Acta Materialia 103 (2016) 89–102

Contents lists available at ScienceDirect

Acta Materialia

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

Stochastic microstructure characterization and reconstruction via supervised learning

Ramin Bostanabad [a], Anh Tuan Bui [b], Wei Xie [c], Daniel W. Apley [b, *], Wei Chen [a]

[a] Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA
[b] Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208, USA
[c] Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, NY 12180, USA

ARTICLE INFO

Article history:
Received 13 July 2015
Received in revised form
22 September 2015
Accepted 23 September 2015
Available online xxx

Keywords:
Stochastic microstructure
Characterization and reconstruction
Supervised learning
Statistical equivalency

ABSTRACT

Microstructure characterization and reconstruction have become indispensable parts of computational materials science. The main contribution of this paper is to introduce a general methodology for practical and efficient characterization and reconstruction of stochastic microstructures based on supervised learning. The methodology is general in that it can be applied to a broad range of microstructures (clustered, porous, and anisotropic). By treating the digitized microstructure image as a set of training data, we generically learn the stochastic nature of the microstructure via fitting a supervised learning model to it (we focus on classification trees). The fitted supervised learning model provides an implicit characterization of the joint distribution of the collection of pixel phases in the image. Based on this characterization, we propose two different approaches to efficiently reconstruct any number of statistically equivalent microstructure samples. We test the approach on five examples and show that the spatial dependencies within the microstructures are well preserved, as evaluated via correlation and lineal-path functions. The main advantages of our approach stem from having a compact empirically-learned model that characterizes the stochastic nature of the microstructure, which not only makes reconstruction more computationally efficient than existing methods, but also provides insight into morphological complexity.

1. Introduction

To date, considerable research has been conducted towards computational discovery and design of advanced materials. Many of these works are focused on a class of materials that are composed of multiple distinct constituents/phases and have an underlying stochastic behavior. This class is termed random heterogeneous materials and is ubiquitous in science and engineering (i.e., polymer nanocomposites) as well as the nature (i.e., sandstone) [1]. We consider the problem of developing a general methodology for characterization and reconstruction of a broad range of random heterogeneous microstructures. The major challenges with this problem are twofold: (1) How to efficiently and accurately quantify (characterize) the stochastic nature of high dimensional data embedded in the material morphology and (2) how to use this characterization to generate (reconstruct) virtual microstructure samples that are statistically equivalent, preserving as much of the inherent stochasticity as possible. In this paper, we address this problem via a supervised learning approach in which we first fit a flexible model to the high dimensional data and then employ the fitted model for reconstruction. It should be mentioned that by "microstructure" we mean a structure whose microscopic features are smaller than the characteristic length-scale of the macroscopic sample but larger than the molecular spatial arrangements [2]. We note, however, that the concept is relative and the scales can be transcended (see for example [3–5]).

Recent advances in imaging techniques [6–9] have enabled the collection of digital structural information at various scales. The need for computational characterization and reconstruction of the collected data for understanding the role of structure in processing-structure-property linkage is highlighted in the literature [10–20]. They provide the means for building an ensemble of representative volume elements (RVE's) or statistical volume elements (SVE's), which is used for estimating materials properties (see Refs. [10,11] for discussion on RVE's and SVE's). Broadly, we have classified related prior works on characterization and reconstruction into

* Corresponding author.
E-mail address: Apley@northwestern.edu (D.W. Apley).

http://dx.doi.org/10.1016/j.actamat.2015.09.044
1359-6454/

three categories based on the reconstruction procedure: (1) *Optimization*, (2) *Random Field*, and (3) *Texture Synthesis* and the closely-related *multiple-point statistics* (there are some works that do not fall into the aforementioned categories though. See Refs. [21,22]). We note that throughout the paper we use the term "image" to refer to a microstructure sample, either the original or reconstructed, because the samples are represented as images (e.g., a two-phase microstructure sample is represented as a binary black/white image, where the color corresponds to the morphological phase at that pixel location).

In the first category, the reconstructed image is iteratively adjusted (i.e., optimized) so as to minimize an appropriately defined cost function that measures the statistical differences between the original image and the reconstructed one. The choice of the cost function depends on the characterization scheme. One set of approaches [1,23–34] characterize the material structure via various correlation functions (i.e., two-point, lineal-path, and two-point cluster). In reconstruction, first an initial random image with the same volume fraction[1] (VF) for all the phases/constituents as in the original image is generated. Then, its pixels are iteratively swapped via some heuristic optimization algorithm (i.e., simulated annealing [27,32,34] or genetic algorithm [35,36]) to reduce some cost function that measures the differences between the correlation functions of the original image and those of the reconstructed one (as most microstructures cannot be characterized solely by one specific correlation function, usually multiple of them are incorporated into the cost function). Although several improvements in the pixel-swapping heuristics have been developed [26,33,37–40], the optimization is still computationally expensive and prohibitive for reconstructing large/anisotropic images. In addition, correlation functions are infinite dimensional in a sense that they do not (at least explicitly) represent the results for a specific set of descriptors (i.e. average particle size or nearest neighbor). Although several analytical expressions have been developed in the literature to relate the correlation functions to the physical descriptors of the structure, they are problem-dependent (see Refs. [23,27,41–46] for some examples considering two-point correlation function).

A second set of optimization-based approaches uses physical descriptor characteristics of the original image to characterize the microstructure [41,47–50] and the optimization process aims to preserve as many important descriptor characteristics as possible. The choice of physical descriptors depends on the materials and properties of interest. For example, because transport processes in particulate heterogeneous systems are sensitive to nearest neighbor distances between particles [2], one might choose nearest neighbor distances between particles as one of the descriptors and attempt to match its distributional characteristics (such as the mean and variance) in the original and reconstructed samples. This method sensibly characterizes the topological features and hence can be used for design [17] purposes (for instance, one can adjust the distribution of nearest neighbors and investigate how it would affect the material property). However, it requires image analysis (for extracting the characteristics of the descriptors from the original image), and one needs to define/choose the appropriate descriptors beforehand. In addition, extension of the methodology to reconstruct microstructures with irregular inclusions or those that are dense and multiphase is nontrivial. The associated computational costs are generally lower than the previous approach but they will increase if a pixel/voxel moving scheme [41] is required for matching some of the descriptor (i.e., nearest neighbor).

In the second category, random fields (RFs) are utilized for material (in particular porous media) characterization and reconstruction [51–54]. These methods are faster than correlation function-based ones and model the microstructure phase by level-cutting a relatively simple Gaussian RF. They rely heavily on correlation functions, in the sense that the Gaussian RF model is fitted by matching its correlation functions to those of the training image. The methodology is usually restricted to bi-phase isotropic structures based on only two-point correlation function and hence lacks high versatility accuracy for microstructures that cannot be characterized solely by their two-point correlation function [27]. In Refs. [53,55,56] RF's were integrated with an optimization-based approach by post-processing an RF-based reconstructed image to reduce the differences between correlation functions. Although this approach improves the accuracy (in matching correlation functions), it is still subject to the limitations discussed above.

Recently, texture synthesis methods that were originally developed for computer graphics problems [57–59] have been applied to material characterization and reconstruction. This approach assumes that the microstructure behaves as a stationary Markov random field (MRF) (see Sec. 2.1 for further discussion on this property). In this approach, no characterization is done and an image is reconstructed pixel-by-pixel (voxel in 3D) in a specific order (i.e. raster scan). Each pixel's value in the reconstructed image is found by searching for the pixel (or a set of pixels) in the original image whose neighboring pixels best match the neighbors of the pixel to be generated. Different methods differ in their choice of neighborhood geometry, definition of similarity, and search method. Sundaraghavan [60] uses a non-causal neighborhood (see Sec. 2.2.1) to reconstruct a 3D structure based only on three 2D images taken along orthogonal directions and validates the results by comparing correlation functions of the original and reconstructed images. Liu et al. [13] use the texture synthesis methodology developed in Refs. [57], but instead of choosing the single pixel with the best-matched neighborhood, they randomly choose from among a set of pixels with closely-matched neighborhoods (as in Ref. [59]). They reconstruct multi-phase 2D/3D structures via a causal neighborhood (see Sec. 2.2.2) and show various material characteristics (i.e. correlation functions and Minkowski functionals) are well preserved. Texture synthesis based methods are applicable to isotropic, anisotropic, and multi-phase materials.

Works similar to texture synthesis can also be found in the earth science literature [61–65] where characterization and reconstruction of porosity in geological structures (such as soil and sandstone) is of particular interest. The methodology used in these works is based on multiple-point statistics, characterizing the structure by calculating and storing the conditional probabilities of finding a specific phase at a pixel, given the phases of a particular configuration of neighboring pixels. Like the texture synthesis approach, the multiple-point statistics approaches implicitly characterize the microstructure by exhaustive enumeration of all possible phase combinations for all possible neighborhood configurations that have occurred in the training image. Reconstruction is accomplished pixel-by-pixel, also similar to the texture synthesis reconstruction, by searching for the training neighborhood that best matches that of the pixel being reconstructed and subsequently sampling from the conditional probability for that neighborhood. Different methods vary in their choice of neighborhood geometry, search method, and reconstruction order (i.e., random or raster scan). Wu et al. [61] use two small causal neighborhoods, a two-pixel neighborhood for boundary pixels and a five-pixel neighborhood for the other pixels, for characterization. Hajizadeh et al. [63,64] and Okabe et al. [64] use the algorithm developed in Ref. [65] and a non-causal neighborhood for characterization. In Refs. [61,63,64], the original image must be stationary and the geometry and size of the neighborhood are determined manually.

[1] Area Fraction (AF) in 2D. We use the common acronym VF throughout the paper but the meaning is clear from the context.

Our fundamental idea is to treat the microstructure characterization problem as a data-driven supervised learning (aka machine learning or statistical learning) one. Specifically, using the microstructure image as the training dataset, we fit a supervised learning model to predict the phase of an image pixel as a function of the phases of a (suitably large) neighborhood of its surrounding pixels. In supervised learning parlance, the response variable is the phase of a pixel, and the predictor variables are the phases of some neighborhood of surrounding pixels. The fitted supervised learning model can be viewed as a predictive model representing the conditional distribution of each pixel's phase, given its neighbors' phases. This set of conditional distributions embodied by the supervised learning model provides an implicit characterization (Sec. 2.1) of the full joint distribution of all the pixels within the image, which is the most complete and generic statistical characterization possible. It also provides a computationally efficient means of generating statistically equivalent reconstructed microstructures (Sec. 2.2). In theory, any supervised learning (Sec. 2.3) method can be used but we focus on classification trees because of their computational efficiency, interpretability, and suitability for handling categorical variables (pixel phases). It should be noted that fitting classification trees as we do is fundamentally different than using a tree structure to exhaustively enumerate all of the phase combinations and neighborhood configurations that have occurred in the training image, as is done in some of the multiple-point statistics [63] or texture synthesis [13] methods. In Sec. 3, we test the algorithm on five different examples and illustrate that the model can characterize and reconstruct a wide range of stochastic microstructures (e.g., isotropic or anisotropic).

The major contribution of this paper is to use, for the first time to the best of the authors' knowledge, a direct supervised learning approach for the problem of microstructure characterization and reconstruction. Desirable aspects of this approach are that (like texture synthesis and the multiple-point statistics methods) it is more flexible and generic than descriptor-based approaches, being applicable to microstructures that are isotropic or anisotropic and with randomly varying, irregularly-shaped inclusions; and (unlike texture synthesis and the multiple-point statistics methods) it also results in a compact model which characterizes the stochastic nature of the microstructure and is learned in a completely data-driven manner. Texture synthesis methods involve no fitted model to characterize the microstructure, and reconstruction is done by exhaustively searching for similar neighborhoods in the original image. The multiple-point statistics methods similarly involve no fitted model, other than exhaustive enumeration of every phase combination for every neighborhood configuration that has occurred at least once in the training image. Having a compact model entails a number of advantages including significant computational efficiency when reconstructing new samples and a means of comparing different microstructures that provides insight into the material's morphology. Moreover, the model-based supervised learning framework has the potential to be extended in a number of directions, which are mentioned in Sec. 4.

2. Supervised learning approach for microstructure characterization and reconstruction

2.1. Overview of the approach and assumptions

Let $\boldsymbol{X}$ denote the collection of pixels in the original (training) microstructure image of size n_1 (rows) $\times$ n_2 (columns). The pixels in $\boldsymbol{X}$ are ordered, and each one is a categorical variable that indicates the phase at that spatial location. While the proposed algorithm is applicable to multi-phase microstructures, in this paper we limit ourselves to bi-phase materials for notational and illustrative simplicity. Therefore, the elements in $\boldsymbol{X}$ are the binary variables, $X_{ij}\in\{0,1\}$ for $i=1,2,\ldots,n_1$ and $j=1,2,\ldots,n_2$. $\boldsymbol{X}$ can be thought of as a random sample from its underlying full joint distribution, denoted by $f(\boldsymbol{X})$. From this perspective, in order to reconstruct a new but statistically equivalent image $\boldsymbol{Y}$ (of any size) we must learn $f(\boldsymbol{X})$ from the training image and use it for reconstruction. With no assumptions, it is obviously impossible to estimate an extremely high dimensional distribution like $f(\boldsymbol{X})$ from only a single realization (the training image).

To overcome this difficulty, we assume that the random microstructure can be modeled as a form of stationary MRF, which is also assumed in the texture synthesis-based works [13,60]. Intuitively, the Markovian assumption means the following: Given a sufficiently large neighborhood $\boldsymbol{N}_{ij}$ surrounding pixel X_{ij} (see Fig. 1), there is no additional information contained in the *remaining* pixels of $\boldsymbol{X}$ that could further improve the predictability of X_{ij}. Let $\boldsymbol{X}^{(-ij)}\equiv\{X_{mn}:(m,n)\neq(i,j),\ m=1,2,\ldots,n_1,\ n=1,2,\ldots,n_2\}$ denote the set of all pixels in $\boldsymbol{X}$ excluding X_{ij}. Mathematically, the stationary MRF assumptions are:

- Locality: $f(X_{ij}|\boldsymbol{X}^{(-ij)})=f(X_{ij}|\boldsymbol{N}_{ij})$ for a sufficiently large neighborhood $\boldsymbol{N}_{ij}$.
- Stationarity: $f(X_{ij}|\boldsymbol{N}_{ij})$ does not depend on pixel location (i,j).

In the preceding, it should be noted that the conditional distribution $f(X_{ij}|\boldsymbol{N}_{ij})$ is a *Bernoulli* distribution with an event probability (the event being defined as $X_{ij}=1$) that is some function of the pixel values in $\boldsymbol{N}_{ij}$. Essentially, the goal of the supervised learning step is to learn this predictive function by fitting an appropriate supervised learning model to the training image. The appropriate size of $\boldsymbol{N}_{ij}$ is typically on the order of the largest topological feature in $\boldsymbol{X}$, although this also can be learned from the data and the algorithm is not significantly sensitive to it (we discuss this further in Sec. 2.3 and Sec. 3.5). Regarding the stationarity assumption, we temporarily ignore any boundary effects and discuss them in Sec. 2.4.

Fig. 2 is a flowchart of the basic procedure to reconstruct a new image $\boldsymbol{Y}$ that is statistically equivalent to $\boldsymbol{X}$. The training data, to which the supervised learning model for $f(X_{ij}|\boldsymbol{N}_{ij})$ is fitted, consists of the paired observation $(X_{ij},\ \boldsymbol{N}_{ij})$ for $i\in\{1,2,\ldots,\ n_1\}$ and $j\in\{1,2,\ldots,n_2\}$. In Sec. 2.3, we discuss details on the choice of $\boldsymbol{N}_{ij}$ and the

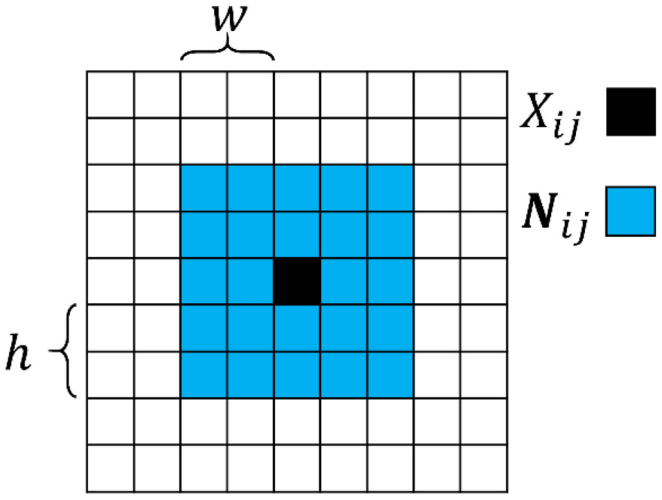

Fig. 1. A square (non-causal) neighborhood $\boldsymbol{N}_{ij}$ of X_{ij} with size $w=h=2$. The response pixel and those within $\boldsymbol{N}_{ij}$ are color-coded as, respectively, black and blue (the colors do not represent phase values. The reader is referred to the web version of this article for interpretation of the colors).

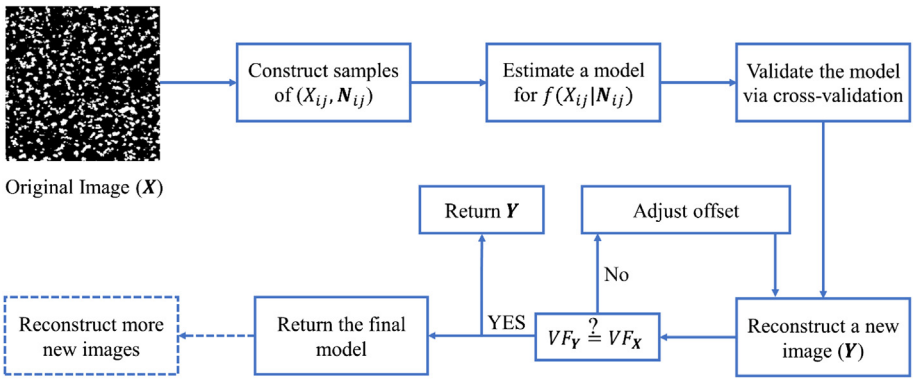

Fig. 2. Flowchart of the supervised learning approach for microstructure characterization and reconstruction.

supervised learning algorithm for empirically estimating $f(X_{ij}|\boldsymbol{N}_{ij})$. Given the fitted supervised learning model for $f(X_{ij}|\boldsymbol{N}_{ij})$, various methods can be used for generating a statistically equivalent reconstructed image $\boldsymbol{Y}$ as a random sample from a joint distribution consistent with $f(\boldsymbol{X})$. In Sections 2.2.1 and 2.2.2 we discuss, respectively, non-causal and causal approaches for accomplishing the reconstruction. In Sec. 2.4 details regarding initialization and boundary effects are discussed.

2.2. Reconstruction

We present two approaches – non-causal and causal – for generating reconstructed images, each having their respective advantages.

2.2.1. Non-causal approach

The non-causal reconstruction procedure is based on Gibbs sampling [66]. Gibbs sampling is a general statistical technique for drawing a random sample $\boldsymbol{Y}$ from some multivariate distribution $f(\boldsymbol{Y})$. Gibbs sampling is useful when it is difficult to directly sample from $f(\boldsymbol{Y})$, but much easier to sample from the conditional distributions $f(Y_{ij}|\boldsymbol{Y}^{(-ij)})$ for each element Y_{ij} of $\boldsymbol{Y}$. It is an iterative procedure, in which one starts with some initial $\boldsymbol{Y}$. Sequentially, each element Y_{ij} is generated anew from $f(Y_{ij}|\boldsymbol{Y}^{(-ij)})$. When the last element of $\boldsymbol{Y}$ is generated, one starts over and sequentially generates each element of $\boldsymbol{Y}$ again from $f(Y_{ij}|\boldsymbol{Y}^{(-ij)})$, using the most recently generated $\boldsymbol{Y}^{(-ij)}$. The procedure iterates, generating $\boldsymbol{Y}$ as many times as needed, until a form of statistical convergence is reached. After convergence, the generated $\boldsymbol{Y}$ in the last iteration can be viewed as a random sample from $f(\boldsymbol{Y})$.

For our application, by the stationary MRF assumption, the conditional distributions $f(Y_{ij}|\boldsymbol{Y}^{(-ij)})=f(Y_{ij}|\boldsymbol{N}_{ij})$ are independent of the pixel index (i,j) (hence, the same model for $f(Y_{ij}|\boldsymbol{N}_{ij})$ can be used when generating pixel values at every spatial location in the image), and the model $f(Y_{ij}|\boldsymbol{N}_{ij})$ is learned directly in the supervised learning stage. Since $f(Y_{ij}|\boldsymbol{N}_{ij})$ is a Bernoulli distribution with event probability that depends on the neighborhood pixel values, sampling from the conditionals $f(Y_{ij}|\boldsymbol{N}_{ij})$ is straightforward. The following is pseudocode for non-causal reconstruction of a microstructure image $\boldsymbol{Y}$ (of any specified size $s_1 \times s_2$) based on Gibbs sampling. The algorithm assumes that $f(Y_{ij}|\boldsymbol{N}_{ij})$ has already been learned in the supervised learning stage (Sec. 2.3).

1. Start with an initial image $\boldsymbol{Y}^{(0)}$ of size $m_1 \times m_2$, where $m_1 > s_1$ and $m_2 > s_2$ (see Sec. 2.4 for initialization)
2. For $k=1,2,\ldots, K$
   a. Set $\boldsymbol{Y}^{(k)}=\boldsymbol{Y}^{(k-1)}$
   b. For $i=1,2,\ldots, m_1$ and $j=1,2,\ldots, m_2$
      i. Extract the neighborhood $\boldsymbol{N}_{ij}^{(k)}$ from $\boldsymbol{Y}^{(k)}$
      ii. Use the fitted supervised model to predict the Bernoulli parameter $p_{ij} = f(Y_{ij}|\boldsymbol{N}_{ij}^{(k)})$ and generate Y_{ij}~Berourlli(p_{ij}).
      iii. Use the newly generated phase value Y_{ij} to update the corresponding pixel in $\boldsymbol{Y}^{(k)}$
3. To avoid boundary inaccuracies (see Sec. 2.4), retain the central $s_1 \times s_2$ portion of $\boldsymbol{Y}^{(K)}$ as the reconstructed image.

In the preceding reconstruction algorithm, we generate the pixels in a raster scan order, as illustrated in Fig. 3. In the k^{th} iteration, not every pixel in $\boldsymbol{N}_{ij}^{(k)}$ is available when Y_{ij} is close to the boundaries (see Fig. 3(a)). We refer to this as *missing data* throughout the paper. The missing data effect causes reconstruction inaccuracies near the boundaries of the reconstructed image that do not disappear even when K becomes large. Therefore, our strategy is to reconstruct a larger image than required and discard the boundary regions, choosing the central part as the final reconstructed image (see Sec. 2.4 for more details).

A potential drawback of the non-causal reconstruction approach is its computational cost, as K might need to be large. This is especially true if $\boldsymbol{Y}$ is large and we need to reconstruct multiple realizations. The value of K primarily depends on (1) how much the initial image differs from the original image and (2) the morphological complexity of the microstructure. Our studies suggest that K usually must be more than 50. Next, we introduce a causal reconstruction approach that avoids this issue.

2.2.2. Causal approach

The joint distribution $f(\boldsymbol{X})$ can be written in the factorial form as:

$$\begin{aligned} f(\boldsymbol{X}) &= f(X_{11}) f(X_{12}|X_{11}) f(X_{13}|X_{11}, X_{12}) \cdots f\left(X_{n_1 n_2}|X_{11}, X_{12}, \ldots, X_{n_1(n_2-1)}\right) \\ &= f\left(X_{11}\left|\boldsymbol{X}^{(<11)}\right.\right) f\left(X_{12}|\boldsymbol{X}^{(<12)}\right) f\left(X_{13}|\boldsymbol{X}^{(<13)}\right) \cdots f\left(X_{n_1 n_2}|\boldsymbol{X}^{(<n_1 n_2)}\right) \end{aligned} \quad (1)$$

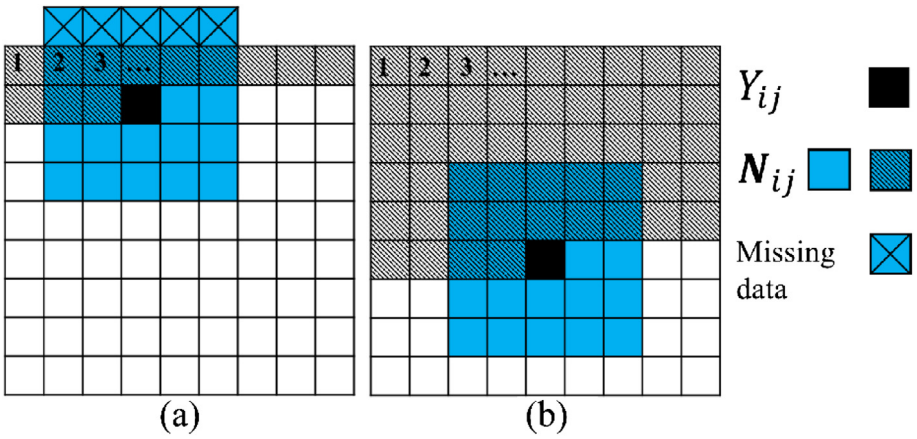

Fig. 3. Non-causal reconstruction with neighborhood size $w=h=2$ when pixel (i,j) to be reconstructed is (a) close to the boundary, (b) at the central part. As the numbers indicate, the image is reconstructed in a raster scan order and the pixels are color-coded to distinguish between the ones that are reconstructed in the current iteration (patterned) and those that are reconstructed in the previous iteration (white). The colors do not represent phase values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $\boldsymbol{X}^{(<ij)}$ denotes the set of all the pixels in $\boldsymbol{X}$ ordered before X_{ij}. If we have all of the conditional distributions in Eq. (1), we can use this to sequentially generate the pixels in a new random draw $\boldsymbol{Y}$ from the same distribution $f(\boldsymbol{X})$ as follows. First generate Y_{11} from $f(Y_{11}|\boldsymbol{Y}^{(<11)})$; then, conditional on Y_{11}, generate Y_{12} from $f(Y_{12}|\boldsymbol{Y}^{(<12)})$ and so on. Notice that conditional distributions of Y_{ij} (for $i=1,2,\ldots, n_1$ and $j=1,2,\ldots, n_2$) in the right-hand side of Eq. (1) only depend on the pixels previously generated. Thus, the causal approach reconstructs a new image $\boldsymbol{Y}$ via a *single pass* over the pixels in $\boldsymbol{Y}$. This is a major computational advantage over the non-causal approach, which must make multiple passes over the pixels in $\boldsymbol{Y}$.

Since the conditional distributions $f(Y_{ij}|\boldsymbol{Y}^{(<ij)})$ are required for reconstruction, we need to estimate them via the original image. By applying the locality assumption of MRF, we have $f(Y_{ij}|\boldsymbol{Y}^{(<ij)}) = f(Y_{ij}|\boldsymbol{Y}^{(<ij)}\cap\boldsymbol{N}_{ij})$, where the required size of the neighborhood $\boldsymbol{N}_{ij}$ is not necessarily the same as for the non-causal approach. Let $\boldsymbol{M}_{ij}\equiv\{\boldsymbol{Y}^{(<ij)}\cap\boldsymbol{N}_{ij}\}$ denote the causal neighborhood of Y_{ij} as illustrated in Fig. 4. Although the number of pixels in $\boldsymbol{N}_{ij}$ is fixed, that of $\boldsymbol{Y}^{(<ij)}$ depends on how close pixel index (i,j) is to the boundary. Thus, the number of pixels in $\boldsymbol{M}_{ij}$ (and hence its geometry) is different near the boundaries. One strategy would be to fit a collection of supervised learning models for $f(Y_{ij}|\boldsymbol{M}_{ij})$ for causal neighborhoods $\boldsymbol{M}_{ij}$ of various sizes. That is, fit a supervised learning model to predict $f(Y_{ij}|Y_{i,j-1})$, fit a second supervised learning model to predict $f(Y_{ij}|\{Y_{i,j-1},Y_{i,j-2}\})$, etc., up to the largest size causal neighborhood when $\boldsymbol{N}_{ij}$ no longer extends beyond the boundary of the image. Such a strategy would yield a more exact implementation of reconstruction sampling via Eq. (1) and would eliminate any inaccuracies due to boundary approximations. However, fitting multiple supervised learning models would be cumbersome and computationally expensive. Consequently, in light of the similarity between this boundary issue and the boundary issue for the non-causal reconstruction approach, we simplify the implementation and fit only a single supervised learning model for $f(Y_{ij}|\boldsymbol{M}_{ij})$ with the largest size $\boldsymbol{M}_{ij}$, as illustrated in Fig. 5. Inaccuracies near the boundary of the reconstructed image are handled as in the non-causal approach, by generating a larger $\boldsymbol{Y}$ than needed and discarding the boundary (and by using the initialization strategy outlined in Sec. 2.4).

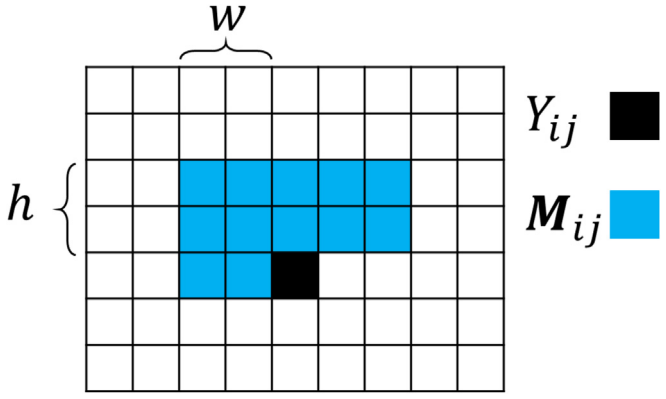

Fig. 4. A causal neighborhood of pixel Y_{ij} with size $w=h=2$. The color-code is the same as that in Fig. 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The following pseudocode summarizes the main steps for reconstructing an image of arbitrary size $s_1\times s_2$ via the causal approach. We note that only the boundaries of the initial $\boldsymbol{Y}$ are needed in Step 1, due to the causal manner in which the pixels are generated.

1. Start with an initial image $\boldsymbol{Y}$ of size $m_1\times m_2$, where $m_1>s_1$ and $m_2>s_2$ (see Sec. 2.4 and Fig. 7 for boundary definition and initialization).
2. For $i=h+1, h+2, \ldots, m_1$ and $j=w+1, w+2,\ldots, m_2-w$
   a. Use the fitted supervised model to predict the Bernoulli parameter $p_{ij}=f(Y_{ij}|\boldsymbol{M}_{ij})$ and generate Y_{ij}~Berouli(p_{ij})
   b. Use the newly generated Y_{ij} to update the corresponding pixel in $\boldsymbol{Y}$
3. Pick the central part of $\boldsymbol{Y}$ with size $s_1\times s_2$ as the new image (see Fig. 7).

2.3. Supervised learner

To capture the complex spatial stochastic dependencies of the pixel phases represented via the conditional distribution $f(X_{ij}|\boldsymbol{N}_{ij})$ (or $f(X_{ij}|\boldsymbol{M}_{ij})$ for the causal method), we have found that a non-parametric classification tree [67,68] is particularly well suited to serve as the supervised learner. The automatic construction of classification trees from training data dates back to the early 1960s [69], but they did not receive widespread attention until the 1980s [69]. An example tree is shown in Fig. 6, in which the root node is

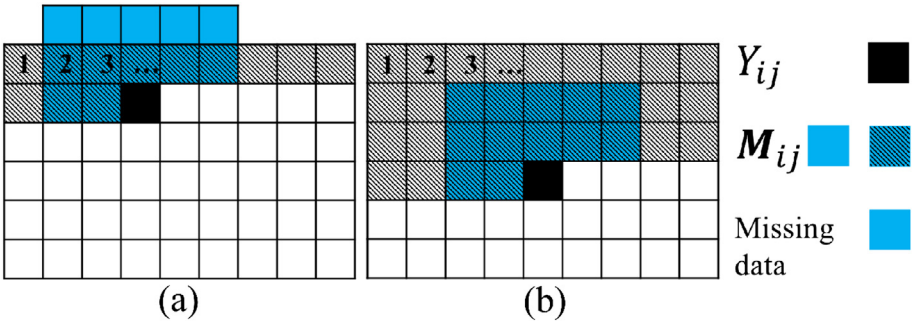

Fig. 5. Causal reconstruction with neighborhood size $w=h=2$ when pixel (i,j) to be reconstructed is (a) close to the boundary, (b) at the central part. As we choose the largest size $\boldsymbol{M}_{ij}$ for reconstructing all the pixels, missing data issue would occur near boundaries. The pixels are color-coded to distinguish the reconstructed ones (patterned pixels) from the pixels that are not yet reconstructed (white). The colors do not represent phase values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

located at the top, terminal (leaf) nodes are located at the bottom, and interior nodes are in between the root and the leaves. Each non-leaf node represents a split at one specific value for one particular predictor variable. In short, classification trees partition the predictor-variable space into regions. Each region corresponds to one of the leaves and is defined by the sequence of splits from the root node to that leaf node. Given a tree that has been fitted to a set of training data (which is automatically done using commercial software), the response for a new observation is predicted by traversing down the tree (from the root node to the correct leaf node) via the sequence of splits that correspond to the set of predictor variables for the new observation. Each leaf stores a predicted class probability for the response variable, and this class probability is estimated as the sample fraction of training observations with response values in that class, out of the training observations with predictor values that fell into that leaf. In our case, each observation corresponds to one pixel (say pixel (i,j)), and X_{ij} and $\boldsymbol{N}_{ij}$ are the response variable and the set of predictor variables, respectively, for that observation. Trees are perhaps the most natural supervised learner for handling categorical variables, and our response and predictor variables are comprised entirely of categorical variables (pixel phase values). Trees are also highly interpretable and, as shown in Sec. 3, very computationally efficient to either fit or make predictions with. Classification tree algorithms, like any other supervised learning algorithm, have tuning parameters that can be adjusted to increase the predictive power (e.g., the number of hidden layers in a neural network or the maximum polynomial degree in regression models). An advantage of our approach is that we can simply use cross-validation (CV) to select all tuning parameters of the supervised learning algorithm in order to best approximate $f(X_{ij}|\boldsymbol{N}_{ij})$. In this section, we will describe how to fit a tree to best approximate $f(X_{ij}|\boldsymbol{N}_{ij})$ in the non-causal approach. For the causal approach, the procedure is nearly identical. For simplicity, we illustrate with a neighborhood size of $w=h=1$ and explain how to empirically find the most appropriate w and h at the end of the section.

Recall that by the MRF stationarity assumption, the model $f(X_{ij}|\boldsymbol{N}_{ij})$ does not depend on pixel index (i,j), and it represents the probability distribution of the predicted phase value at a pixel location, given the phase values of the neighboring pixels. From the collection of pixels in the original $n_1\times n_2$ image $\boldsymbol{X}$, first we build a two-dimensional array of data having $T=(n_1-2h)\times(n_2-2w)$ rows, where each row is comprised of the "response" variable X_{ij} and the "predictor" variables in $\boldsymbol{N}_{ij}$ corresponding to one pixel X_{ij}. Denote this array of training data by $D\equiv\{(X_{ij},\boldsymbol{N}_{ij}),\ i=h+1,h+2,\ldots, n_1-h;\ j=w+1,w+2,\ldots,n_2-w\}$. Notice that the boundary pixels in the training image, for which a full sized neighborhood is not available, are not used as response variables in the training data D. To illustrate, because $w=h=1$ in Fig. 6, there are 8 pixels (predictor variables) in $\boldsymbol{N}_{ij}$ (denoted by P1, P2, …, P8). After building D, we considered various off-the-shelf tree fitting algorithms (R [70,71], MATLAB [72], and Python [73]) to fit a classification tree for predicting the probability $p_{ij}=f(X_{ij}|\boldsymbol{N}_{ij})$ that the material state at (i,j) is 1, given its neighbor $\boldsymbol{N}_{ij}$. Although similar performance can be achieved with any of the above software, we used Python for our examples because it was computationally more efficient.

The common procedure for fitting a tree is to first overgrow (overfit) it and then prune it to the optimal size via CV [74,75]. CV is perhaps the most widely used nonparametric method for estimating the predictive power of a model and can be used to optimize the tuning parameters and default settings of any supervised learning model.

Inaccuracies in the fitted supervised learning model, coupled with the reconstruction algorithm, can sometimes result in reconstructed images having different VFs than the original image. We have found that the overall performance of the reconstruction approach can sometimes be improved by the following simple empirical adjustment to the fitted tree to match the VFs of the original and reconstructed images. Each leaf node in the tree specifies a predicted probability that $X_{ij}=1$. Since the fitted probabilities p_{ij} in the leaf nodes are Bernoulli parameters, the corresponding Bernoulli standard deviation is $\sqrt{p_{ij}(1-p_{ij})}$. In light of this, our strategy for matching VFs is to adjust the probability of every leaf node via

$$p_{ij}^{adjusted}=p_{ij}+c\sqrt{p_{ij}\left(1-p_{ij}\right)}, \tag{2}$$

here the offset parameter c is chosen empirically in the following manner: We first fit a tree to the original image and then use it to reconstruct an image. If the VF of the images match, no offset is needed (i.e., $c=0$). However, if the VF of the reconstructed image is higher (lower) than that of the original image, a small negative (positive) value is assigned to c and this reconstruction/adjustment process is carried out again until the VFs match. Typically, only small values of c are needed. We recommend starting with small values of c (i.e., $c=\pm 0.001$) and increasing-decreasing it as needed.

The geometry (shape and size) of the neighborhood depends on the structural features (i.e. minimum size of the inclusions) as well as the reconstruction approach (causal or non-causal). We determine the optimum neighborhood size in a data-driven manner via CV by starting with a relatively large neighborhood and shrinking the size down until the CV error of the fitted model does not further decrease. A nice feature of our approach (see Sec. 3.5) is that its

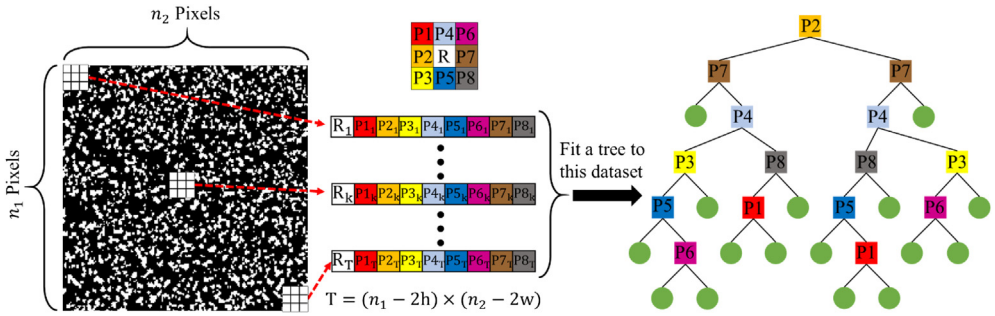

Fig. 6. Illustrative example for fitting a classification tree to a sample binary microstructure. First, the user chooses an initial size for the neighborhood (here w=h=1) to scan the original image with it. The result of scanning is a training dataset (a 2*D* numeric matrix) of size T (rows)×9 (columns). Rows in the dataset are constructed by (1) putting the neighborhood on the image, (2) recording the phase values observed at each pixel in the neighborhood, and (3) rearranging the order of stored values and inserting them in the dataset. At this point, the user feeds the training dataset to an off-the-shelf tree fitting algorithm (we use Scikit package in Python) to obtain the fitted tree. The tree fitting algorithm automatically finds and splits the predictors (represented with color-coded squares and denoted by P1 through P8) to best capture the patterns/stochasticity in the dataset. As the neighborhood size is rather small in this case, all the pixels are used as predictors in the tree. Once the model is fitted, the user can proceed to reconstruct a new image. For predicting the response of a new observation in general tree-based modeling applications, the user provides the fitted tree with the predictor values for the observation (arranged in a row) and the tree predicts the probabilities for each response class. The predicted value in our case, is the probability of X_{ij}=1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

performance is not highly sensitive to the neighborhood size, within limits: If the chosen size is larger than that of the optimum neighborhood, the tree will automatically exclude the unimportant predictors from the model. If, however, the chosen size is somewhat smaller than that of the optimum neighborhood the tree can partially compensate for this by building a more complex supervised learner. Here, complexity refers to the number of terminal nodes and not the number of predictors used in the tree.

2.4. Boundary effects and initialization

For both the non-causal and causal reconstructions, as mentioned earlier we have an issue of missing data at the boundaries (Figs. 3(a) and 5(a)) because the neighborhoods extend beyond the image. To minimize the impact of inaccuracies due to this boundary issue, we reconstruct a larger image than required and choose the central part as the new image. Fig. 7 illustrates this for the causal approach. The patterned blue region of size $m_1 \times m_2$ represents the initial reconstructed image (from Step 1 of the pseudocode) and is reconstructed using the fitted model

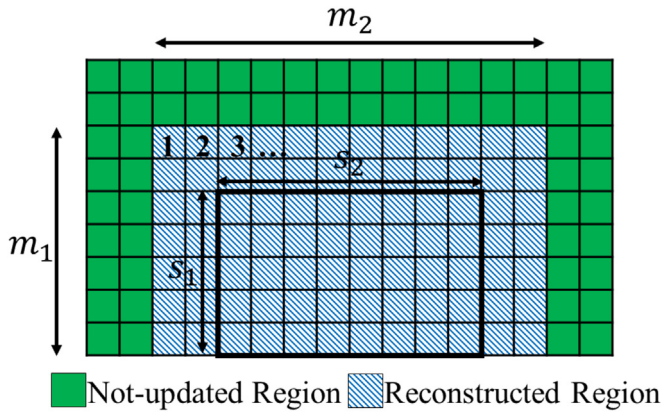

Fig. 7. Avoiding the missing data issue in the causal approach: For reconstructing an image with the desired size of $s_1 \times s_2$, a larger image of size $m_1 \times m_2$ is reconstructed. The green region is added so as the boundary pixels of the patterned blue region have a full size neighborhood. The same method can be used for the non-causal approach except that the blue and green regions will also be added to the bottom of the black box. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for $f(X_{ij}|\boldsymbol{M}_{ij})$. The green region is added to the exterior of the blue region so that its boundary pixels would not have missing data in their neighborhoods. Since the green region is not updated, the boundary pixels of the blue region may be adversely affected. To minimize the adverse effects, we choose the central part of the blue region of size $s_1 \times s_2$ (shown with a black box) as the final reconstructed image. The differences $m_1 - s_1$ and $m_2 - s_2$ depend on how we construct the green region: if the green region is pure black or pure white (all pixels set to 0 or all pixels set to 1), the differences might need to be on the order of twice the size of the neighborhood ($(m_1 - s_1, m_2 - s_2) = (2h, 4w)$). However, if it is constructed using the original image (i.e., by splicing together strips from the original image), the differences can be set to $(h, 2w)$.

We note that, if desired, one could potentially handle the boundary issues by forcing the reconstructed image to possess periodic boundaries. However, because reconstruction using our approach is very computationally efficient, we have taken the simpler (and we think more robust) approach described above that generates a larger image than is needed and discards the boundary.

In the non-causal reconstruction, the choice of the initial image $\boldsymbol{Y}^{(0)}$ directly influences the boundary effects as well as the number of iterations required for convergence of the Gibbs sampler to a steady-state sampling distribution. For the causal approach, the choice for the initial boundary of $\boldsymbol{Y}$ (the green region in Fig. 7) influences how large m_1 and m_2 should be to minimize the boundary effects. We investigated the performance of our approach with different initial images: (1) pure white/black, (2) pure random, and (3) generated periodically by splicing copies of the original image side-by-side (or splicing strips end-to-end for the boundary for the causal approach). Our studies indicate that the latter produces better results. Specifically, for the non-causal approach, it not only alleviates adverse boundary effects, but also requires fewer iterations for convergence of the Gibbs sampler, since $\boldsymbol{Y}^{(0)}$ would be statistically closer to something drawn from $f(\boldsymbol{Y})$. For the causal approach, it mitigates the boundary effects. This choice of initial image also demonstrates that the algorithm can learn the randomness of the microstructure and automatically inject it into the reconstructed image.

3. Results and discussion

In this section, we conduct five examples (EXs) to demonstrate the proposed approach. The microstructure for each EX has unique

features that the approach is able to characterize and reconstruct quite well. We assess performance by comparing the two-point ($S_2(r)$) and two-point cluster ($C_2(r)$) correlation functions as well as the lineal-path ($L(r)$) function of the original and reconstructed images (see the Appendix for implementation details). All the functions are calculated for the white phase and compared in terms of L_2 norm. Specifically, we set the longest pairwise distance (maximum length of the thrown line segment; see the Appendix) to 100 pixels in all the evaluations. We emphasize that in all of the examples the results were not sensitive to the neighborhood size for the range of neighborhood sizes that we considered. In section 3.5, we illustrate the effect of neighborhood size on the characterization and reconstruction results.

3.1. EX1: clustered isotropic microstructure

The first example is a bi-phase microstructure with 20.04% silica in rubber matrix. After preprocessing the image using techniques such as contrast adjustment [76], median filtering [76], Gaussian filtering and thresholding, the resulting binary image is shown in Fig. 8(a). The inclusions form small clusters and possess neither a particular geometry nor any apparent regular spatial distribution pattern other than a stochastic spatial behavior.

We used the causal approach and fitted a tree supervised learner to the training image in Fig. 8(a) to estimate $f(X_{ij}|\boldsymbol{M}_{ij})$. This model can be subsequently used to reconstruct as many statistically equivalent microstructure samples as desired, two representative samples of which are shown in Fig. 8(b) and (c). Details on the fitted parameters, computational costs, and statistical evaluations are given in Table 1 and Table 2. In this example, we set the neighborhood size to $w=h=5$. As the neighborhood size is quite small, all the predictors are found to be important (retained in the model) and the tree has many leaves (see the last two columns of Table 1).

Visual inspection of the images in Fig. 8 indicates that the algorithm captures the stochasticity of the original structure. As the first two rows of Table 1 show, the computational cost (the sum of Char. and Rec. costs) is quite small, and this is of particular interest if a batch of images are reconstructed. In particular, the reconstruction is done almost instantaneously. In addition, Fig. 9(a) and (b) show that the correlation and lineal-path functions of the original and reconstructed images agree, from which we conclude that the supervised learner captures the short-range spatial dependencies in the microstructure quite well (see Sec. 3.5 for an example with long-range correlations). Table 2 lists summary measures of the overall differences in the VF and correlation functions for each reconstructed realization. In this example, the neighborhood size is quite close to the range after which the correlations die out (see Fig. 9(b)), although this relationship does not necessarily have to hold (e.g., in a first-order autoregressive time series, the correlations may decay very slowly, while a neighborhood size of only one suffices for characterization). We further investigate the link between the effective correlation range and neighborhood size in EX5.

3.2. EX2: polymer nanocomposite

Fig. 10(a) illustrates the microstructure of interest in this example; a dielectric nanocomposite with 1.43% VF of silica. This structure is of particular interest as (1) the low VF of the secondary phase means the supervised learner must learn the morphology with scarce data on one phase, and (2) the stationary MRF assumptions may not be entirely satisfied at least not for the scale of the training image (i.e., the local VF changes quite dramatically).

We used the causal approach and set the neighborhood size to ($h=w=5$). The reconstruction results, shown in Fig. 10(b) and (c), are visually appealing, and the statistical evaluations (see Fig. 11) indicate that they agree well with the original structure. We note that, as in EX1, the fitted tree in EX2 finds all the predictors in the neighborhood to be important but has far less leaves (240 vs. 917) and is much less complex than the tree in EX2. These findings are in line with the fact that the microstructure in Fig. 9 is visually more complex than that in Fig. 10.

3.3. EX3: perfectly geometric inclusions

In this example we apply our approach to a structure with circular inclusions (radius of $r=5$ pixels). The inclusions are placed randomly and have a prescribed minimum nearest neighbor distance of $d=5$ pixels. In the ideal case, the algorithm is expected to automatically identify not only the geometry of the inclusions but also the constraint on their spatial distribution.

We used the causal approach and set the neighborhood size to ($h=w=15$). Had we not known r and d, we simply would have had to choose a large enough neighborhood to make sure all necessary features are captured. As the VF of the first reconstructed image was slightly above that of the original structure, we set the offset parameter to -0.001. Fig. 12(b) and (c) illustrate the reconstructed structures and indicate, along with statistical evaluations in (Fig. 13), that the supervised learner has effectively learned both r and d. We note, however, that sometimes the reconstructions might include an irregular inclusion that is not a perfect circle, although usually quite close (i.e. Fig. 12(c)). We address this issue in Sec. 4.

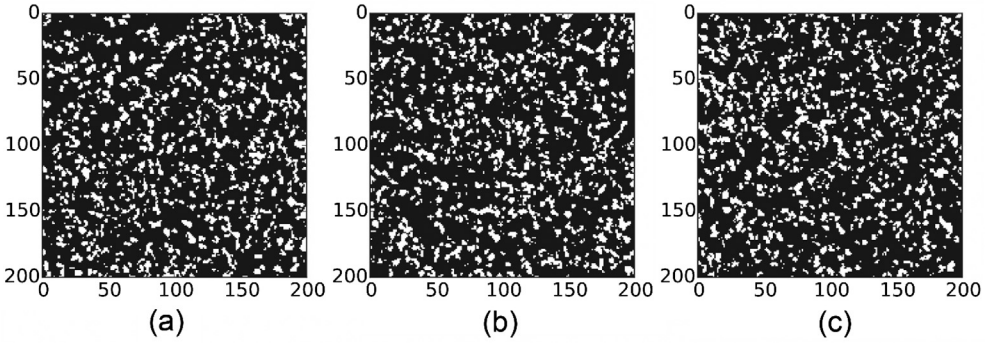

Fig. 8. EX1: (a) The original training image with 20.04% silica in a rubber matrix, (b, c) Two different reconstructed images representing two random realizations generated from the model that is fitted to the training image in (a). The numbers are pixel indices, and the images are reconstructed via the causal approach.

Table 1
The fitted parameters and computational costs for all the examples. The characterization (Char.) cost includes the total cost of fitting the tree. The reconstruction (Rec.) cost is associated with reconstructing the patterned blue region in Fig. 7. The last two columns enumerate the number of leaves and predictors (out of all the initial predictors) in the fitted tree. See Fig. 4 for the definition of window size.

Example	Window size	Char. cost (s)	Rec. cost (s)	Offset	Leaf count	Retained predictors
1, Realization 1	(5, 5)	0.31	0.49	0	917	60/60
1, Realization 2	–	–	0.48	0	–	–
2, Realization 1	(5, 5)	0.46	0.43	0	240	60/60
2, Realization 2	–	–	0.43	0	–	–
3, Realization 1	(15, 15)	2.43	0.44	− 0.001	176	123/480
3, Realization 2	–	–	0.45	− 0.001	–	–
4, Realization 1	(15, 15)	2.38	0.84	0	369	238/480
4, Realization 2	–	–	0.86	0	–	–
5, Small window	(5, 5)	0.26	0.49	0	473	60/60
5, Large window	(20, 20)	3.79	0.67	0	344	273/840

Table 2
L_2 norm errors in point correlation and lineal-path functions, as well as errors in VF between the reconstructed images and the corresponding original ones.

Example	$\lvert VF_{Original} - VF_{Rec.} \rvert$	$\Delta S_2(r)$	$\Delta C_2(r)$	$\Delta L(r)$
1, Realization 1	0.000%	1.04%	0.63%	0.57%
1, Realization 2	0.002%	1.65%	1.79%	1.93%
2, Realization 1	0.00%	4.62%	2.65%	2.48%
2, Realization 2	0%	4.76%	2.97%	1.39%
3, Realization 1	0.000%	2.87%	2.25%	1.82%
3, Realization 2	0.001%	2.92%	1.44%	1.32%
4, Realization 1	0.000%	3.42%	3.77%	4.51%
4, Realization 2	0.000%	1.64%	2.78%	4.21%
5, Small window	0.003%	7.1%	1.88%	2.92%
5, Large window	0.000%	5.51%	3.18%	4.47%

3.4. EX4: porous medium

The porous microstructure of interest (Fig. 14(a)) is a slice of Fontainebleau sandstone (Fig. 6 of [26]) with pores occupying 21.93% of the volume. We used the causal approach and, as the general size of the pores was rather large, set the neighborhood size to (h=w=15). Details of the fitted model are summarized in Table 1.

The reconstructed porous structures are illustrated in Fig. 14(b) and (c) and compared to the original one in Fig. 15 (details in Table 2). As it can be observed the statistical equivalency is quite well preserved. We note that, as Fig. 15(b) illustrates, the correlations almost die out beyond r=20 pixels. Although this distance is quite close to the neighborhood size we chose for this example, the results would be similar had we chosen a moderately larger neighborhood (see EX5).

3.5. EX5: anisotropic structure

To test the performance of the algorithm on anisotropic microstructures and investigate whether the reconstruction order (i.e. the raster scan direction) affects the results, we diagonally stretched an isotropic structure to create an anisotropic structure with VF of 34.53% (Fig. 16(a)). As in the previous examples, we used

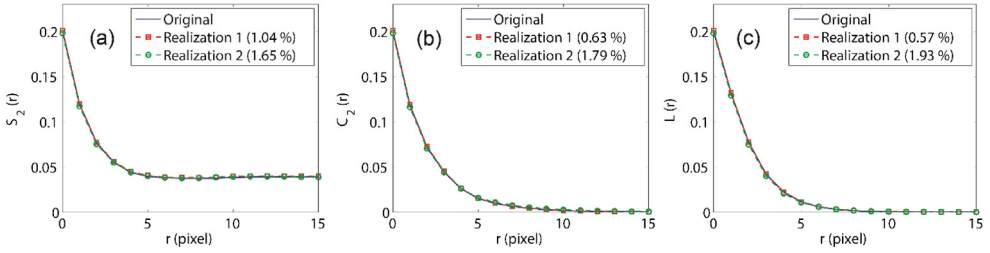

Fig. 9. (a) Two-point correlation, (b) two-point cluster correlation, and (c) lineal-path functions for the original and two reconstructed images in EX1. The numbers in parentheses indicate the L_2 norm errors.

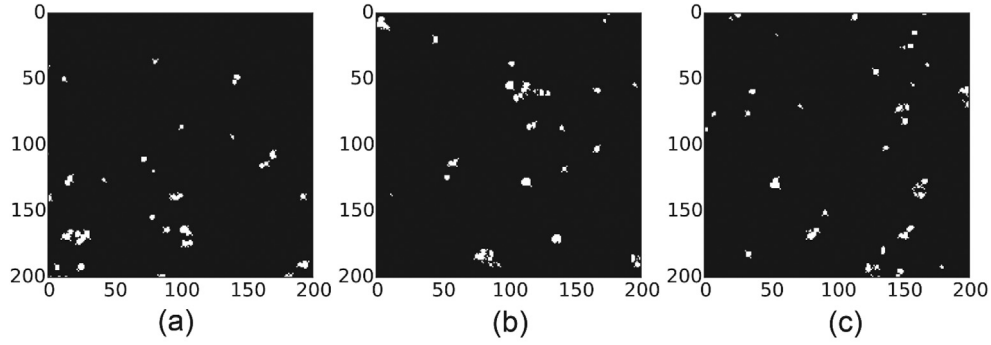

Fig. 10. EX2: (a) The polymer nanocomposite with 1.43% of silica, (b, c) Two different reconstructed images generated from the model that is fitted to the training image in (a). The numbers are pixel indices, and the images are reconstructed via the causal approach.

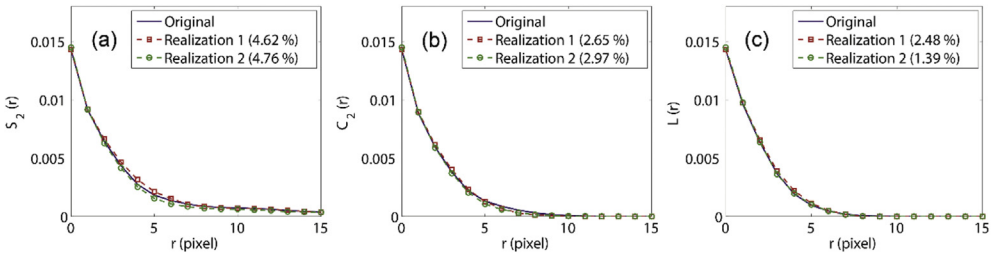

Fig. 11. (a) Two-point correlation, (b) two-point cluster correlation, and (c) lineal-path functions for the original and two reconstructed images in EX2. The numbers in parentheses indicate the L_2 norm errors.

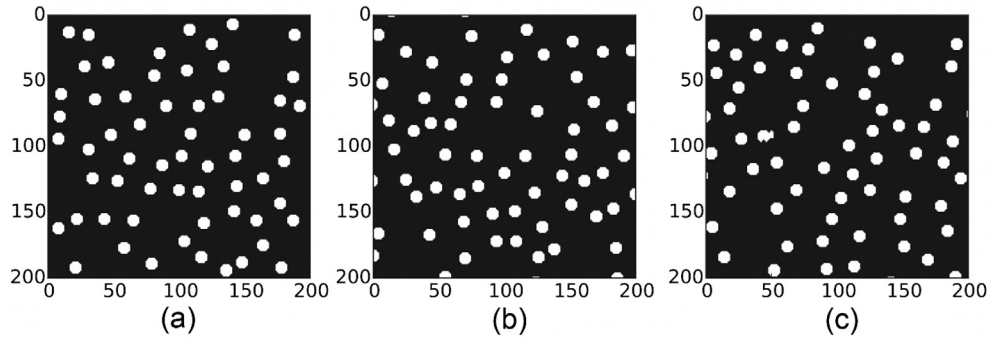

Fig. 12. EX3: (a) The original image with 10.35% VF, (b, c) Two different reconstructed images. The numbers are pixel indices, and the images are reconstructed via the causal approach.

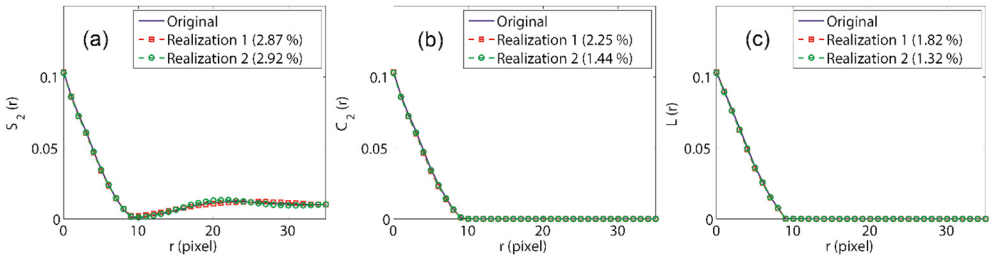

Fig. 13. (a) Two-point correlation, (b) two-point cluster correlation, and (c) lineal-path functions for the original image and the realizations in EX3. The numbers in parentheses indicate the L_2 norm errors.

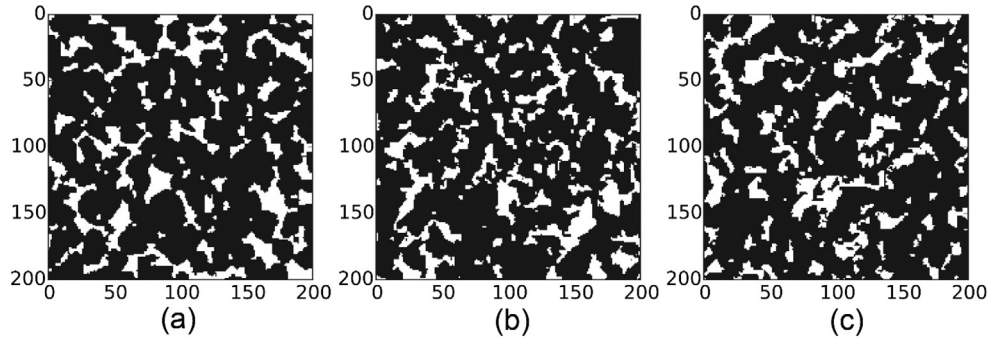

Fig. 14. EX4: (a) The original porous structure with 21.93% pores, (b, c) Two different reconstructed images. The numbers are pixel indices, and the images are reconstructed via the causal approach.

the causal approach, but this time we fitted two models with different neighborhood sizes: a small one (h=w=5) and a large one (h=w=20).

The reconstruction results are illustrated in Fig. 16(b) and (c) and compared to the original microstructure in Fig. 17 (comparison is along the direction of anisotropy; the other diagonal direction had similar results). As these results show, the algorithm is not significantly sensitive to the neighborhood size and performs almost

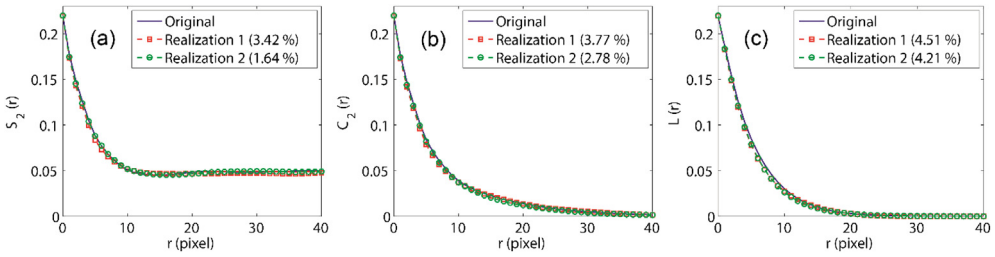

Fig. 15. (a) Two-point correlation, (b) two-point cluster correlation, and (c) lineal-path functions for the original image and the realizations in EX4. The numbers in parentheses indicate the L_2 norm errors.

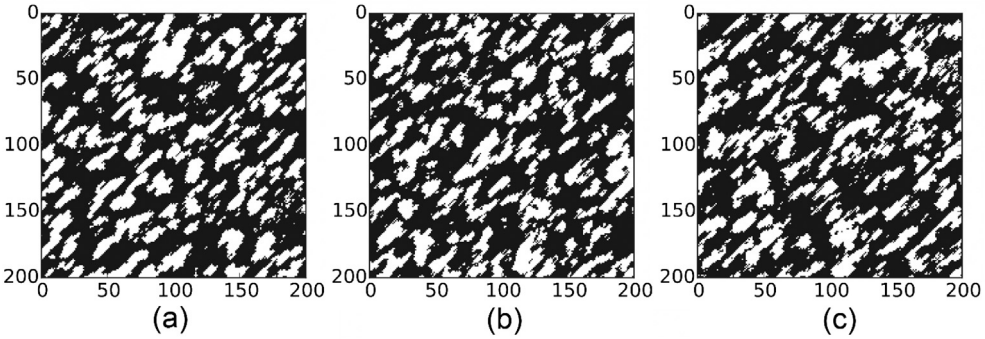

Fig. 16. EX5: (a) The original diagonally anisotropic image with VF=34.53%, (b) the reconstructed image with neighborhood size (h=w=5), and (c) the reconstructed image with neighborhood size (h=w=20). The numbers are pixel indices, and the images are reconstructed via the causal approach.

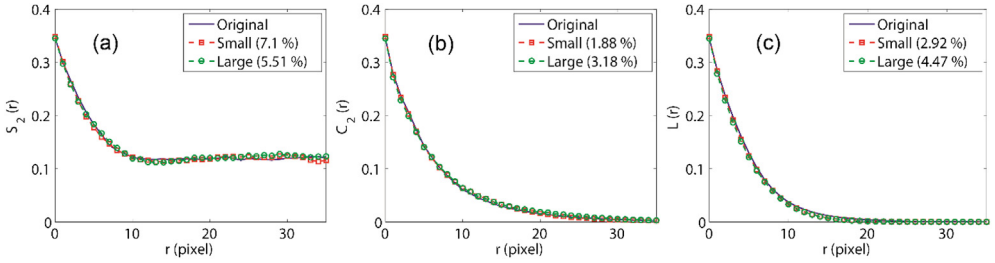

Fig. 17. Unidirectional (diagonal) (a) two-point correlation, (b) two-point cluster correlation, and (c) lineal-path functions for the original and reconstructed images in EX5. The numbers in parentheses indicate the L_2 norm errors.

equally well in both cases. To explain this, we note that (1) while the tree with small neighborhood has retained all the predictors (60 out of 60), the tree with large neighborhood has retained only about one third of the predictors (273 out of 840), and (2) the tree with the smaller neighborhood is more complex and has more leaves (473 vs. 374). These points indicate that the small neighborhood has resulted in the tree being grown larger to partially compensate for the (possible) shortage in the number of predictors.

Although Fig. 17 indicates that the statistical equivalency is quite well preserved, a closer look at Fig. 16 shows that the white phase appears to have a slightly different long-range connectivity in the reconstructed images than in the original one, and there is some noise in the results. We speculate that more complex characteristics like connectivity potentially can be better preserved if one were to use an extension of our approach that includes additional user-defined variables and rules (e.g., related to connectivity) as auxiliary predictor variables in the supervised learning model. Allowing additional predictors to be included is one advantage of the general supervised learning approach. We briefly discuss this extension in Sec. 4.

4. Conclusion and future works

Microstructure characterization and reconstruction of statistically equivalent samples are of major importance in computational materials engineering. The primary goal of this paper is to develop, for the first time to the best of our knowledge, a supervised learning approach for the characterization and reconstruction of a broad range of stochastic microstructures. The approach begins by converting a digitized microstructure image into a set of training data, to which a supervised learning model is fitted to learn the complex stochastic spatial behavior and dependencies of the pixel phases. This relatively compact model is subsequently used to reconstruct any number of statistically equivalent microstructure samples. We have developed two variants of the approach, one causal and one non-causal, which involve single-pass and multi-pass reconstruction schemes, respectively.

We have demonstrated that the fitted supervised learning model provides an implicit representation of the full joint distribution $f(\boldsymbol{X})$ of the phase values of all the pixels in the microstructure image. In theory, the joint distribution provides the most complete and generic representation of the microstructure nature, from

which all other stochastic properties can be derived. In practice, the quality of this implicit representation of the joint distribution depends on the ability of the supervised learning model to capture the conditional distributions $f(X_{ij}|\boldsymbol{N}_{ij})$, as well as on the validity of the locality and stationarity MRF assumptions. In our examples, which cover a range of stochastic materials with different characteristics (clustered, porous, and anisotropic), the reconstruction results (see Figs. 8–17) indicate that the classification tree supervised learning model did a reasonable job of learning the stochastic microstructure behavior, and the reconstruction algorithm did a reasonable job of preserving it in the reconstructed images.

We believe that the main advantages of our approach stem from having a compact yet generic model. This not only makes the approach computationally efficient, but also provides insight into material's structure. As we showed, different models can be compared (in terms of the number of leaves or predictors) with each other to demonstrate the similarities/differences in stochastic spatial pixel dependencies in different microstructures.

The tree-fitting and reconstruction costs depend on the size and complexity of the original image. In general, the larger the desired size of the reconstructed image, the more expensive it will be to reconstruct it (the complexity of the tree also affects the reconstruction cost but not as strongly as the desired size). However, for a fitted model, the reconstruction cost increases linearly with the number of pixels in the reconstructed image and hence is quite manageable even for large structures. The tree-fitting (characterization) cost is directly (and nonlinearly) related to the morphology and increases as the microstructure becomes more complex (as larger neighborhoods and more complex trees are required in these cases). The size of the original image also affects the tree-fitting cost as it (along with the neighborhood size) determines the size of the dataset. However, for general supervised learning applications, trees are widely regarded as being computationally efficient to fit and scaling up nicely to large size datasets.

We believe the foundation of our approach is novel and has many potential extensions that are currently under investigation. The focus in this paper is on 2D reconstructions of bi-phase materials. Extension to multi-phase is straightforward and only requires a supervised learning method that can handle a categorical response with more than two categories, which trees (and many other supervised learners) can do automatically. Extension to 3D reconstruction is conceptually straightforward if 3D training images (e.g., a stack of 2D slices) are available. If we only have 2D training images, extension to 3D reconstruction is more challenging and nontrivial, although we anticipate using ensemble/voting methods of supervised learning to combine different predictive models learned from a collection of 2D training images. In addition, although our method can be applied to microstructures with perfectly geometric (e.g., perfectly circular or ellipsoidal) inclusions, it typically will reconstruct inclusions that have approximately, but not exactly, the same perfect geometry (see Fig. 12(c)). For such cases, we anticipate a hybrid approach in which supervised learning is used to learn the spatial stochastic distribution of the particle centroids (and size, orientation, and other characteristics, if relevant), and the reconstruction phase is used to generate the particle centroids, with the known geometric information being used to fully generate the particles.

Finally, in our supervised learning model, the predictor variables were comprised entirely of individual pixel phases, and no user-defined predictors were incorporated into the model. However, the approach can be extended by including any additional physically meaningful predictors. For example, the average phase values over meaningful sets of pixels can be regarded as a single predictor or an additional binary rule (e.g., one that predicts the probability of the phase value if all the immediate neighbor pixels are in phase 0) can be learned and subsequently used in the reconstruction. These extensions are of particular interest for enabling the algorithm to better capture the long-range connectivity.

Acknowledgments

The authors are grateful to an anonymous referee for many insightful comments that have helped to improve the paper. This work is supported by the U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD) award 70NANB14H012, National Science Foundation Award No. CMMI-1265709, and the Air Force Office of Scientific Research (AFOSR) Award No. FA9550-12-1-0458.

Appendix. Correlation and lineal-path functions

Following the notation introduced in Sec. 2.1, we have:

$$X_{ij} = \begin{cases} 1 & \text{if } ij \in \text{phase1} \\ 0 & \text{otherwise} \end{cases},$$

here, ij is the pixel index and determines its location within the image. Denoting this location by the vector $\boldsymbol{r}$, the two-point correlation function for phase i can be defined as:

$$S_2^{(i)}(\boldsymbol{r}_1, \boldsymbol{r}_2) = X(\boldsymbol{r}_1)X(\boldsymbol{r}_2)$$

where the angular brackets denote the expectation operator. $S_2^{(i)}$ can be thought of as the probability of tossing a line on X and having both its ends land on phase i. If X is statistically homogeneous and isotropic, $S_2^{(i)}$ will only depend on the distance between the two points ($S_2^{(i)}(\boldsymbol{r}_1, \boldsymbol{r}_2) = S_2^{(i)}(\Delta\boldsymbol{r}_{12}) = S_2^{(i)}(|\Delta\boldsymbol{r}_{12}|)$). Hence, for a homogeneous and isotropic material, $S_2^{(i)}$ has a simplified formulation and can be efficiently calculated (i.e. via FFT [77]).

If the aforementioned two assumptions are not satisfied, other methods (i.e. Monte Carlo) need to be used. For example, for the anisotropic structure in EX5, we use an MC sampling procedure as follows: A diagonally-aligned line with length l_k ($0 \le k \le K$) is first randomly thrown on X for a total of N times. Next, the number of times that the thrown line has both its ends in phase i is calculated and then divided by N. This process is done for all k and the result is plotted in Fig. 17 (we chose the maximum length as half of the original image size). We note that the above procedure needs to be done in all directions but we limit ourselves to a diagonal one since it has the strongest correlations.

Two-point cluster correlation function ($C_2^{(i)}$) is similar to $S_2^{(i)}$ but it requires the end points of the thrown line to be in the same cluster. For an anisotropic structure (i.e., EX5), the described MC method, with the addition of a constraint, can be used. For isotropic structures, we use the pixel–pixel distance histogram as follows: First, the distance between any two pixels that belong to one cluster and have phase i is calculated and a histogram of the distances is built. Then, this process is done for any two pixels in the image to build another histogram. The ratio of the histograms gives the cluster correlation function.

Lineal-path function ($L^{(i)}$) is also similar to $S_2^{(i)}$ in that it captures the probability of randomly throwing a line on X and having the whole line land on phase i. In all of our EXs, we chose one direction (vertical and diagonal directions for, respectively, isotropic and anisotropic) for calculating $L^{(i)}$ by using the histogram method explained above except that this time the constraint requires the whole line to be in phase i.

References

[1] S. Torquato, Statistical description of microstructures, Annu. Rev. Mater. Res. 32 (2002) 77–111.
[2] S. Torquato, Random Heterogeneous Materials: Microstructure and Macroscopic Properties, Springer, 2002.
[3] I. Szapudi, Introduction to Higher Order Spatial Statistics in Cosmology. Data Analysis in Cosmology, Springer, 2009, pp. 457–492.
[4] M. Yuan, L.-S. Turng, Microstructure and mechanical properties of microcellular injection molded polyamide-6 nanocomposites, Polymer 46 (2005) 7273–7292.
[5] F.A. Marín, R.H. Wechsler, J.A. Frieman, R.C. Nichol, Modeling the Galaxy Three-Point Correlation Function, Astrophys. J. 672 (2008) 849.
[6] J. Kastner, B. Plank, A. Reh, D. Salaberger, C. Heinzl, Advanced X-ray tomographic methods for quantitative characterisation of carbon fibre reinforced polymers, in: Proc. of 4th International Symposium on NDT in Aerospace. Augsburg, Deutschland, 2012.
[7] R.R. Edelman, S. Warach, Magnetic resonance imaging, N. Engl. J. Med. 328 (1993) 708–716.
[8] M. Kwiecien, I. Macdonald, F. Dullien, Three-dimensional reconstruction of porous media from serial section data, J. Microsc. 159 (1990) 343–359.
[9] L. Salvo, P. Cloetens, E. Maire, S. Zabler, J. Blandin, J.-Y. Buffière, W. Ludwig, E. Boller, D. Bellet, C. Josserond, X-ray micro-tomography an attractive characterisation technique in materials science, Nucl. Instrum. Meth. B 200 (2003) 273–286.
[10] S.R. Niezgoda, D.M. Turner, D.T. Fullwood, S.R. Kalidindi, Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics, Acta Mater. 58 (2010) 4432–4445.
[11] S.R. Niezgoda, Y.C. Yabansu, S.R. Kalidindi, Understanding and visualizing microstructure and microstructure variance as a stochastic process, Acta Mater. 59 (2011) 6387–6400.
[12] C. Ward, Materials genome initiative for global competitiveness, in: 23rd Advanced Aerospace Materials and Processes (AeroMat) Conference and Exposition: Asm, 2012.
[13] X. Liu, V. Shapiro, Random heterogeneous materials via texture synthesis, Comput. Mater. Sci. 99 (2015) 177–189.
[14] H. Xu, M.S. Greene, H. Deng, D. Dikin, C. Brinson, W.K. Liu, C. Burkhart, G. Papakonstantopoulos, M. Poldneff, W. Chen, Stochastic reassembly strategy for managing information complexity in heterogeneous materials analysis and design, J. Mech. Des. 135 (2013) 101010.
[15] G.B. Olson, Designing a new material world, Science 288 (2000) 993–998.
[16] G.B. Olson, Computational design of hierarchically structured materials, Science 277 (1997) 1237–1242.
[17] C.M. Breneman, L.C. Brinson, L.S. Schadler, B. Natarajan, M. Krein, K. Wu, L. Morkowchuk, Y. Li, H. Deng, H. Xu, Stalking the materials genome: a data-driven approach to the virtual design of nanostructured polymers, Adv. Funct. Mater. 23 (2013) 5746–5752.
[18] D.T. Fullwood, S.R. Niezgoda, B.L. Adams, S.R. Kalidindi, Microstructure sensitive design for performance optimization, Prog. Mater. Sci. 55 (2010) 477–562.
[19] M. Committee on Chemical Engineering Frontiers: Research Needs and Opportunities; Commission on Physical Sciences, and Applications, Division on Engineering and Physical Sciences; National Research Council. Frontiers in Chemical Engineering: Research Needs and Opportunities, National Academies, 1988.
[20] N.R.C. Committee on Integrated Computational Materials Engineering, Integrated Computational Materials Engineering: a Transformational Discipline for Improved Competitiveness and National Security, National Academies Press, 2008.
[21] G. Povirk, Incorporation of microstructural information into models of two-phase materials, Acta Metallurgica Materialia 43 (1995) 3199–3206.
[22] V. Sundararaghavan, N. Zabaras, Classification and reconstruction of three-dimensional microstructures using support vector machines, Comput. Mater. Sci. 32 (2005) 223–239.
[23] Y. Liu, M. Steven Greene, W. Chen, D.A. Dikin, W.K. Liu, Computational microstructure characterization and reconstruction for stochastic multiscale material design, Comput. Aided Des. 45 (2013) 65–76.
[24] D.T. Fullwood, S.R. Niezgoda, S.R. Kalidindi, Microstructure reconstructions from 2-point statistics using phase-recovery algorithms, Acta Mater. 56 (2008) 942–948.
[25] S. Torquato, G. Stell, Microstructure of two-phase random media. I. The n-point probability functions, J. Chem. Phys. 77 (1982) 2071–2077.
[26] Y. Jiao, F. Stillinger, S. Torquato, Modeling heterogeneous materials via two-point correlation functions. II. Algorithmic details and applications, Phys. Rev. E 77 (2008) 031135.
[27] C. Yeong, S. Torquato, Reconstructing random media, Phys. Rev. E 57 (1998) 495.
[28] M.D. Rintoul, S. Torquato, Reconstruction of the structure of dispersions, J. Colliod Interf. Sci. 186 (1997) 467–476.
[29] Y. Jiao, F. Stillinger, S. Torquato, Modeling heterogeneous materials via two-point correlation functions: basic principles, Phys. Rev. E 76 (2007) 031110.
[30] J. Quintanilla, Microstructure and properties of random heterogeneous materials: a review of theoretical results, Poly Eng. Sci. 39 (1999) 559–585.
[31] D. Li, M. Tschopp, M. Khaleel, X. Sun, Comparison of reconstructed spatial microstructure images using different statistical descriptors, Comput. Mater. Sci. 51 (2012) 437–444.
[32] C. Yeong, S. Torquato, Reconstructing random media. II. Three-dimensional media from two-dimensional cuts, Phys. Rev. E 58 (1998) 224.
[33] L.M. Pant, S.K. Mitra, M. Secanell, Stochastic reconstruction using multiple correlation functions with different-phase-neighbor-based pixel selection, Phys. Rev. E 90 (2014) 023306.
[34] Y. Jiao, F. Stillinger, S. Torquato, A superior descriptor of random textures and its predictive capacity, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 17634–17639.
[35] B. Collins, K. Matous, D. Rypl, Three-dimensional reconstruction of statistically optimal unit cells of multimodal particulate composites, Int. J. Multiscale Com. 8 (2010).
[36] N.C. Kumar, K. Matouš, P.H. Geubelle, Reconstruction of periodic unit cells of multimodal random particulate composites using genetic algorithms, Comput. Mater. Sci. 42 (2008) 352–367.
[37] W.B. March, K. Czechowski, M. Dukhan, T. Benson, D. Lee, A.J. Connolly, R. Vuduc, E. Chow, A.G. Gray, Optimizing the computation of n-point correlations on large-scale astronomical data, in: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, IEEE Computer Society Press, 2012, p. 74.
[38] T. Tang, Q.-z Teng, X.-h. He, D. Luo, A pixel selection rule based on the number of different-phase neighbours for the simulated annealing reconstruction of sandstone microstructure, J. Microsc. 234 (2009) 262–268.
[39] R. Piasecki, W. Olchawa, Speeding up of microstructure reconstruction: I. Application to labyrinth patterns, Model. Simul. Mater. Sci. Eng. 20 (2012) 055003.
[40] D. Chen, Q. Teng, X. He, Z. Xu, Z. Li, Stable-phase method for hierarchical annealing in the reconstruction of porous media images, Phys. Rev. E 89 (2014) 013305.
[41] H. Xu, Y. Li, C. Brinson, W. Chen, A descriptor-based design methodology for developing heterogeneous microstructural materials system, J. Mech. Des. 136 (2014) 051007.
[42] P. Debye, A. Bueche, Scattering by an inhomogeneous solid, J. Appl. Phys. 20 (1949) 518–525.
[43] P.B. Corson, Correlation functions for predicting properties of heterogeneous materials. I. Experimental measurement of spatial correlation functions in multiphase solids, J. Appl. Phys. 45 (1974) 3159–3164.
[44] H. Garmestani, S. Lin, B. Adams, S. Ahzi, Statistical continuum theory for large plastic deformation of polycrystalline materials, J. Mech. Phys. Solids 49 (2001) 589–607.
[45] S. Torquato, Necessary conditions on realizable two-point correlation functions of random media, Ind. Eng. Chem. Res. 45 (2006) 6923–6928.
[46] P.B. Corson, Correlation functions for predicting properties of heterogeneous materials. II. Empirical construction of spatial correlation functions for two-phase solids, J. Appl. Phys. 45 (1974) 3165–3170.
[47] A. Tewari, A. Gokhale, Nearest-neighbor distances between particles of finite size in three-dimensional uniform random microstructures, Mater. Sci. Eng. A 385 (2004) 332–341.
[48] S. Holotescu, F. Stoian, Prediction of particle size distribution effects on thermal conductivity of particulate composites, Materialwiss. Und Werkst. 42 (2011) 379–385.
[49] A. Al-Ostaz, A. Diwakar, K.I. Alzebdeh, Statistical model for characterizing random microstructure of inclusion–matrix composites, J. Mat. Sci. 42 (2007) 7016–7030.
[50] L. Karasek, M. Sumita, Characterization of dispersion state of filler and polymer-filler interactions in rubber-carbon black composites, J. Mat. Sci. 31 (1996) 281–289.
[51] J.A. Quiblier, A new three-dimensional modeling technique for studying porous media, J. Colliod Interf. Sci. 98 (1984) 84–102.
[52] M. Grigoriu, Random field models for two-phase microstructures, J. Appl. Phys. 94 (2003) 3762–3770.
[53] M. Talukdar, O. Torsaeter, M. Ioannidis, J. Howard, Stochastic reconstruction, 3D characterization and network modeling of chalk, J. Petroleum Sci. Eng. 35 (2002) 1–21.
[54] P. Levitz, Off-lattice reconstruction of porous media: critical evaluation, geometrical confinement and molecular transport, Adv. Colloid Interfacac 76 (1998) 71–106.
[55] Z. Jiang, W. Chen, C. Burkhart, Efficient 3D porous microstructure reconstruction via Gaussian random field and hybrid optimization, J. Microsc. 252 (2013) 135–148.
[56] T. Tang, Q.-z. Teng, X.-h. He, A hybrid reconstruction method of sandstone from 2D section image, in: Neural Networks and Signal Processing, 2008 International Conference on, IEEE, 2008, pp. 342–347.
[57] L.-Y. Wei, M. Levoy, Fast texture synthesis using tree-structured vector quantization, in: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, ACM Press/Addison-Wesley Publishing Co, 2000, pp. 479–488.
[58] A.A. Efros, W.T. Freeman, Image quilting for texture synthesis and transfer, in: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, ACM, 2001, pp. 341–346.
[59] A.A. Efros, T.K. Leung, Texture synthesis by non-parametric sampling, in: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, IEEE, 1999, pp. 1033–1038.
[60] V. Sundararaghavan, Reconstruction of three-dimensional anisotropic microstructures from two-dimensional micrographs imaged on orthogonal

planes, Integr. Mater Manuf. Innov. 3 (2014) 1–11.
[61] K. Wu, N. Nunan, J.W. Crawford, I.M. Young, K. Ritz, An efficient Markov chain model for the simulation of heterogeneous soil structure, Soil Sci. Soc. Am. J. 68 (2004) 346–351.
[62] A. Elfeki, M. Dekking, A Markov chain model for subsurface characterization: theory and applications, Math. Geol. 33 (2001) 569–589.
[63] A. Hajizadeh, A. Safekordi, F.A. Farhadpour, A multiple-point statistics algorithm for 3D pore space reconstruction from 2D images, Adv. Water Resour. 34 (2011) 1256–1267.
[64] H. Okabe, M.J. Blunt, Pore space reconstruction using multiple-point statistics, J. Petroleum Sci. Eng. 46 (2005) 121–137.
[65] S. Strebelle, Conditional simulation of complex geological structures using multiple-point statistics, Math. Geol. 34 (2002) 1–21.
[66] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, Bayesian Data Analysis, CRC Press, 2013.
[67] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, R. Tibshirani, The Elements of Statistical Learning, Springer, 2009.
[68] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, 1984.
[69] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
[70] R.C. Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[71] T. Therneau, B. Atkinson, B. Ripley, rpart: Recursive Partitioning and Regression Trees, 2014.
[72] M.U.s. Guide, The Mathworks, Inc., Natick, MA 5, 1998, p. 333.
[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[74] E. Alpaydin, Introduction to Machine Learning, MIT Press, 2014.
[75] B. Efron, G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation, Am. Statistician 37 (1983) 36–48.
[76] K.R. Castleman, Digital Image Processing, Prentice Hall, 1995.
[77] D. Fullwood, S. Kalidindi, S. Niezgoda, A. Fast, N. Hampson, Gradient-based microstructure reconstructions from distributions using fast Fourier transforms, Mater. Sci. Eng. A 494 (2008) 68–72.