins to include important details of the material over a multitude of material structure length scales (from the macroscale to the atomistic).

5 Conflict of Interest

Authors Brad L. Boyce and Remi Dingreville were employed by Sandia National Laboratories. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest."

6 Author Contributions

These Authors have contributed equally to this work: SRK, MB, BLB, RD.

7 Funding

MB and SRK acknowledge support from NSF DMREF Award# 2119640. This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under Contract No. DE-NA0003525. The views of the U.S. Department of Energy or the U.S. Government.

8 References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (pp. 265-283).
[2] Adams, B. L., Kalidindi, S., & Fullwood, D. T. (2012). *Microstructure sensitive design for performance optimization*. Butterworth-Heinemann.
[3] Allison, J., Backman, D., & Christodoulou, L. (2006). Integrated computational materials engineering: a new paradigm for the global materials profession. Jom, 58(11), 25-27.
[4] Azim, S., & Aggarwal, S. (2014). Hybrid model for data imputation: using fuzzy c means and multi layer perceptron. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 1281-1285). IEEE.
[5] Banko, L., Lysogorskiy, Y., Grochla, D., Naujoks, D., Drautz, R., & Ludwig, A. (2020). Predicting structure zone diagrams for thin film synthesis by generative machine learning. Communications Materials, 1(1), 1-10.
[6] Belianinov, A., Ievlev, A. V., Lorenz, M., Borodinov, N., Doughty, B., Kalinin, S. V., et al. (2018). Correlated materials characterization via multimodal chemical and functional imaging. ACS nano, 12(12), 11798-11818.
[7] Brough, D. B., Wheeler, D., & Kalidindi, S. R. (2017). Materials knowledge systems in python—a data science framework for accelerated development of hierarchical materials. Integrating materials and manufacturing innovation, 6(1), 36.
[8] Hu, C., Martin, S., Dingreville, R., (2022). Accelerating phase-field predictions via recurrent neural networks learning the microstructure evolution in latent space. Computer Methods in Applied Mechanics and Engineering. Under review.
[9] Castillo, A. R., & Kalidindi, S. R. (2021). Bayesian estimation of single ply anisotropic elastic constants from spherical indentations on multi-laminate polymer-matrix fiber-reinforced composite samples. Meccanica, 56(6), 1575-1586.
[10] Castillo, A. R., Venkatraman, A., & Kalidindi, S. R. (2021). Mechanical Responses of Primary-α Ti Grains in Polycrystalline Samples: Part II—Bayesian Estimation of Crystal-Level Elastic-Plastic Mechanical Properties from Spherical Indentation Measurements. Integrating Materials and Manufacturing Innovation, 10(1), 99-114.
[11] Castillo, A., & Kalidindi, S. R. (2019). A Bayesian framework for the estimation of the single crystal elastic parameters from spherical indentation stress-strain measurements. Frontiers in Materials, 6, 136.
[12] Cecen, A., Fast, T., & Kalidindi, S. R. (2016). Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure. Integrating Materials and Manufacturing Innovation, 5(1), 1-15.
[13] Coleman, S. P., Sichani, M. M., & Spearot, D. E. (2014). A computational algorithm to produce virtual X-ray and electron diffraction patterns from atomistic simulations. Jom, 66(3), 408-416.
[14] Correa-Baena, J. P., Hippalgaonkar, K., van Duren, J., Jaffer, S., Chandrasekhar, V. R., Stevanovic, V., et al. (2018). Accelerating materials development via automation, machine learning, and high-performance computing. Joule, 2(8), 1410-1420.
[15] de Oca Zapiain, D. M., Stewart, J. A., & Dingreville, R. (2021). Accelerating phase-field-based microstructure evolution predictions via surrogate models trained by machine learning methods. npj Computational Materials, 7(1), 1-11.
[16] de Pablo, J. J., Jackson, N. E., Webb, M. A., Chen, L. Q., Moore, J. E., Morgan, D., et al. (2019). New frontiers for the materials genome initiative. npj Computational Materials, 5(1), 1-23.

[17] Dingreville, R., Karnesky, R. A., Puel, G., & Schmitt, J. H. (2016). Review of the synergies between computational modeling and experimental characterization of materials across length scales. Journal of materials science, 51(3), 1178-1203.
[18] Fullwood, D. T., Niezgoda, S. R., Adams, B. L., & Kalidindi, S. R. (2010). Microstructure sensitive design for performance optimization. Progress in Materials Science, 55(6), 477-562.
[19] Fullwood, D. T., Niezgoda, S. R., Adams, B. L., & Kalidindi, S. R. (2010). Microstructure sensitive design for performance optimization. Progress in Materials Science, 55(6), 477-562.
[20] Generale, A. P., & Kalidindi, S. R. (2021). Reduced-order models for microstructure-sensitive effective thermal conductivity of woven ceramic matrix composites with residual porosity. Composite Structures, 274, 114399.
[21] Ghoreishi, S. F., & Allaire, D. (2019). Multi-information source constrained Bayesian optimization. Structural and Multidisciplinary Optimization, 59(3), 977-991.
[22] Gil, Y., & Selman, B. (2019). A 20-year community roadmap for artificial intelligence research in the US. arXiv preprint arXiv:1908.02624.
[23] Gomberg, J. A., Medford, A. J., & Kalidindi, S. R. (2017). Extracting knowledge from molecular mechanics simulations of grain boundaries using machine learning. Acta Materialia, 133, 100-108.
[24] Gongora, A. E., Xu, B., Perry, W., Okoye, C., Riley, P., Reyes, K. G., et al. (2020). A Bayesian experimental autonomous researcher for mechanical design. Science advances, 6(15), eaaz1708.
[25] Ghosh, S., & Groeber, M. A. (2020). Developing virtual microstructures and statistically equivalent representative volume elements for polycrystalline materials. *Handbook of Materials Modeling: Methods: Theory and Modeling*, 1631-1656.
[26]Groeber, M. A., & Jackson, M. A. (2014). DREAM. 3D: a digital representation environment for the analysis of microstructure in 3D. *Integrating materials and manufacturing innovation*, *3*(1), 56-72.
[27] Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. Trends in Chemistry, 1(3), 282-291.
[28] Hase, F., Roch, L. M., Kreisbeck, C., & Aspuru-Guzik, A. (2018). Phoenics: a Bayesian optimizer for chemistry. ACS central science, 4(9), 1134-1145.
[29] Herr, J. E., Koh, K., Yao, K., & Parkhill, J. (2019). Compressing physics with an autoencoder: Creating an atomic species representation to improve machine learning models in the chemical sciences. The Journal of chemical physics, 151(8), 084103.
[30] Horstemeyer, M. F. (2009). Multiscale modeling: a review. Practical aspects of computational chemistry, 87-135.
[31] Jenks, C., Lee, N., Lewis, J., Kagan, C., Nealey, P., Braun, P., et al. (2020). Basic Research Needs for Transformative Manufacturing (Report). USDOE Office of Science (SC).
[32] Kalidindi, S. R. (2015). Hierarchical materials informatics: novel analytics for materials data. Elsevier.
[33] Kalidindi, S. R. (2020). Feature engineering of material structure for AI-based materials knowledge systems. Journal of Applied Physics, 128(4), 041103.
[34] Kalidindi, S. R., & De Graef, M. (2015). Materials data science: current status and future outlook. Annual Review of Materials Research, 45, 171-193.
[35] Kalidindi, S. R., Khosravani, A., Yucel, B., Shanker, A., & Blekh, A. L. (2019). Data infrastructure elements in support of accelerated materials innovation: ELA, PyMKS, and MATIN. Integrating Materials and Manufacturing Innovation, 8(4), 441-454.
[36] Kapteyn, M. G., Pretorius, J. V., & Willcox, K. E. (2021). A probabilistic graphical model foundation for enabling predictive digital twins at scale. Nature Computational Science, 1(5), 337-347.

[37] Kaundinya, P. R., Choudhary, K., & Kalidindi, S. R. (2021). Machine learning approaches for feature engineering of the crystal structure: Application to the prediction of the formation energy of cubic compounds. Physical Review Materials, 5(6), 063802.
[38] Khosravani, A., Caliendo, C. M., & Kalidindi, S. R. (2020). New insights into the microstructural changes during the processing of dual-phase steels from multiresolution spherical indentation stress–strain protocols. *Metals*, *10*(1), 18.
[39] Khosravani, A., Thadhani, N., & Kalidindi, S. R. (2021). Microstructure Quantification and Multiresolution Mechanical Characterization of Ti-Based Bulk Metallic Glass-Matrix Composites. *JOM*, 1-11.
[40] Kunka, C., Shanker, A., Chen, E. Y., Kalidindi, S. R., & Dingreville, R. (2021). Decoding defect statistics from diffractograms via machine learning. npj Computational Materials, 7(1), 1-9.
[41] LaBonte, T., Martinez, C., & Roberts, S. A. (2019). We Know Where We Don't Know: 3D Bayesian CNNs for Credible Geometric Uncertainty. arXiv preprint arXiv:1910.10793.
[42] Lassila, O., & Swick, R. R. (1998). Resource description framework (RDF) model and syntax specification.
[43]Latypov, M. I., Toth, L. S., & Kalidindi, S. R. (2019). Materials knowledge system for nonlinear composites. Computer Methods in Applied Mechanics and Engineering, 346, 180-196.
[44] Lim, J., Perullo, C. A., Milton, J., Whitacre, R., Jackson, C., Griffin, C., et al. (2021). The EPRI Gas Turbine [43] Digital Twin–a Platform for Operator Focused Integrated Diagnostics and Performance Forecasting. In
[45] Margaria, T. and Schieweck, A., 2019, December. The digital thread in industry 4.0. In International Conference on Integrated Formal Methods (pp. 3-24). Springer, Cham
[46] Marquis de Laplace, P. S. (1814). A Treatise Upon Analytical Mechanics: Being the First Book of the Mécanique Céleste. H. Barnett.
[47] Marshall, A., & Kalidindi, S. R. (2021). Autonomous Development of a Machine-Learning Model for the Plastic Response of Two-Phase Composites from Micromechanical Finite Element Models. JOM, 1-11.
[48] Matouš, K., Geers, M. G., Kouznetsova, V. G., & Gillman, A. (2017). A review of predictive nonlinear theories for multiscale modeling of heterogeneous materials. Journal of Computational Physics, 330, 192-220.
[49] McDowell, D. L., & LeSar, R. A. (2016). The need for microstructure informatics in process–structure–property relations. MRS Bulletin, 41(8), 587-593.
[50] McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. W3C recommendation, 10(10), 2004.
[51] Mei, H., Haider, M. F., Joseph, R., Migot, A., & Giurgiutiu, V. (2019). Recent advances in piezoelectric wafer active sensors for structural health monitoring applications. Sensors, 19(2), 383.
[52] Millan-Espitia, N., Mohan, S., Pilchak, A. L., & Kalidindi, S. R. (2021). Mechanical Responses of Primary-α Ti Grains in Polycrystalline Samples: Part I—Measurements of Spherical Indentation Stress–Strain Curves. Integrating Materials and Manufacturing Innovation, 10(1), 82-98.
[53] Morgado, J. F., Ghedini, E., Goldbeck, G., Hashibon, A., Schmitz, G. J., Friis, J., & Anne, F. (2020). Mechanical Testing Ontology for Digital-Twins: a Roadmap Based on EMMO. In SeDiT@ ESWC.
[54] National Science and Technology Council (US). (2011). Materials genome initiative for global competitiveness. Executive Office of the President, National Science and Technology Council.
[55] Niederer, S. A., Sacks, M. S., Girolami, M., & Willcox, K. (2021). Scaling digital twins from the artisanal to the industrial. Nature Computational Science, 1(5), 313-320.
[56] Niezgoda, S. R., Kanjarla, A. K., & Kalidindi, S. R. (2013). Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. Integrating Materials and Manufacturing Innovation, 2(1), 54-80.

[57] Niezgoda, S. R., Yabansu, Y. C., & Kalidindi, S. R. (2011). Understanding and visualizing microstructure and microstructure variance as a stochastic process. Acta Materialia, 59(16), 6387-6400.
[58] Nikolaev, P., Hooper, D., Webber, F., Rao, R., Decker, K., Krein, M., et al. (2016). Autonomy in materials research: a case study in carbon nanotube growth. npj Computational Materials, 2(1), 1-6.
[59] Panchal, J. H., Kalidindi, S. R., & McDowell, D. L. (2013). Key computational modeling issues in integrated computational materials engineering. Computer-Aided Design, 45(1), 4-25.
[60] Pandita, P., Bilionis, I., & Panchal, J. (2019). Bayesian optimal design of experiments for inferring the statistical expectation of expensive black-box functions. Journal of Mechanical Design, 141(10).
[61] Paulson, N. H., Priddy, M. W., McDowell, D. L., & Kalidindi, S. R. (2017). Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics. Acta Materialia, 129, 428-438.
[62] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 8026-8037.
[63] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
[64] Pendleton, I. M., Cattabriga, G., Li, Z., Najeeb, M. A., Friedler, S. A., Norquist, A. J., et al. (2019). Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. MRS Communications, 9(3), 846-859.
[65] Polonsky, A. T., & Pandey, A. (2021). Advances in Multimodal Characterization of Structural Materials. JOM, 1-2.
[66] Raj, P., & Surianarayanan, C. (2020). Digital twin: the industry use cases. In Advances in Computers (Vol. 117, No. 1, pp. 285-320). Elsevier.
[67] Robertson, A. E., & Kalidindi, S. R. (2021). Digital representation and quantification of discrete dislocation networks. arXiv preprint arXiv:2101.03925.
[68] Robertson, A. E., & Kalidindi, S. R. (2021). Efficient Generation of Anisotropic N-Field Microstructures From 2-Point Statistics Using Multi-Output Gaussian Random Fields. Available at SSRN 3949516.
[69] Rocklin, M. (2015, July). Dask: Parallel computation with blocked algorithms and task scheduling. In Proceedings of the 14th python in science conference (Vol. 130, p. 136). Austin, TX: SciPy.
[70] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference (Vol. 57, p. 61).
[71] Singh, V., & Willcox, K. E. (2018). Engineering design with digital thread. AIAA Journal, 56(11), 4515-4528.
[72] Solomou, A., Zhao, G., Boluki, S., Joy, J. K., Qian, X., Karaman, I., et al. (2018). Multi-objective Bayesian materials discovery: Application on the discovery of precipitation strengthened NiTi shape memory alloys through micromechanical modeling. Materials & Design, 160, 810-827.
[73] Sorkin, B. C., Betz, J. M., & Hopp, D. C. (2020). Toward FAIRness and a User-Friendly Repository for Supporting NMR Data.
[74] Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). JSON-LD 1.0. W3C recommendation, 16, 41.
[75] Takhtaganov, T., & Müller, J. (2018). Adaptive Gaussian process surrogates for Bayesian inference. arXiv preprint arXiv:1809.10784.

[76] Talapatra, A., Boluki, S., Duong, T., Qian, X., Dougherty, E., & Arróyave, R. (2018). Autonomous efficient experiment design for materials discovery with Bayesian model averaging. Physical Review Materials, 2(11), 113803.
[77] Tao, F., Zhang, M., Liu, Y., & Nee, A. Y. (2018). Digital twin driven prognostics and health management for complex equipment. Cirp Annals, 67(1), 169-172.
[78] The Minerals, Metals & Materials Society (TMS), Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering (Pittsburgh, PA: TMS, 2017). Electronic copies available at www.tms.org/mdistudy
[79] The Minerals, Metals & Materials Society (TMS), Modeling Across Scales: A Roadmapping Study for Connecting Materials Models and Simulations Across Length and Time Scales (Warrendale, PA: TMS, 2015). Electronic copies available at www.tms.org/multiscalestudy.
[80] Torquato, S., & Haslach Jr, H. W. (2002). Random heterogeneous materials: microstructure and macroscopic properties. Appl. Mech. Rev., 55(4), B62-B63.
[81] Torquato, S., & Stell, G. (1982). Microstructure of two-phase random media. I. The n-point probability functions. The Journal of Chemical Physics, 77(4), 2071-2077.
[82] Turbo Expo: Power for Land, Sea, and Air (Vol. 84966, p. V004T09A009). American Society of Mechanical Engineers.
[83] Ullo, S. L., & Sinha, G. R. (2020). Advances in smart environment monitoring systems using IoT and sensors. Sensors, 20(11), 3113.
[84] Voigt, S. P., & Kalidindi, S. R. (2021). Materials graph ontology. Materials Letters, 295, 129836.
[85] Weaver, J. S., Khosravani, A., Castillo, A., & Kalidindi, S. R. (2016). High throughput exploration of process-property linkages in Al-6061 using instrumented spherical microindentation and microstructurally graded samples. Integrating Materials and Manufacturing Innovation, 5(1), 192-211.
[86] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.
[87] Xiang, Z., Fan, M., Tovar, G. V., Trehem, W., Yoon, B. J., Qian, X., et al. (2021). Physics-constrained Automatic Feature Engineering for Predictive Modeling in Materials Science. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 12, pp. 10414-10421).
[88] Xie, R., Chen, M., Liu, W., Jian, H., & Shi, Y. (2021). Digital Twin Technologies for Turbomachinery in a Life Cycle Perspective: A Review. Sustainability, 13(5), 2495.
[89] Yang, Z., Yabansu, Y. C., Al-Bahrani, R., Liao, W. K., Choudhary, A. N., Kalidindi, S. R., & Agrawal, A. (2018). Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets. Computational Materials Science, 151, 278-287.
[90] Zaccaria, V., Stenfelt, M., Aslanidou, I., & Kyprianidis, K. G. (2018). Fleet monitoring and diagnostics framework based on digital twin of aero-engines. In Turbo Expo: Power for Land, Sea, and Air (Vol. 51128, p. V006T05A021). American Society of Mechanical Engineers.
[91] Zeb, S., Mahmood, A., Hassan, S. A., Piran, M. D., Gidlund, M., & Guizani, M. (2021). Industrial digital twins at the nexus of nextG wireless networks and computational intelligence: A survey. arXiv preprint arXiv:2108.04465.