Full length article

Extraction of reduced-order process-structure linkages from phase-field simulations

Yuksel C. Yabansua, Philipp Steinmetzc, Johannes Hötzer c, d, Surya R. Kalidindi a, b, *, Britta Nestler c, d

a George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
b School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
c Institute for Applied Materials, Karlsruhe Institute of Technology (KIT), Haid-und-Neu-Str. 7, 76131 Karlsruhe, Germany
d Institute of Materials and Processes, Karlsruhe University of Applied Sciences, Moltkestrasse 30, 76133 Karlsruhe, Germany

A R T I C L E   I N F O
Article history:
Received 24 June 2016
Received in revised form 26 September 2016
Accepted 28 October 2016
Available online 11 November 2016

A B S T R A C T
Phase-field simulations have achieved notable success in capturing characteristic details of microstructure evolution in directional solidification of ternary eutectic alloys. In spite of the impressive advances in high performance computations, phase-field simulations for most practical problems in materials design are resource intensive because of the need to incorporate multiple physical fields over large-scale three-dimensional domains. There is, therefore, a need to learn and capture the underlying materials knowledge embedded in the results produced by such expensive simulations, and facilitating an easy transferability to new problems of interest. This paper demonstrates the viability of extracting the salient process-structure linkages from phase-field simulations, while casting them in forms amenable for a rapid and efficient exploration of a relatively large process space. The presented framework is based on low dimensional representation of material structure obtained through principal component analysis (PCA) of 2-point spatial correlations.

© 2016 Acta Materialia Inc. Published by Elsevier Ltd. All rights reserved.

1. Introduction

New and improved materials have been at the core of almost all of the advanced technologies introduced throughout human history. However, the pace of materials innovation lags significantly behind the rate of innovation in many other fields. Today, introducing a new material takes around 20 years from its discovery to its deployment in a commercial product [1–4]. The central impediment comes from the need to understand and utilize the important relationships between the material structure at multiple length scales, its evolution through different manufacturing processes, and the associated macroscale properties. These are generally referred as process-structure-property (PSP) linkages [5–17]. The conventional approaches used for extracting the important PSP linkages needed in materials innovation have proven to be slow, expensive and effort intensive. Integrated Computational Materials Engineering (ICME) [4] paradigm and Materials Genome Initiative (MGI) [1] propose to dramatically accelerate material innovation through utilization of data science tools in conjunction with multiscale experiments and multiscale simulations.

Phase-field simulations have played a pivotal role in simulating the microstructure evolution in several phenomena such as thin film growth [18], grain growth [19,20], recrystallization [21], and directional solidification [22–24]. In this study, we focus on building process-structure linkages from phase-field simulations of directionally solidified ternary eutectic alloys. During the directional solidification of a eutectic system, two or more solid phases evolve simultaneously from the melt. In ternary eutectic alloys, three phases form, resulting in various spatial patterns, depending on the process parameters. The overall (effective) mechanical properties are indeed influenced by these spatial patterns [25–27]. Therefore, it is of interest to quantify the influence of the process parameters on the evolving microstructure. One alloy of scientific interest is the ternary eutectic system Al-Ag-Cu [28–32], due its favorable physical properties. As an example, the influence of sedimentation in directional solidification is being studied in experiments conducted on the international space station (ISS) as...
part of the Solidification along a Eutectic path in Ternary Alloys (SETA) program [33]. Phase-field models have been used successfully for simulating these experiments. Several investigations of the pattern formation process in the Al-Ag-Cu system have been conducted with the phase-field method [24,34–38], which include three-dimensional simulations [34] as well as large-scale simulations [24,35]. The influence of the melt concentration in the vicinity of the ternary eutectic point on the pattern formation was also investigated [36]. Although the influence of the process parameters on the microstructure pattern formation has been studied in Al-Alg-Cu, there do not yet exist low cost surrogate models (also called metamodels or linkages) capturing these relationships.

A rigorous quantification of the microstructure is central to extracting a high value process-structure linkage. Although a number of measures such as volume fraction [39], average grain size [40], number of nearest neighbors [24], and pair correlations [41] have been employed in literature to quantify the material structure, only n-point spatial correlations (also referred as n-point statistics) [42–44] give the most systematic and complete information about the material internal structure. For example, one-point statistics reflect the probability density of finding a local state of interest at any single location selected in the representative volume of the material microstructure. At the next level, two-point statistics describe the probability density of finding local states \( h \) and \( h' \) separated by a prescribed vector \( r \). In prior work [14,37,38,45,46], two-point statistics were demonstrated to be successful in objective classification of microstructure. Higher-order statistics can be defined in a similar manner. However, the number of measures included in these representations (which reflects the dimensionality of structure representation) increases exponentially. Hence, a reduced-order representation of microstructure is required. For this purpose, principal component analysis (PCA) is used [14,37,38,45,46]. PCA essentially involves a linear transformation of the coordinate frame, where the available data could be viewed to capture the salient differences in the datapoints generating the calibration datasets (red box), (ii) reduced order representation of the microstructure (green box), and (iii) building the process-structure linkages (blue box). In many respects, this workflow has the same main components as the workflow presented in our prior work for establishing structure-property linkages [14]. One of the main benefits of such workflow templates is that they provide modularity (for example, the individual components in Fig. 1 can employ any of the different options available for that component), while promoting higher levels of interoperability needed for standardization, automation, and scale-up. Briefly, these templated workflows start by identifying (or specifying) the process parameters of interest and their ranges. This information is then utilized to generate a suitable calibration dataset using a previously established approach in the domain of practice (in the present case, these are the phase-field models [24,37]). After generating the calibration dataset, the microstructure is quantified through spatial statistics and this information is projected onto a reduced-order space to obtain a low dimensional representation. In the final step, the process space and the reduced order representation of microstructure are mapped to arrive at the needed process-structure linkages through machine learning approaches and cross validation techniques.

2. Phase-field model and calibration data sets

2.1. Phase-field model

To model the microstructure evolution of ternary eutectics during directional solidification, a thermodynamically consistent phase-field model based on the Grand potential approach is employed [47,48]. The evolution equations for the \( N \) phase-fields \( \phi_a \) in the vector \( \phi \) are calculated from the total energy in the system. Based on the Grand potential functional [47,48] and applying an Allen-Cahn type approach, the evolution equations of the form

\[
\frac{\partial \phi_a}{\partial t} = -\mathbf{\tau} \frac{\partial G}{\partial \phi_a} - \frac{1}{\mathbf{\tau}} \frac{\partial \phi_a}{\partial t} - \sum_{\beta=1}^{N} \frac{h_{la}(\phi)}{\partial \phi_a} \partial_{\phi_a} - \lambda, \quad \lambda = \frac{1}{N} \sum_{a=1}^{N} \mathbf{r}_{1a} + \mathbf{r}_{2a}.
\]

are derived with the Lagrange multiplier \( \lambda = \frac{1}{N} \sum_{a=1}^{N} \mathbf{r}_{1a} + \mathbf{r}_{2a} \) to ensure the constraint \( \sum_{a=1}^{N} \phi_a = 1 \).

The interface kinetics is described by the parameter \( \tau \), coupling the different time scales in the simulation. The form of the interface is modeled by the gradient energy density \( \mathcal{G} \) as well as the potential energy \( \mathcal{E} \), and its thickness is related to \( \epsilon \). The driving forces for the phase transitions in Eq. (1) are comprised in the differences of the Grand potentials. These potentials \( \psi_\beta \) are interpolated with \( h_\beta \) [49].

The evolution of the temperature \( T \) during the directional solidification process, is calculated analytically according to the simplified condition

\[
\frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \left( T_0 + G(z - vt) \right) = -G(v),
\]

where \( T_0 \) is the base temperature, \( G \) is the gradient, \( z \) is the spatial position in the growth direction, \( v \) is the velocity, and \( t \) is the time. The evolution of the chemical potentials \( \mu \) is derived from the energy functional
The scalar dynamic CALPHAD databases has been described in earlier work [55]. Linkages. (For interpretation of the references to colour in this datasets. The green components address the reduced order representation of microstructure. The blue components address the establishment of the desired process-structure linkages. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The advantages of coupling phase-field simulations to thermodynamic databases CALPHAD databases [51–53], we determined 800 × 800 as the optimum domain (grid) size parallel to the growth front for achieving both computational efficiency as well as statistically similar volume elements (SSVEs). Thus, all simulations here were assigned the same domain size of 800 × 800 cells parallel to the growth front.

Different combinations of the velocity of temperature gradient and the liquid concentration are considered in the phase-field models to produce the datasets used in this study for the calibration and validation of the desired process-structure linkages. The concentration of Al in the liquid is kept the same for all simulations. Hence, the simulations contain only two independent process parameters: (i) velocity of the temperature gradient, and (ii) concentration of Ag in the melt (or Cu due to redundancy). Although only two process parameters were included in this study for demonstrating the viability of establishing the process-structure linkages through data science methods, it is emphasized that the framework is quite general and extensible for including other process parameters (such as the temperature gradient) as more data becomes available. Selected microstructures (cross sections parallel to growth front or normal to the direction of solidification) from a set of 37 different total phase-field simulations are shown in Fig. 2. The simulations in the top row refer to the same concentration values but different velocity values. For higher velocity values, the chains (phases Al2Cu and Al3Cu) become thicker and the volume fraction of red phase (AI) decreases. In the bottom row, the results illustrate the patterns with similar velocities but different concentrations. It can be seen that as we increase the concentration of Ag in the liquid, the volume fraction of red phase (Al) increases. Also the chains become shorter and thinner. Even though examining these simulation results in a qualitative manner gives an insight about the effect of process parameters on the microstructure, a rigorous quantification of microstructure is necessary for linking these process parameters to microstructures.

\[
g_{\text{c}}(\mathbf{c}, \mathbf{T}) = \langle \mathbf{c}, \mathbf{Z}_a(T)\mathbf{c} \rangle + \langle \mathbf{c}, \xi_a(T) \rangle + X_a(T),
\]

where \(\langle \cdot, \cdot \rangle\) denotes a scalar product [54].

The main focus of this paper is to build process-structure linkages for ternary eutectic alloys by applying machine learning and other data science approaches on the results aggregated from phase-field simulations. In prior work on quantification of ternary eutectic microstructures [37], we determined 800 × 800 as the optimum domain (grid) size parallel to the growth front for achieving both computational efficiency as well as statistically similar volume elements (SSVEs). Thus, all simulations here were assigned the same domain size of 800 × 800 cells parallel to the growth front.

Different combinations of the velocity of temperature gradient and the liquid concentration are considered in the phase-field models to produce the datasets used in this study for the calibration and validation of the desired process-structure linkages. The concentration of Al in the liquid is kept the same for all simulations. Hence, the simulations contain only two independent process parameters: (i) velocity of the temperature gradient, and (ii) concentration of Ag in the melt (or Cu due to redundancy). Although only two process parameters were included in this study for demonstrating the viability of establishing the process-structure linkages through data science methods, it is emphasized that the framework is quite general and extensible for including other process parameters (such as the temperature gradient) as more data becomes available. Selected microstructures (cross sections parallel to growth front or normal to the direction of solidification) from a set of 37 different total phase-field simulations are shown in Fig. 2. The simulations in the top row refer to the same concentration values but different velocity values. For higher velocity values, the chains (phases Al2Cu and Al3Cu) become thicker and the volume fraction of red phase (Al) decreases. In the bottom row, the results illustrate the patterns with similar velocities but different concentrations. It can be seen that as we increase the concentration of Ag in the liquid, the volume fraction of red phase (Al) increases. Also the chains become shorter and thinner. Even though examining these simulation results in a qualitative manner gives an insight about the effect of process parameters on the microstructure, a rigorous quantification of microstructure is necessary for linking these process parameters to microstructures.

\[
\frac{\partial \mathbf{c}}{\partial t} = \left[ \sum_{a=1}^{N} h_a(\phi) \mathbf{Z}_a^{-1} \right]^{-1} \left( \nabla \cdot (\mathbf{M}(\phi, \mathbf{c}, \mathbf{T}) \nabla \mathbf{c} \right)
\]

\[
- \sum_{a=1}^{N} h_a(\phi) \mathbf{Z}_a^{-1} \left( \mathbf{J}_{al}(\phi, \mathbf{c}, \mathbf{T}) \right)
\]

\[
\xi_a(T) = \begin{bmatrix}
-A_a + C_a & C_a \\
C_a & B_a + C_a
\end{bmatrix},
\]

\[
X_a(T) = C_a + F_a.
\]
3. Reduced order representation of microstructure

The selection of suitable microstructure measures plays an important role in establishing reliable, robust, process-structure-property relationships. As noted earlier, most of the currently used approaches in microstructure rely on an intuitive selection of the microstructure measures [24,60,61]. Even though, these measures can provide valuable insights into the role of processing/manufacturing history on the microstructure and its effect on property, they do not lead to a systematic, broadly applicable approach that could potentially be templated and automated.

3.1. Microstructure function

The quantification of microstructure starts with the representation of spatial distribution of the local states in the representative volume of the material microstructure. Microstructure function [44,62], \( m(x,h) \), denotes the probability of finding a distinct local state, \( h \), at a prescribed position \( x \) in the microstructure. The local state, \( h \), can represent several variables that distinguish the local attributes of the material such as the local chemical composition [63], the local crystal lattice orientation [9,10], and/or the local curvature [64], among other variables. The corresponding velocity and concentration values are given on top of each microstructure.

\[
\begin{align*}
\nu &= 0.0525 \\
c &= 0.237
\end{align*}
\]

\[
\begin{align*}
\nu &= 0.0659 \\
c &= 0.237
\end{align*}
\]

\[
\begin{align*}
\nu &= 0.0844 \\
c &= 0.237
\end{align*}
\]

\[
\begin{align*}
\nu &= 0.0774 \\
c &= 0.237
\end{align*}
\]

\[
\begin{align*}
\nu &= 0.0787 \\
c &= 0.2385
\end{align*}
\]

\[
\begin{align*}
\nu &= 0.0780 \\
c &= 0.2430
\end{align*}
\]

Fig. 2. Microstructures obtained from phase-field simulations and used for the data analysis framework. The simulations in the top row correspond to fixed concentrations and different velocities, whereas the simulations in the bottom row have similar velocity values but different concentration values. Microstructures are extracted as cross-sections normal to the solidification direction. The corresponding velocity and concentration values are given on top of each microstructure.

3.2. Statistical representation of microstructure

When a microstructure is considered as a stochastic process [45,46], then every micrograph or three dimensional data set should be treated as one potential outcome of this process. The major drawback in utilizing microstructure function directly to describe the structural features lies in the lack of a natural origin for the spatial variable \( x \) in the microstructure function. If we invoke translational invariance of microstructures, then only the relative...
Fig. 3. Statistical representation of an exemplary Al-Ag-Cu ternary eutectic alloy microstructure obtained via phase-field simulations. (a) Cross section of the simulated microstructure under directional solidification conditions, (b) autocorrelation corresponding to red phase and (c) crosscorrelation of red and blue phases. The images (d) and (e) show the retained statistics as 2-D plots after the truncation to the region with the dashed black lines in (b) and (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The considered material system has three distinct local states, shown as red (Al), blue (Al2Cu) or green (Ag2Al) in Fig. 3. For a three-phase material, the total number of possible two-point correlations are nine (resulting from the permutation of h and h’ in Eq. (7)) [37,38,45,46,65]. Three of these correlations are auto-correlations (with h = h’) and the remaining six correlations are cross-correlations (with h ≠ h’). In our prior work on ternary eutectic alloys [37], the autocorrelation of red phase and the crosscorrelation of red and blue phases were studied because the red and blue phases exhibited the largest ranges in the volume fractions (when comparing all of the microstructures included in this study). A similar approach is followed in this paper. For the statistical representation of microstructures in this study, autocorrelations of red and blue phases and the crosscorrelation between them were computed and analyzed. However, other correlations can be included in future studies as needed.

A cross section of a phase-field simulation for directional solidification can be seen in Fig. 3(a). The domain size parallel to the growth front is 800 × 800. The diagrams in Fig. 3(b) and (c) display the three dimensional surface plots of the autocorrelation of red phase and the crosscorrelation of red and blue phases for this microstructure. Attention must be paid to the x and y axes of the plots which represent the x and y components of the vector space where two-point statistics are defined. The number of two-point statistics for autocorrelation (or crosscorrelation) is S which is equal to the number of pixels in the microstructure (S = 800^2 = 640000). The center of the plot in autocorrelation (corresponds to r = 0 vector) displays the volume fraction (in this case, for the red phase). For crosscorrelation, the value at the center of the plot is simply zero since two distinct phases cannot exist in the same cell. In other words, the probability of finding red and blue phases in the same cell is zero.

Even though the entire set of two-point statistics is calculated in a very computationally efficient manner through DFT operations, the statistics corresponding to large vectors do not contain any salient information on the microstructure. The surface plots of autocorrelation and crosscorrelation are shown in parts (b) and (c) of Fig. 3, respectively. The surface plot of autocorrelation in Fig. 3(b) shows that the statistics decay to a value equal to the square of volume fraction of red phase and fluctuate around it. In the surface plot of crosscorrelation in Fig. 3(c), the corresponding asymptotic value is the product given by the volume fractions of blue and red phases. The regions of small fluctuations for both autocorrelation and crosscorrelation can be identified between the dashed and solid black lines in the surface plots shown in parts (b) and (c) of Fig. 3. Hence, the two-point statistics can be conveniently truncated without a significant loss of information. The dashed black lines in surface plots (parts (b) and (c) Fig. 3) bound the regions representing the retained statistics. A two dimensional representation of the retained autocorrelation and crosscorrelation plots can be seen in parts (d) and (e) of Fig. 3, respectively. The same truncation size is employed for the autocorrelation of blue phase as well. For this study, the truncation level is determined as vector size of 200 cells in both x and y directions. The number of dimensions for the statistical representation of microstructure is decreased from 3^640000 = 1920000 to 3^401^2 = 482403. We remark that the
factor 3 is due to utilizing two autocorrelations and one cross-correlation for the statistical representation of microstructure.

3.3. Reduced order representation of microstructure

Although the statistical representations presented in previous section capture efficiently the salient features of microstructures, it spans an unwieldy 35 dimensional space. Even after truncation of large vectors, the dimensionality of the microstructure representation is still too large for building practically useful process-structure linkages. In prior work \cite{14,37,38,45,46}, principal component analysis (PCA) has been employed successfully for building PSP linkages as well as classification of microstructures. PCA transforms high dimensional data onto a new orthogonal coordinate system with components ordered from highest to lowest variance. The most important advantage of employing PCA is in the reduction of the dimensions required to represent the data in the new space. In many applications, the statistical representation of microstructure can be visualized in the principal component space with only a handful of components.

In this study, two-point statistics of microstructure are projected onto principal component space to obtain a low dimensional representation. Two-point statistics of the kth microstructure image in a selected ensemble can be written in principal component space as

\[
 f_r^{(k)} = \sum_{i=1}^{\text{min}(K,R)} a_i^{(k)} \varphi_{ir} + \bar{f}_r
\]

where \( R \) is the total number of dimensions in the retained statistics of microstructure and \( K \) is the total number of realizations (or microstructures) in the ensemble. \( \varphi_{ir} \) are the components of the orthogonal set of directions found by PCA (or basis vectors) and \( a_i^{(k)} \) are principal component weights (or scores) for kth microstructure in the ensemble. \( \bar{f}_r \) represents ensemble averaged two-point statistics. PCA performs a distance preserving orthogonal decomposition on each data point (in this case the microstructure statistics) where each point is represented by a series summation. Each term in the series is a product of a principal component weight or PC score (\( a_i \)) and basis vector (\( \varphi_{ir} \)). These terms can be found from a singular value decomposition (SVD) on the data matrix \( X \) where rows and columns represent microstructures and dimensions of their statistical representation, respectively (\( X = U \Sigma V^T \)). The basis vectors, \( \varphi_{ir} \) and PC scores, \( a_i \), correspond to \( V \) and \( U \Sigma \), respectively.

On the other hand, the eigenvalues, \( b_i \), are related to the singular values (\( b_i = \Sigma_i^2 \)) due to the fact that \( \Sigma \) is a diagonal matrix and represent the importance of each principal component \cite{66,67}. The first principal component captures the highest variance in the ensemble (corresponds to the first and largest eigenvalue). By retaining only a relatively small number of components in the principal component space, an objective reduced order representation of microstructure statistics can be written as:

\[
 f_r^{(k)} = \sum_{i=1}^{K} a_i^{(k)} \varphi_{ir} + \bar{f}_r
\]

Once only a selected number of principal components are retained in the reduced order representation, denoted by \( a_i^{(k)} \) (where \( i = 1, 2, ...K \)) and simply referred as PC scores or weights. These provide salient measures of the microstructure and are denoted as PCi with \( i = 1, 2, ...K \) representing measures of one selected microstructure. The physical interpretation of the PC scores is quite challenging as they represent the strength of a fairly large dimensional spatial pattern in the microstructure. This will be discussed in more detail in the next section.


In this section, the data science framework employed for building the desired process-structure linkages is introduced and demonstrated (see Fig. 1).

4.1. Quantification of microstructures

For the present study, cross-sections at regular intervals of 20 cells in the growth direction are extracted from each phase-field simulation to capture the time evolution of the microstructure. For each simulation, this produced a set of 211 microstructures. Each cross-section taken perpendicular to the growth direction contained the phase information in 800^2 cells. Since there are 37 different simulations (for different combinations of the values of the process parameters), the complete ensemble of microstructures includes 211 \( \times \) 37 microstructures, each with 800^2 cells. Next, the two-point statistics of the microstructures in this ensemble are computed. PCA is performed on the full ensemble of the collected two-point statistics. As described earlier, the set of protocols provides a low dimensional representation of each microstructure. The variances associated with the first five principal components are shown in Fig. 4. Each bar correspond to the individual variance of its respective principal component and the black line represents the cumulative variance of the principal components. The first three principal components account for more than 95% of the total variance in the ensemble. This value can be even increased to almost 97% by retaining five components. In other words, 95% of the information about the statistical representation of the microstructures in the aggregated ensemble for this study can be defined with just 3 numbers. This represents a dramatic dimensionality reduction from \( 3 \times 401^2 = 482403 \) to just three, demonstrating the power and utility of the proposed data-driven workflows.

Fig. 5 displays the first 3 basis vectors for the autocorrelation of red phase (top row), autocorrelation of blue phase (middle row) and the crosscorrelation of red and blue phases (bottom row). The plots of each row are scaled to the same limits indicated by the colorbars on the right. From the small range of values in the first basis vectors for autocorrelations (top and middle left plots of Fig. 5), it can be seen that PC1 is dominated by the volume fraction red and blue phases. This is also in accordance with

Fig. 4. Scree-plot for the first 5 principal components. Each bar shows the individual variance captured by its respective principal component and black line reflects the cumulative variance.
the first basis vector for crosscorrelation where the entire field is very close to the value 0. The second and third basis vectors (middle and right columns) show that they capture increasingly more complex spatial patterns in the microstructure. As an example, in the second basis vector for the crosscorrelation (bottom middle plot of Fig. 5), the region shown with dark blue color has negative values. This indicates that microstructures with large positive PC2 weights exhibit lower tendency to be occupied by red and blue phases at the head and tail of the short vectors corresponding to these blue regions. Many similar insights can be deduced from the information embedded in the basis vector representation.

The changes in the values of the first three PC weights over the growth height (i.e., the direction perpendicular to the solidification front) for a selected set of phase-field simulations are illustrated in Fig. 6. Each line represents the microstructure evolution in one phase-field simulation. The simulations in Fig. 6 (a) and (b) are color-coded according to the non-dimensionalized velocity values shown on the right side of the plots. For all simulations, the microstructure experiences dramatic changes in PC1 along the growth direction. Since PC1 is dominated by the volume fraction of red and blue phases, these changes correspond to the phase transformations in the early stages of the simulations. The changes in PC1, PC2 and PC3 decrease substantially as the microstructure evolution approaches a steady state. The effect of the boundary conditions on microstructure evolution and steady state growth has been evaluated in our prior work [37].

Parts (a) and (b) of Fig. 6 are color-coded according to the velocity values. Even though both process parameters have influence on all principal components, it can be seen that PC1 and PC2 values are strongly influenced by the velocity, while for PC3, the influence of concentration is more significant than velocity. As the velocity value used in simulations is increased, the final microstructures with higher velocity values get lower PC1 and higher PC2 values. On the other hand, the effect of concentration can be seen to be more dominant on PC3 scores in Fig. 6 (c), where the curves are color-coded by the concentration values provided in the legend. The final microstructures produced by simulations with higher concentration values get lower PC3 scores than the microstructures with lower concentration values. In other words, PC1 and PC2 scores exhibit decreasing and increasing trends, respectively, with increasing velocities. A similar comment can be made for the trend of increasing PC3 scores with increasing concentration values. These trends can be seen more significantly in the distribution of the final microstructures of all 37 simulations shown in Fig. 7. Each point is associated with a concentration and a velocity value which are not necessarily the same for all points. However, part (a) corresponds to a PC1 vs PC2 plot color-coded according to velocity values, whereas part (b) depicts a PC1 vs PC3 plot color-coded
according to concentration values. The location of the final microstructures with respect to aforementioned trends can be clearly seen in these plots. The trends seen readily in these plots attest to the strength of the workflows used here for objective and hierarchical identification of salient microstructure features. Most importantly, the microstructure features are independent and uninformed of the classification process (i.e., unsupervised classification).

Fig. 6. Microstructure evolution evaluated by the first three principal components. The curves in (a) and (b) record the change of PC1 and PC2 over growth height. Both plots are color-coded with respect to the velocity values. The change of PC3 over the growth height in (c) is color-coded with respect to mole-fraction concentration values.

Fig. 7. Distribution of final microstructures of all 37 simulations. (a) PC1 vs PC2 plot color-coded according to the velocity values. (b) PC1 vs PC3 plot color-coded according to the concentration values.

Fig. 8. Final microstructures of a selected set of phase-field simulations. The microstructures are color-coded with respect to velocity values. The simulations are conducted with the same mole fraction of Ag in the liquid (≈0.237).
4.2. Process-structure linkages

Fig. 8 visualizes 16 simulations projected in the first two components of the PC space. All of these simulations have the same concentrations, but different velocities. The points with the same color indicate simulations with the same velocity and the same concentration values, but with different initial microstructures. We observe that the final predicted microstructures within each cluster are very close to each other. This indicates that the domain size determined from our prior work [37] renders the final microstructures independent from the initial microstructures. We take advantage of this independence in formulating the linkages between the process parameters and reduced order representation of the final microstructures.

The process parameters, i.e., velocity and concentration, are treated as inputs to the desired process-structure linkages, while the outputs refer to the truncated PC scores of the final microstructure. The linkages are built through regression between the process parameters (inputs) and the PC scores of the final microstructures (output). The PC scores are obtained from the reduced order representation of final microstructures aggregated from phase-field simulations. We seek here a spectral representation of the linkages between the inputs and the outputs, where the mathematical forms are expressed as a weighted series using a suitable orthogonal basis. Since the initial microstructure does not seem to influence significantly the final microstructure, this parameter is not included in the model. The desired reduced order model for process-structure linkages is expressed as

$$
\mathbf{q}_i = \sum_{l=0}^{L} \sum_{k=0}^{L} A_{i,k}^l P_{l,k}(v^*) P_{l,k}(c^*) + \epsilon_i,
$$

where $\mathbf{q}_i$ represents the score of principal component indexed by $i$ and the spectral coefficients $A_{i,k}^l$ are obtained by fitting Eq. (10) to the available data (extracted from phase-field simulations) through ordinary least squares (OLS) method. $P_{l,k}$ in Eq. (10) represent the orthonormal Legendre polynomial basis over the interval [-1,1]. Hence, the velocity and concentration variables must be normalized to fit in this interval. The normalized velocity $v^*$ and concentration $c^*$ can be expressed as

$$
v^* = \frac{2v - v_{\text{min}} - v_{\text{max}}}{v_{\text{max}} - v_{\text{min}}}, \quad c^* = \frac{2c - c_{\text{min}} - c_{\text{max}}}{c_{\text{max}} - c_{\text{min}}},
$$

where subscripts min and max denote the respective minimum and maximum values.

5. Discussion of reduced order models

The process-structure linkages of interest to the present study are captured by Eq. (10); the coefficients $A_{i,k}^l$ efficiently capture the knowledge associated with such linkages. However, different levels of truncation in the spectral series will result in different levels of accuracy and robustness of the extracted process-structure linkages. In order to systematically explore these linkages, the truncation levels are varied and the error in the model is recorded. The truncation levels correspond directly to the degree of polynomials employed in the linkage, and the error in the model is defined as the absolute difference between the original value from the phase-field simulation and the predicted value from the reduced-order model in Eq. (10). Truncation level $L$ determines the polynomial used in the model according to the constraint $L_e + L_c \leq L$. Fig. 9 summarizes the average errors for all three principal components of the microstructures of interest identified earlier. From left to right, columns contain the bar plots for first, second and third principal components. The top row represents the average of the absolute errors over all 37 simulations for a selected principal component score. If we take the blue bars in an average error plot for the first principal component as an example, each bar is calculated from the average of the absolute errors between the first principal scores of the original and predicted values for a selected polynomial truncation level used in the model. The same approach is followed in the yellow bars except the absolute errors are obtained through leave-one-out-cross validation (LOOCV). On the other hand, bottom row shows standard deviations in the errors for both cases (i.e., with and without LOOCV). If we examine only the errors in the fit (blue bars), as expected they always decrease as we increase the polynomial truncation level in the spectral series. However, with LOOCV, we do see that the inclusion of additional terms in the series leads to overfit (i.e., the predictions are sensitive to leaving out one of the data points).

The approach shown in Fig. 9 grants the advantage of selecting the best model for each principal component. From the averages and standard deviations of the errors calculated with LOOCV, we can see that the differences between the original and the predicted points increase after certain truncation level in the polynomials. For PC1, the error measures of the models developed with and without LOOCV are in very good agreement. The average error measure drops slightly when we switch from $L = 1$ to $L = 2$. However, the standard deviation increases slightly with LOOCV. After $L = 2$, both the average and standard deviation of the errors worsen for LOOCV. Hence, a robust and reliable linkage for PC1 can be obtained with $L = 2$. Similarly, the linkages for both PC2 and PC3 can be found with $L = 3$. The coefficients for the established linkages for the first three principal components are presented in Table 1.

The LOOCV implicitly validates the process-structure linkages. In order to further validate the performance of these process-structure linkages, we define and adopt an error measure for the final predicted microstructure. Based on the earlier discussion, it is suggested that a euclidean distance measure of the differences in the two-point statistics of two microstructures is a good measure of this error. One of the most important advantages of employing the PCA is that this error measure is best evaluated in the PCA space. Note that this is an orthogonal linearly transformed space that preserves the distance measures. Consequently, the distance measure, $D$, can be expressed as [38,45,68]:

$$
D = \sqrt{\sum_{i=1}^{R_e} (a_i^{(1)} - a_i^{(2)})^2},
$$

where $a_i^{(1)}$ and $a_i^{(2)}$ are ith PC weights of first and second microstructures, respectively. $R_e$ denotes the number of principal components retained in the reduced order representation of the microstructure. In this study, only the first three principal components are retained in building process-structure linkages, i.e., $R_e = 3$. The scores of the three principal components for all 37 microstructures are calculated according to the models given in Table 1. Then, the predicted scores are plugged into Eq. (12) to find the distances between the original and predicted representations of the microstructure. The mean and the standard deviations are found to be $\mu(D) = 1.856$ and $\sigma(D) = 1.386$, respectively. On the other hand, the minimum and maximum distances are 0.467 and 6.296. To assess and understand the degree of error involved, these error values should be compared to the values of the PC scores in Figs. 7 and 8. In order to facilitate such an in-depth assessment of error, PC1 vs PC2 plots are shown in Fig. 10 for selected simulated microstructures and their corresponding predicted
other for distances larger than the average. On the other hand, the agreement between the original and predicted values of the PC scores for the microstructure in the middle left plot of Fig. 10. The prediction for the microstructure in the bottom right has a noticeable error. These results elucidate that the surrogate model built as process-structure linkages provides excellent results within the ensemble of simulation results. Another criterion beyond accuracy is the computational efficiency. Once the reduced order representation of the final microstructures is obtained through PCA, the calibration of the surrogate model takes only a few seconds considering the fact that the model for each principal component has around 10 variables. The computational cost of using the metamodel produced in this study is extremely small (almost zero computational cost). Hence the metamodel provides a practical approach for reducing the number of computationally intensive phase-field simulations.

Even though the predictions are in very good agreement with the original points in principal component space, one should remember that the comparisons are in the reduced-order space. To validate the data science approaches, we need to compare the statistics of the original microstructure, retained statistics in the PCA representations, and the statistics for the predicted microstructures. The results of such a comparison can be seen in Fig. 11 for the microstructure in the middle left plot of Fig. 10. The prediction corresponds to an example with an average (distance) error. Top, middle, and bottom rows reflect the statistics for autocorrelation of red phase, autocorrelation of blue phase, and the cross-correlation between red and blue phases, respectively. Each row is scaled to the same color limits displayed at the right of the predictions. The second column shows the retained statistics of the original microstructure using the truncated PCs. In other words, the statistics of the original microstructure is reconstructed from principal component space by using the scores of only the first three principal components. The comparisons between the first and second row demonstrate the consequence of the truncation of principal components corresponding to $i > 3$. The third column shows the predicted statistics. After the predictions of the scores of microstructures. Although PC3 values are not shown in these plots (three dimensional plots make it difficult to visually compare), it is noted that the predictions of PC3 are either more accurate or as good as the predictions of PC1 and PC2. Red and blue points in each plot refer to the original and predicted values of the PC scores for the selected microstructures, respectively. The process parameters and the distance between the original and predicted points are shown on the top of each plot. The two predictions in the top row have very small errors. It can be seen that the original and predicted points are almost indistinguishable from each other. In the middle row, the plot on the right provides a prediction with a distance measure very close to the average calculated from all 37 microstructures. The original and predicted points are again in good agreement with each other. In the bottom row, two microstructures with a distance larger than the average and with maximum distance (worst prediction) are drawn on the left and right plots, respectively. We observe that the points are still very close to each other for distances larger than the average. On the other hand, the

<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{1,0}$</td>
<td>1.543</td>
<td>−1.414</td>
</tr>
<tr>
<td>$A_{1,1}$</td>
<td>11.183</td>
<td>−0.518</td>
</tr>
<tr>
<td>$A_{1,2}$</td>
<td>−0.696</td>
<td>1.469</td>
</tr>
<tr>
<td>$A_{1,3}$</td>
<td>−0.323</td>
<td>−0.035</td>
</tr>
<tr>
<td>$A_{1,0}$</td>
<td>−36.089</td>
<td>2.845</td>
</tr>
<tr>
<td>$A_{1,1}$</td>
<td>−1.847</td>
<td>−4.844</td>
</tr>
<tr>
<td>$A_{1,2}$</td>
<td>−1.145</td>
<td>−2.124</td>
</tr>
<tr>
<td>$A_{1,0}$</td>
<td>−1.657</td>
<td>4.618</td>
</tr>
<tr>
<td>$A_{1,1}$</td>
<td>1.165</td>
<td>1.427</td>
</tr>
<tr>
<td>$A_{1,0}$</td>
<td>1.427</td>
<td>−0.251</td>
</tr>
</tbody>
</table>

Fig. 9. Bar plots for the error between the original and predicted PC scores of the microstructures. From left to right, columns represent the first, second and third PC scores. Top and bottom rows contain the bar plots for the mean and standard deviations of the errors, respectively. Yellow and blue bars are for the results obtained with and without LOOCV, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Fig. 10. PC1 vs PC2 plot of a selected set of microstructures used in the model. Red and blue points show the original and predicted points, respectively. The axes are scaled to the same limits shown in part (a) of Fig. 7 to emphasize the performance of the approach. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 11. Comparison of the statistics of original and predicted microstructures for the prediction in the middle left plot of Fig. 10. From top to bottom, rows represent the autocorrelation of red phase, autocorrelation of blue phase and the crosscorrelation between red and blue phases. From left to right, the columns represent the full statistics of original microstructure, retained statistics of original microstructure after PCA truncation, and the predicted statistics. The axes labels and values for all 2-point statistics plots are the same as applied in Fig. 3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
the first 3 principal components, they are plugged in Eq. (9) to recover the statistics of prediction. From the first and second columns, it can be seen that both the full as well as the retained statistics of the original microstructure have very close values in the short vector region (i.e., near the center of the plots), but there is some degradation of information for the longer vectors. Keeping in mind the huge dimensionality reduction we accomplished (we used only three PC scores), this is quite remarkable. Of course, one can opt to use more PC scores (e.g., using 5 PC scores instead of 3 PC scores) in order to retain more information on the spatial correlations in the quantification of the microstructure. The comparison between the first and second columns provides quantitative insights into the retained information about the microstructure after the dimensionality reduction, and can be used to fine tune the truncation levels employed in the PCA.

The comparison of the retained statistics of the original microstructure and the predicted statistics is indeed much more impressive. It should be noted that the metamodel was indeed calibrated only to the PCA truncated statistics. When we look at the second and third columns, both the short vector range and the long vector range are very well captured. Also the patterns in the statistics (the ring with dark red color in crosscorrelation, the double rings with dark blue color in the autocorrelation of the plane phase, etc.) exist in both columns with almost the same shape. This could prove the accuracy of the surrogate model and moreover the strength of reduced order representation methods in capturing the most significant information about the microstructure with only a handful of numbers.

Additional points can be made about the extraction of process-structure linkages pursued in this study. 1) The presented model is independent of the initial microstructure, as can be concluded from the random orientation of the chains occurring in the final microstructures. The small discrepancies due to utilization of different initial microstructure (Fig. 8) can be further reduced by increasing the domain size. 2) The model demonstrates the computational efficiency and accuracy of the data analytics approach. More simulations with a wider range of velocities and concentration parameters are planned for future work to generalize the approach for a broader window of application. 3) The workflow followed in Fig. 1 is a templated workflow that can be applied with any selected set of process parameters used in most phase-field simulations. 4) The errors attributed to the small number of principal components can be further reduced by generating a richer calibration set (i.e., higher number of simulations) and by retaining a higher number of principal components in the reduced representation of the microstructure. 5) The predicted PC scores of the microstructure can be used to reconstruct a microstructure using a microstructure reconstruction algorithm [65,69–72].

6. Conclusion

Reduced-order process-structure linkages have been extracted from phase-field simulations through novel data science approaches. Reduced-order representation of the microstructure through two-point statistics and PCA granted the advantages of identifying the salient features of microstructure and of evaluating the effect of process parameters. By reducing the high dimensional space spanned by two-point statistics, we were able to establish computationally efficient surrogate models with high reusability. It should be noted that even though executing multiple phase-field models incurs significant computational expense, this is a one-time computational cost. Once the surrogate model is formulated, the statistical representation of the microstructure can be predicted with minimum effort for new process parameters of interest.

Acknowledgments

YCY and SRK acknowledge support from AFOSR award FA9550-12-1-0458. PS thanks for the funding through the German Research Foundation project NE 822/14-2. JH and BN acknowledge the financial support by the MWK cooperative graduate school on microstructure analysis and process evaluation. We also would like to thank the Leibniz Supercomputing Centre in Munich for computational resources, provided as part of a Large Scale Gauss Project and Martin Bauer for helping to implement the solver in the massive parallel framework WALBERLA.

References
