Integr Mater Manuf Innov (2017) 6:54–68
DOI 10.1007/s40192-017-0088-1

RESEARCH

Process-Structure Linkages Using a Data Science Approach: Application to Simulated Additive Manufacturing Data

Evdokia Popova[1] · **Theron M. Rodgers**[2] · **Xinyi Gong**[3] · **Ahmet Cecen**[4] · **Jonathan D. Madison**[5] · **Surya R. Kalidindi**[1,3,4]

Received: 27 September 2016 / Accepted: 17 November 2016 / Published online: 13 March 2017

Abstract A novel data science workflow is developed and demonstrated to extract process-structure linkages (i.e., reduced-order model) for microstructure evolution problems when the final microstructure depends on (simulation or experimental) processing parameters. This workflow consists of four main steps: data pre-processing, microstructure quantification, dimensionality reduction, and extraction/validation of process-structure linkages. Methods that can be employed within each step vary based on the type and amount of available data. In this paper, this data-driven workflow is applied to a set of synthetic additive manufacturing microstructures obtained using the Potts-kinetic Monte Carlo (kMC) approach. Additive manufacturing techniques inherently produce complex microstructures that can vary significantly with processing conditions. Using the developed workflow, a low-dimensional data-driven model was established to correlate process parameters with the predicted final microstructure. Additionally, the modular workflows developed and presented in this work facilitate easy dissemination and curation by the broader community.

Keywords PSP linkages · Workflows · Microstructure quantification · Additive manufacturing · Monte Carlo simulation

Introduction

Acceleration in the rate of material development and deployment has been the focus of several recent efforts in current literature (e.g., [1–6]). In this regard, multiscale modeling and simulation has been identified as a key enabler [7–11], because of its potential to dramatically reduce time and effort expended in experimentation. However, there is now an increasing recognition that this alone cannot bring about the desired acceleration in material development. There is a critical need for the development and deployment of a suitable supporting data infrastructure that efficiently integrates closed-loop iterations between experimental and multiscale modeling/simulation efforts. This need is being addressed by a new cross-disciplinary field known as materials data science and informatics [1, 3, 12–20].

A foundational element of a data science approach is a versatile framework that enables capture, aggregation,

✉ Surya R. Kalidindi
surya.kalidindi@me.gatech.edu

Evdokia Popova
evdokia.popova@me.gatech.edu

Theron M. Rodgers
trodger@sandia.gov

Xinyi Gong
xinyigong@gatech.edu

Ahmet Cecen
ahmetcecen@gatech.edu

Jonathan D. Madison
jdmadis@sandia.gov

[1] Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[2] Computational Materials & Data Science, Sandia National Laboratories, PO Box 5800, MS-1411, Albuquerque, NM 87185, USA

[3] School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[4] School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[5] Material Mechanics, Sandia National Laboratories, PO Box 5800 MS-0889, Albuquerque 87185, NM, USA

curation, dissemination, and re-use of high-value knowledge by communities of researchers. In materials innovation efforts, this knowledge is chiefly desired in the form of process-structure-property (PSP) linkages at length (and time) scales relevant to the material systems of interest [1, 12, 18, 21–25]. For multiscale material modeling efforts, this would imply the development of formal data science approaches for distilling re-usable PSP linkages from ensembles of simulation datasets. Figure 1 describes this strategy schematically. The top row of this figure describes the typical workflow explored by computational materials scientists in developing PSP linkages. Typically, this entails the use of highly sophisticated physics in conjunction with numerical algorithms and, as a result, incurs substantial computational cost. This cost can present a major impediment to materials innovation efforts, as one would often need to run these simulations a large number of times under varying inputs. This is precisely where a data science approach offers many advantages. As suggested in Fig. 1, one can learn from previously accumulated numerical datasets and extract the embedded linkages between inputs and simulated outputs. In the context of multiscale materials phenomena, this learning is most efficiently carried out in a mathematically rigorous framework for PSP linkages, while taking full advantage of legacy knowledge, advanced statistics, and machine learning techniques. As described in Fig. 1, one of the central benefits of adding a data science component to the overall workflow is that it produces a very practical (low computational cost) approach to solving inverse problems that lie at the core of all materials innovation efforts. This is mainly because the PSP linkages are usually cast as metamodels (or surrogate models) that allow easy inversion due to their relatively simple mathematical reduction.

A central impediment in the implementation of the approach described in Fig. 1 comes from a lack of validated and broadly adopted frameworks for the rigorous quantification of hierarchical material structures or microstructure. Microstructure plays a central role in the formulation of PSP linkages and is often an important input and/or output. Furthermore, microstructure can often require a higher dimensional representation compared to other variables involved in the PSP linkages. From a practical viewpoint, it becomes essential to seek suitable reduced-order representations of material structure and use them in formulating PSP linkages. Traditionally, this dimensionality reduction has been performed by materials scientists based on intuition or insight of the materials phenomena studied. As a specific example, one might quantify polycrystalline microstructures using grain size or shape distributions, and possibly orientation and misorientation distributions, when studying their plastic response. However, such approaches have not yet identified a common set of low-dimensional measures that can be universally applied across diverse material systems for identification of a majority of material response characteristics. This, however, is a key element in the formulation of re-usable, high-value, material knowledge systems.

Emerging toolsets in materials data science and informatics have demonstrated tremendous promise in addressing some of the key challenges described above. It is now possible to generate a large ensemble of datasets (inputs and outputs) from a simulation toolset and publicly share these with the broader scientific community in an open-access data repository [20]. Once this is accomplished, it is possible to engage the broader scientific community in the extraction of the embedded knowledge of these datasets. If this activity is guided in a suitable framework for PSP linkages, it could lead to accelerated and robust curation of the knowledge, while simultaneously ensuring the highest levels of access, sharing, and dissemination for re-use.

The main goal of this work is to explore the viability of the concepts and philosophies described above with an example demonstrator focused on process-structure (P-S) linkages with a view toward additive manufacturing. Additive manufacturing (AM) is a rapidly growing field of advanced materials processing [26, 27]. Process improvements in recent years have enabled the creation of near-fully dense parts with sophisticated geometries that are unobtainable using traditional manufacturing techniques [28]. While AM has seen significant adoption as a prototyping and small-batch production tool, the science behind AM part creation is complex and only partially understood. Variations in factors such as powder composition,

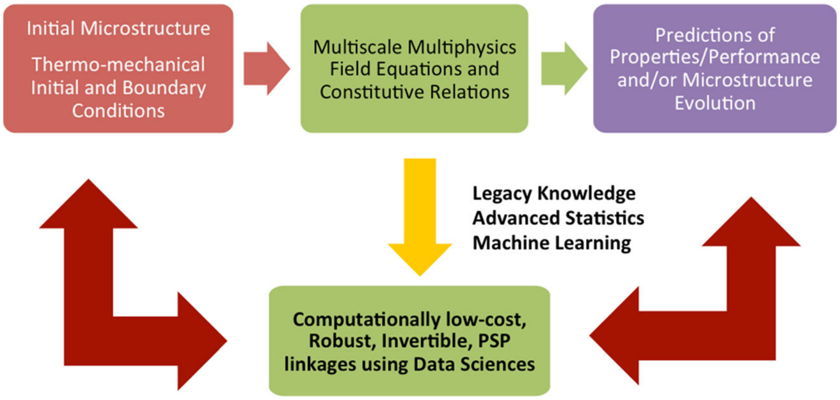

Initial Microstructure
Thermo-mechanical Initial and Boundary Conditions
Multiscale Multiphysics Field Equations and Constitutive Relations
Predictions of Properties/Performance and/or Microstructure Evolution
Legacy Knowledge
Advanced Statistics
Machine Learning
Computationally low-cost, Robust, Invertible, PSP linkages using Data Sciences

Fig. 1 A schematic description of the workflow typically employed in current computational materials science efforts (*top row*) and how this can be augmented with a data science approach to recover computationally low-cost, high-value PSP linkages of interest to materials innovation efforts [12] (Color figure online)

processing technique, and component shape can result in dramatically different microstructures and material properties. Additionally, microstructure can vary significantly even within a single as-built part. The interplay between the length scales of AM builds and those of processing (e.g., localized melt pool size and shape) presents new challenges in the analysis and prediction of microstructure-sensitive performance characteristics. Furthermore, irregular component geometries and material anisotropies create compounded difficulties for traditional analysis methods [29].

Among the many processing variables of interest, beam power density and scan pattern stand out as relatively dominant factors. Power density is directly controlled by beam parameters (spot size, power, scan rate, etc.), but is also indirectly influenced by the scan pattern used to construct the build. Together, power density and scan pattern greatly influence both the overall microstructure and the local microstructural variations [30–32]. Although a number of experimental and simulation studies are underway [27, 33–35] to quantify the P-S relationships in AM, the opportunity for advanced data analysis has also been recognized [27, 35–38]. The multiscale heterogeneity present throughout a solidified AM build would suggest that a rigorous, quantitative, and statistical analysis is essential to achieve high-fidelity success in the realm of qualification for significant industrial or high-consequence applications [2].

A Monte Carlo Potts model has been employed successfully to simulate grain growth [39], recrystallization [40], electron beam welding [40], and AM processing [41], and has demonstrated a remarkable qualitative agreement with experimental data. The simulation method yields predictions of three-dimensional (3-D) polycrystalline microstructures under a variety of scenarios and has even been demonstrated to couple effectively with additional models for the inclusion of additional physics [41]. With recent advances in computational infrastructure, it is now possible to conduct a large number of simulations to generate an aggregate dataset composed of thousands of individual simulations, where input parameters are systematically varied to cover specific ranges of interest. While extracting re-usable P-S linkages in the form of low-cost surrogate models from these datasets is a nontrivial task, this paper will address this task using emerging toolsets of materials data science and informatics.

Additive Manufacturing Simulation Dataset

A user subroutine was created for the SPPARKS kinetic Monte Carlo (kMC) simulation suite [42] to approximate multiple passes of a localized heat source during AM processing. The adaptation utilizes a modified Potts-Monte Carlo [43] approach to simulate grain growth during directional solidification. In this study, we do not seek to exhaustively describe the simulation approach but will offer a few salient specifics to provide a cursory understanding of the synthetic microstructure generation. The reader is advised to always ensure the accuracy and reliability of the physics-based simulation toolsets before embarking on the calibration and extraction of reduced-order models of interest. For the simulation suite used here, this effort was carried out elsewhere and the reader is directed to refs. [44, 45] for details regarding the approach and its validation against experimental results. Only a few essential points and modifications will be discussed here. A collection of sites on a cubic lattice compose the simulation domain, in which each site is assigned a “spin” to identify its membership to a certain grain identified by that spin. The physical arrangement of similar and dissimilar spins defines the grain structure and total energy of the simulation. Simulation time is expressed in Monte Carlo steps (MCS). One MCS corresponds to an attempted Monte Carlo spin flip at each neighbor of every lattice site. Although the exact relationship to physical time is difficult to define, they are related by a constant factor [43]. A local heat source is rastered through the domain using a prescribed pattern. To simulate melting, a site’s spin is randomized when it is located within the “melt pool” of the heat source. Resolidification and grain growth occur in the heat-affected zone (HAZ) surrounding the melt pool. Elongated grains grow in the direction of the maximum thermal gradient, and an anisotropic polycrystalline microstructure is produced bearing the history of the scanning strategy and the inherent size and shape of the heat source. In the present study, the melt pool was rastered across each layer in four parallel passes, with each pass alternating direction by 180°. This was repeated for 4 layers of deposition, resulting in 16 passes of the simulated heat source. Additionally, the melt pool used in this study was comprised of a half-ellipsoid shape and is reproduced here for convenience in Fig. 2.

The approach described here allows for rapid exploration of varying simulation conditions and the use of relatively large simulation domains (300 × 300 × 200 elements) at low computational costs. The kMC simulations are non-dimensional but include an implied length scale resulting from the shape of the molten zone and the height of the remelt layers as both determine significant amounts of the resulting microstructure’s arrangement. In the simulations presented, a molten zone of 60, 70, 80, or 90 sites corresponds to physical dimension of 0.3, 0.35, 0.4, or 0.45 mm, respectively. These layer-by-layer structures with limited remelting of prior layers are consistent with the low-power experimental validation comparisons of [46] presented in [45]. A total of 1799 microstructures (each corresponding to a different combination of process parameters) were generated on a Linux-computing cluster to comprise the ensemble dataset for this study. In comparison, state-of-the-art thermofluid, multiphysics, simulations of AM processes are generally capable of simulating only a single pass under a similar computational cost [35].

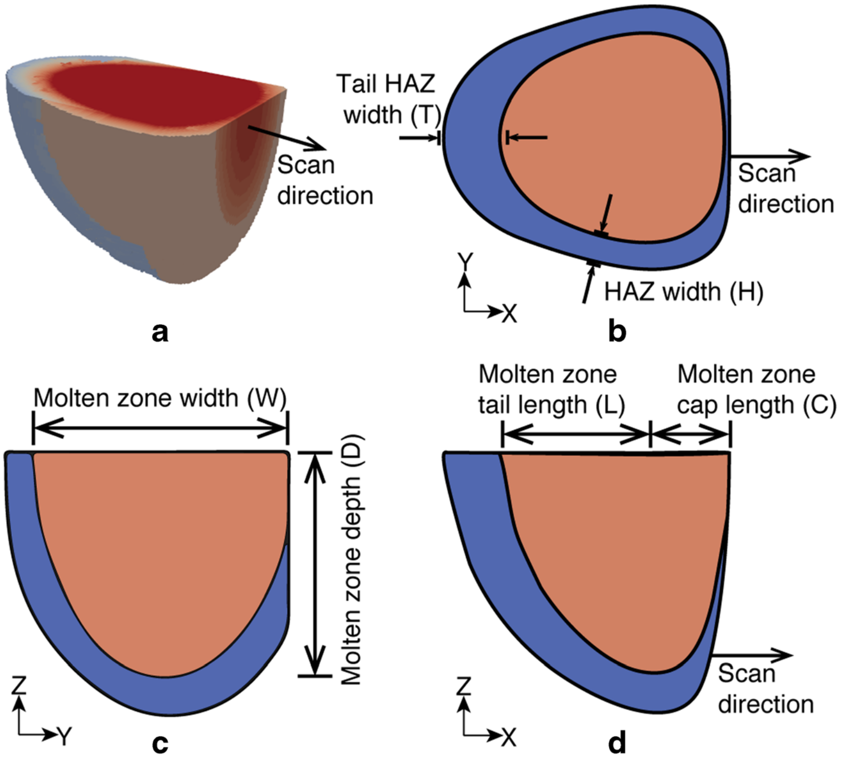


Fig. 2 **a** Idealized 3D melt pool with temperature gradient profile used for kMC synthetic microstructure generation. **b**–**d** Orthogonal cross-section schematics of melt pool; molten zone is depicted in *orange* and HAZ is shown in *blue* [41] (Color figure online)

The input simulation parameters used to generate the ensemble dataset were selected to mimic processing parameters found in metal AM techniques, and are listed in Table 1. While several parameters were varied during the study, an identical number of layers and passes per layer were used across all cases. The domain size and hatch spacing between scans were also maintained constant. The values of the simulation parameters were selected to span an experimentally relevant range, but were not intended to be exhaustive. The relative variation in the ensemble dataset is illustrated by the four-image composite of simulated microstructures shown in Fig. 3a, along with their corresponding process parameter set in Fig. 3b. Three orthogonal cross sections are shown in Fig. 3a for each simulation, where the arrows indicate in-plane scan directions, while circles with a cross or dot denote inward or outward out-of-plane scan directions with respect to the page. Two scan patterns between successive layers were studied. The first used a uniform pattern across all layers (i.e., parallel build, simulations 1 and 2 in Fig. 3) whereas the second rotated the raster pattern by 90° between each successive layer, (i.e., cross hatch, simulations 3 and 4 in Fig. 3). Each simulation produced a microstructure with unique grain size distributions and varying directional anisotropies.

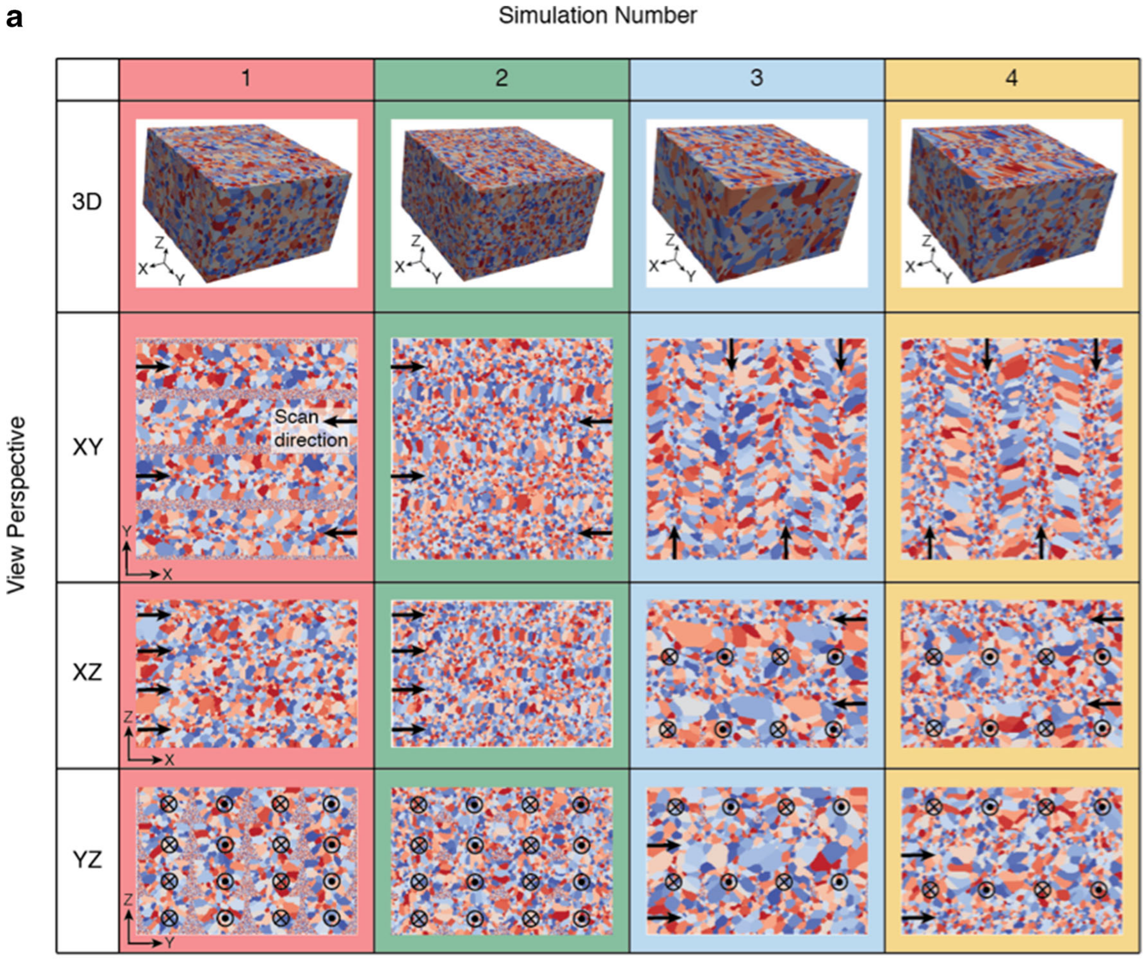
Table 1 The range of simulation conditions used in the study

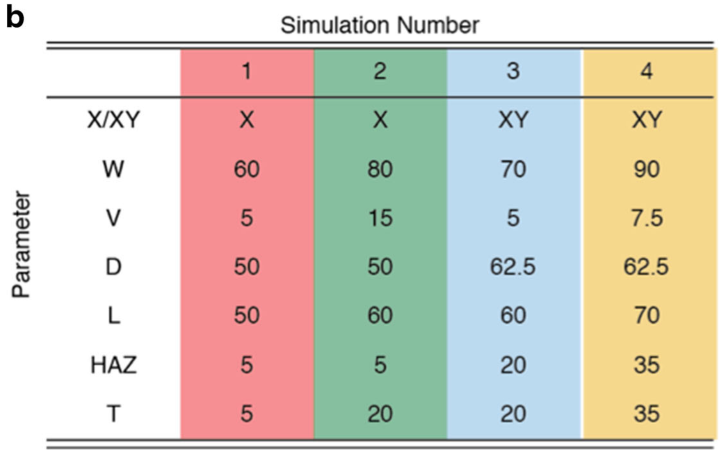
Variable	Values explored
(*X*/*XY*) Scan pattern	Parallel (*X*) or cross hatch (*XY*)
(*W*) Molten zone width (lattice sites)	60, 70, 80, 90
(*V*) Velocity (sites/Monte Carlo step)	2.5, 5, 7.5, 10, 15
(*D*) Molten zone depth (sites)	50, 62.5
(*L*) Molten zone tail length (sites)	50, 60, 70
(HAZ) Heat-affected-zone width (sites)	5, 20, 35
(*T*) Tail heat-affected-zone length (sites)	5, 20, 35

Data Science Workflow for Extracting Process-Structure Linkages

Building on prior work [12, 47], a generalized four-step workflow was designed for establishing P-S linkages and is shown in Fig. 4. This workflow has been designed to serve as a generic template that is applicable to the broad class of microstructure evolution phenomena that are likely to be studied by a variety of techniques (these could include modeling techniques such as phase-field models [48], cellular automata [49], and level-set methods [50] or experimental techniques such as X-ray computed tomography [51, 52]). The main steps are listed in blue boxes. The white accompanying boxes show specific methods and/or procedures that might be employed in that step.

The first step in the workflow is a pre-processing step aimed at ensuring quality and consistency of the dataset. While the identification of the phases, boundaries, or other features of interest in simulated data is trivial in most cases, experimental data often requires segmentation of images to properly identify a given feature of interest. As needed, one




b

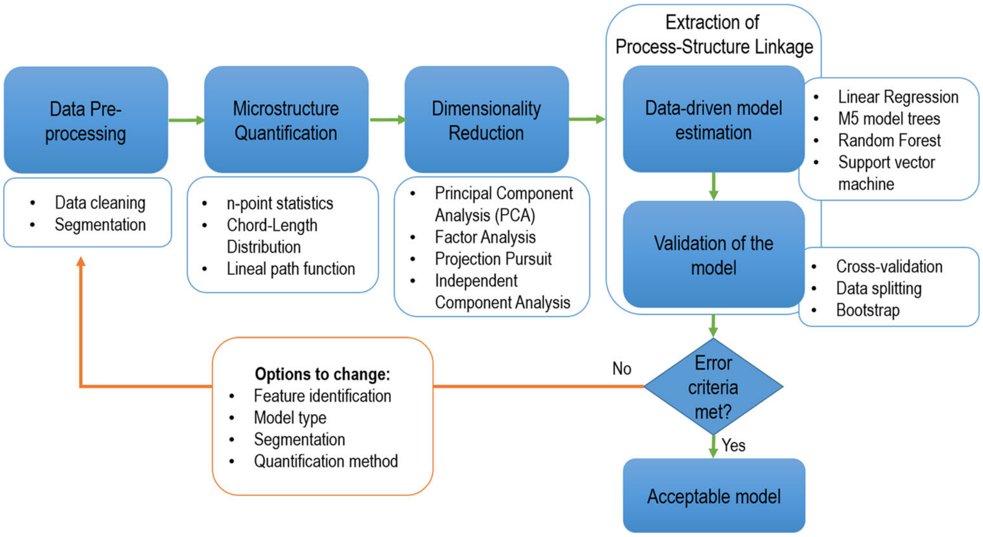
Parameter	Simulation Number 1	2	3	4
X/XY	X	X	XY	XY
W	60	80	70	90
V	5	15	5	7.5
D	50	50	62.5	62.5
L	50	60	60	70
HAZ	5	5	20	35
T	5	20	20	35

Fig. 3 **a** Orthogonal views of four synthetic microstructures with the scan direction indicated for each pass, and **b** corresponding simulation parameters (Color figure online)

might set a criterion to eliminate spurious or questionable data (e.g., the data that does not conform to known physics). In this step, the inputs (process parameters) are also clearly associated with the outputs (microstructure data).

In the second step, microstructures are quantified to obtain salient statistical measures of microstructures. In a data science approach, it is desirable to capture a very large set of measures at this stage. Consequently, it is preferable to adopt a microstructure quantification framework that allows one to increase systematically the numbers of potential features included in the analyses. In this regard, the framework of *n*-point spatial correlations [12, 53, 54] offers tremendous promise

Fig. 4 Workflow template for establishing process-structure linkages using a data science approach. Blue boxes describe general steps while corresponding white boxes list example methods that can be used within a step (Color figure online)



because of its scalability (ability to define an infinite number of microstructural features), organization (value of n can start with one and increase systematically), and available access to efficient computational toolsets [55, 56]. Another option for this step includes lineal path functions [57] or chord-length distributions [58, 59] that provide information about shape and size distribution of a specific feature of interest.

The third step in the workflow focuses on reducing the dimensionality of microstructure representation using data science approaches. Some of the established dimensionality reduction techniques include principal component analysis [12], factor analysis [60], projection pursuit [61], and independent component analysis [62], among others. These methods are designed to reduce dataset dimensions, while losing only the smallest amounts of information. The use of dimensionality reduction leads to savings in both computational time and storage, and leads to identification of salient features that can be used to establish models. For example, in prior work [12], PCA has proven to be remarkably efficient in producing high-value, low-order, representations of microstructures that are ideally suited to establishing PSP linkages in a broad variety of material systems.

The last step of the workflow focuses on establishing and validating a reliable and robust P-S linkage. This step typically involves an iterative process of model selection. The first part of this step requires establishing a model using a variety of machine learning techniques ranging from simple regression to sophisticated M5 model trees [63] and support vector machines [64]. It is important to recognize that the models developed are indeed dependent on the available data. Therefore, the model itself can change as one adds more data. Validation of the model established in this step is typically performed using accuracy estimation methods. Cross-validation [65] has been found to be quite effective in avoiding overfitting of the data to the model. Data splitting is another validation method in which each ensemble dataset is generally split into calibration and test subsets. Data splitting was shown to be an effective technique, where collection of new validation data is avoided [66]. In this step, a model selection is accomplished iteratively based on the optimization of error parameters. Error metrics therefore play an important role in the model selection process. Popular choices have included various combinations and variants of the mean of absolute error (MAE), the standard deviation of error (SDE), the coefficient of correlation (R), and the explained variance (R^2) [56]:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}\right| \tag{1}$$

$$\mathrm{SDE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left|y_i - \hat{y}\right| - \mathrm{MAE}\right)^2} \tag{2}$$

$$R = \frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)\left(\hat{y_i} - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 \sum_{i=1}^{N}\left(\hat{y_i} - \bar{y}\right)^2}} \tag{3}$$

where y_i denotes the value of the output variable from the actual data, y is the corresponding predicted value from the statistical model produced, and N is the number of data points. The R^2 value is a metric of goodness of fit, where a value closer to one indicates a superior fit.

After obtaining a data-driven model, errors are checked, and if they do not satisfy the error criteria, a new iteration in model building is launched (see Fig. 4). It is, however, important to identify which step contributed most to the unreliable model. If this insight is available, suitable modifications can be implemented in any step of the workflow in the next iteration. For instance, one might select a different model learning algorithm or identify new features using a different dimensionality reduction technique. The modular nature of the workflow shown in Fig. 4 allows one to explore a very large number of potential models in highly computationally efficient toolkits [67, 55] before settling on the best model for the phenomena studied.

Suitable error criteria for an acceptable model should be defined or set by the user for any practical implementation of the workflow shown in Fig. 4. These criteria are likely to be highly dependent on the intended purpose of the reduced-order model extracted using this workflow. In most MGI or ICME applications, a materials designer is likely to use the reduced-order models for rapid screening of a large design space under consideration. Therefore, the requirements for accuracy should be based on obtaining reliable guidance for meaningful down-selection of the design choices. Note also that all data sources (in the present case, simulation codes employed to generate an ensemble of microstructures) inherently exhibit certain (often non-negligible) uncertainty (or inaccuracy) that can be attributed to the numerous approximations and idealizations employed. Therefore, it would be unwise to establish an error criterion that exceeds the inherent uncertainty in the data source.

Case Study: Application to Additive Manufacturing Datasets

The workflow discussed in Fig. 4 provides a generalized template to extract a P-S linkage from a collection of data points, where each data point includes both the final microstructure (measured or simulated) and the process parameters associated with it. In this section, we demonstrate the application of this workflow to analyses of the additive manufacturing simulation dataset described in “Additive Manufacturing Simulation Dataset.”.

The first step in the workflow is a data check to ensure that the data points are reliable and consistent. The additive manufacturing dataset described in “Additive Manufacturing Simulation Dataset” has been made publicly accessible [68] and consists of 1799 individual synthetic microstructures derived from simulations performed with varying AM processing parameters. A check of the data revealed some of the downloaded data to be corrupt (could not be opened), and a small number of microstructures showed unusually large grains that typically extended in length over the entire domain (in one direction). These instances were considered as outliers and eliminated from the analyses presented here. The total data for analyses reported in this paper consisted of 1599 structures.

The second step of the workflow addresses microstructure quantification. As mentioned earlier, this step is central to the extraction of transferrable materials knowledge needed in multiscale materials modeling efforts [69]. Although it is possible to select a number of different measures for the quantification of the microstructure in this work, the most logical choice here would be chord length distributions (CLDs). This is because the main microstructure feature of interest is the grain size and shape distributions and their anisotropy. Also, chord lengths connect directly to plastic properties of interest through established models such as the Hall-Petch models [70]. Furthermore, the computational cost of computing the CLDs is substantially low (order of the number of voxels in the microstructure); this is a significant criterion for this work as we intend to analyze a very large ensemble of microstructures. A chord is defined as a line segment within the microstructure contained within a single grain whose endpoints lie at grain boundaries. Relatedly, chord length is defined as the length of any such chord. CLD quantifies the probability of finding a chord of a specified length within a microstructure. CLDs can be resolved directionally [59] by treating chords exclusively in only one direction. Figure 5a illustrates the sampling of representative chords in X and Y directions in a voxelized microstructure, where the length of each chord is indicated by its color (shorter chords are blue and longer chords are yellow). In this plot (as well as the case study presented later), the microstructures are digitized and the length of a chord is simply taken as the number of pixels composing each chord. It should be noted that the chords in the edge grains are not included in the analyses (see Fig. 5a). Figure 5b shows corresponding CLDs resolved along X and Y directions of the micrograph. The broader CLD in the Y direction indicates that the grains are elongated in the Y direction compared to the X (see also Fig. 5a).

The variation in directionally resolved CLDs along the three reference orthogonal directions, among the four synthetic structures in Fig. 3, is presented in Fig. 6. The color designations of each microstructure in Fig. 3 are continued in the line colors of the corresponding distributions in Fig. 6. Solid or dotted lines within Fig. 6 correspond to differing (orthogonal) directions. As can be observed, there is a drastic difference between the CLDs in the volume fraction (i.e., frequency) of chords whose length is of the size of one voxel. This particular statistic relates to the fraction of sites (i.e., voxels) within the virtual microstructure retaining their initial (unique) site identifiers, presumably due to the absence of a sufficiently strong interaction with the heat source. Physically, this lack of interaction is due to the combined effects of the molten zone geometry and the overlap distance between successive passes of the heat source. The experimental analog of these synthetic phenomena is commonly referred to in the AM community as “lack-of-fusion” defects that result in porosity and regions of unmelted powder inclusions [27, 38, 51]. In Fig. 6, it is seen that the “lack-of-fusion” regions are somewhat equiaxed (i.e., no significant sensitivity to any directions associated with the CLDs for each microstructure), and decrease monotonically in extent from 25% to 5% for cases 1, 2, and 3, respectively, and were not observed at all in case 4. This is consistent with the increase in melt pool size, as one goes from case 1 to case 4.

The second most apparent variation between CLDs shown in Fig. 6 is in regard to the value of the highest frequency

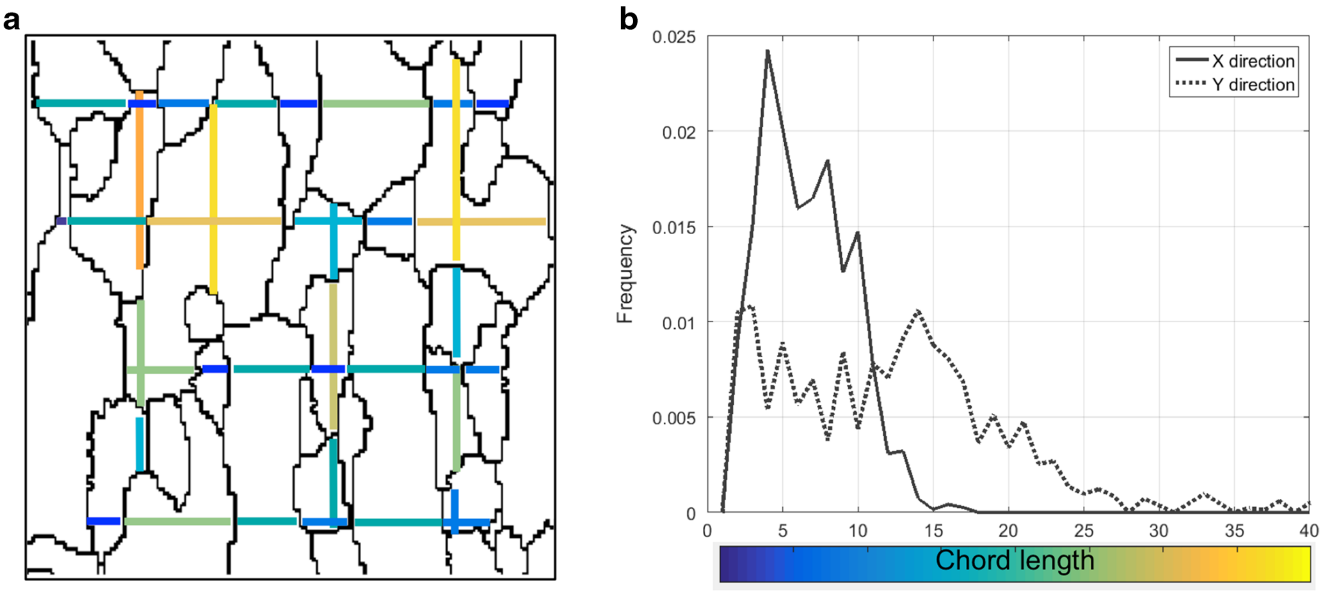


Fig. 5 Illustration of chord length distributions: **a** micrograph with example chords drawn in *X* and *Y* directions; **b** chord-length distributions in *X* and *Y* directions for the micrograph in **a** (Color figure online)

(corresponding to the most populous chords), excluding chord lengths of the size of one voxel. In general, it is seen that the chord length corresponding to the highest frequency is around 5 voxel lengths, but the frequency varies from 15% to 7% for the four cases shown. It should be noted that higher frequencies for the peak of the distribution generally correspond to narrower distributions (as each distribution is normalized such that the sum of the frequencies adds to one), implying that the grains within the microstructure are more similar to one another in both size and shape. Additionally, a slightly larger variation between the CLDs resolved in all three directions was observed for cases 1 and 2, in comparison to cases 3 and 4. This can be attributed to the fact that cases 3 and 4 implemented a crosshatching scan pattern, which is expected to produce more isotropic grain structures. Most interestingly, the tails of the distributions (capturing the decay in the distributions) vary significantly for the different microstructures, and are likely influenced by the changes in the size of the molten zone. In general, the parallel build pattern exhibits a sharper decay (narrower distribution of grain sizes) compared to the cross-hatching build pattern.

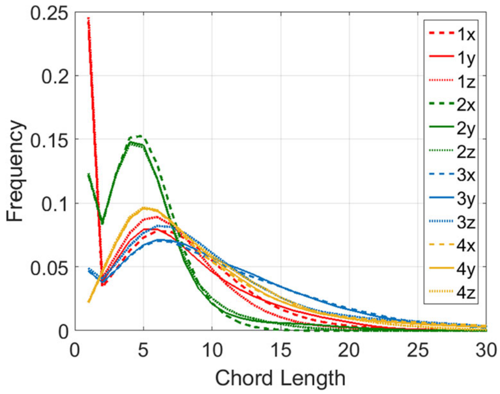


Fig. 6 Chord length distributions in three orthogonal reference directions (*X*, *Y*, and *Z*) for the four synthetic microstructures shown in Fig. 3 (Color figure online)

While some, but not all, experimental AM processing conditions can produce columnar grains that extend over several build layers, builds of this type would certainly produce heavily skewed *Z*-direction CLDs in comparison to those of the *X* and *Y* directions. However, in the simulations presented here, a maximum of no more than a 20% sublayer remelting was imposed. This was done to reduce the propensity for overwhelmingly biased *Z*-directional CLDs and produce microstructures which are in effect more reminiscent of powder-fed processes; e.g., directed energy deposition (DED) or laser-engineered net shaping (LENS) AM techniques. These processes often create builds with larger layer heights and significantly less remelting of prior layers than those of powder-bed systems [26].

As mentioned previously, CLDs are computed in each orthogonal direction (*X*, *Y*, and *Z*) and are then concatenated one after the other in a specific sequence to produce a large feature vector for each microstructure. The largest possible chord could, in theory, be equal to the dimensions of a microstructure in a given direction producing 300, 300, and 200 chord length statistics (in the *X*, *Y*, and *Z* directions, respectively). However, the maximum chord lengths in the ensemble of 1599 microstructures studied were identified to be 210, 203, and 90 voxels in *X*, *Y*, and *Z* directions, respectively. The CLDs in each direction were therefore truncated at these levels for all microstructures studied. The three CLDs for each microstructure were then concatenated to produce one large feature vector of 503 chord length statistics (the sum of 210, 203, and 90 chord length statistics obtained for each microstructure). It is unwieldy to utilize such high-dimensional representations in the practical extraction of P-S linkages. Therefore, a dimensionality reduction is performed as a next step of the workflow using principal component analysis (PCA).

PCA is a data-driven linear transformation to a new orthogonal framework that captures variance in a dataset with the minimum number of dimensions [18]. Although a number of options exist for dimensionality reduction, PCA was chosen here because it offers the following benefits: (i) it is a distance-preserving transformation, which allows a highly accurate and low-cost computation of a difference measure between any two microstructures using just the low-dimensional representations, (ii) it provides an orthogonal basis for representing the microstructure statistics which should lead to robust representations of process-structure-property (PSP) linkages, (iii) easy access to highly efficient computational toolsets for computing PCA on large datasets [67, 71, 72], (iv) a remarkable ability to recover the original high-dimensional microstructure statistics with only a handful of PC scores as long as the eigenvectors found in the PCA are stored [47], and (v) prior success in establishing robust PSP linkages in a wide range of multiscale materials phenomena [47, 73, 74]. Consequently, for each microstructure indexed by m, its feature vector (set of three chord length distributions representing a total of 503 chord length statistics) denoted by CLD_m can be approximately decomposed into a linear combination of basis vectors (called principal components) and weights (i.e., PC scores) [74] such that

$$\mathrm{CLD}_m \approx \sum_j^N \alpha_j^m A_j + A_0 \tag{4}$$

where A_j denotes the jth principal component (each A_j is a vector of 503 statistics), A_0 (also a vector of 503 statistics) is the mean CLD of all the microstructures in the dataset, and α_j^m denotes the jth PC score (for the microstructure indexed by m). More importantly, N denotes the truncation level in the orthogonal decomposition, and is selected based on the variance captured by the principal components. Since PCA prioritizes the (orthogonal) principal components in such a way that they sequentially explain the next greatest variance in the dataset, the value of N is typically very small. For the present case study, the percentage of variance explained by each principal component is depicted in Fig. 7. It is seen that the first four principal components explain well over 99% of all variance in the dataset. This is indeed a significant dimensionality reduction (from 503 to just 4).

Although the PCA described above results in dramatic and objective (data-driven) dimensionality reduction, it is also remarkable in its ability to reconstruct the original CLD. An example of this is demonstrated in Fig. 8. The PC scores for a randomly chosen 801th microstructure (i.e., $m = 801$) are $\alpha_j^{801} = (-0.0743, -0.1741, -0.0553, 0.0061)$. Each plot in Fig. 8 represents a systematic reconstruction of the original data point using one PC score at a time using Eq. (4). Note that the contributions drop dramatically as we add the higher principal components. The final reconstructed CLD (using four PC scores) and the original CLD of the microstructures are highlighted in the green box. The mean error between the reconstructed and the original CLD was 5.96e−4.

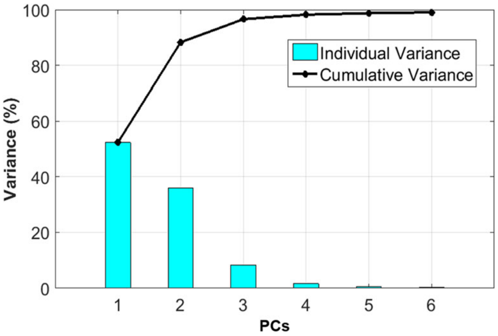


Fig. 7 Accumulated variance captured by the different principal components (PCs) retained in the reduced-order representation (Color figure online)

The ability to reconstruct the original data point is mainly due to the use of the stored principal components (basis vectors of the orthogonal decomposition in Eq. (4)). These basis vectors carry embedded information on the main differences between the individual data points and the ensemble mean in their original dimensionality. Figure 9 provides the plots of the basis vectors for the first four principal components obtained in the present case study. Each of the basis vector plots provides a “fingerprint” of the changes it will induce in the CLD. For example, an increase in the first PC score would significantly reduce the short chords and increase the medium-sized chords (~7 to ~25 pixels long). Similarly, an increase in the second PC score would significantly reduce the very short chords (about 1–2 pixels long) and increase the short chords (~2 to ~10 pixels long). It is also seen that the first three PC scores do not carry much information related to the anisotropy (differences in the CLDs in the three directions included in the analyses for each microstructure); only the fourth PC begins to carry some of this information.

Once the reduced-order representations are established, the next step of the workflow is to build a model using machine learning methods. The reduction of dimensionality from 503 to 4 has significantly reduced the difficulty associated with this step. The P-S linkage of interest in the present case study was extracted using a regression technique. Regression typically consists of four primary steps [75]: (1) defining dependent (output) and independent (input) variables, (2) identifying the form of the function (linear, parabolic, exponential,

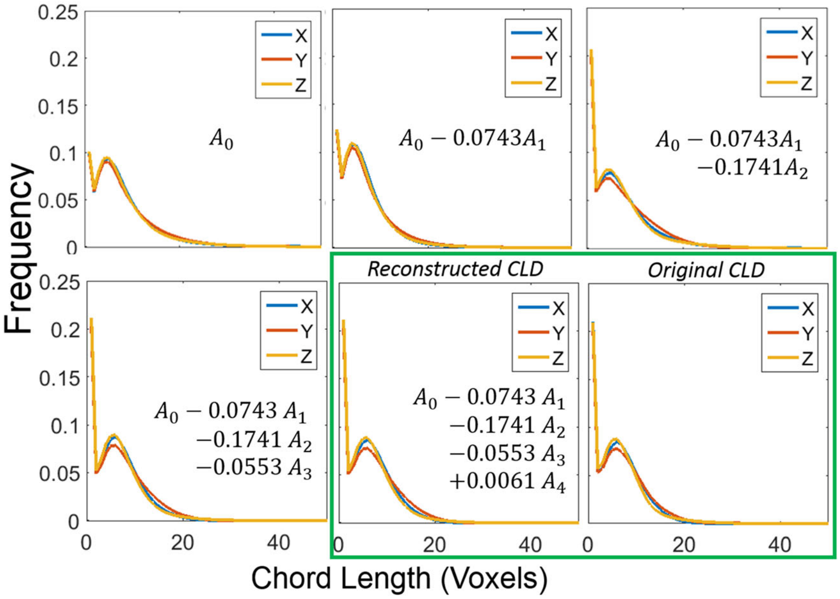


Fig. 8 Demonstration of the PC decomposition of CLD using the first four PC scores. Final reconstructed CLD and the original CLD are compared in the green frame (Color figure online)

etc.), (3) computing the regression function, and (4) performing error analysis.

In the present case study, the multivariate polynomial regression was used to establish a surrogate model for the PC scores of the final microstructures (treated as outputs of the model) as a function of the processing parameters identified in Table 1 (treated as inputs). Although a large number of choices exist, the polynomial regression was used as it is readily accessible in commonly used analytics packages [71, 76], computationally cheap, and has been seen to provide robust PSP linkages in prior work [73]. In many ways, multivariate polynomial regression is the simplest model form to explore as a first choice in model building. The goal here is to learn from the aggregated simulation datasets previously generated using the SPPARKS code [77], so that one may make predictions for new combinations of process parameters in the future without even having to run a SPPARKS simulation. The model being built takes the form:

$$\alpha_j = f_j(T, V, W, D, L, \mathrm{HAZ}) \qquad (7)$$

where α_j is the jth PC score, T is temperature, V is scanning velocity, W is melt pool width, D is melt pool depth, L is melt pool length, and HAZ is the width of the heat-affected zone. Once the PC scores are predicted, the CLD can be reconstructed as shown earlier in Fig. 8, using the stored values of A_j (including A_0).

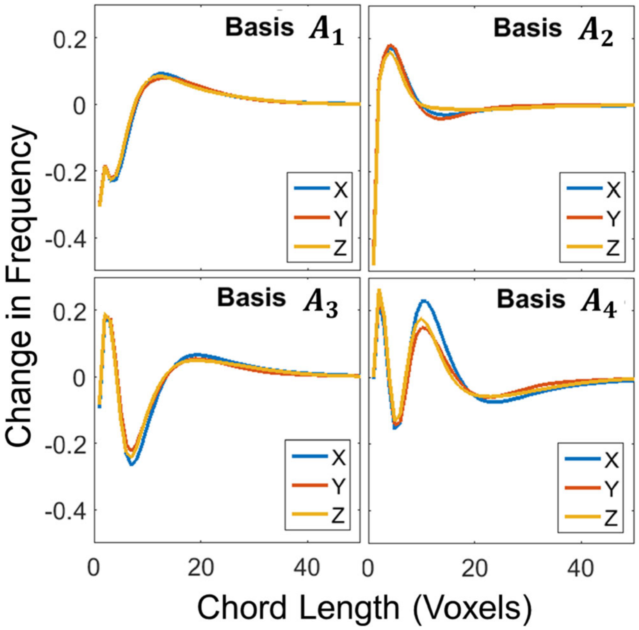


Fig. 9 Basis functions associated with first four PCs. *Vertical axes* represent the change in frequency from the mean value A_0 in Eq. 4 (Color figure online)

The models explored in this work were evaluated for accuracy using both a data splitting approach [66] and a leave-one-out cross validation (LOOCV). Data splitting allows an unbiased evaluation of the model for new inputs that were not utilized in the model development. For this purpose, the dataset is divided into non-overlapping calibration (training) and validation (test) sets. (Note that "calibration set" and "training set" as well as "validation set" and "test set" are used interchangeably in this work.) More specifically, the data points corresponding to values of variables $V = 7.5$ and $W = 70$, comprising a total of 684 synthetic structures, were selected as the validation set. The remaining dataset of 915 structures composed the calibration set used to build the models for each α_j. Note that the validation set was excluded even in the dimensionality reduction step. Thus, the validation conducted here is a validation of the entire workflow, including all the choices made for microstructure quantification (CLDs), dimensionality reduction (PCA), and model forms (multivariate polynomials). Although one can implement a number of other strategies to obtain the split between calibration and validation datasets, the above strategy was preferred in this study due to its ability to evaluate critically the model predictions for new inputs not included in the model building effort.

However, the simple data splitting strategy described above would often be inadequate in extracting a robust reduced-order model. This is because there is a danger that the model is likely over-fitted in the model building process. This often occurs because all measures of error typically reduce, when the numbers of parameters in the model are increased. It is therefore important to implement some means of a cross-validation along with the model building trials. For example, since there is only a limited set of distinct values for most input variables employed in this work (see Fig. 3b), indiscriminate use of higher-order polynomials would result in rank deficiency in the regression step. Suitable restrictions were placed on polynomial terms that would be explored in the model building process. In order to specifically avoid over-fitting, leave-one-out cross validation (LOOCV) was performed within the calibration set, when the model is trained using 914 data points and the error is computed for the data point set aside, then the process is repeated 915 times. Error estimates are therefore computed both with and without LOOCV for a very large set of multivariate polynomial combinations within the set constraints. A model was then selected based on the combination of lowest MAE and highest R^2 of the built model, and its cross-validation (CV) errors were also checked to ensure there is no overfitting, e.g., the CV errors are within the same range of errors as the built model. Table 2 summarizes these error measures for each accepted model for four PCs and their CV. The error measures presented in this work have been normalized to aid in their interpretation. It was decided to use maximum distance between data points (range of original data) as the normalization factor. As a final step, the best models identified in the model building step are then used to predict the outputs for the validation set (this contains the 684 data points that were hard-split from the full dataset). The errors computed for the validation set are summarized in Table 3.

The truncation level of PCs in the model as well as the degree of polynomial were varied to arrive at an optimized data-driven model. After numerous iterations between the steps of the workflow, it was identified that the first four principal components provided the best balance between the accuracy of the model and the number of features used. It is somewhat remarkable that specific values of acceptable error in the present study were not pre-determined. Rather, the specific models that exhibited the lowest errors (computed using the measures defined earlier) for the validation set, were identified and selected.

Table 2 Error metric values of the acceptable models for PC1, PC2, PC3, and PC4

PC	R^2	CV—R^2	MAE	CV-MAE	MAESDE	CV-MAESDE
PC_1	0.9945	0.9934	0.0118	0.0129	0.0106	0.0118
PC_2	0.9776	0.9729	0.0290	0.0317	0.0250	0.0278
PC_3	0.9292	0.9153	0.0426	0.0466	0.0308	0.0337
PC_4	0.8766	0.8529	0.0517	0.0563	0.0392	0.0430

Table 3 Error metric values of the validation of the models for PC1, PC2, PC3, and PC4

PC	Test—R^2	Test—MAE	Test—MAESDE
PC_1	0.9884	0.0168	0.0124
PC_2	0.9487	0.0378	0.0277
PC_3	0.9048	0.0440	0.0337
PC_4	0.7349	0.0688	0.0523

Resulting acceptable third-order polynomial models consisted of over 70 terms and coefficients. Although all 70+ terms of the model were used in this work, the authors acknowledge that optimization can be performed on the model to arrive at a more compact form with a smaller number of terms. Removing polynomial terms based on their coefficient decimals (e.g., smaller coefficients) resulted in increased mean error of predicting PC1 scores for test set from 0.0032 to 0.0132. Simply eliminating one term at a time also did not improve the results. Therefore, better optimization techniques are needed if one would like to obtain a model with fewer number of terms. However, since the computational cost of the reduced-order model produced in this work is minimal, there is no significant benefit to such pruning of the model terms.

Figure 10 shows parity plots of model fit for the first four PC scores and provides a visual depiction of the model accuracy relative to each PC score for both the training and test datasets. The diagonal straight line in the parity plots depicts a perfect match between actual data and predictions obtained using the surrogate model. Populations closer to the line are indicative of a more accurate prediction. The models for the first two PC scores show higher accuracy predictions than for the third and fourth PC scores. As discussed earlier, the contributions from the third and fourth PCs were generally significantly less than those of the first two PCs (see Fig. 8). Therefore, although the accuracies of the models for these two PC scores are lower than that of PC1 and PC2, they are still acceptable as they only provide a secondary or tertiary tuning of the reconstructed CLDs.

The P-S linkage for the simulated additive manufacturing microstructures presented here consists of a large table of coefficients of polynomials, as well as the basis functions and the mean value A_0. These tables are not presented here due to their size; however, the authors are willing to share the results upon request.

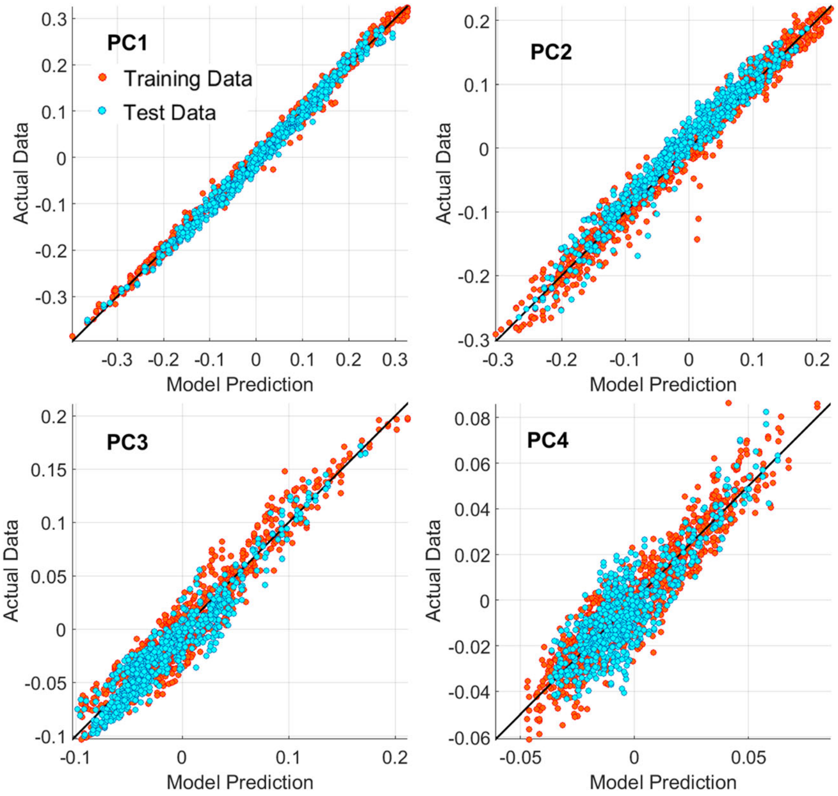


Fig. 10 PC model prediction vs. actual data for PC1, PC2, PC3, and PC4 values. *Red and blue circles* are calibration and validation data points, respectively (Color figure online)

In order to demonstrate the predictive capability of the model produced in this work, the reconstructed CLDs from the predicted PC scores were compared directly with the original CLDs. The distribution of the mean absolute errors from such comparisons is shown in Fig. 11. The highest normalized mean absolute error was 0.28, while the normalized mean error was less than 0.10 on average. The original and predicted CLDs for these two cases are shown as inserts in Fig. 11. Overall, it is shown that the data-driven models produce excellent predictions for the majority of data points evaluated. It

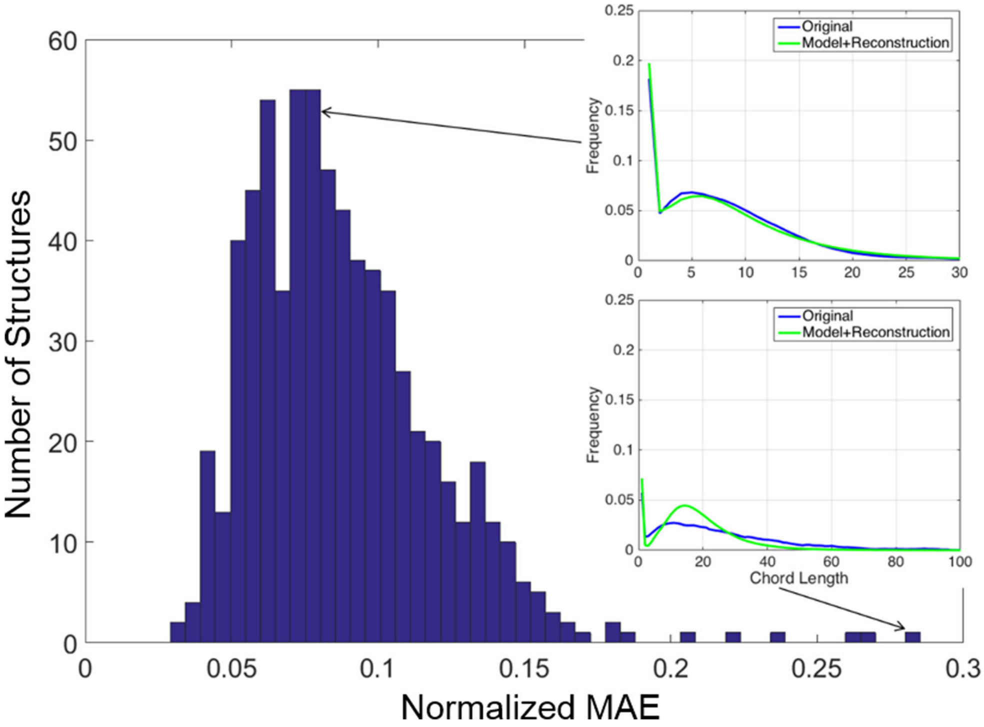


Fig. 11 Histogram of normalized mean of absolute error in reconstructed CLDs for microstructures in the test dataset. *Insets* show comparisons of original and reconstructed CLDs for microstructures with a typical mean average error (*top*) and the maximum error (*bottom*) (Color figure online)

should also be noted that these CLDs can also be used to reconstruct an actual microstructure. While not pursued here, such reconstructions have been successfully demonstrated in literature [78–80].

From Fig. 11, it is evident that the normalized MAE for some of the predicted CLDs in the test dataset were significant (in excess of 0.25). It was observed that the CLDs for these test cases were significantly outside the range of the CLDs used in the calibration dataset. This is most conveniently visualized in PC space. Figure 12 shows a scatter plot of all calibration and validation structures in the space of the first two PCs. Each point in this plot corresponds to one CLD. The red points indicate the training set. The points in the test set are colored based on the CLD reconstruction error using the MAE distribution shown in Fig. 11, e.g., the color bar in Fig. 12 corresponds to MAE values in Fig. 11. It can be clearly seen from this plot that the highest error occurs in the predictions that represent extrapolations from the calibration set. One should therefore be cognizant of this limitation of the reduced-order models and pay particular attention in the selection of the training dataset. Indeed, one of the main advantages of the protocols employed here is that the low-dimensional representation of the microstructure using only a few PC scores can provide this guidance.

In this work, our focus was exclusively on building a reduced-order model for the process-structure linkage. In prior work, we have demonstrated the viability of employing the same overall strategy for structure-property linkages [47]. Because of the use of a consistent framework for microstructure quantification and its low-dimensional representation in both classes of linkages, it should now be possible to establish interoperable process-structure and structure-property linkages. These reduced-order PSP linkages are central to the realization of the ambitious goals set forth in the MGI and which is implicitly necessary in ICME frameworks. This is because the reduced-order PSP linkages are the only practical way forward for conducting a rapid screening of extremely large design spaces (i.e., strategies for inverse solutions scanning large spaces). Keeping in mind that the main requirement in such efforts is objective (data-driven) guidance in down-selection of the design space, the authors offer reduced-order PSP linkages are the only practical way forward. Of course, one must keep in mind the limitations on the expected accuracy of these models, and develop and implement strategies to continuously refine and improve the reduced-order models with new data (both new simulations and new experiments). Indeed, the reduced-order models can serve as a natural bridge between the modeling and experimental efforts identifying not only new opportunities with high potential payoff (e.g., improved properties or performance) but also providing objective guidance on where (and how much) effort should be expended (e.g., improving fidelity mainly in the input ranges that lead to the desired changes in the microstructure).

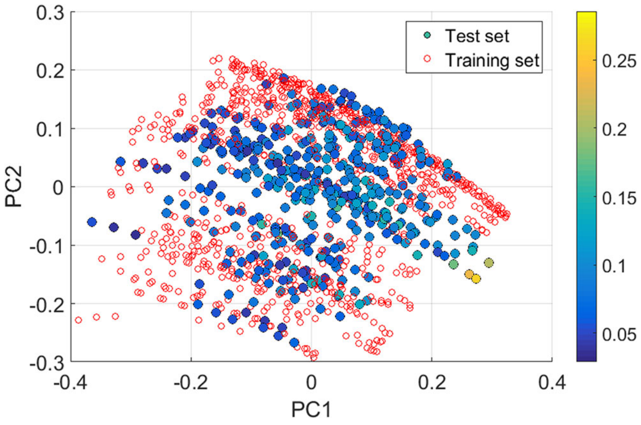


Fig. 12 Scatter plot of PC1 and PC2 of CLDs: training set is shown in *red color*; test set is colored based on the error in the CLD reconstruction (*color bar* corresponds to MAE values from Fig. 11) (Color figure online)

Conclusions

A novel workflow template is presented to extract process-structure linkages in microstructure evolution problems through the utilization of advanced data science techniques. The presented workflow is scalable and expandable and can be applied to a broad variety of microstructure evolution datasets. This workflow consists of four modular steps: (1) data pre-processing, (2) microstructure quantification, (3) dimensionality reduction, and (4) extraction and validation of process-structure linkages. Each step of the workflow allows selection and utilization of readily accessible codes from a large library of repositories.

The application of this template to quantify and predict synthetic additive manufacturing microstructures has been demonstrated. A publicly available set of simulated additive manufacturing microstructures has been created and shared to support exploration of AM processing parameters and the resultant grain-scale microstructural arrangements. The dataset consisted of 1599 unique microstructures and would have been extremely difficult to analyze effectively and comprehensively with conventional materials science approaches. Using the data-science approach presented here, chord length distribution calculations, principal component analysis, and multivariate polynomial regression were combined to produce a reliable reduced-order model, which was also cross-validated.

Although the process-structure linkage obtained here using a data science approach showed excellent results, the goal of this work was to establish a generic workflow to extract process-structure linkage for microstructure evolution problems. While the methods used in this case study are specific for the datasets presented, they can be altered to suit a variety of investigations and data types. Additionally, this workflow can be fully automated. This test case has demonstrated that exploration of process-structure linkages can be conducted most efficiently by exploiting modern data science-based workflows, the central feature of which is their automated consideration of a very large number of regression fits leading to a selection of surrogate models that meet the defined error and validation criteria.

Acknowledgments Evdokia Popova and Surya R. Kalidindi would like to acknowledge support from NIST grant 70NANB14H191. Xinyi Gong acknowledges support from NSF award 1435237. Ahmet Cecen acknowledges support from AFOSR award FA9550-12-1-0458. Sandia National Laboratories is a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

1. McDowell DL, Kalidindi SR (2016) The materials innovation ecosystem: a key enabler for the materials genome initiative. MRS Bull 41(04):326–337
2. Drosback M (2014) Materials Genome Initiative: Advances and Initiatives. JOM. 66: 334–335
3. Breneman CM, Brinson LC, Schadler LS, Natarajan B, Krein M, Wu K, & Xu H (2013) Stalking the materials genome: a data-driven approach to the virtual design of nanostructured polymers. Adv Funct Mater 23(46):5746–5752
4. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S et al (2013) Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. APL Mater 1(1): 011002
5. Holdren JP (2011) Materials genome initiative for global competitiveness. National Science and Technology Council OSTP. Washington, USA
6. Kalidindi SR, Medford AJ, McDowell DL (2016) Vision for data and informatics in the future materials innovation ecosystem. JOM 68(8):2126–2137
7. Panchal JH, Kalidindi SR, McDowell DL (2013) Key computational modeling issues in integrated computational materials engineering. Comput Aided Des 45(1):4–25
8. Pollock TM, Allison JE, Backman DG et al (2008) Integrated computational materials engineering: a transformational discipline for improved competitiveness and national security. Washington DC, The National Acamedies Press
9. Schmitz GJ, Prahl U (2014) ICMEg—the Integrated Computational Materials Engineering Expert Group—a new European coordination action. Integr Mater Manuf Innov 3(1):2
10. Spanos G, Allison J, Cowles B, Deloach J, Pollock T (2013) Integrated Computational Materials Engineering (ICME): implementing ICME in the aerospace, automotive, and maritime industries, Tech. rep., The Minerals, Metals & Materials Society (TMS)
11. Voorhees P and G Spanos (2015) 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)
12. Kalidindi SR (2015) Hierarchical materials informatics: Novel analytics for materials data. Elsevier
13. Krein MP, Natarajan B, Schadler LS et al (2012) Development of materials informatics tools and infrastructure to enable high throughput materials design. MRS Online Proceedings Library. **1425**: doi:10.1557/opl.2012.57.
14. Peurrung L, Ferris K, Osman T (2007) The materials informatics workshop: theory and application. JOM 59(3):50
15. Cebon D, Ashby MF (2006) Engineering materials informatics. MRS Bull 31(12):1004–1012
16. Liu Z-K, Chen L-Q, Rajan K (2006) Linking length scales via materials informatics. JOM 58(11):42–50
17. Rajan K (2005) Materials informatics. Mater Today 8(10):38–45
18. Kalidindi SR (2015) Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. Int Mater Rev 60(3):150–168
19. Kalidindi SR and De Graef M (2015) Materials data science: current status and future outlook. Annu Rev Mater Res 45:171–193
20. Kalidindi SR, Brough DB, Li S, Cecen A, Blekh AL, Congo FYP, et al (2016) Role of materials data science and informatics in accelerated materials innovation. MRS Bull 41(8):596–602
21. McDowell DL, Olson GB (2008) Concurrent design of hierarchical materials and structures. Sci Model Simul 15:207–240
22. Olson GB (2000) Pathways of discovery designing a new material world. Science 228(12):933–998
23. Olson GB (1997) Computational design of hierarchically structured materials. Science 277(29):1237–1242
24. Olson GB (1997) Systems design of hierarchically structured materials: advanced steels. J Computer-Aided Mater Des 4:143–156
25. McDowell DL, Panchal J, Choi HJ, Seepersad C, Allen J, et al (2009). Integrated design of multiscale, multifunctional materials and products. Butterworth-Heinemann
26. Frazier WE (2014) Metal additive manufacturing: a review. J Mater Eng Perform 23(6):1917–1928
27. Seifi M, Salem A, Beuth J, Harrysson O, et al (2016) Overview of materials qualification needs for metal additive manufacturing. JOM 68(3):747–764
28. Brackett, D., I. Ashcroft, and R. Hague (2011) Topology optimization for additive manufacturing. In Proceedings of the Solid Freeform Fabrication Symposium. Austin, TX
29. Holesinger TG, Carpenter JS, Lienert TJ, Patterson BM, Papin PA, Swenson H & Cordes NL (2016). Characterization of an aluminum alloy hemispherical shell fabricated via direct metal laser melting. JOM, 68(3), 1000-1011
30. Thijs L, Verhaeghe F, Craeghs T, Van Humbeeck J, & Kruth JP (2010) A study of the microstructural evolution during selective laser melting of Ti–6Al–4V. Acta Mater 58(9):3303–3312
31. Dehoff RR, Kirka MM, Sames WJ, Bilheux H, Tremsin AS, Lowe LE, & Babu SS (2015) Site specific control of crystallographic grain orientation through electron beam additive manufacturing. Mater Sci Technol 31(8):931–938
32. Niendorf T, Leuders S, Riemer A, Brenne F, Tröster T, Richard HA, & Schwarze D (2014) Functionally graded alloys obtained by additive manufacturing. Adv Eng Mater 16(7):857–861
33. Martukanitz R, Michaleris P, Palmer T, DebRoy T, Liu ZK, Otis R et al (2014) Toward an integrated computational system for describing the additive manufacturing process for metallic materials. Addit Manuf 1-4:52–63
34. Witherell P, Feng S, Simpson TW, Saint John DB et al (2014) Toward metamodels for composable and reusable additive manufacturing process models. J Manuf Sci Eng 136(6):061025
35. King WE, Anderson AT, Ferencz RM, Hodge NE, Kamath C, Khairallah SA, & Rubenchik AM (2015) Laser powder bed fusion additive manufacturing of metals; physics, computational, and materials challenges. Appl Phys Rev 2(4):041304
36. Huang Y, Leu MC, Mazumder J, & Donmez A (2014) Additive manufacturing: current state, future potential, gaps and needs, and recommendations. J Manuf Sci Eng 137(1):014001
37. Regli W, Rossignac J, Shapiro V & Srinivasan V (2016) The new frontiers in computational modeling of material structures. Comput Aided Des 77:73–85

38. Kamath C (2016) Data mining and statistical inference in selective laser melting. Int J Adv Manuf Technol 86(5–8):1659–1677
39. Garcia AL, Tikare V, Holm EA (2008) Three-dimensional simulation of grain growth in a thermal gradient with non-uniform grain boundary mobility. Scr Mater 59(6):661–664
40. Madison JD, Tikare V, Holm EA (2012) A hybrid simulation methodology for modeling dynamic recrystallization in UO 2 LWR nuclear fuels. J Nucl Mater 425(1):173–180
41. Tikare V, Hernandez-Rivera E, Madison JD, Holm EA, Patterson BR, & Homer ER (2013) Hybrid models for the simulation of microstructural evolution influenced by coupled, multiple physical processes, Brigham Young University, Provo, UT; Sandia National Laboratories (SNL-NM), Albuquerque, NM (United States)
42. Plimpton S, Battaile C, Chandross M, Holm L, Thompson A, Tikare V, & Slepoy A (2009) Crossing the mesoscale no-man's land via parallel kinetic Monte Carlo. Sandia National Laboratory
43. Holm EA, Battaile CC (2001) The computer simulation of microstructural evolution. JOM 53(9):20–23
44. Rodgers TM, J Madison, and V Tikare (2016) Predicting mesoscale microstructural evolution in electron beam welding. JOM (5): 1419–1426
45. Rodgers TM, J Madison, and V Tikare (2016) Simulation of metal additive manufacturing microstructures using kinetic Monte Carlo. Computational Materials Science - submitted for review
46. Parimi LL, Ravi GA, Clark D, & Attallah MM (2014) Microstructural and texture development in direct laser fabricated IN718. Mater Charact 89:102–111
47. Gupta A, Cecen A, Goyal S, Singh AK, & Kalidindi SR (2015) Structure–property linkages using a data science approach: application to a non-metallic inclusion/steel composite system. Acta Mater 91:239–254
48. Steinmetz P, Yabansu YC, Hötzer J, Jainta M, Nestler B, & Kalidindi SR (2016) Analytics for microstructure datasets produced by phase-field simulations. Acta Mater 103:192–203
49. Raabe D (2002) Cellular automata in materials science with particular reference to recrystallization simulation. Annu Rev Mater Res 32(1):53–76
50. Bernacki M, Resk H, Coupez T (2009) Finite element model of primary recrystallization in polycrystalline aggregates using a level set framework. Model Simul Mater Sci Eng 17(6):064006
51. Cunningham R, Narra SP, Ozturk T, Beuth J, & Rollett AD (2016) Evaluating the effect of processing parameters on porosity in electron beam melted Ti-6Al-4V via synchrotron X-ray microtomography. JOM 68(3):765–771
52. Rama P, Liu Y, Chen R, Ostadi H et al. (2010) An X-ray tomography based lattice Boltzmann simulation study on gas diffusion layers of polymer electrolyte fuel cells. J Fuel Cell Sci Technol 7(3):031015
53. Kalidindi SR, Niezgoda SR, Salem AA (2011) Microstructure informatics using higher-order statistics and efficient data-mining protocols. JOM 63(4):34–41
54. Adams BL, Kalidindi S, Fullwood DT (2013) Microstructure-sensitive design for performance optimization. Butterworth-Heinemann
55. Wheeler D, Brough D, Fast T, Kalidindi S, & Reid A (2014) PyMKS: Materials Knowledge System in Python (Figshare, 2014). doi:10.6084/m9.figshare.1015761
56. Agrawal A, Deshpande PD, Cecen A, Basavarsu GP, Choudhary AN & Kalidindi SR (2014) Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. Integr Mater Manuf Innov 3(8):1–19
57. Lu B, Torquato S (1992) Lineal-path function for random heterogeneous materials. Phys Rev A 45(2):922–929
58. Torquato S, Lu B (1993) Chord-length distribution function for two-phase random media. Phys Rev E 47(4):2950
59. D Turner SN, Kalidindi SR (2016) Efficient computation of the angularly resolved chord length distributions and lineal path functions in large microstructure datasets. Modelling and Simulation in Materials Science and Engineering doi:10.1088/0965-0393/24/7/075002.
60. Mardia KV, Kent JT, Bibby JM (1980) Multivariate analysis (probability and mathematical statistics). Academic Press, London
61. Fodor IK (2002) A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory 9:1–18
62. Hyvärinen, A. (1999) Survey on independent component analysis. Neural Computing Surveys 2(4):94–128
63. Quinlan JR (1992) Learning with continuous classes. In 5th Australian joint conference on artificial intelligence. Singapore
64. Hearst MA, Dumais ST, Osuna E, Platt J, & Scholkopf B (1998) Support vector machines. IEEE Intell Syst Appl 13(4):18–28
65. Shao J (1993) Linear model selection by cross-validation. J Am Stat Assoc 88(422):486–494
66. Snee RD (1977) Validation of regression models: methods and examples. Technometrics 19(4):415–428
67. Pedregosa F, Varoquaux G, Gramfort et al. (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12(Oct):2825–2830
68. Rodgers T (2015) Exploration of process-structure linkages in simulated additive manufacturing microstructures. Harvard Dataverse V1. doi:10.7910/DVN/KJMK9Z
69. Kalidindi SR, Gomberg JA, Trautt ZT & Becker CA (2015) Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets. Nanotechnology 26(34):344006
70. Armstrong RW, Codd I, Douthwaite RM, & Petch NJ (1962) The plastic deformation of polycrystalline aggregates. Philos Mag 7(73):45–58
71. Team RC (2013) R: a language and environment for statistical computing 2013 (Global Biodiversity Information Facility, Copenhagen, Denmark)
72. Berthold MR, Cebron N, Dill F, Gabriel TR et al. (2009) KNIME—the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD Explor Newsl 11(1):26–31
73. Çeçen A, Fast T, Kumbur EC, & Kalidindi SR (2014) A data-driven approach to establishing microstructure–property relationships in porous transport layers of polymer electrolyte fuel cells. J Power Sources 245:144–153
74. Niezgoda SR, Yabansu YC, Kalidindi SR (2011) Understanding and visualizing microstructure and microstructure variance as a stochastic process. Acta Mater 59(16):6387–6400
75. Sinha P (2013) Multivariate polynomial regression in data mining: methodology, problems and solutions. Int J Sci Eng Res 4(12):962–965
76. Jones E, Oliphant T, Peterson P (2015) SciPy: Open source scientific tools for Python, 2001. URL http://www.scipy.org . 73: p. 86
77. Plimpton S, Thompson A, Slepoy A (2012) SPPARKS kinetic Monte Carlo simulator. http://spparks.sandia.gov/
78. Schlüter S, Vogel H-J (2011) On the reconstruction of structural and functional properties in random heterogeneous media. Adv Water Resour 34(2):314–325
79. Yeong CLY, Torquato S (1998) Reconstructing random media. Phys Rev E 57(1):495–506
80. Worlitschek J, Hocker T, Mazzotti M (2005) Restoration of PSD from chord length distribution data using the method of projections onto convex sets. Part Part Syst Charact 22(2):81–98