Applied Mathematical Modelling 88 (2020) 589–603

Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

# A data-driven surrogate model to rapidly predict microstructure morphology during physical vapor deposition

Elizabeth Herman[a,b], James A. Stewart[b], Rémi Dingreville[b,*]

[a] Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA
[b] Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, NM 87185, USA

## ARTICLE INFO

*Article history:*
Received 24 March 2020
Revised 4 June 2020
Accepted 10 June 2020
Available online 10 July 2020

*Keywords:*
Process-structure relationships
Physical vapor deposition
Microstructure reconstruction
Surrogate model
Phase-field

## ABSTRACT

Here, we present a surrogate model that rapidly predicts the microstructures of a binary-alloy thin film during physical vapor deposition. This surrogate model is constructed and trained from a data set produced by phase-field simulations of physical vapor deposition. It relies on a statistical representation of the microstructure, principal component analysis, polynomial chaos expansion, and a microstructure-reconstruction algorithm to estimate the microstructure as a function of the deposition parameters and properties of the materials being deposited. This protocol, exercised on a simplified physical vapor deposition model, demonstrates the efficacy of the surrogate model to rapidly predict a broad class of microstructures as a function of deposition conditions with good accuracy relative to high-fidelity models. The considerable computational gain from the surrogate model compared to the detailed phase-field approach highlights the importance of pursuing such approaches, especially when used for producing parameter-microstructure maps for rapid and accurate predictions of the microstructure. As such, this surrogate model can be used to guide the choice of deposition conditions and materials being deposited to fabricate functional thin films with targeted microstructures.

## 1. Introduction

Numerous nanostructured thin films are used as coatings to serve specific mechanical, thermal, magnetic, or optical functionalities [1–3]. The properties of these thin films depend directly on the underlying microstructure resulting from the synthesis process. The synthesis process – typically physical vapor deposition (PVD) or chemical vapor deposition (CVD) – is complex, due to competitions between the depositing vapor flux and surface transport effects, which govern the formation of specific microstructure morphologies [4–6]. The complexity and high-dimensionality of the processing input space separates conceptual understanding of microstructure formation from accurate and rapid predictions of these microstructures, which can be used to rationally design synthesis protocols to fabricate reliable and functional thin films. Building a surrogate model that rapidly links process to microstructure is central to realizing “designer” microstructures and “made-to-order” coatings.

The challenges associated with selecting PVD processing parameters to achieve specific microstructures are at least two-fold. The first challenge is related to the high-dimensionality of the parameter space defining the deposition conditions and materials being deposited. For example, deposition parameters for magnetron sputtering include background pressure,

---

[*] Corresponding author.
*E-mail address:* rdingre@sandia.gov (R. Dingreville).

https://doi.org/10.1016/j.apm.2020.06.046

gas composition, temperature, and sputtering power. Similarly, material properties relevant for deposition encompass surface diffusion, bulk diffusion, and stoichiometry of the atomic species being deposited. Early examples of linking deposition conditions to microstructure include the work by Movchan and Demchishin [7] and Thornton [8]. They developed structure-zone diagrams (SZDs) as a low-dimensional diagrammatic relationship to link a few deposition processing conditions (*e.g.*, background pressure and substrate temperature) to simple microstructural features (*e.g.*, grain size and shape). These early models were created largely based on experimental observations and have proven to be useful in predicting the formation and growth of single-element thin films [5,9]. Due to their phenomenological nature, their low-dimensional representation of the process-structure relationship, and their specificity to given ‘simple’ materials, SZDs have narrow and limited predictability. This is especially true for complex compositions beyond single-element materials for which the input space defining the deposition conditions and materials is much larger. The second challenge is concerned with the non-unique mapping between deposition conditions and the resulting microstructure. Given the inherent variability and randomness that can be found in a microstructure for given deposition conditions [10–12], obtaining exact process-structure linkages for PVD prove to be impractical.

Numerous high-fidelity simulation methods using atomistic, kinetic Monte Carlo, or phase-field methods [6,13–15] predict the microstructural evolution during PVD quite well. For example, we recently developed a phase-field model for simulating PVD of binary-alloy thin films [6] (see brief description of the model in Appendix A). While this model accurately predicts experimentally observed microstructures across a broad class of immiscible materials systems and deposition conditions, it is computationally expensive. This phase-field model is not well suited to rapidly explore the large number of combinations of processing parameters that would be needed to optimize a targeted microstructure. Conversely, while not specifically focused on the PVD process-structure linkage, a large portion of the literature has been centered around the development of surrogate models that efficiently and quickly link process, structure, and properties for a wide variety of materials and fabrication processes [16–26]. Such linkage is constructed using a range of different techniques including, regression models [20–22,25], Gaussian processes [26], convolutional neural networks [19], or deep-learning methods [27]. All of these methods rely on a statistical representation of the microstructure to construct a surrogate model that can be used to identify process-structure-property relations. A surrogate model refers to a simplified model designed to represent and substitute a computationally-expensive, high-fidelity model. For example, Yabansu et al. [24] presented a workflow to extract process-structure linkages from phase-field simulations of the microstructure evolution of ternary eutectics during solidification. Their methodology used two-point statistics, principal component analysis (PCA), and a spectral representation to link the temperature-gradient to the solidification of random microstructures. This protocol was shown to work well for random microstructures exhibiting a linear response to changes in the process input space. As identified by Yabansu et al. [24], this methodology needed to be demonstrated for a wide range of solidification rates and eutectic concentrations. In our present case, this would correspond to deposition rate and phase fraction. Predicting microstructures that are achievable during PVD as a function of the deposition parameters presents, however, an additional challenge related to the fact that deposition conditions can yield a much broader range of microstructure morphologies [3,6,28].

This study expands upon the above body of work and addresses the aforementioned shortcomings by introducing a surrogate model that rapidly links PVD parameters to the deposited microstructure. For the sake of demonstrating the overall workflow of the method, we use a simplified version of our phase-field model for PVD [6] by considering deposition rate ($\dot{d}$) and phase fraction ($f_c$) of the material being deposited as the only processing parameters. The surrogate model is trained on a data set obtained from phase-field simulations and relies on a combination of a statistical representation of microstructures, PCA, and polynomial chaos expansion. This protocol can be combined with a microstructure-reconstruction algorithm to visualize the results in a meaningful (non-statistical representation) way. It rapidly estimates the microstructure morphology as a function of $\dot{d}$ and $f_c$. As illustrated in Fig. 1, the construction of this phase-field-informed, surrogate model for PVD comprises three steps. The first step consists of performing high-fidelity phase-field simulations to generate a data set of different types of microstructure morphologies as a function of $\dot{d}$ and $f_c$. The second step quantifies these microstructures by using two-point statistics and reducing the dimensional representation of these microstructures by performing PCA on the two-point statistics. This second step is designed to capture the variance in the microstructure representation for a broad range of processing conditions while using a reduced set of microstructural descriptors (see Section 2.1). The third step uses polynomial chaos expansion on the PCA representation to map the predicted two-point statistics back to the processing parameters (see Section 2.2). In Section 3, we show results and notable predictions for various types of microstructure morphologies as compared to high-fidelity phase-field simulations. The two-point statistics between the predicted and true microstructures are shown to be in good agreement. These results illustrate the robustness and efficiency of the protocol for predicting the structure from deposition conditions. In Section 4, we discuss the computational efficiency of the surrogate model approach as compared to the high-fidelity phase-field model. We also discuss some of the challenges of combining our surrogate model with a microstructure-reconstruction algorithm to visualize the results. Our comparisons of microstructures obtained via high-fidelity phase-field simulations and reconstructed from the surrogate model highlight (i) the deficiencies of the non-unique mapping, and (ii) the inaccuracies resulting from the microstructure reconstruction.

## 2. Construction of a surrogate model for physical vapor deposition

Mathematically, we denote $\boldsymbol{S}_{\mathrm{PVD}}$ as the phase-field-informed, surrogate model for PVD that maps $\mathbb{R}^{n_p} \rightarrow \mathbb{R}^N$, where $N$ is the number of grid points in the $n_x \times n_y$ domain of interest. The input of the surrogate model are $n_p$ processing parameters

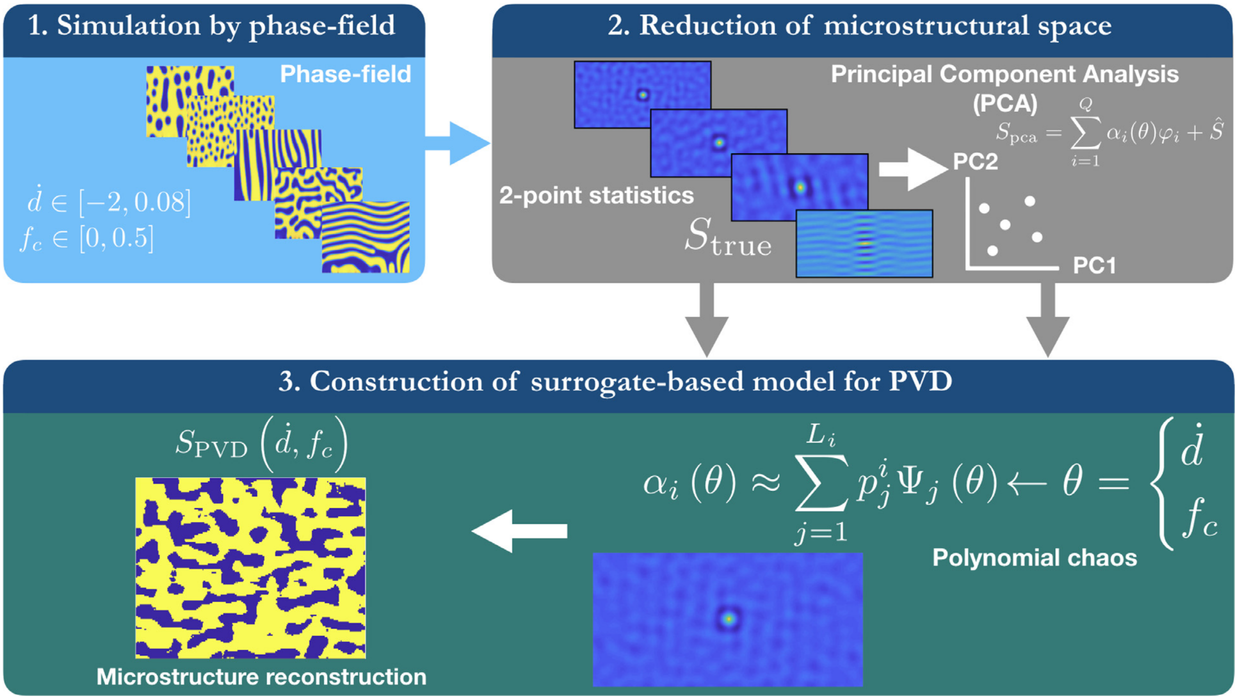
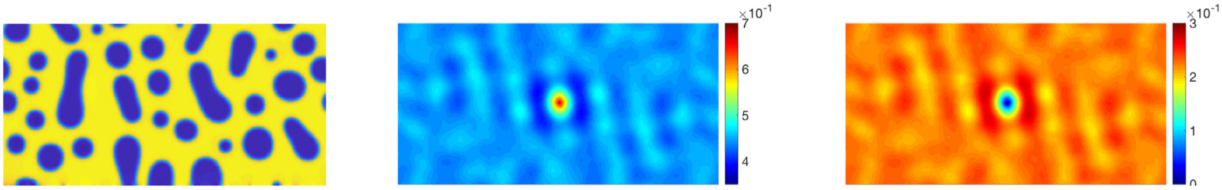


**Fig. 1.** Overview of the workflow to construct a surrogate model for predicting microstructure morphologies during simulated physical vapor deposition. The first step consists of running direct numerical phase-field simulations for a broad range of deposition rate ($\dot{d}$) and phase fraction ($f_c$). The second step reduces the microstructural input space by generating two-point statistics of the microstructures and performing a principal component analysis on it. The third step consists of using polynomial chaos to link processing parameters ($\dot{d}$, $f_c$) to predicted two-point statistics. A microstructure reconstruction algorithm is used to reconstruct the microstructure from the predicted two-point statistics.



(a) (b) (c)

**Fig. 2.** An example of the transformation of a (a) two-phase microstructure morphology realization to its two-point statistical representation: (b) the auto-correlation function $\boldsymbol{S}_2^{(1,1)}$ for the yellow phase and (c) the cross-correlation function $\boldsymbol{S}_2^{(1,2)}$ for the yellow and blue phases. Here, the color yellow indicates phase 1 and the color blue indicates phase 2. For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

described by a vector denoted as $\boldsymbol{\theta} \in \mathbb{R}^{n_p}$. The output of the surrogate model is a two-point statistics representation of the microstructure $\boldsymbol{m}(\boldsymbol{\theta})$. The two-point statistics are computed for the $n_x \times n_y$ microstructure morphology that is described by using a discretized microstructure function $\boldsymbol{m}(\boldsymbol{\theta}) \in \mathbb{R}^N$, where each entry represents a local state (*i.e.*, a material phase) of the microstructure at a given location $\boldsymbol{x}_i$ for $i = 1, \ldots, N$. For example, in Fig. 2(a), a two-phase microstructure is represented on a uniform grid using a binary scalar (colored as yellow and blue) to indicate the local phase. The reconstructed microstructure from the predicted two-point statistics takes the form, $\boldsymbol{m}\big(\boldsymbol{\theta}\big) = \mathcal{G}\big[\boldsymbol{S}_{\text{PVD}}\big(\boldsymbol{\theta}\big)\big]$, where the function $\mathcal{G}$ represents a microstructure-reconstruction algorithm.

The surrogate model, $\boldsymbol{S}_{\text{PVD}}$, is constructed by using a set of $N_{\text{sim}}$ phase-field simulations that simulate the growth of a binary-alloy thin film as a function of deposition conditions. Phase-field simulations are performed on a uniform, finite-difference computational grid of size $N = n_x \times n_y$. Each phase-field simulation is independent and corresponds to a microstructure morphology realization, $\omega_k, k \in N_{\text{sim}}$, obtained by sampling each element $\theta_i$ of $\boldsymbol{\theta}$ from a distribution. In the present case, we have chosen a uniform distribution, *i.e.*, $\theta_i \sim \mathcal{U}(\theta_i^{\text{min}}, \theta_i^{\text{max}})$ for $i = 1, \ldots, n_p$. We assume that, for a given set of process parameters, the PVD process is a stochastic process and each microstructure realization, $\omega_k$, produced by a PVD phase-field simulation is a realization of this stochastic process. The idea of using realizations of a stochastic process to represent microstructures and their associated properties is not new [29–32]. The central concept is that each microstructure realization comes from some probability configurational space. The correlation in the microstructure morphology resulting from the distribution of processing parameters $\boldsymbol{\theta}$ is captured within that probability configurational space, and can be represented by a functional form of the type $\boldsymbol{S}_{\text{PVD}}$.

### 2.1. Statistical representation of microstructures

Let us first consider the statistical representation of microstructures for fixed processing conditions $\boldsymbol{\theta}$. We are primarily concerned with characterizing the correlation function of the spatially dependent composition field, $c(\boldsymbol{x}, t)$, for thin films grown to a given thickness, $t$. In the remainder of the manuscript, we fix $t$ to suppress the dependency on the thickness of the thin film, such that $c(\boldsymbol{x}, t)$ is equivalently represented by $c(\boldsymbol{x}, \omega_k)$. For a given microstructure realization, $\omega_k$, we use a compositional indicator function, $I^{(i)}(\boldsymbol{x}, \omega_k)$, to identify the dominant phase at a location $\boldsymbol{x}$ within a microstructure, such that

$$I^{(i)}(\boldsymbol{x}, \omega_k) = \begin{cases} 1, & \text{if } c(\boldsymbol{x}, \omega_k) > 0 \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

The indicator function is a distribution function that indicates which of the $i$ phases is the representative phase at location $\boldsymbol{x}$ [29,30]. Note that, in our case, the range of the field variable $c$ is $-1 \leq c \leq 1$; thus motivating our use of 0 as the cutoff to “binarize” the microstructure data.

The mathematical form used to quantify a microstructure utilizes statistical correlation functions [30,31,33]. Although not considered here, other statistical representations include nearest-neighbor distributions [34] and lineal-path functions [35]. The two-point cross-correlation function, $S_2^{(i,j)}$, is defined as the expectation of the product $I^{(i)}(\boldsymbol{x}_1)I^{(j)}(\boldsymbol{x}_2)$, *i.e.*,

$$S_2^{(i,j)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle I^{(i)}(\boldsymbol{x}_1) I^{(j)}(\boldsymbol{x}_2) \rangle. \tag{2}$$

The auto-correlation function $S_2^{(i,i)}$ can be interpreted as the conditional probability density that two points at positions $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are found to be in phase $i$. The reader is referred to Torquato and Haslach [33] for a detailed discussion on microstructural correlation functions. Since we assume that the microstructures of interest are static, $S_2^{(i,i)}$ will only depend on the distance between two points, *i.e.*, $S_2^{(i,i)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = S_2^{(i)}(\boldsymbol{r})$, where $\boldsymbol{r} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ for any vector $\boldsymbol{r}$ randomly placed in the microstructure. Hence, in this form, the two-point statistics resembles a convolution operator and can be efficiently computed using Fast Fourier Transform (FFT) [24,30–32] as applied to the finite-difference discretized scheme. By implementing the FFT computational procedure, we make two assumptions about our microstructure: (i) the microstructure is periodic, and (ii) the vector, $\boldsymbol{r}$, is discretized with the same scheme as the discretized domain.

Because the microstructures of interest are comprised of two phases, there are four two-point statistics: two auto-correlations functions $\boldsymbol{S}_2^{(1,1)}$ and $\boldsymbol{S}_2^{(2,2)}$, and two cross-correlation functions $\boldsymbol{S}_2^{(1,2)}$ and $\boldsymbol{S}_2^{(2,1)}$. Fig. 2 illustrates the transformation of a microstructure (Fig. 2(a)) to its auto- and cross-correlation functional representations (Fig. 2(b) and (c), respectively). Given the properties of the Fourier transform, only one of these four correlation functions is necessary to fully describe the microstructure [36]. Here, we choose $\boldsymbol{S}_2^{(1,1)} = \boldsymbol{S}_2 \in \mathbb{R}^N$ as the independent correlation function to completely describe the microstructure.

#### 2.1.1. Reduced-dimensional representation of the microstructure

The correlation function for a given microstructure realization, $\omega_k$, quantifies the microstructure morphology in a probabilistic sense and cannot be readily used to map the processing parameters to the microstructure. Presently, every element in the discretized auto-correlation function, $\boldsymbol{S}_2$, depends on the process parameters, $\boldsymbol{\theta}$. To decrease the number of entries that depend on $\boldsymbol{\theta}$, we employ PCA to obtain a reduced-dimensional representation of the microstructure spatial statistics. PCA produces a truncated orthonormal representation of a microstructure, denoted as $\boldsymbol{S}_{\text{pca}}$. This representation ensures the maximum variance of the auto-correlation function is retained among all $N_{\text{sim}}$ microstructures in a lower dimensional space. The basis vectors of this space, $\boldsymbol{\varphi}_i$, for $i = 1 \ldots N$, are called principal components (PC), and the weights, $\alpha_i$, are called PC scores [37]. The basis vectors of the lower dimensional space are computed by forming the sample covariance matrix of $\boldsymbol{S}_2(\boldsymbol{\theta}^{(k)}), k = 1, \ldots, N_{\text{sim}}$, and computing its eigenvectors. The eigenvectors are the principal components and the eigenvalues are the variances corresponding to the basis vectors, $\boldsymbol{\varphi}_i$. The sample covariance matrix is computed with all $N_{\text{sim}}$ two-point statistics of the microstructure realizations generated for the PVD processing input space sampled. The PCA representation, $\boldsymbol{S}_{\text{pca}} : \mathbb{R}^{n_p} \to \mathbb{R}^N$ of a given microstructure is given by,

$$\boldsymbol{S}_{\text{pca}}(\boldsymbol{\theta}) = \sum_{i=1}^{Q} \alpha_i(\boldsymbol{\theta}) \boldsymbol{\varphi}_i + \hat{\boldsymbol{S}}, \tag{3}$$

where $Q$ is the number of PC directions retained. The term $\hat{\boldsymbol{S}}$ represents the sample mean of the two-point statistics, $\boldsymbol{S}_2(\boldsymbol{\theta}^{(k)})$, for $k = 1, \ldots, N_{\text{sim}}$. The summation term represents the variance of the microstructure with respect to that same ensemble.

In Eq. (3), instead of having $\boldsymbol{\theta}$ depend on $N$ entries in $\boldsymbol{S}_2$, this lower dimensional representation, $\boldsymbol{S}_{\text{pca}}$, only requires $Q < N$ entries to describe this dependency.

### *2.2. Mapping of PVD parameters to microstructure*

The coefficients in Eq. (3), $\alpha_i(\boldsymbol{\theta}^{(k)})$, are what characterize a given microstructure $k$. The mapping between the processing parameters and the microstructure "$k$" is accomplished by determining a functional form that maps $\alpha_i(\boldsymbol{\theta}^{(k)})$ back to $\boldsymbol{\theta}^{(k)}$. Since we assumed that the processing parameters are sampled from a probability distribution, we use polynomial chaos expansion [38–40].

A polynomial chaos expansion is a way of representing an arbitrary random variable (in this case the microstructural descriptors $\alpha_i(\boldsymbol{\theta})$) as a function of another random variable with a given distribution (*i.e.*, the processing parameters $\boldsymbol{\theta}$), by describing the relationship as a polynomial expansion. The coefficients of the PCA expansion, $\alpha_i(\boldsymbol{\theta})$, are expressed as a linear combination of orthogonal polynomial basis functions, $\Psi_j(\boldsymbol{\theta})$, such that,

$$\alpha_i(\boldsymbol{\theta}) \approx \sum_{j=1}^{L_i} p_j^i \Psi_j(\boldsymbol{\theta}), \tag{4}$$

where $p_j^i$ are the deterministic coefficients of the expansion, truncated at some polynomial order $L_i$. The advantage of expressing the functional relationship between $\alpha_i(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ by an orthonormal multivariate basis lies in the rapidly decaying spectrum properties of the polynomial chaos expansion coefficients, $p_j^i$ [39]. Hence, in the context of our goal to rapidly predict microstructure morphologies during PVD, fewer $\Psi_j$ basis functions are needed than if they were not orthogonal. In addition, the orthogonality of the basis functions results in the fast computation of global sensitivity measures [38].

The choice of an orthogonal polynomial family, $\Psi_j$, depends on the (known or assumed) distribution of $\boldsymbol{\theta}$, since the orthogonality of $\Psi_j$ is directly derived from the probability space it is defined on (and therefore, the probability density functions associated with the distribution of $\boldsymbol{\theta}$) [38–40].

Recall we have assumed $\theta_i$ for $i = 1, \ldots, n_p$ are mutually independent random variables with uniform distributions. To ensure orthogonality, $\Psi_i$ are chosen to be the multi-dimensional Legendre polynomials on the interval $[-1, 1]$. Since each process parameter, $\theta_i$, is assumed to be defined in the interval $[\theta_i^{\min}, \theta_i^{\max}]$, we map the process parameters to the interval $[-1, 1]$ by defining $\widetilde{\boldsymbol{\theta}}$ to be the normalized processing parameters vector computed as,

$$\widetilde{\theta}_i = \frac{2\theta_i - \theta_i^{\min} - \theta_i^{\max}}{\theta_i^{\max} - \theta_i^{\min}} \qquad i = 1, \ldots, n_p. \tag{5}$$

The multi-dimensional basis is then constructed by using the one-dimensional Legendre polynomials and a multi-index. For more information on constructing the multi-dimensional basis, the reader is referred to the textbook by Smith [38].

Having defined the basis functions, the polynomial expansion coefficients, $p_j^i$, are then computed by constructing a linear system of equations for each pair $(\boldsymbol{\theta}^{(k)}, \alpha_i(\boldsymbol{\theta}^{(k)}))$. To find the coefficients in Eq. (4), we solve $Q$ overdetermined linear systems of the form

$$\begin{bmatrix} \alpha_i(\boldsymbol{\theta}^{(1)}) \\ \alpha_i(\boldsymbol{\theta}^{(2)}) \\ \vdots \\ \alpha_i(\boldsymbol{\theta}^{(N_{\text{sim}})}) \end{bmatrix} = \begin{bmatrix} \Psi_0(\widetilde{\boldsymbol{\theta}}^{(1)}) & \Psi_1(\widetilde{\boldsymbol{\theta}}^{(1)}) & \cdots & \Psi_{L_i}(\widetilde{\boldsymbol{\theta}}^{(1)}) \\ \Psi_0(\widetilde{\boldsymbol{\theta}}^{(2)}) & \Psi_1(\widetilde{\boldsymbol{\theta}}^{(2)}) & \cdots & \Psi_{L_i}(\widetilde{\boldsymbol{\theta}}^{(2)}) \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_0(\widetilde{\boldsymbol{\theta}}^{(N_{\text{sim}})}) & \Psi_1(\widetilde{\boldsymbol{\theta}}^{(N_{\text{sim}})}) & \cdots & \Psi_{L_i}(\widetilde{\boldsymbol{\theta}}^{(N_{\text{sim}})}) \end{bmatrix} \begin{bmatrix} p_1^i \\ p_2^i \\ \vdots \\ p_{L_i}^i \end{bmatrix}, \text{ for } i = 1, \ldots, Q. \tag{6}$$

We determine the truncation level, $L_i$, by performing $k$-fold cross validation, with $k = 10$, to ensure that the number of basis functions are not an overfit to the data vector $\boldsymbol{\alpha}(\boldsymbol{\theta}^{(k)})$ for $k = 1, \ldots, N_{\text{sim}}$. Our choice of $L_i$ is the truncation level which results in the smallest mean error between the true and predicted $\boldsymbol{\alpha}(\boldsymbol{\theta})$ for the validation set.

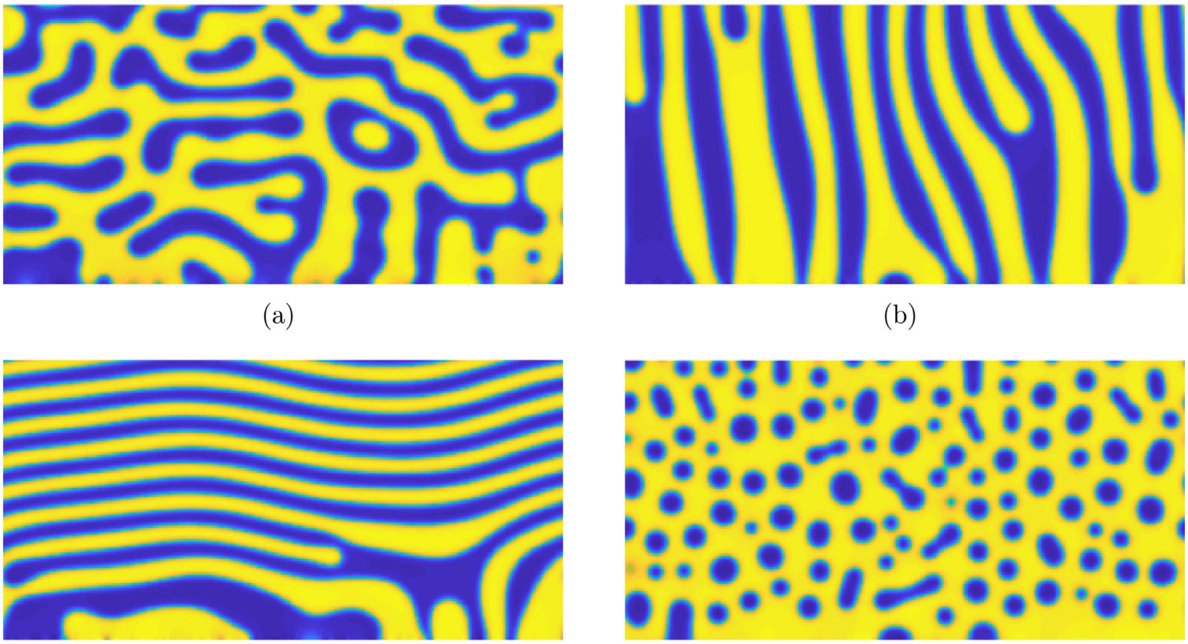
The complete surrogate model predicting the two-point statistics of the microstructure as a function of the processing parameters is therefore finalized by using the polynomial chaos expansion (Eq. (4)) as the means to estimate each $\alpha_i(\boldsymbol{\theta})$ in the PCA representation of the microstructure (Eq. (3)). Hence, the surrogate PVD model $\boldsymbol{S}_{\text{PVD}} : \mathbb{R}^{n_p} \to \mathbb{R}^N$ takes the form,

$$\boldsymbol{S}_{\text{PVD}}(\boldsymbol{\theta}) = \hat{\boldsymbol{S}} + \sum_{i=1}^{Q} \sum_{j=1}^{L_i} p_j^i \Psi_j(\widetilde{\boldsymbol{\theta}}) \boldsymbol{\varphi}_i. \tag{7}$$

## 3. Results: predicting microstructures during PVD via a phase-field-informed surrogate model

We now employ the protocol discussed in Section 2 on a simplified PVD model by considering the deposition rate, $\dot{d}$, and the phase fraction of alloy constituents, $f_c$, as the only processing parameters, such that $n_p = 2$. Even for this low-dimensional input space, the achievable microstructures exhibit significant variations in their morphologies. As illustrated in Fig. 3, there are (at least) four classes of microstructures obtained from the simulated PVD [6]:

- a random concentration modulation (RCM) corresponding to an interconnected bicontinous network of phase $i$ and phase $j$ (Fig. 3(a)),



(a) (b)

(c) (d)

**Fig. 3.** Example of four classes of microstructures obtained from the simulated PVD. (a) Random concentration modulation, (b) vertical concentration modulation, (c) horizontal concentration modulation, (d) nanoprecipitate concentration modulation.

- a lateral and a vertical concentration modulation (LCM/VCM) corresponding to laminate-like structures alternating phase $i$ and $j$ with lateral/vertical orientations with respect to the substrate (Fig. 3(b) and (c) respectively), and
- a nanoprecipitate concentration modulation (NPCM) corresponding to the dispersion of small precipitates of phase $i$ or $j$ into a matrix of phase $j$ or $i$ respectively (Fig. 3(d)).

In the remainder of this section, we describe the results for each step of the protocol to build the surrogate model for PVD. This includes (i) obtaining a representative ensemble of microstructures synthesized through simulated PVD (Section 3.1), (ii) quantifying these microstructures using two-point statistics and performing PCA to reduce their dimensional representation (Section 3.2), and (iii) calibrating the process-structure linkage (Section 3.3). Note that, even though executing $N_{\text{sim}}$ phase-field simulations in the first step incurs significant computational cost, this is a one-time computational cost.

### 3.1. Phase-field model simulations

The ensemble of representative microstructures obtained with simulated PVD has been generated using a simplified version of our phase-field model. This model simulates the growth of a thin film during PVD as a function of deposition conditions and materials being deposited. Although we only consider two processing parameters, namely the deposition rate and phase fraction, the protocol described in Section 2 is not restricted to these two processing parameters. In fact, the full processing input space of our PVD phase-field model includes: four vapor deposition parameters (deposition rate, phase fraction, vapor density, and vapor diffusivity), four phase kinetics parameters (bulk and surface mobilities for each species), and seven elastic properties parameters (elastic moduli and compositional coefficient of expansion). A brief description of the central features of this model is provided in Appendix A, along with the necessary parameters and boundary conditions considered for the simulations.

For the simplified version of the phase-field model, we consider a two-dimensional domain as illustrated in Fig. 4. The entire computational domain is denoted as $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, and the boundary of the domain as $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \cup \Gamma_4$. The subdomain $\mathcal{D}_1$ corresponds to where the vapor is being transported (from top to bottom, *i.e.*, $\dot{d} < 0$) onto the thin film. The thin film is defined as the subdomain $\mathcal{D}_2$. Since the thin film is assumed to grow in the positive vertical direction, $\mathcal{D}_2$ increases in the vertical direction as the thin film grows through time.

The field variables in our PVD model are time- and path-dependent; however, in this work, we are only interested in characterizing the microstructure of the thin film once it has reached a prescribed thickness, $t_f$. While the spatial domain is discretized into a uniform mesh of 256 × 256 grid points, we choose the final microstructure to be 256 × 128 grid points in size (i.e., $t_f = 128$) for construction of the surrogate model. To obtain a representative data set of achievable microstructures synthesized during PVD, each of the two processing parameters, $\dot{d}$ and $f_c$, are sampled independently from the following

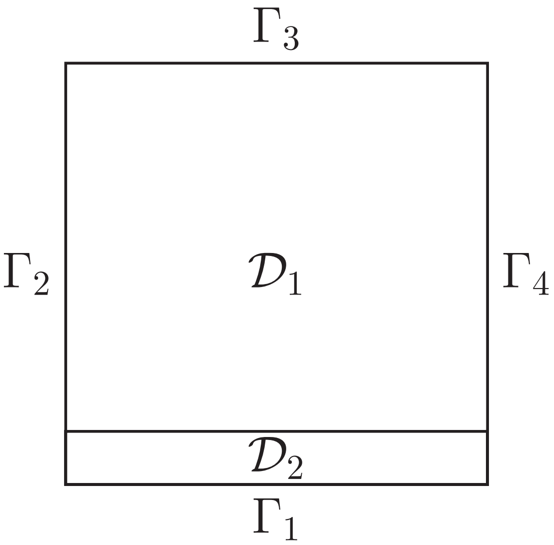


**Fig. 4.** Description of computational domain used for the PVD phase-field model.

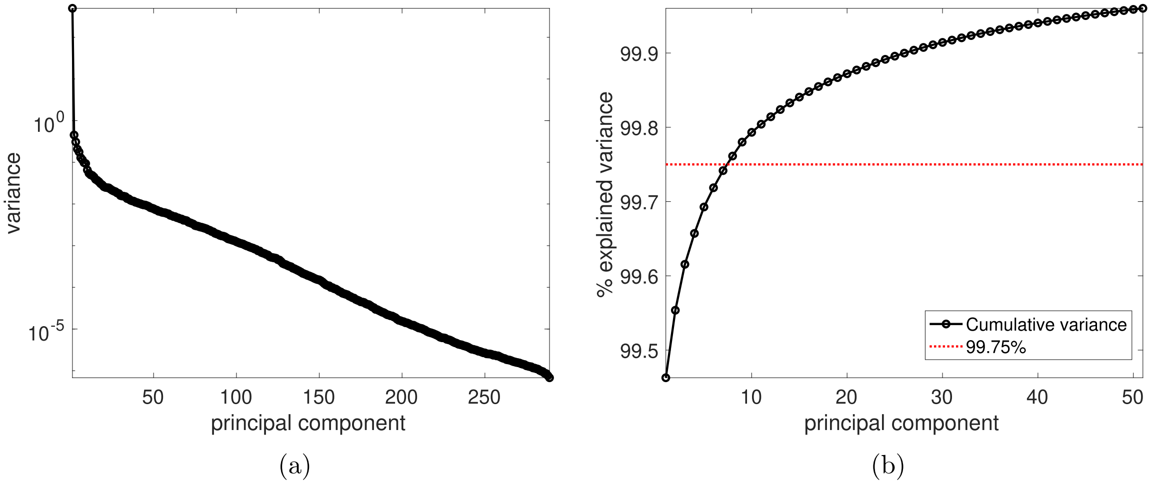


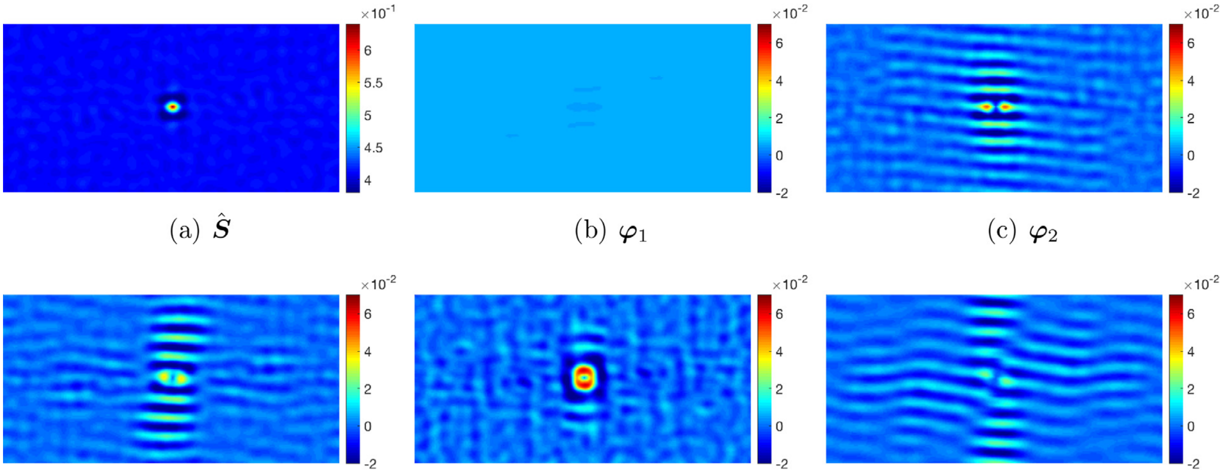
**Fig. 5.** Principal component analysis on the two-point statistics of the microstructure dataset: (a) individual variance for the principal directions, and (b) cumulative variance as a function of the number of principal directions included.

distributions: $\dot{d} \sim \mathcal{U}(-2.0, -0.08)$, and $f_c \sim \mathcal{U}(0.0, 0.5)$. The distribution of $\dot{d}$ is sampled at equally spaced intervals of 0.03. The sampled phase fractions are chosen to correspond to 50/50, 55/45, 60/40, 65/35, and 73/27 phase fractions in the vapor. A total of 325 processing parameter conditions have been sampled to generate 325 different microstructure morphologies. These 325 samples are then randomly split into a training set, consisting of 90% of the data, and a testing set, containing the remaining 10% of the data. The training set is the only data used in the construction of the surrogate model. In our error analyses, we compute the error of the predicted microstructure statistics for both the training and testing data sets.

### *3.2. Principal component analysis of microstructures*

Once the ensemble of microstructures has been generated, we perform PCA on the microstructures' two-point statistics in the training data set to reduce the dimensionality description of these microstructures as described in Section 2.1.1. Fig. 5(a) shows the proportion of total variance for the principal directions, and Fig. 5(b) shows the total explained variance in the ensemble of microstructures as a function of the number of principal directions included. This figure illustrates that, when the microstructures are represented by their two-point statistics, only 8 basis functions are needed to capture over 99.75% of the variance in the data set. Thus, these 8 basis functions provide an accurate low-rank representation of the final microstructures of interest.

Fig. 6 depicts the resulting average two-point statistics over all microstructures, $\hat{\boldsymbol{S}}$ (Fig. 6(a)), along with the first five basis elements (Fig. 6(b)–(f)), $\boldsymbol{\varphi}_i$, for $i = 1 \ldots 5$, which were computed as a result of the PCA process. First, notice the first basis



(a) $\hat{\boldsymbol{S}}$ (b) $\varphi_1$ (c) $\varphi_2$

(d) $\varphi_3$ (e) $\varphi_4$ (f) $\varphi_5$

**Fig. 6.** Basis elements $\boldsymbol{\varphi}_i$ representing the two-point statistics space as described in Eq. (3). The average (a) and first five basis elements ((b)–(f)) are illustrated in this figure. The actual two-point statistics of the microstructure can be approximated by a linear combination of these basis functions.

**Table 1**
Error associated with using 8 basis functions to approximate the high-fidelity two-point statistics from the phase-field samples of the training and testing set.

| Error | Min | Max | Mean | Standard deviation |
|---|---|---|---|---|
| $\mathrm{err}_1(\boldsymbol{\theta}_{\mathrm{train}})$ | 0.0030 | 0.0690 | 0.0146 | 0.0123 |
| $\mathrm{err}_1(\boldsymbol{\theta}_{\mathrm{test}})$ | 0.0030 | 0.0441 | 0.0123 | 0.0091 |

element (Fig. 6(b)) in the two-point statistics space. This element is dominated by the phase fraction at the center of $S_2(\boldsymbol{r})$. The basis elements corresponding to smaller variances (Fig. 6(c) and (f)) are more oscillatory and capture some complexity of the morphology and long-range behavior in the microstructure. Since we only keep 8 basis elements, the most oscillatory aspects of two-point statistics are eliminated from the reduced representation.

After using the microstructure training data set to compute $\boldsymbol{\varphi}_i$, for $i = 1, \ldots, 8$, we estimate the error in the approximation of the two-point statistics of the microstructure from the dimensionality reduction process, $\boldsymbol{S}_{\mathrm{pca}}$, with respect to the two-point statistics of the true microstructure, $\boldsymbol{S}_{\mathrm{true}}$, by using the following error metric

$$\mathrm{err}_1 = \frac{\|\boldsymbol{S}_{\mathrm{true}}(\boldsymbol{\theta}) - \boldsymbol{S}_{\mathrm{pca}}(\boldsymbol{\theta})\|_2}{\|\boldsymbol{S}_{\mathrm{true}}(\boldsymbol{\theta})\|_2}. \tag{8}$$

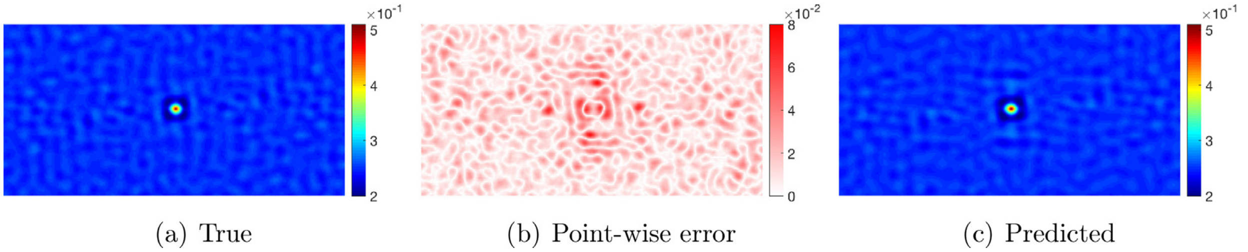
Table 1 shows the minimum, maximum, mean, and standard deviation of $\mathrm{err}_1$, for both the training and testing set. It is seen that, on average, the error of using this smaller space as a representation of the data results in less than a 2% error for both the training and testing sets when compared to the high-fidelity phase-field model, confirming the predictiveness and robustness of the PCA representation.

### 3.3. Mapping of deposition conditions to microstructure representation

The previous step captured the variance in the spatial composition of a given thin film through the basis functions $\boldsymbol{\varphi}_i$ and the associated PC scores $\alpha_i$. We now approximate the relationship between the process parameters $\boldsymbol{\theta}$ and $\alpha_i(\boldsymbol{\theta})$ in Eq. (4) by using (i) the training set of two-point statistics, (ii) the linear expansion in the space spanned by the 8 basis functions in Eq. (3), and (iii) the linear system in Eq. (6). To determine the truncation level $L_i$ for each coefficient, we perform 10-fold cross validation to evaluate the surrogate model for the limited data sets. The truncation level is chosen to produce the smallest average error in the validation sets.

Fig. 7 illustrates the comparison between the predicted two-point statistics obtained from the full surrogate model, $\boldsymbol{S}_{\mathrm{PVD}}$, as defined in Eq. (7), and the actual two-point statistics for a given microstructure, $\boldsymbol{S}_{\mathrm{true}}$, from the set of testing parameters. In this example, the deposition rate was chosen as being $\dot{d} = -1.85$ and the phase fraction as $f_c = 0$ corresponding to a 50/50 distribution of phases in the thin film. The measure of error shown in Fig. 7(b) quantifies the relative point-wise error between the predicted and true auto-correlation function and is defined as,

$$\mathrm{err}_{\mathrm{pt}}(\boldsymbol{\theta}) = \frac{|\boldsymbol{S}_{\mathrm{true}}(\boldsymbol{\theta}) - \boldsymbol{S}_{\mathrm{PVD}}(\boldsymbol{\theta})|}{\boldsymbol{S}_{\mathrm{true}}(\boldsymbol{\theta})}, \tag{9}$$



(a) True (b) Point-wise error (c) Predicted

**Fig. 7.** Example prediction using the surrogate model: (a) true two-point statistics $\boldsymbol{S}_{\text{true}}$, and (c) the predicted two-point statistics $\boldsymbol{S}_{\text{PVD}}$. The relative point-wise error $\text{err}_{\text{pt}}$ is shown in panel (b).

**Table 2**
Error associated with approximating $\boldsymbol{\alpha}(\boldsymbol{\theta})$ using polynomial chaos expansion ($\text{err}_2$), and (ii) error incurred by using the surrogate model ($\text{err}_3$).

| Error | Min | Max | Mean | Standard deviation |
|---|---|---|---|---|
| $\text{err}_2(\boldsymbol{\theta}_{\text{train}})$ | 0.0920 | 5.7683 | 0.8109 | 1.0067 |
| $\text{err}_2(\boldsymbol{\theta}_{\text{test}})$ | 0.1106 | 2.8706 | 0.5296 | 0.5065 |
| $\text{err}_3(\boldsymbol{\theta}_{\text{train}})$ | 0.0032 | 0.1197 | 0.0206 | 0.0223 |
| $\text{err}_3(\boldsymbol{\theta}_{\text{test}})$ | 0.0032 | 0.0771 | 0.0154 | 0.0139 |

where all operations are point-wise. With errors $\lesssim 0.1\%$, the PVD surrogate model, $\boldsymbol{S}_{\text{PVD}}$, successfully reproduces the two-point statistics for microstructures in the training and testing sets.

In order to quantify the error associated with using the polynomial chaos expansion to approximate $\alpha(\boldsymbol{\theta})$, we define two error metrics. The first error metric, $\text{err}_2(\boldsymbol{\theta})$, measures the error in the vector of PC scores $\boldsymbol{\alpha}(\boldsymbol{\theta})$ for a microstucture computed with process parameters, $\boldsymbol{\theta}$, such that,

$$\text{err}_2(\boldsymbol{\theta}) = \|\boldsymbol{\alpha}_{\text{true}}(\boldsymbol{\theta}) - \boldsymbol{\alpha}_{\text{predict}}(\boldsymbol{\theta})\|_2. \tag{10}$$

The second error metric, $\text{err}_3(\boldsymbol{\theta})$, evaluates the total error in the approximation of the two-point statistics of the microstructure as predicted by the surrogate model, Eq. (7), with respect to the two-point statistics of the true microstructure. This error is defined as,

$$\text{err}_3(\boldsymbol{\theta}) = \frac{\|\boldsymbol{S}_{\text{true}}(\boldsymbol{\theta}) - \boldsymbol{S}_{\text{PVD}}(\boldsymbol{\theta})\|_2}{\|\boldsymbol{S}_{\text{true}}(\boldsymbol{\theta})\|_2}. \tag{11}$$

Table 2 summarizes the performance of the surrogate model for both the training data set and testing data set. On average, the error in the surrogate model ($\text{err}_3$) is less than $\sim 2\%$ for both the training and testing sets.

This error analysis highlights the fact that the largest source of error in our workflow stems from using the polynomial chaos expansion (see $\text{err}_2$ in Table 2). Since we are using continuous polynomials to construct the linkage between the process parameters and the PCA space, we are assuming that the data is also continuously dependent on these parameters. While this is impossible to determine without prior knowledge about the stochastic process, in our problem, this may not be the case. We find that the process parameters which yield the largest error in the prediction of $\boldsymbol{S}_{\text{PVD}}(\boldsymbol{\theta})$ occurs when the output of the random vector varies vastly from a subtle perturbation in the choice of process parameters. This indicates that the basis functions used in the polynomial expansion do not appropriately capture the rapid changes in microstructure morphologies incurred by small variations in the deposition conditions, and the data may therefore not be continuously dependent on the deposition parameters. Possible alternative methods to more accurately relate the process parameters to $\boldsymbol{\alpha}(\boldsymbol{\theta})$ include using Galerkin projection [38] or machine learning techniques (deep neural networks or convolutional neural networks) [19] to train the surrogate model with randomly sampled parameters (instead of equi-spaced process parameter values).

## 4. Discussion

### 4.1. Computational efficiency of the surrogate model

The results in Section 3 not only illustrate the good accuracy relative to the high-fidelity phase-field model by which our surrogate model predicts the broad class of microstructures as a function of deposition conditions, it is also computationally efficient. The main computational cost is a one-time cost incurred during the training of the surrogate model since it necessitates the execution of $N_{\text{sim}}$ phase-field simulations to generate a data set of different types of microstructure morphologies as a function of the deposition parameters. The computational cost associated with the high-fidelity phase field simulations depends on (i) the size of the system and (ii) the deposition rate (*i.e.*, the slower the deposition rate the longer the simulation). For example, a PVD phase-field simulation, on a 256 × 256 domain, and for a (fast) deposition rate $\dot{d} = -1.85$ as shown in Fig. 7 requires approximately 15 mins on 72 processors (2 nodes with 36 processors per node) on a high-performance computer. Conversely, our trained PVD surrogate model takes less than a second for the same prediction. In fact, once trained, the computational time associated with the PVD surrogate model is independent of the deposition rate. Such computational gain by several orders of magnitude between the data-driven and the detailed phase-field model highlights the importance of pursuing this approach, especially when used for producing parameter-microstructure maps and not predicting the actual microstructure. In other words, surrogate models are particularly suited to rapidly explore the large number of combinations of processing parameters that would be needed to optimize the deposition conditions for a targeted microstructure.

### *4.2. Microstructure reconstruction*

At this stage of the surrogate-model construction, an estimated representation of the two-point statistics is sufficient to map the deposition conditions to the microstructure. It is, however, practical to obtain a visual estimation of the actual microstructure morphology from its statistical interpretation in the form: $\boldsymbol{m}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \mathcal{G}\big[\boldsymbol{S}_{\text{PVD}}\big(\boldsymbol{\theta}\big)\big]$, where $\mathcal{G}$ represents the phase recovery operation. Reconstructing a microstructure from its two-point statistics is a non-trivial and ill-posed problem [41]. Since the auto-correlation is obtained by computing the magnitude of the Fourier transform of $\boldsymbol{S}_2$, the phase information associated with the microstructure is lost. Additionally, the true microstructure is not uniquely defined by its two-point statistics. Similar microstructures can yield the same auto-correlation function, and therefore be statistically identical. Thus, any reconstruction of a microstructure from its two-point statistics may be a translation and/or an inversion of the original microstructure. Here, we discuss some of these issues by utilizing a microstructure reconstruction (phase-recovery) algorithm. The term "phase-recovery", as used in this section, refers to the recovery of the complex argument (or phase) of the FTT of the microstructure, and not the physical phase of a material.

#### *4.2.1. Phase-recovery algorithm*

Let us start with the description of the phase-recovery algorithm. Let $\boldsymbol{I}^{(1)}$ be the discretized function in Eq. (1), which represents the volume fraction of phase 1 at each spatial location $\boldsymbol{x}$. Then, as mentioned in Section 2.1, the auto-correlation function $\boldsymbol{S}_2$ of phase 1 is computed by using FFT such that,

$$\boldsymbol{S}_2 = \frac{1}{N}\mathcal{F}^{-1}\left(\overline{\mathcal{F}(\boldsymbol{I}^{(1)})}\mathcal{F}(\boldsymbol{I}^{(1)})\right), \tag{12}$$

$\mathcal{F}$ and $\mathcal{F}^{-1}$ represents the discrete Fourier transform and its inverse, respectively, and $\overline{\mathcal{F}}$ denotes the complex conjugate of $\mathcal{F}$. Note that the transformed function $\mathcal{F}(\boldsymbol{I}^{(1)})$ can be expressed as,

$$\mathcal{F}(\boldsymbol{I}^{(1)}) = |M| \exp{(\boldsymbol{i}\beta)}, \tag{13}$$

where $|M|$ is the complex modulus or amplitude, and $\beta$ is the complex argument or phase. Eq. (12) can alternatively be expressed as,

$$\boldsymbol{S}_2 = \mathcal{F}^{-1}\left(|M|^2\right). \tag{14}$$

Since the phase information is lost in the computation of the two-point statistics in Eq. (14), we employ a variation of the Gerchberg–Saxton phase recovery algorithm [42,43]. This algorithm was originally developed for optical applications, but was later adapted to microstructure reconstructions [30]. This algorithm alternates from the Fourier space to the real (or signal) space with the following steps:

1. *Compute the amplitude from the two-point statistics:* $|M_{\text{true}}| = \sqrt{N\mathcal{F}(\boldsymbol{S}_2)}$.
2. *Guess a trial microstructure:* choose a random initial microstructure $\boldsymbol{I}_0^{(1)}$ with magnitude $|M^{(0)}|$ and phase $\beta^{(0)}$ with the same volume fraction as the true solution.
3. *Replace the amplitude of the trial microstructure with that of the true microstructure:* use the phase of $\boldsymbol{I}_0^{(1)}$ and update the microstructure by

$$\boldsymbol{I}_{\text{trial}}^{(1)} = \mathcal{F}^{-1}\left(\frac{|M_{\text{true}}|}{|M_0|}\mathcal{F}(\boldsymbol{I}_0^{(1)})\right).$$

4. *Accept the trial microstructure entries:*

$$\boldsymbol{I}_{\text{next}}^{(1)} = \begin{cases} \boldsymbol{I}_{\text{trial}}^{(1)} & \text{if } \boldsymbol{I}_{\text{trial}}^{(1)} \geq \varepsilon \\ 0 & \text{if } \boldsymbol{I}_{\text{trial}}^{(1)} < \varepsilon \end{cases},$$

   where $\varepsilon$ is determined to ensure a consistent volume fraction.
5. *Iterate:* set $\boldsymbol{I}_0^{(1)} = \boldsymbol{I}_{\text{next}}^{(1)}$ and repeat steps 3–5 until convergence is achieved.

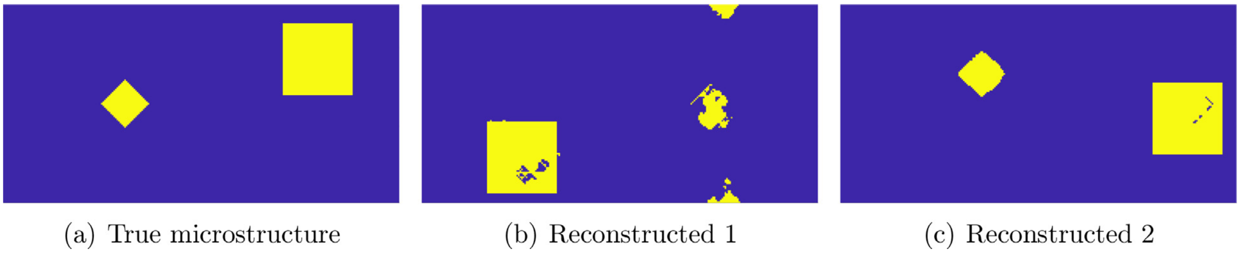

(a) True microstructure (b) Reconstructed 1 (c) Reconstructed 2

**Fig. 8.** Example of phase-recovery algorithm for microstructure reconstruction for different initial trial microstructures. (a) True microstructure, (b) reconstructed microstructure 1 with one initial trial microstructure $\boldsymbol{m}^{(\text{trial},1)}$, (c) reconstructed microstructure 2 with another initial trial microstructure $\boldsymbol{m}^{(\text{trial},2)}$.

In Fig. 8, we show two examples of reconstructed microstructures for different initial trial microstructures. There are no guarantees that this algorithm converges to a global minimum. It has been shown, however, that, (i) this algorithm is error-reducing, such that $\| |M^{(\text{true})}| - |M^{(0)}| \|_2$ is non-increasing, and (ii) that a global minimum exists [43]. Multiple variations have been proposed as an alternative to step 4 in order to take advantage of additional information about the signal or magnitude to produce hybrid algorithms [44,45]. Not only do some of these variations utilize specific information about the problem, there are also variations which address the fact that it is possible for this algorithm to stagnate. Since the microstructure reconstruction is a non-convex problem, there are many local minima for the algorithm to get stuck in. A viable approach to overcome this drawback would be to combine the above described phase recovery algorithm with a gradient-based approach for example.

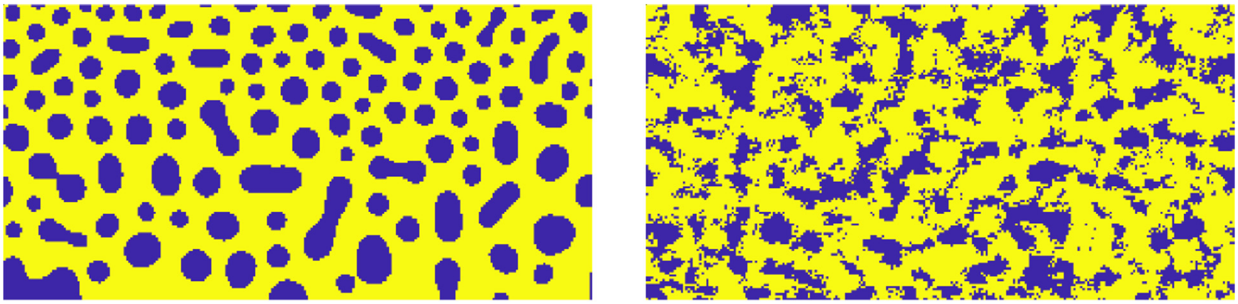
It should be noted that the present reconstruction algorithm reconstructs the microstructure solely based on $\boldsymbol{S}_2$. However, $\boldsymbol{S}_2$ cannot uniquely characterize the microstructure. Additional statistical functions requirements could be added to the above workflow to improve the microstructure reconstruction. For example, Jiao et al. [46,47] demonstrated an improvement of the reconstruction of a two-phase microstructure by including both the microstructure auto-correlation function and the two-point cluster-correlation function. The two-point cluster-correlation function quantifies the probability of finding two positions within phase $i$ that are at a distance '$\boldsymbol{r}$' and in the same cluster. Adding the two-point cluster-correlation function to the microstructure reconstruction algorithm is however computationally expensive (especially if the microstructure contains spatial directionality to it or for large microstructures). Such computational cost may outweigh the benefits of including two-point cluster-correlation functions in the context of constructing a rapid surrogate model for process-structure relationships.

### *4.2.2. Comparison of phase-field and surrogate model microstructures*

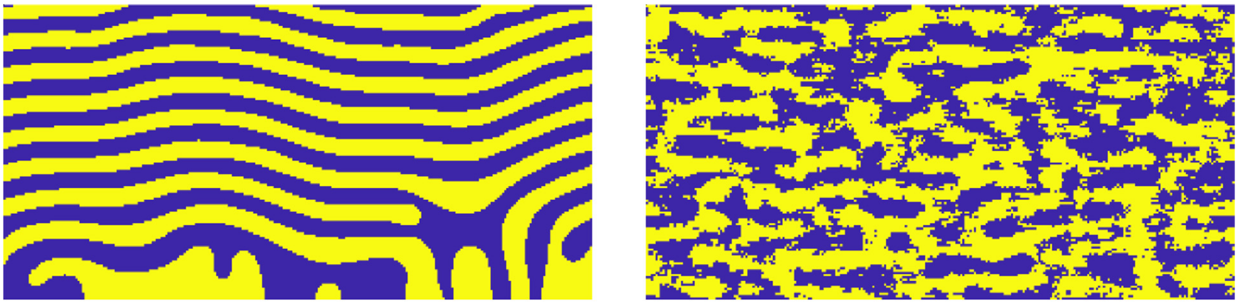
Using the microstructure-reconstruction methodology described above, we employ the phase-recovery algorithm to reconstruct the microstructure predicted by $\boldsymbol{S}_{\text{PVD}}(\boldsymbol{\theta})$ for direct comparison with those obtained by high-fidelity phase-field simulations. We run the phase-recovery algorithm with ten random initial microstructure and take the reconstructed microstructure with the best error. The phase-recovery algorithm is stopped if the change in the 3rd to last step does not differ from the current solution by $10^{-16}$, or if the number of iterations exceeds 150,000.

Fig. 9 shows examples of microstructure reconstructions for NPCM and VCM microstructures, respectively. In these two examples, we used our surrogate model, $\boldsymbol{S}_{\text{PVD}}(\boldsymbol{\theta})$, to compute the predicted two-point statistics, and then used these predicted two-point statistics to compute a reconstruction of the microstructure. While the reconstructed microstructures have more 'noise' than the microstructures simulated by the high-fidelity phase-field simulations, the general features of the microstructure are easily recognizable, *i.e.*, isolated precipitates for NPCM, and interconnected and horizontal domains for the VCM. Also, the reconstructed microstructures are in good agreement for similar deposition conditions as those observed experimentally [3,28], bolstering the predictability of the surrogate model.

Fig. 10 shows the relative error between the two-point statistical representation of the true microstructure and the two-point statistical representation from the reconstructed microstructure. Comparison of the two-point statistical representation demonstrates that the reconstructed microstructures are statistically equivalent to the true microstructures. However, some accuracy is lost due to, not only using surrogate model to predict the microstructure, but also employing the reconstruction algorithm. Predictions of the microstructure morphology tend to be better for NCPM and RCM microstructures as compared to LCM and VCM microstructures. One of the explanations behind this loss of accuracy is due to the composition of the ensemble of microstructures used to train the surrogate model. Indeed, the set of LCM and VCM microstructures constitutes only a small portion of this ensemble as compared to NPCM and RCM microstructure. This is due to the fact that deposition conditions yielding this LCM and VCM microstructure is narrower than those leading to NPCM and RCM microstructures. A more balanced set of microstructures may yield better results. For example, we could use importance sampling [48] to sample the deposition parameters corresponding to the transition from one type of microstructure morphology to another. Another possibility would be to smooth-out the reconstructed microstructures with just of few steps within the high-fidelity phase-field model.

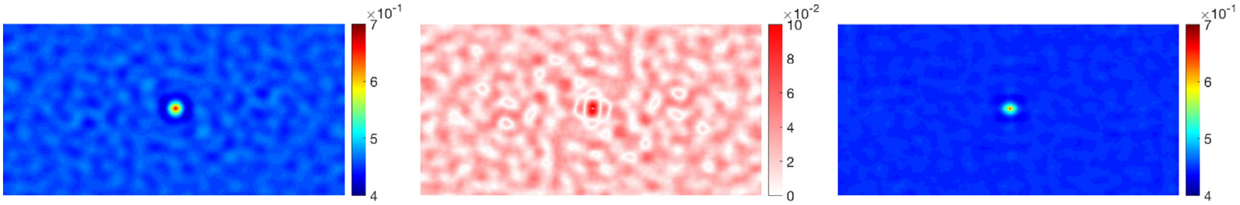


(a) NPCM phase-field (b) NPCM surrogate model

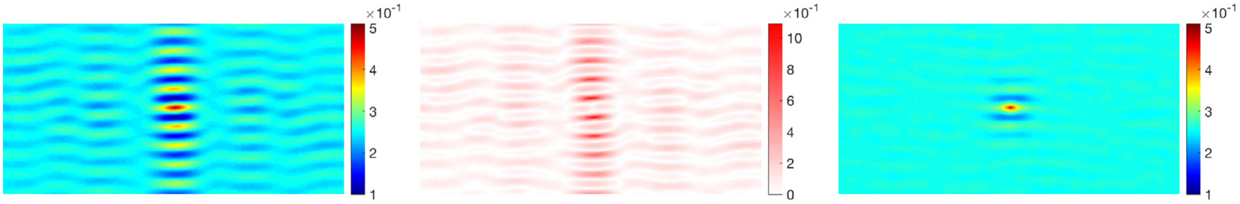


(c) VCM phase-field (d) VCM surrogate model

**Fig. 9.** Example microstructure reconstructions: (a) and (c) true microstructures as simulated by high-fidelity phase-field, (b) and (d) reconstructed microstructure using the phase-recovery algorithm on $\boldsymbol{S}_{\mathrm{PVD}}(\boldsymbol{\theta})$.



(a) NPCM true (b) NPCM point-wise error (c) NPCM reconstructed



(d) VCM true (e) VCM point-wise error (f) VCM reconstructed

**Fig. 10.** Comparison of microstructure two-point statistics: (a) and (d) true two-point statistics, $\boldsymbol{S}_{\mathrm{true}}$ for NPCM and VCM microstructures respectively, and (c) and (f) reconstructed microstructure $\boldsymbol{S}_{\mathrm{PVD}}$. The relative local error between true and predicted is illustrated in (b) and (e) for NPCM and VCM respectively.

## 5. Conclusion

In summary, our results show the efficiency of a phase-field-informed, surrogate model to rapidly predict microstructure morphologies from PVD. This surrogate model is trained against high-fidelity phase-field simulations for PVD of binary-alloy thin films. The key features composing the surrogate model include a reduced-dimensional representation of the microstructure via PCA, and a mapping back to the deposition conditions via polynomial chaos expansion. Major takeaways include:

1. Results for the two-point statistics of microstructures obtained by phase-field and the two-point statistics predicted by our surrogate model are in good agreement. This finding emphasizes the robustness and efficacy of the overall protocol.
2. Through an error analysis of both the PCA and polynomial chaos expansion, we show that using polynomial chaos expansion in the surrogate model results in the largest loss of accuracy for the prediction of the microstructure of microstructure morphologies as a function of deposition conditions. This is explained by the fact that the basis functions used in the polynomial expansion do not appropriately capture the rapid changes in microstructure morphologies incurred by small variations in the deposition conditions. Hence, we recommend to use importance sampling to build the training set of microstructures in order to better capture the transitions from one type of microstructure to another as the deposition conditions change. Additionally, we also recommend to use alternative techniques to polynomial chaos expansion to map the microstructure representation back to the deposition conditions.
3. The computational efficiency of the overall protocol makes it an attractive solution for rapidly producing parameter-microstructure maps that can be used to optimize deposition conditions.
4. While not necessarily needed in the process of creating the process-structure linkage, our results on microstructure reconstruction from the two-point statistics highlight issues associated with the non-uniqueness of the reconstruction process itself, and the associated decrease in the accuracy of the predictions. The inclusion of higher-order correlation functions to describe the microstructure may help solve this issue.

Overall, our findings provide new insights into the construction of a meaningful computer method to predict achievable microstructure morphologies during PVD. This is important because building a methodology that rapidly links processing conditions to the final microstructure for PVD is paramount to rationally design fabrication protocols to reliably synthesize functional thin films.

Moving forward, several improvements to the present protocol are warranted. As mentioned earlier, the polynomial chaos expansion approach used to map the microstructure representation back to the processing input space could be replaced by Galerkin projections or neural networks to train the surrogate model. Another interesting future direction is to extend this protocol to a broader range of deposition conditions and materials properties of the materials of the depositing species in order to develop a comprehensive SZD. Finally, the inclusion of experimental data within the training data set of microstructure would undoubtedly enrich the classes of microstructure prediction. This process would be increasingly important when introducing complexity in the material composition, or time-dependent deposition routes.

## Data availability

The raw simulation data required to reproduce these findings can be made available upon request. The most current version of the phase-field code used to generate the data, along with the MATLAB scripts use for the microstructure reconstruction, can be requested from the corresponding author (rdingre@sandia.gov) and is distributed at the discretion of Sandia National Laboratories through the Center for Integrated Nanotechnologies (CINT) User Program.

## Acknowledgments

This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy. This work was also supported by a Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in this article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

## Appendix A. Phase-field PVD model and numerical solution

The phase-field model used to simulate PVD is described in details elsewhere [6]. Below, we present the central features of this model. This model accounts for major aspects of the PVD process and has been shown to capture experimentally observed concentration modulations. Two conserved order-parameters, $\phi$ and $c$, are introduced to describe the structural and compositional ordering of the system as the thin film grows. The field variable $\phi$ distinguishes between the solid and vapor phases ($\phi = \pm 1$) and is used as the means of thin-film growth. The field variable $c$ distinguishes between the equilibrium components of the binary alloy species ($c = \pm 1$). Neglecting any elastic effects, the total free-energy of the system is taken as a functional of these field variables and their gradients, with the following form:

$$\mathcal{F} = \int \left\{ f_\phi + \frac{\kappa_\phi^2}{2}(\nabla\phi)^2 + s(\phi)\left( f_c + \frac{\kappa_c^2}{2}(\nabla c)^2 \right) \right\} \mathrm{d}\Omega. \tag{A.1}$$

Here, $\kappa_\phi$ and $\kappa_c$ are the gradient energy coefficients. The interpolation function, $s(\phi) = (1+\phi)^2/4$, ensures that compositional contributions are eliminated within the vapor phase. The first two terms in Eq. (A.1) describe energetics between the solid and the vapor, while the terms in parentheses describe the phase-ordering kinetics within the growing thin film. The double-well functions, $f_\phi$ and $f_c$, describe the equilibrium free-energy contributions of the solid-vapor and alloy component phases, respectively.

The equations of motion for thin-film growth and composition evolution are determined through minimization of the free energy in Eq. (A.1) according to the Cahn–Hilliard equation. For the composition evolution, the equation of motion is given as

$$\frac{\partial c}{\partial t} = \nabla \cdot \left[ M_c(\phi, c) \nabla \frac{\delta \mathcal{F}}{\delta c} \right], \tag{A.2}$$

where $M_c(\phi,\ c)$ is the structurally and compositionally dependent mobility function. This mobility function captures the relative difference between bulk and surface mobilities and allows for a smooth transition between phase A and phase B, and their respective bulk ($M_i^{\text{Bulk}}$) or surface ($M_i^{\text{Surf}}$) mobilities.

To describe thin-film growth, a source term is incorporated in the Cahn–Hilliard equation. The evolution of the solid-vapor field variable, $\phi$, is then given as

$$\frac{\partial \phi}{\partial t} = \nabla \cdot \left[ M(\phi) \nabla \frac{\delta \mathcal{F}}{\delta \phi} \right] + \nabla \phi \cdot (\rho \mathbf{v}). \tag{A.3}$$

The first term in Eq. (A.3) captures arbitrary surface morphology formation and surface diffusion effects (i.e., surface smoothening). The second term couples the thin-film evolution to the incident vapor flux and acts as the source term for interfacial growth and surface roughening.

Transport of the incident vapor flux towards the thin-film surface is modeled via the convection–diffusion equation,

$$\frac{\partial \rho}{\partial t} = \nabla \cdot [D_\rho \nabla \rho] - \nabla \cdot (\rho \mathbf{v}) - \nabla \phi \cdot (\rho \mathbf{v}) \tag{A.4}$$

where $D_\rho$ is the mass diffusivity of the vapor. The first two terms in Eq. (A.4) describe the diffusive and ballistic transport (strength and direction) of the depositing vapor towards the thin-film surface. The last term in Eq. (A.4) couples thin-film growth dynamics to the incident vapor dynamics and acts as a sink that removes vapor that has been converted to solid.

As described in Stewart and Dingreville [6], Eqs. (A.2)–(A.4) are numerically solved using the finite-difference method with second-order central difference stencils for all spatial derivatives. Numerical integration of the equations is performed using the explicit Euler method. All 2D simulations in this work are performed on a uniform mesh of 256 × 256 grid points with dimensionless numerical and physical parameters where $\Delta x = \Delta y = 1$ and $\Delta t = 10^{-2}$. The fixed model parameters $M_\phi$, $\omega_\phi$, $\kappa_\phi$, $\omega_c$, and $\kappa_c$ are taken to be 1, while $D_\rho = 5 \times 10^{-3}$. The initial configuration utilized in this work is that of an initially flat film that is two grid-points thick and spans the entire $x$-axis such that growth occurs along the $+y$-axis. Within this initial film $\phi = 1$ and $\rho = 0$ while $\phi = -1$ and $\rho = 1$ everywhere else (*i.e.*, in the vapor). The composition field within the simulation domain is randomly populated via sampling a truncated Gaussian distribution between $-1$ and 1 with a standard deviation of 0.35 and means of 0.00, 0.10, 0.20, 0.30 and 0.45 to generate nominally 50/50, 55/45, 60/40, 65/35, and 73/27 phase fraction distributions, respectively. Parallel to the film surface, periodic boundary conditions are applied to the $\phi$, $c$, and $\rho$ fields. At the bottom and top of the domain, Dirichlet boundary conditions are applied to $\phi$ and $\rho$ such that $\phi(y_{\text{min}}) = 1,\ \phi(y_{\text{max}}) = -1,\ \rho(y_{\text{min}}) = 0$, and $\rho(y_{\text{max}}) = \rho = 1$ while no-flux conditions are applied to $c$ at $y_{\text{min}}$ and $y_{\text{max}}$.

The coupled equations comprising the phase-field PVD model in Eqs. (A.2)–(A.4) have been implemented in Sandia's in-house multi-physics phase-field modeling code: Mesoscale Multiphysics Phase-Field Simulator (MEMPHIS). Verification and validation of this phase field are in good agreement with a suite of benchmark problems [49].

## References

[1] A. Chen, Z. Bi, C.-F. Tsai, J. Lee, Q. Su, X. Zhang, Q. Jia, J. MacManus-Driscoll, H. Wang, Tunable low-field magnetoresistance in $(La_{0.7}Sr_{0.3}MnO_3)_{0.5}$:$(ZnO)_{0.5}$ self-assembled vertically aligned nanocomposite thin films, Adv. Funct. Mater. 21 (13) (2011) 2423–2429, doi:10.1002/adfm.201002746.

[2] J. Ma, J. Hu, Z. Li, C. Nan, Recent progress in multiferroic magnetoelectric composites: from bulk to thin films, Adv. Mater. 23 (9) (2011) 1062–1087, doi:10.1002/adma.201003636.

[3] Y. Cui, B. Derby, N. Li, A. Misra, Design of bicontinuous metallic nanocomposites for high-strength and plasticity, Mater. Des. 166 (2019) 107602, doi:10.1016/j.matdes.2019.107602.

[4] J. Thornton, The microstructure of sputter-deposited coatings, J. Vac. Sci. Technol. A 4 (6) (1986) 3059–3065, doi:10.1116/1.573628.

[5] C. Thompson, Structure evolution during processing of polycrystalline films, Annu. Rev. Mater. Sci. 30 (2000) 159–190, doi:10.1146/annurev.matsci.30.1.159.

[6] J. Stewart, R. Dingreville, Microstructure morphology and concentration modulation of nanocomposite thin films during simulated physical vapor deposition, Acta Mater. 188 (2020) 181–191, doi:10.1016/j.actamat.2020.02.011.

[7] B. Movchan, A. Demchishin, Structure and properties of thick condensates of nickel, titanium, tungsten, oxides, and zirconium dioxide in vacuum, Fiz. Metal. Metalloved. 28 (1969) 653–660.

[8] J. Thornton, High rate thick film growth, Annu. Rev. Mater. Sci. 7 (1) (1977) 239–260, doi:10.1146/annurev.ms.07.080177.001323.

[9] A. Anders, A structure zone diagram including plasma-based deposition and ion etching, Thin Solid Films 518 (15) (2010) 4087–4090, doi:10.1016/j.tsf.2009.10.145.

[10] A. Matthews, Titanium nitride PVD coating technology, Surf. Eng. 1 (2) (1985) 93–104, doi:10.1179/sur.1985.1.2.93.

[11] G. Mauer, A. Hospach, R. Vaßen, Process development and coating characteristics of plasma spray-PVD, Surf. Coat. Technol. 220 (2013) 219–224, doi:10.1016/j.surfcoat.2012.08.067.

[12] U. Schulz, S. Terry, C. Levi, Microstructure and texture of EB-PVD TBCs grown under different rotation modes, Mater. Sci. Eng. A 360 (1–2) (2003) 319–329, doi:10.1016/S0921-5093(03)00470-2.

[13] L. Dong, R. Smith, D. Srolovitz, A two-dimensional molecular dynamics simulation of thin film growth by oblique deposition, J. Appl. Phys. 80 (10) (1996) 5682–5690, doi:10.1063/1.363621.

[14] J. Cho, S. Terry, R. LeSar, C. Levi, A kinetic monte carlo simulation of film growth by physical vapor deposition on rotating substrates, Mater. Sci. Eng. A 391 (1–2) (2005) 390–401, doi:10.1016/j.msea.2004.09.015.

[15] Y. Lu, C. Wang, Y. Gao, R. Shi, X. Liu, Y. Wang, Microstructure map for self-organized phase separation during film deposition, Phys. Rev. Lett. 109 (8) (2012) 086101, doi:10.1103/PhysRevLett.109.086101.

[16] S. Kalidindi, S. Niezgoda, A. Salem, Microstructure informatics using higher-order statistics and efficient data-mining protocols, JOM 63 (4) (2011) 34–41, doi:10.1007/s11837-011-0057-7.

[17] D. McDowell, S. Ghosh, S. Kalidindi, Representation and computational structure-property relations of random media, JOM 63 (3) (2011) 45–51, doi:10.1007/s11837-011-0045-y.

[18] J. Smith, W. Xiong, W. Yan, S. Lin, P. Cheng, O. Kafka, G. Wagner, J. Cao, W. Liu, Linking process, structure, property, and performance for metal-based additive manufacturing: computational approaches with experimental support, Comput. Mech. 57 (4) (2016) 583–610, doi:10.1007/s00466-015-1240-4.

[19] A. Cecen, H. Dai, Y. Yabansu, S. Kalidindi, L. Song, Material structure-property linkages using three-dimensional convolutional neural networks, Acta Mater. 146 (2017) 76–84, doi:10.1016/j.actamat.2017.11.053.

[20] A. Khosravani, A. Cecen, S. Kalidindi, Development of high throughput assays for establishing process- structure-property linkages in multiphase polycrystalline metals: application to dual-phase steels, Acta Mater. 123 (2017) 55–69, doi:10.1016/j.actamat.2016.10.033.

[21] M. Latypov, S. Kalidindi, Data-driven reduced order models for effective yield strength and partitioning of strain in multiphase materials, J. Comput. Phys. 346 (2017) 242–261, doi:10.1016/j.jcp.2017.06.013.

[22] N. Paulson, M. Priddy, D. McDowell, S. Kalidindi, Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics, Acta Mater. 129 (2017) 428–438, doi:10.1016/j.actamat.2017.03.009.

[23] L. Xia, B. Raghavan, P. Breitkopf, Towards surrogate modeling of material microstructures through the processing variables, Appl. Math. Comput. 294 (2017) 157–168, doi:10.1016/j.amc.2016.08.056.

[24] Y. Yabansu, P. Steinmetz, J. Hötzer, S. Kalidindi, B. Nestler, Extraction of reduced-order process-structure linkages from phase-field simulations, Acta Mater. 124 (2017) 182–194, doi:10.1016/j.actamat.2016.10.071.

[25] J. Stewart, A. Kohnert, L. Capolungo, R. Dingreville, Design and analysis of forward and reverse models for predicting defect accumulation, defect energetics, and irradiation conditions, Comput. Mater. Sci. 148 (2018) 272–285, doi:10.1016/j.commatsci.2018.02.048.

[26] J. Jung, J. Yoon, H. Park, J. Kim, H. Kim, An efficient machine learning approach to establish structure-property linkages, Comput. Mater. Sci. 156 (2019) 17–25, doi:10.1016/j.commatsci.2018.09.034.

[27] Z. Yang, X. Li, C. Brinson, A. Choudhary, W. Chen, A. Agrawal, Microstructural materials design via deep adversarial learning methodology, J. Mech.Des. 140 (11) (2018) 111416, doi:10.1115/1.4041371.

[28] M. Powers, B. Derby, R.E. Shaw A., A. Misra, Microstructural characterization of phase-separated co-deposited Cu–Ta immiscible alloy thin films, J. Mater. Res. (2020) 1–12, doi:10.1557/jmr.2020.100.

[29] B. Adams, X. Gao, S. Kalidindi, Finite approximations to the second-order properties closure in single phase polycrystals, Acta Mater. 53 (2005) 3563–3577, doi:10.1016/j.actamat.2005.03.052.

[30] D. Fullwood, S. Niezgoda, S. Kalidindi, Microstructure reconstructions from 2-point statistics using phase-recovery algorithms, Acta Mater. 56 (2008) 942–948, doi:10.1016/j.actamat.2007.10.044.

[31] S. Niezgoda, Y. Yabansu, S. Kalidindi, Understanding and visualizing microstructure and microstructure variance as a stochastic process, Acta Mater. 59 (2011) 6387–6400, doi:10.1016/j.actamat.2011.06.051.

[32] B. Adams, S. Kalidindi, D. Fullwood, Microstructure Sensitive Design for Performance Optimization, 1st, Butterworth-Heinemann, Newton, MA, USA, 2012.

[33] S. Torquato, H. Haslach Jr, Random heterogeneous materials: microstructure and macroscopic properties, Appl. Mech. Rev. 55 (4) (2002) B62–B63, doi:10.1115/1.1483342.

[34] S. Torquato, Nearest-neighbor statistics for packings of hard spheres and disks, Phys. Rev. E 51 (4) (1995) 3170, doi:10.1103/PhysRevE.51.3170.

[35] B. Lu, S. Torquato, Lineal-path function for random heterogeneous materials, Phys. Rev. A 45 (2) (1992) 922, doi:10.1103/PhysRevA.45.922.

[36] S. Niezgoda, D. Fullwood, S. Kalidindi, Delineation of the space of 2-point correlations in a composite material system, Acta Mater. 56 (18) (2008) 5285–5292, doi:10.1016/j.actamat.2008.07.005.

[37] C. Suh, A. Rajagopalan, X. Li, K. Rajan, The application of principal component analysis to materials science data, Data Sci. J. 1 (2002) 19–26, doi:10.2481/dsj.1.19.

[38] R. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2013.

[39] B. Debusschere, H. Najm, P. P. Pébay, O. Knio, R. Ghanem, O. Le Maître, Numerical challenges in the use of polynomial chaos representations for stochastic processes, SIAM J. Sci. Comput. 26 (2004) 698–719, doi:10.1137/S1064827503427741.

[40] M. Gerritsma, J.-B. Van der Steen, P. Vos, G. Karniadakis, Time-dependent generalized polynomial chaos, J. Comput. Phys. 229 (22) (2010) 8333–8363, doi:10.1016/j.jcp.2010.07.020.

[41] R. Bostanabad, Y. Zhang, X. Li, T. Kearney, L. Brinson, D. Apley, W. Liu, W. Chen, Computational microstructure characterization and reconstruction: review of the state-of-the-art techniques, Prog. Mater. Sci. 95 (2018) 1–41, doi:10.1016/j.pmatsci.2018.01.005.

[42] G. Yang, B. Dong, B. Gu, J. Zhuang, O. Ersoy, Gerchberg–Saxton and Yang–Gu algorithms for phase retrieval in a nonunitary transform system: a comparison, Appl. Opt. 33 (2) (1994) 209–218, doi:10.1364/AO.33.000209.

[43] Y. Shechtman, Y. C. Eldar, H. Cohen O.and Chapman, J. Miao, M. Segev, Phase retrieval with application to optical imaging: a contemporary overview, Signal Process. Mag. 32 (2015) 87–109, doi:10.1109/MSP.2014.2352673.

[44] J. Fienup, Reconstruction of an object from modulus of its Fourier transform, Opt. Lett. 3 (1978) 27–29, doi:10.1364/OL.3.000027.

[45] S. Mukherjee, C. Seelamantula, An iterative algorithm for phase retrieval with sparsity constraints: application to frequency domain optical coherence tomography, in: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 553–556, doi:10.1109/ICASSP.2012.6287939.

[46] Y. Jiao, F. Stillinger, S. Torquato, Modeling heterogeneous materials via two-point correlation functions: basic principles, Phys. Rev. E 76 (3) (2007) 031110, doi:10.1103/PhysRevE.76.031110.

[47] Y. Jiao, F. Stillinger, S. Torquato, Modeling heterogeneous materials via two-point correlation functions. II. Algorithmic details and applications, Phys. Rev. E 77 (3) (2008) 031135, doi:10.1103/PhysRevE.77.031135.

[48] C. Clark, Importance sampling in Monte Carlo analyses, Oper. Res. 9 (5) (1961) 603–620, doi:10.1287/opre.9.5.603.

[49] R. Dingreville, J. Stewart, E. Chen, Benchmark Problems for the Mesoscale Multiphysics Phase Field Simulator (MEMPHIS), Technical Report SAND2020-4015, Sandia National Laboratories, Albuquerque, NM (United States), 2020, doi:10.2172/1615889.