Integrating Materials and Manufacturing Innovation (2024) 13:120–154
https://doi.org/10.1007/s40192-023-00340-4

THEMATIC SECTION: HARNESSING THE POWER OF MATERIALS DATA

MICRO2D: A Large, Statistically Diverse, Heterogeneous Microstructure Dataset

Andreas E. Robertson[1] · Adam P. Generale[1] · Conlain Kelly[2] · Michael O. Buzzy[2] · Surya R. Kalidindi[1,2]

Received: 13 November 2023 / Accepted: 28 December 2023 / Published online: 12 February 2024

Abstract
The availability of large, diverse datasets has enabled transformative advances in a wide variety of technical fields by unlocking data scientific and machine learning techniques. In Materials Informatics for Heterogeneous Microstructures capitalization on these techniques has been limited due to the extreme complexity of generating or curating sizeable heterogeneous microstructure datasets. Historically, this difficulty can be attributed to two main hurdles: quantification (i.e., measuring microstructure diversity) and curation (i.e., generating diverse microstructures). In this paper, we present a framework for curating large, statistically diverse mesoscale microstructure datasets composed of 2-phase microstructures. The framework generates microstructures which are statistically diverse with respect to their n-point statistics—the primary emphasis is on diversity in their 2-point statistics. The framework's foundation is a proposed set of algorithms for synthesizing salient 2-point statistics and neighborhood distributions. We generate statistically diverse microstructures by using the outputs of these algorithms as inputs to a statistically conditioned Local-Global Decomposition generation procedure. Finally, we demonstrate the proposed framework by curating MICRO2D, a diverse, large-scale, and open source heterogeneous microstructure dataset comprised of 87, 379 2-phase microstructures. The contained microstructures are periodic and 256×256 pixels. The dataset also contains salient homogenized elastic and thermal properties computed across a range of constituent contrast ratios for each microstructure. Using MICRO2D, we analyze the statistical and property diversity achievable via the proposed framework. We conclude by discussing important areas of future research in microstructure dataset curation.

Keywords Big Data · 2-point statistics · Heterogeneous Microstructures · Diffusion-based Deep Learning · Local-Global Decompositions · Dataset Curation

Introduction

Over the last decade and a half, researchers have used data science, machine learning, and deep learning techniques to make tremendous advances on a wide variety of challenging problems [1–7]. These advances are overwhelmingly clustered in domains where the needed training datasets can be readily curated (e.g., natural language processing [2, 8], computer vision [3, 9–12], recommendation systems [13], translation [14]). More recently, these techniques are slowly being adopted in the sciences and engineering. For example, advances in bioinformatics [15–19] have rapidly accelerated our understanding of genomics—these methods were instrumental in the development of the COVID19 vaccines [17]. However, capitalization on these transformative techniques in the sciences and engineering is bottlenecked by the expense of curating training datasets. For example, DeepMind's breakthrough protein folding algorithm, AlphaFold [15], was trained on the World Wide Protein Data Bank—a database of almost two hundred thousand *experimental measurements* of proteins [17]. This database has taken nearly fifty years to curate. Better methods are critically needed for rapidly curating such big datasets.

Over the last decade, national initiatives such as the Materials Genome Initiative [20] have fostered new research directions which leverage data science and machine learning to accelerate the design [21–30], discovery [31–33] and manufacturing [34–36] of engineering materials. However, limited and irregular materials data has remained the largest

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

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

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

roadblock to progress (e.g., [37]), even as research in this area has proliferated. A wide variety of data generation and data infrastructure initiatives have arisen in response [38–47]. In particular, significant efforts have been made for atomistic [38, 48, 49] and polymeric systems [50–52]. However, similar progress at the mesoscale—generating datasets comprised of diverse heterogeneous microstructures—has been absent [37, 45, 47, 53, 54]. Statistically diverse datasets at this lengthscale are difficult to curate directly due to challenges in quantifying these systems and the difficulty of diversely sampling these quantifying measures. Currently, two potential options exist for such a direct approach: deep learning methods and those based on microstructural statistics. Deep learning methods have demonstrated the capacity to construct highly expressive and easily sampled latent spaces—seemingly ideal for generating datasets [55–58]. However, the out-of-distribution instability of these methods means that these spaces are largely limited to containing statistically similar microstructures to the deep learning algorithm's original training data. As a result, they act as a source of generating statistically similar data to what is already available, precluding exploratory ability. In contrast, statistical methods—such as n-point statistics—provide a learning-free, stable quantification theoretically encompassing the full space of microstructures [40, 44, 59–64]. However, this pathway presents the conjugate challenge; it is difficult to perform statistically conditioned microstructure generation as well as to uniformly sample the space of microstructure statistics—particularly for salient higher order spatial statistics, i.e., 2-point statistics [22, 62–67]. As a result, this approach has been historically limited to generating statistically diverse datasets with respect to mean-field (i.e., first-order) statistics [39, 40, 68, 69]. These datasets are limited by their lack of diversity and control over the spatial arrangement of their features [40, 70]—an important characteristic only quantified by higher-order statistical measures such as 2-point statistics [60, 61, 71]. Recently, significant progress has even been made on 2-point statistics conditioned microstructure generation [63–66]. However, the second requirement—diversely sampling the statistical space—remains elusive when conditioning on 2-point statistics because of the space's complex boundaries and high dimensionality. Without the ability to systematically quantify and generate arbitrary microstructure data (in a manner similar to what is possible for lower length-scale systems—such as atomistics [38, 48, 49]), efforts have been limited to pursuing microstructure generation efforts via process modeling [31, 72, 73], introducing an additional complex nonlinear linkage.

Prior efforts in the process driven generation of diverse heterogeneous microstructures can be classified broadly into two dominant categories. The first category is experimental data [41, 42, 45, 47, 53, 74, 75]; several large campaigns have attempted to directly collate experimental datasets [45, 47, 53]. However, such efforts are limited by the complexity of data collection—experimental samples are expensive to synthesize, complicated to image, and must be carefully segmented before usage. This triad often curtails the size and diversity of such datasets. A common shortcut is extracting multiple images from the same material system, resulting in visually diverse but not statistically diverse microstructure datasets (e.g., [55, 57, 76]). The second category is simulated data [44, 54, 72, 73, 77, 78]. Again, the complexity of synthesizing heterogeneous microstructure data caps the achievable diversity. Prominent examples utilize parametric models specialized to mimic specific material systems [54]. Altogether, both methods highlight the limitation of taking a process-centered approach to dataset generation: the diversity in the dataset is directly limited by the diversity of the generating process.

In this paper, we propose a novel framework for directly generating statistically diverse, heterogeneous, mesoscale microstructure datasets of 2-phase composites from the joint quantifying spaces of 2-point spatial statistics and neighborhood distributions. An important goal of this paper is to propose a data curation method and provide an open-source dataset that will support ongoing microstructure informatics efforts (e.g., [37, 72, 76, 79–83]). We focus on periodic representative mesoscale systems since this type of data is used extensively in ongoing efforts, such as Process-Structure–Property modeling [31, 72, 83–85]. Additionally, we selected to modulate these specific microstructure features because it is well established that many salient microstructure properties are highly sensitive to these features. In particular, we preferentially focus on 2-point statistics because of the absence of existing methods to produce second order diverse datasets in the literature and their well documented importance [86–96]. Via a direct approach, we are able to synthesize a wide dataset with uniform representation across a large, representative section of the space of 2-point statistics. The framework involves three components. First, proposal: this algorithm, inspired by the spectral mixture concept [97, 98], synthesizes and proposes potential 2-point statistics. Next, filtering: this algorithm sub-samples the proposed 2-point statistics to recover a sparse, uniform coverage of the quantifying space of 2-point statistics. Finally, generation: a microstructure dataset is sampled from the candidate 2-point statistics dataset using Local–Global Decomposition (LGD)-based generative models [64]. Notably, the proposed approach incorporates both the candidate spatial statistics as well as a wide diversity of local neighborhoods effectively expanding the diversity of the generated dataset. Qualitatively, the diversity in the neighborhood distributions allows us to incorporate local features mimicking several salient material classes (e.g., fiber composites [99] and nickel-based superalloy [94]). The variation in the 2-point

statistics modulates the spatial patterning of these individual features—a variation which can have a significant impact on the microstructure's homogenized properties [86–96]. In this paper, we demonstrate an initial application of the proposed framework by synthesizing MICRO2D, a large, open-source and statistically diverse 2-phase microstructure dataset. The contained microstructures are periodic and 256×256 pixels. We analyze the dataset's microstructural and property diversity. Our analysis demonstrates that pursuing diversity with respect to 2-point statistics automatically achieves diversity with respect to a wide variety of material properties. We design and distribute this dataset with the intention of supporting the budding microstructure informatics community. Finally, we outline some significant areas of future work in the continuing development of microstructure datasets.

Background

The proposed framework relies and expands upon several important topics in Materials Informatics and statistical modeling. For clarity we briefly introduce the notation adopted throughout the paper. Vector-valued quantities are demarcated in bold, $\boldsymbol{a}$. Quantities with spatial dependency, such as spatially resolved functions, are demarcated using a subscript for discrete quantities or a spatial dependency for continuous quantities: a_s and $a(\boldsymbol{x})$, respectively. Components of vector-valued quantities are indexed using a superscript, $a^\beta = \boldsymbol{a} \cdot \boldsymbol{e}^\beta$, where $\boldsymbol{e}^\beta$ is the β-basis vector. To avoid confusion, when necessary, exponents will be applied outside of parentheses, $(a)^2$. Finally, summations will always be written explicitly using the summation operator and are never implied by repeated indices.

2-Point Statistics

Our express aim in this paper is to systematically synthesize a statistically diverse heterogeneous microstructure dataset for 2-phase composites. In this context, "diversity" is inherently defined only with respect to the selected statistic. As a result, a set of stable and highly expressive statistics is highly desirable. 2-point statistics are a powerful, flexible, and analytic microstructure quantification paradigm that has been utilized in a wide variety of Materials Informatics frameworks at various lengthscales; such as the development of analytic [100–105] and learned [86–96] advanced structure–property homogenization models, process-structure linkages [34, 35, 106], discovery [31, 91], microstructure sensitive design [21, 24, 107, 108], and inverse problems for nondestructive testing [109]. The value of these statistics arises from their direct development in statistical continuum mechanics [60, 101, 110]—guaranteeing their sensitivity to many material properties and processes—and their sensitivity to the spatial arrangement of salient microstructure features [59, 71, 111]. Importantly, these expressive statistics contain other important microstructure statistics, such as mean-field measures [71]. As a result, diversity with respect to the 2-point statistics represents significant diversity in the microstructure space.

The 2-point statistics can be efficiently computed via the following discrete Fourier transform expression [61].

$$f_r^{\beta\gamma} = \frac{1}{S}\mathcal{F}^{-1}[\mathcal{F}[m_s^\beta]_t^* \mathcal{F}[m_s^\gamma]_t]_r \tag{1}$$

Here, $\mathcal{F}[\cdot]$ and $\mathcal{F}^{-1}[\cdot]$ are the Fast Fourier Transform operation and its inverse, respectively, and $(\cdot)^*$ represents conjugation. m_s^β is the discrete microstructure function – a discretized representation of the microstructure—for material state β and voxel s [61, 63]. The significant limitation of the space of 2-point statistics is its high dimensionality and complex constraints. For quantification and analysis, the MKS framework overcomes this high dimensionality by extracting salient low-dimensional representations of a dataset of 2-point statistics using Principal Component Analysis (PCA) [112] (e.g., [87, 88, 92]). PCA is used because it is a distance-preserving [112] dimensionality reduction technique.[1] This analysis technique will be used extensively in Sect. 4. For the purposes of this work, these limitations also complicate sampling and identifying individual 2-point statistics without computing them indirectly via a microstructure and Eq. (1). Analyzing the expression above, Niezgoda et al. [71] delineate several characteristic identities that must be met by a valid set of 2-point statistics. The most important to this work is that the spectrum of a valid autocorrelation—a 2-point statistic between a phase and itself—must be real-valued and positive.

$$\mathcal{F}[f_r^{\beta\beta}]_t \in \mathbb{R}^+ \tag{2}$$

Spectral Mixture Kernels

The first step of the proposed framework proposes a flexible and expressive parameterization that simplifies the identification of novel 2-point statistics. Our proposed parameterization derives from the design of kernels in Gaussian Process Regression (GPR). The expressiveness and flexibility of the covariance kernel directly defines the modeling capacity of GPR models. As a result, this linkage has fostered extensive research efforts focused on carefully designing these kernels. Spectral Mixture Kernels are an extremely

[1] Specifically, PCA is distance preserving only when the entire basis is maintained [112]. However, in practice, truncated PC representations provide useful dimensionality reduction while being approximately distance preserving [87, 88, 92].

expressive variant that learn a problem specific kernel structure by optimizing a distributional mixture model in the kernel function's frequency space. The strategy relies on the following expression [97, 98]:

$$k(\boldsymbol{\tau}) = \int S(s)\exp(2\pi i \boldsymbol{s}^T\boldsymbol{\tau})\mathrm{d}\boldsymbol{s} \tag{3}$$

where $S(\boldsymbol{s})$ is proportional to a valid probability density function[2]—i.e., it is positive and real valued. In their original work, Wilson and Adams [97] parameterize the kernel function by approximating the spectral density, $S(\boldsymbol{s})$, using a Mixture Model composed of symmetrized Gaussians. Other mixture structures have also been proposed [98]. Wilson and Adams [97] argue that this form produces kernels that are dense in the space of all kernels—theoretically justifying the expressiveness of this kernel structure – and provide a series of examples demonstrating their utility on a variety of benchmark problems.

MaxPro Algorithm

The developed parameterization offers a direct method for proposing 2-point statistics. However, it does not guarantee that the proposed statistics efficiently represent (i.e., cover) the statistical space. The MaxPro algorithm [113, 114] provides a general solution to the problem of filtering large datasets. This algorithm identifies a subset of samples from a large candidate dataset such that the subset nearly optimally covers the dataset's space. In this sense, the quality of the coverage of the full space is directly bound by the diversity of the candidate dataset and the greedy structure of the MaxPro algorithm. In practice, empirical observation has shown the later error source to be limited [115]. The MaxPro algorithm sequentially solves the following min–max optimization:

$$\hat{\boldsymbol{x}}_{m+1} = \underset{\boldsymbol{x}\in\mathcal{C}_N\setminus\mathcal{D}}{\operatorname{argmax}} \min_{\hat{\boldsymbol{x}}\in\mathcal{D}} ||\boldsymbol{x}-\hat{\boldsymbol{x}}||_2 \tag{4}$$

Here, $\mathcal{C}_N$ is the full candidate dataset and $\mathcal{D}$ is the current design at step m. The optimization is repeatedly solved, with $\hat{\boldsymbol{x}}_{m+1}$ being added to $\mathcal{D}$ after each solve, until the design contains a desired number of elements.

Local-Global Decomposition Generative Models

The final step in the proposed framework requires an efficient framework for generating microstructures corresponding to the identified 2-point statistics. Recently, we proposed the Local-Global Decomposition (LGD) framework for generating microstructures conditioned on specified combinations of 2-point statistics as well as neighborhood distributions [64]. This probabilistic generative framework provides sampling algorithms that are one to two orders of magnitude faster than alternative options [62, 65, 116–118]. We emphasize that this speed-up is critical for the focus of this paper because of the number of generating operations necessary to build up a large microstructure dataset. The LGD framework can be described by the following approximation of the stochastic microstructure function:

$$p(\boldsymbol{m}_1,\ldots,\boldsymbol{m}_S;\boldsymbol{\mu},\boldsymbol{f}_r) = \mathcal{N}(\hat{\boldsymbol{m}}_1,\ldots,\hat{\boldsymbol{m}}_S;\boldsymbol{\mu},\boldsymbol{f}_r) \prod_{i=1}^{K} p^{cond}(N_i|\hat{N}_i, N_i^c;\Phi^{(3,\ldots)})$$

Here, $\mathcal{N}(\hat{\boldsymbol{m}}_1,\ldots,\hat{\boldsymbol{m}}_S;\boldsymbol{\mu},\boldsymbol{f}_r)$ is a Gaussian Random Field (GRF) over the entire spatial domain that enforces the targeted 1- and 2-point statistics, $\boldsymbol{\mu}$ and $\boldsymbol{f}_r$, respectively [63]. $p^{cond}(N_i|\hat{N}_i, N_i^c;\Phi^{(3,\ldots)})$ is the neighborhood distribution which locally perturbs the output of the GRF to introduce system specific local features, such as sharp phase boundaries, without significantly impacting the targeted 1- and 2-point statistics [63, 64]. Simplifying the full decomposition produces a family of generative models that balance computational efficiency, the need to capture higher-order features in the generated microstructures, and the training requirements. Specifically, it is noted that removing the neighborhood distributions returns the GRF model [63], using deterministic neighborhood distributions produces the filtered GRF [63], and utilizing learned neighborhood distributions yields the full diffusion-based LGD model [64].

Framework

In this section, we outline the proposed framework for constructing 2-phase microstructure datasets and introduce the needed algorithms. The details of the implementation of the overall framework are summarized in Fig. 1. Conceptually, the framework has three main components: two preparation stages in which (1) salient autocorrelations are identified and (2) microstructure neighborhood distributions are selected. Finally, in stage three, we generate microstructures displaying these identified statistics using LGD-based statistically conditioned generation. The autocorrelation—the 2-point statistics of a single phase (in the coming application: the black phase)—is the only 2-point statistics map that must be considered because the microstructures are 2-phase [71].

The framework's first component identifies a diverse dataset of autocorrelations. This component has three substeps. First, parameterization: we propose a

[2] We emphasize the similarity of this requirement to that given by Niezgoda et al. [71] above.

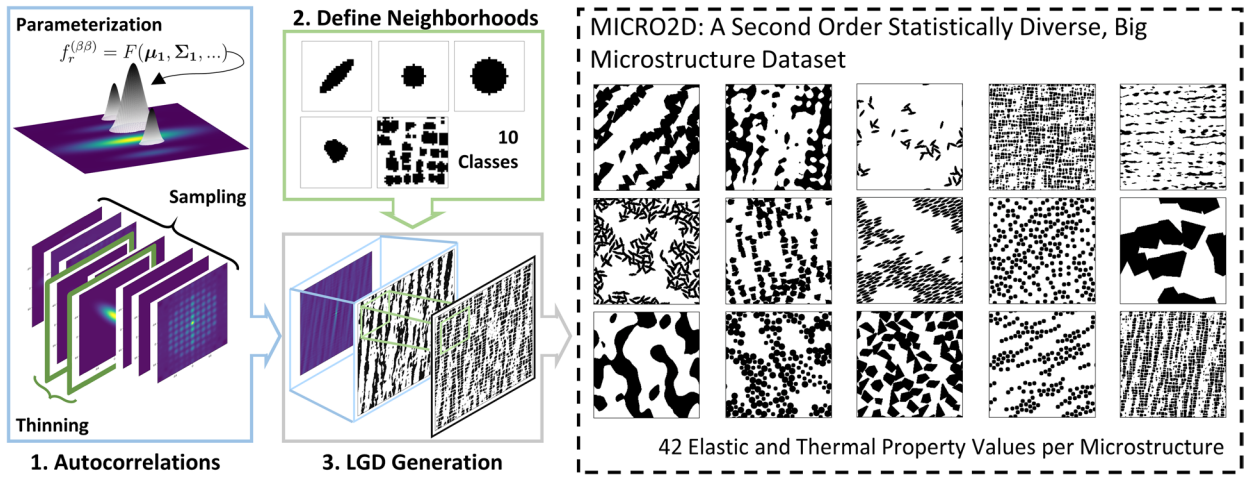

Fig. 1 Visual summary of the proposed framework

flexible parameterization for the autocorrelation function, Sect. 3.1. Second, sampling: we identify an efficient sampling strategy for instantiating parameter values to realize potential autocorrelations, Sect. 4 and 3.1. We use this sampling strategy to sample an extremely large number of potential autocorrelations. Third, thinning: we reduce the initial candidate set into a space-filling subset using the MaxPro algorithm, Sect. 3.2. All steps in the first component are summarized in Algorithm 1.

Algorithm 1 Algorithm summarizing the framework's first component: space-filling autocorrelations.

Define $P = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, (v_f)_i, \alpha_i\}_{i=1,\dots,M} \sim \mathcal{G}$ ▷ Parameter generating process, $\mathcal{G}$, Sec. 4.
Define $\mathcal{H}(\cdot) : P \mapsto f_r$ ▷ Autocorrelation, f_r, parameterization, Eqs. (5)-(10).
Define $\mathcal{C}_N$ ▷ Candidate Autocorrelation Dataset.
Define $\mathcal{D}$ ▷ Final Autocorrelation Dataset.
for $j \leftarrow 1$ **to** N **do** ▷ Sample Candidate Autocorrelations.
  $P_j \sim \mathcal{G}$
  $(f_r)_j \leftarrow \mathcal{H}(P_j)$
  $\mathcal{C}_N \leftarrow (f_r)_j$ ▷ Iteratively augment candidate dataset.
end
Require $O < N$
for $j \leftarrow 1$ **to** O **do** ▷ Extract Final Dataset.
  $(\hat{f}_r)_j = \text{argmax}_{f_r \in \mathcal{C}_\mathcal{N} \setminus \mathcal{D}} \min_{\hat{f}_r \in \mathcal{D}} ||f_r - \hat{f}_r||_2$
  $\mathcal{D} \leftarrow (\hat{f}_r)_j$ ▷ Iteratively augment final dataset.
end

In the second component, we select microstructure neighborhood distributions, Sect. 4. Finally, in the third component, we combine the identified autocorrelations and neighborhood distributions to generate a diverse microstructure dataset using LGD-based generation. The third component is summarized in Algorithm 2. In the remainder of this section, we will focus primarily on the framework's first component since we have exhaustively covered the technical details of statistically conditioned generation previously [63, 64].

Algorithm 2 Algorithm summarizing the framework's third component: generation.

Define $\mathcal{L}(\cdot) : f_r \times \mathbb{N} \mapsto m_s$ ▷ LGD Generator, $\mathcal{L}$, [63, 64] conditioned on autocorrelation f_r and neighborhood, $\mathbb{N}$.
Define $\mathcal{D}$ ▷ Final Autocorrelation Dataset.
Define $\mathcal{N}$ ▷ Set of Candidate Neighborhood Distributions.
Define $\mathcal{M}$ ▷ Microstructure Dataset.
for $f_r \in \mathcal{D}$ **do**
  for $\mathbb{N} \in \mathcal{N}$ **do**
    $m_s \leftarrow \mathcal{L}(f_r, \mathcal{N})$ ▷ Sample Microstructure.
    $\mathcal{M} \leftarrow m_s$ ▷ Add to dataset.
  end
end

Autocorrelation Parameterization

The proposed space-filling procedure for generating diverse autocorrelations revolves around systematically sampling an extremely flexible parameterization of the autocorrelation function. In our prior work on Gaussian Random Field modeling for 2-point statistics conditioned generation, we observed that any 2-point statistics can be readily transformed into a valid set of kernel functions. Here, the validity of the conjugate statement provides a natural pathway for constructing autocorrelations. We propose the following parameterization of the autocorrelation function, heavily inspired by the concept of the Spectral Mixture kernels in Gaussian Process Regression, Sect. 2.2.

$$\hat{k}(\boldsymbol{\tau}) = \sum_{i=1}^{M} \frac{\alpha_i}{2} \left[\phi(\boldsymbol{\tau}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \phi(\boldsymbol{\tau}; -\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right] \tag{5}$$

$$\phi(\boldsymbol{\tau}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(2^k \pi^k |\boldsymbol{\Sigma}|\right)^{-1/2} \exp\left(-0.5(\boldsymbol{\tau} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\tau} - \boldsymbol{\mu})\right) \tag{6}$$

$$\hat{k}_r = \hat{k}(\boldsymbol{\tau}_r) \tag{7}$$

$$k_r = \mathcal{C}^{-1}[\max\left(\mathcal{F}[\hat{k}_r]_t, \epsilon\right)]_r \tag{8}$$

$$\hat{f}_r^{\beta\beta} = k_r + (v_f^{\beta})^2 \tag{9}$$

$$f_r^{\beta\beta} = E_{m_s^{\beta} \sim \mathcal{GRF}\left(\cdot, \hat{f}_r^{\beta\beta}\right)} \left[\frac{1}{S} \mathcal{F}^{-1}[\mathcal{F}[m_s^{\beta}]_t^* \mathcal{F}[m_s^{\beta}]_t]_r \right] \tag{10}$$

Here, $\hat{k}(\boldsymbol{\tau})$ in Eq. (6), is an approximate kernel function constructed via a mixture of symmetric Gaussian distribution. In Eq. (6), superscripts denote exponentiation not indexing. The approximate kernel is parameterized by M, α_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$, the number of mixtures, the mixture weight[3] and mean and covariance of each Gaussian, respectively. Although this parameterization is able to recreate many of the salient features in autocorrelations—e.g., Sect. 4, it contains unacceptable negative frequency values. k_r is a valid discrete kernel function, sampled over the same discrete grid as the microstructure, produced by removing the negative spectral components [92], Eq. (8). In Eq. (8), $\mathcal{C}^{-1}$ is the inverse discrete cosine transform and ϵ is a very small, positive number.[4] $\hat{f}_r^{\beta\beta}$ is the approximate β-phase autocorrelation recovered using the expression identified in Robertson et al. [63]. v_f^{β} is the volume fraction of the β-phase. In the remainder of this work, for compactness, we will drop the β index because the microstructures considered are 2-phase [71]. Critically, a wide diversity of autocorrelations can be systematically constructed by carefully identifying the parameterizing mean vectors, $\boldsymbol{\mu}_i$, and covariance matrices, $\boldsymbol{\Sigma}_i$. Appendix A contains the development of these expressions as well as a careful discussion of the relationship between these expressions and other spectral mixture models.

Although the described parameterization produces valid autocorrelations, it does not guarantee that these autocorrelations could be computed from a possible material system (this includes either an individual microstructure [71, 111]

[3] The mixture weights must sum to 1. In this work, all weights in a single parameterization were set to the same value.

[4] In this work, we set this value to $\epsilon = 10^{-8}$.

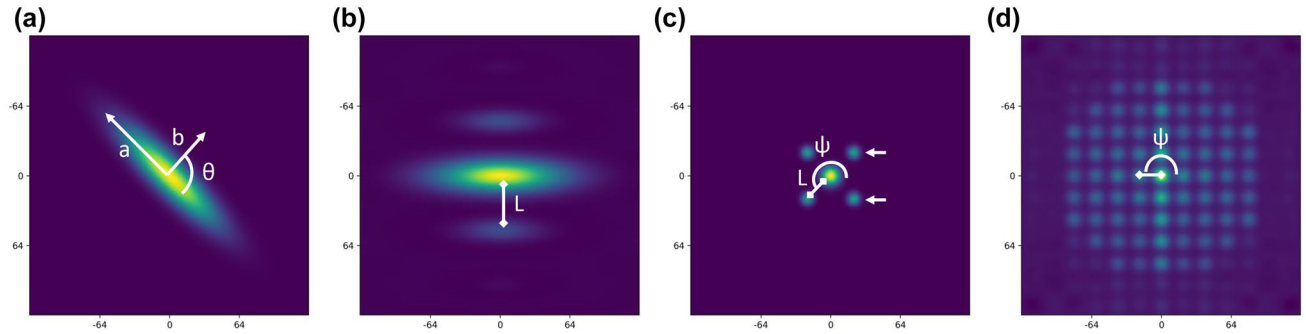

Fig. 2 Visual summary of the four heuristic autocorrelation construction strategies utilized in this paper. White writing identifies each heuristic strategy's salient parameters. As discussed in the main text, almost all parameters included to the left of an individual subfigure are also present in that subfigure (e.g., the white notation in (**a**) also applies to (**b**))

or the average of a set of microstructures [119]). As a result, we impose a final transformation (Eq. (10)); we compute the expected autocorrelation, $f_r^{\beta\beta}$, from samples from the Gaussian Random Field model parameterized by the initial proposed autocorrelation, $\hat{f}_r^{(\beta\beta)}$. This expected autocorrelation is taken as the projection of the proposed autocorrelation into the subspace of autocorrelations associated with 2-phase microstructures and is added to the candidate autocorrelation set. Throughout this work, $N = 20$ GRF samples were used to estimate this average, balancing stability and accuracy against computational demands.

Next, one must identify a set of salient values for the parameters of this approximation in order to cover the space of autocorrelations: M, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, and α_i. An optimal strategy to identify salient values is unclear. While one option is to exhaustively perform this sampling using available procedures [120–122], this process would likely be highly inefficient[5] and repetitive. In Sect. 4, we will propose and utilize a set of expert-guided heuristics for selecting salient parameter values while using this framework to synthesize a diverse 2-phase microstructure dataset.

Space-filling for Autocorrelations

It is likely that the generated candidate set of autocorrelations will not be uniformly spaced regardless of the adopted parameter suggesting procedure.[6] As a result, to complete the space-filling procedure, we distill this initial candidate set, $\mathcal{C}_N$, into a final space-filling design, $\mathcal{D}$, using the MaxPro algorithm [113, 115], Sect. 2.3. Note, this is most effective if the number of candidates in the initial generated candidate set is vastly larger than the desired number of candidates in the final dataset. We reduce the dimensionality of the autocorrelations using Principal Component Analysis (PCA) to accelerate the runtime of the MaxPro algorithm. Unlike standard forward modeling approaches, we retain a high number of principle components to ensure that most fine microstructural details are captured. The number of retained components is application specific. For example, in MICRO2D, we kept 750 components, Sect. 4 and Appendix B. We emphasize that compared with a more information-dense compression algorithm, truncated PCA is preferable due to its approximate distance-preserving property [112].

Even with the established parameterization and space-filling, it is worth noting that defining a sufficient sampling is not simple. Clearly establishing a target for "sufficient diversity" is one of the biggest challenges of curating diverse autocorrelation datasets or, equivalently, of curating microstructure datasets which are diverse with respect to their autocorrelations. This is difficult because the autocorrelation space is extremely high dimensional and is identified by highly complex and high dimensional constraints [71, 108]. Furthermore, much of the space is allocated to high frequency content. These high frequency variations are physically realized as extremely small features and noise; practically, we do not wish to sample these regions when building a dataset because we do not expect them to appreciably vary salient properties. As a result, even identifying a desirable sampling domain, similar to how one might define a minimum and maximum volume fraction, is extremely challenging. Instead, in this work, we have chosen to motivate diverse sampling using two assumed guiding principles. First, large diversity for a set of autocorrelation maps is equivalent to large Euclidean distances between each individual map's coordinates in a representative and efficient orthogonal basis for the autocorrelation space (e.g., a PCA basis). Second,

[5] Empirical observations strongly indicate that large parts of the parameter space are not important for many engineering systems (e.g., [94, 123]). For example, in general, peaks closer to zero, i.e., with $\boldsymbol{\mu}_i$ near zero, are more prevalent and important in real autocorrelations.

[6] This will likely be true even if optimal space filling is accomplished over the parameter space, because of the nonlinear generation transformation step described earlier.

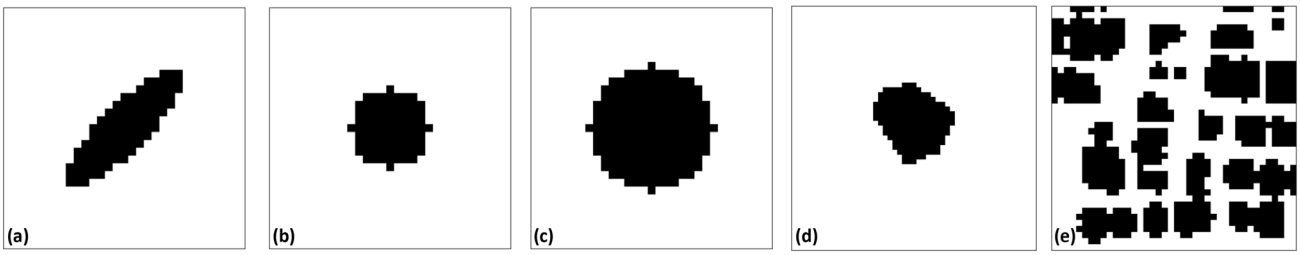

Fig. 3 Visual summary of the foundational neighborhoods used to create the ten neighborhood distributions utilized during generation in this paper

we are primarily interested in autocorrelations which are diverse with respect to the first principle and lack noise. The framework's usage of PCA and the MaxPro algorithm is designed to target the first principle, while the design of the heuristic strategies and the specific mathematics of the proposed parametric approximation guarantees the second (see Appendix A for an analysis of the preference toward low frequency patterns). Even with these principles, identifying whether we achieve a "sufficient diversity" remains a significant, unsolved problem. In the remaining sections, we provide extensive analysis of the achieved diversity via several indirect methods; for example, we compare against other existing datasets and against existing theoretical analysis of the space of 2-point statistics [71]. However, we expect that "sufficiency" will only be established over time by future usage of the framework and the derived dataset and via consensus from the community.

MICRO2D: Second-Order Diversity at Scale

In this section, we will leverage the proposed framework to synthesize a 2-phase heterogeneous microstructure dataset that is statistically diverse with respect to its 2-point statistics. We begin by finalizing the remaining, application specific framework details—i.e., identifying the parameter selection strategy and the local neighborhood distributions for the generative model.

In this work, we adopt a heuristic strategy for parameter selection that aims to approximate and incorporate previously identified salient autocorrelation features [59, 63, 64, 71, 90–92, 94, 95, 111, 123, 124]. In short, the heuristics systematically target the shape and size of peaks close to the center of the autocorrelation. Figure 2 summarizes the four procedures we developed. The simplest, Fig. 2a, targets just the central peak.[7] One, zero-mean mixture is used. The mixture's covariance is parameterized by the length along its major and minor axis, 'a' and 'b' respectively, as well as the rotation of the principal frame with respect to the microstructure's cartesian coordinate system, θ.[8] The second, Fig. 2b, expands this basic parameterization, adding a second mixture that introduces layering [63]. The total parameter count is six: 'a' and 'b' for each mixture, θ for the central mixture—the secondary peak's angle is selected so that the peak remains parallel to the central peak as is shown in the figure, and a distance, 'L', separating the two mixtures. The third, Fig. 2c, continues this pattern adding a second set of secondary peaks. Here, the parameterization is slightly different. All the secondary peaks share a covariance parameterization, ('a_2', 'b_2', 'θ_2'), that is disconnected from the central peak. Additionally, the two secondary peak sets share a distance parameterization, L, measured radially from the center. The secondary peaks are spaced 90-degrees apart. Additionally, ψ controls the rotation of the entire set of secondary peaks around the center. This is a total of eight parameters. The last parameterization, Fig. 2d, extends the secondary peaks out to a predefined cutoff length. Here, the covariance parameters for all peaks were fully coupled and the covariance rotation was set to zero. This leaves a total of four parameters—including the lattice rotation and the spacing. In addition to the identified structural parameters, we varied the volume fraction between (0.0, 0.5]. The remaining volume fractions are simply recoverable by inverting the dataset. We emphasize that the simplicity of placing the Gaussian mixtures to construct each heuristic strategy is one of the major benefits of the proposed autocorrelation parameterization. Additionally, notice that the parameterization leads to an exponential explosion in the number of candidate autocorrelations. For example, the third strategy produces M^9 candidates, where M is the discretization resolution of each parameter. Thinning using MaxPro is necessary.

[7] This approximation is a generalization of PYMKS' standard generative model [125].

[8] The parameterization is numerically implemented as a standard eigenvalue decomposition of the covariance matrix where the eigenvector matrices are the euler rotation matrices.

Table 1 Details of the 10 neighborhood distributions utilized in this study

Class label	Generation parameters	
	Neighborhood	Additional notes
GRF	None	The standard Multi-output Gaussian Random Field model [63]
NBSA	Learned	The hybrid LGD model described in Case Study 1 of Robertson et al. [64]. Here, the model's usage is most similar to Sect. 5.2 in the original work—i.e., the learned neighborhood distribution, e.g., Fig. 3e, is maintained constant while the parameterizing autocorrelations are adjusted. For stability, this class's volume fraction is limited to [0.34, 0.42]
AngEllipse	Prescribed	A single ellipse neighborhood distribution, Fig. 3a. The ellipse is rotated to align with the generating autocorrelation's central peak. The major axis spans 10.5 voxels, and the minor axis ratio is 0.3
RandomEllipse	Prescribed	A mixture neighborhood distribution uniformly combining fifty ellipses, Fig. 3a, with orientation angles equally spaced in [0, 180). Same structural parameters as AngEllipse
VoidSmall	Prescribed	A single circle neighborhood distribution, Fig. 3b. The circle radius is 5.5 voxels
VoidSmallBig	Prescribed	A uniform (i.e., fifty-fifty) mixture of two circle neighborhood distributions, Fig. 3b, c. The circle radii are 5.5 and 8.5 voxels
VoronoiLarge	Prescribed	A uniform mixture of ten voronoi precipitate distributions, Fig. 3d. All ten precipitates are regenerated using a standard voronoi procedure [68] for each sampling. Significant overlap is allowed between placed precipitates. The average precipitate size is approximately 50 voxels
VoronoiMedium	Prescribed	Same as VoronoiLarge—average precipitate size is approximately 25 voxels
VoronoiMediumSpaced	Prescribed	Same as VoronoiMedium. Greater spacing between placed precipitates is enforced
VoronoiSmall	Prescribed	Same as VoronoiLarge—average precipitate size is approximately 14 voxels

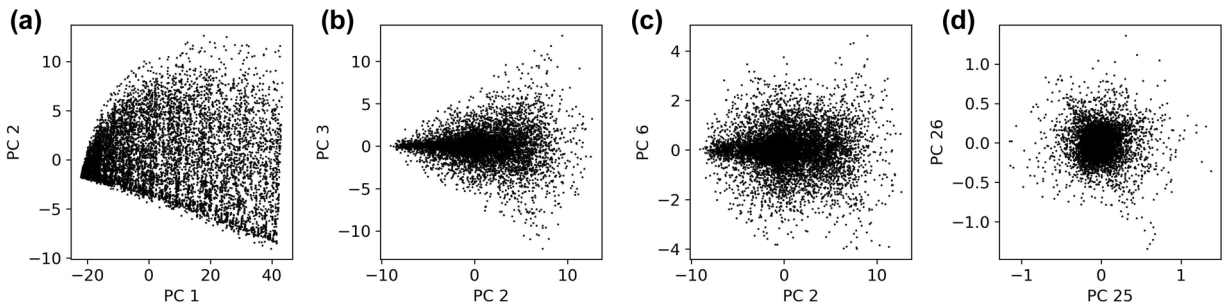

Fig. 4 Four selected principal component projections of the curated autocorrelation dataset

In addition to an identified autocorrelation, Local–Global Decomposition (LGD) generative models are parameterized by an identified neighborhood distribution. Theoretically, these local, neighborhood distributions are a shorthand summary of the higher-order statistics of the targeted material system [64]. Practically, these neighborhood distributions allow us to incorporate important local features and generate microstructures which mimic important material classes (e.g., fiber composites [99]—VoidSmall—and nickel-based superalloy [94]—NBSA). Here, we identify and utilize ten neighborhood distributions. We include a combination of prescribed, filter-based neighborhoods [63] and learned distributions [64]. Unlike in the original work, the prescribed neighborhoods are not necessarily deterministic single geometric shapes, instead, we also used mixtures of shapes. Figure 3 visually summarizes five foundational neighborhood archetypes. These archetypes were combined together to form the ten final neighborhoods. Its important to emphasize that the ellipse, Voronoi grain, and Nickel-Based Superalloy (NBSA), Fig. 3a, d, and e, respectively, are just examples from each archetype class. For example, many different Voronoi grains were sampled and included. Table 1 summarizes the ten neighborhoods. These ten were selected for two reasons. First, they approximate a wide diversity of material classes. Second, they contain some important expected challenges for existing Materials and Microstructure Informatics frameworks.[9] We hope that this ensemble will provide inspiration for future development.

[9] For example, the class 'VoidSmallBig' is nonstationary, breaking the stationarity assumption that accompanies many stochastic quantification frameworks. Similarly, the sharp edges in the Voronoi classes and the small features in the NBSA class will be difficult for localization models [126], in particular those utilizing Fourier filters [127].

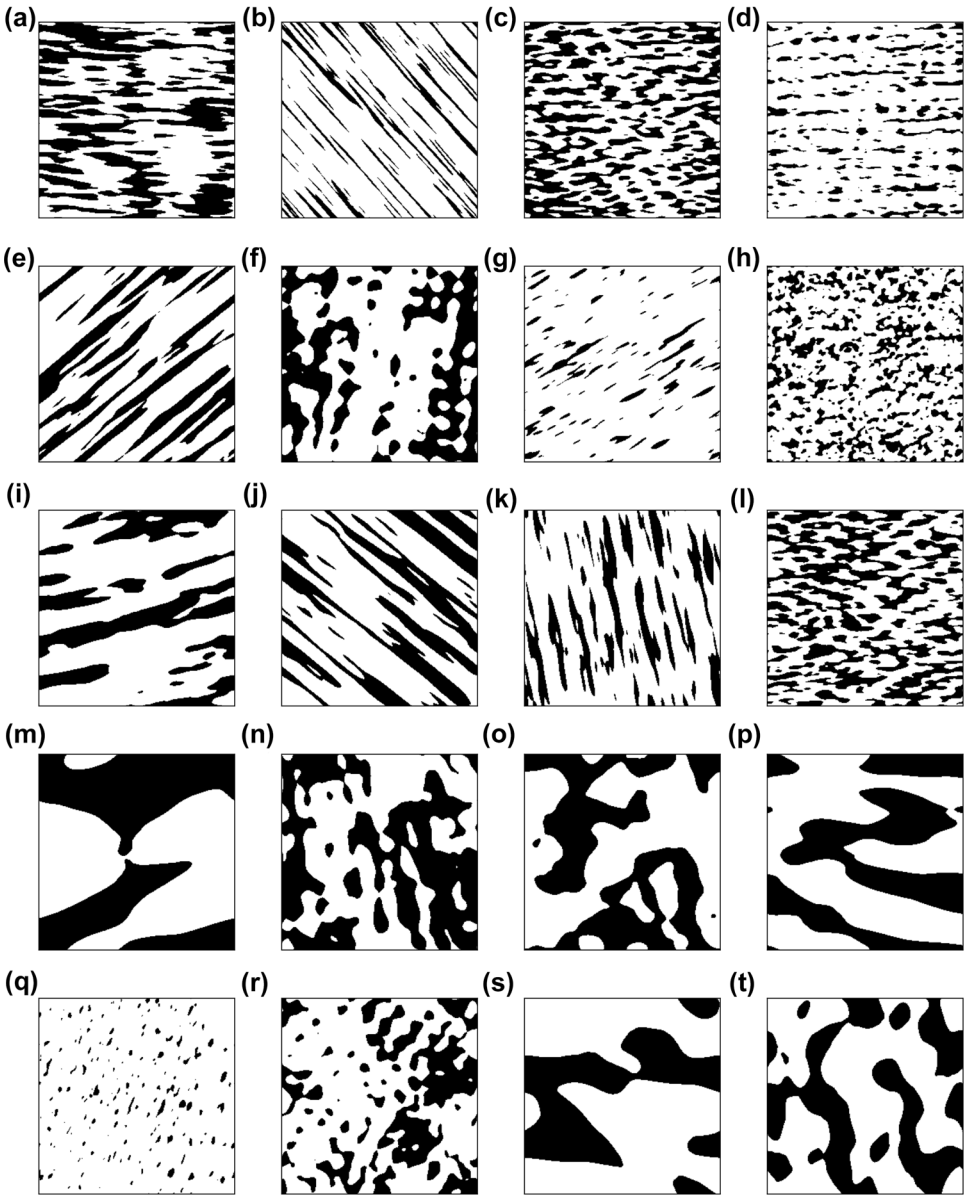

Fig. 5 Illustration of twenty selected microstructures generated using the standard GRF model (i.e., from the GRF class) conditioned with the identified autocorrelations

Initial Autocorrelation Dataset

We generate an initial candidate autocorrelation set comprised of approximately 285, 000 autocorrelations using the four heuristic strategies described previously.[10] Subsequently, we filter this using PCA and the MaxPro algorithm down to 10, 000 final autocorrelations to produce a spacefilling coverage. We retained 750 principal components in the PCA compression for spacefilling. Motivation for this level of truncation can be found in Appendix B. We stratify the filtering by volume fraction and parametric strategy to ensure equal contribution from each.[11] Fig. 4 displays several selected PCA projections of the final autocorrelation dataset. It is important to note that each point in these projections corresponds to an entire autocorrelation map. Importantly, the dataset displays many characteristics of autocorrelation distributions that we expect from both previous study (e.g., [23]) and theoretical analysis (e.g., [111]). For example, the PC 1-PC 2 projection, Fig. 4a, displays the characteristic

[10] The exact parameter values—along with all the code necessary to generate the dataset—can be found in the GitHub repository identified at the end of the paper.

[11] In particular, we noticed that if we did not employ volume fraction stratification the final autocorrelation dataset was strongly skewed toward higher volume fractions. We hypothesize that this is a fingerprint of the spacefilling under the L_2-norm.

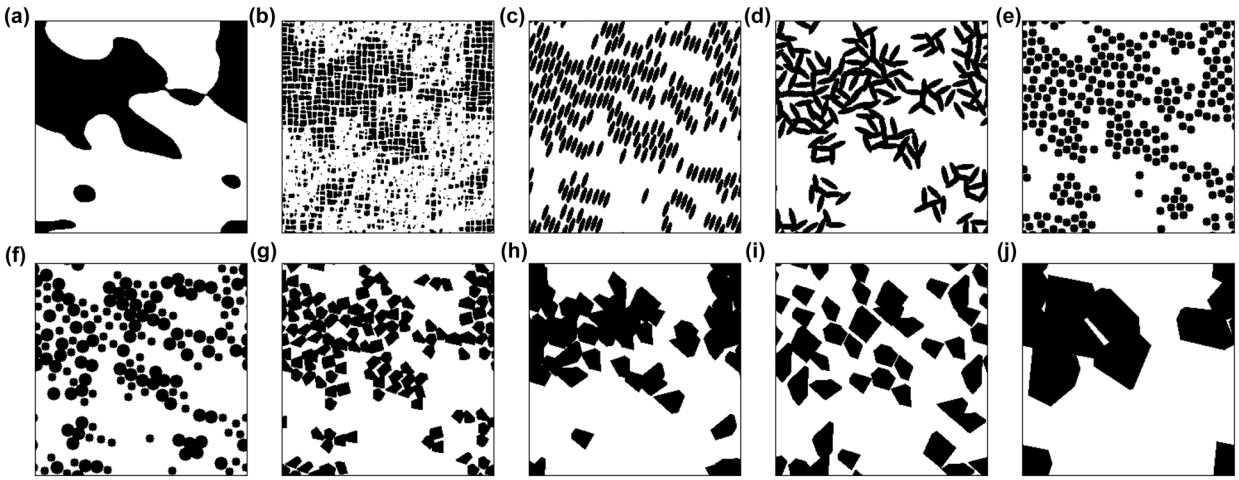

Fig. 6 Illustration of synthetic microstructures generated from a single autocorrelation by varying neighborhood distributions. To make these figures, the seed of the GRF is artificially kept constant

sharp edge along the lower boundary of the distribution. This corresponds to autocorrelations for stochastic microstructure functions with small central features (i.e., where the autocorrelation approximates a dirac delta in real space). Additionally, in almost all projections, the dataset is convex. This indicates that the achieved coverage is largely uniform over a subspace of the 2-point statistics. The reasons for the few exceptions to this and the impact of these exceptions will be discussed in greater detail in Sect. 5.

MICRO2D

Conditioning the LGD generative model using the generated autocorrelation dataset and the ten identified neighborhood distributions, we generate a microstructure dataset comprised of 87, 379 2-phase microstructures.[12] We will refer to this final open-source large microstructure dataset as MICRO2D.[13] Appendix C presents examples of randomly selected images from each microstructure class. The dataset's microstructures are periodic and are represented on a 256×256 pixel grid.[14]

Figure 5 displays a selected variety of spatially diverse microstructures from MICRO2D generated using just the standard GRF model. Beyond visually observing the variety of achievable spatial distributions, this figure also depicts the impact of the different autocorrelation parameterization schemes. For example, Fig. 5s displays an example microstructure generated using the first scheme. It is characterized by highly localized features that lack any persistent long range structure because it was generated using a single mode autocorrelation. Contrasting Fig. 5k against Figs. 5e, i, we see the impact of the addition of the secondary peaks in the second and third strategies compared to just a single central peak in the first. In Fig. 5k the inter-black phase spacing—approximately 28 voxels—is extremely regular and the black phase regions are also regularly aligned within each populated plane. This is a result of the strong statistical coupling introduced by the secondary peak; whenever there is a black phase region, this peak means we expect to find another black phase region aligned with it and separated by the inter-peak spacing [71, 123]. Similarly, this increase in

[12] The total number of microstructures is less than 100, 000 (i.e., $10 \times 10,000$) because several volume fraction and neighborhood combinations resulted in unstable generation, e.g., see NBSA in Table 1.

[13] In the dataset, each class is stored separately to simplify studying subsets of the dataset.

[14] We selected this specific discretization to balance the degree of achievable diversity against practical considerations. This resolution was sufficiently high to allow us to incorporate two important lengthscales: both salient individual features and long range patterns.

Footnote 14 (continued)

However, it is sufficiently low to remain inline with the discretizations preferred by the microstructure informatics community (e.g., in Process-Structure–Property modeling [37, 72, 73, 79, 81–83] and synthetic generation [58, 65, 76, 80, 128]). Additionally, we construct our heuristic strategies to ensure that the chosen discretization is sufficient to represent the generated systems. Primarily, we do this by ensuring that the correlation length of the generated statistics is less than half the domain size and by generating periodic microstructures. It is well established in the micromechanics community that periodic RVEs and SVEs provide highly stable estimates of homogenized properties even using relatively small domains [84, 129]. We note that the proposed framework is not restricted to this discretization and datasets containing smaller, larger, or even 3D microstructures can readily be generated without significantly altering the codebase referenced at the end of this paper. However, more advanced generation strategies will need to be established if one is interested in incorporating more than two feature lengthscales.

structuring in the spatial arrangement occurs again when comparing microstructures from the third strategy with those from the fourth strategy. In Fig. 5c, d—generated using strategy four—repeated, long range order is maintained over the entire microstructure. In contrast, in Fig. 5a, h, and g the structuring is much more localized. This is most evident in Fig. 5a where the relatively regular long range structure is commonly broken by large white phase pockets.

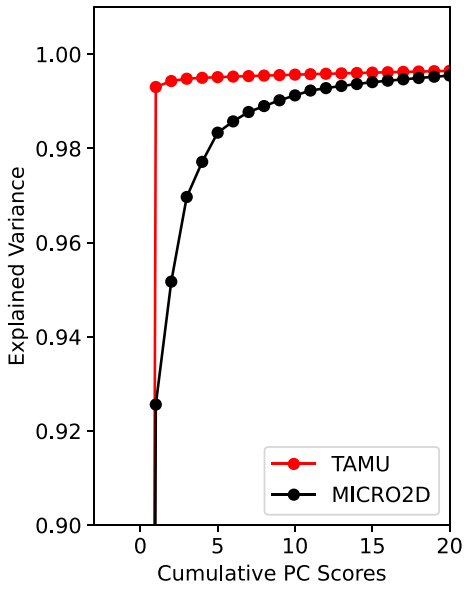

Fig. 7 Comparison of the saturation of the explained variance with increasing number of PC basis vectors between MICRO2D and the TAMU Spinodal dataset. Two separate PC bases are compared

Beyond containing microstructures with diverse spatial arrangements and 2-point statistics, MICRO2D's microstructural diversity also considers local feature diversity because of the local neighborhood distributions used in generation. This means that the dataset contains microstructures whose local features mimic several important material classes (e.g., fiber composites [99]). Figure 6 specifically exemplifies the microstructural diversity that is achieved by just exchanging neighborhood distributions but maintaining a constant autocorrelation. The depicted microstructures are generated using a constant seed to maintain a constant sample from the GRF in the LGD model. We emphasize that the depicted microstructures are simply to illustrate the diversity achievable using neighborhood distributions and are **not** actually contained in MICRO2D. In MICRO2D, the seed is randomized for each sample (i.e., microstructures generated with the same autocorrelations but different neighborhoods will **not** have their neighborhoods placed in the exact same spatial regions.). Clearly, a wide diversity of microstructures is achievable even when restricting generation to a single autocorrelation. For example, this combination produced microstructures with large distinct phases—e.g., Fig. 6a,j—as well as hierarchically clustered microstructures—e.g., Fig. 6f,g. We also see that the dataset is systematically biased to contain microstructures with precipitate like structures reflecting the types of neighborhoods selected.[15]

The magnitude and statistical arrangement of the generated diversity is clearer in PC space. Here, we begin with a standard MKS (i.e., autocorrelation and PCA) analysis [21,

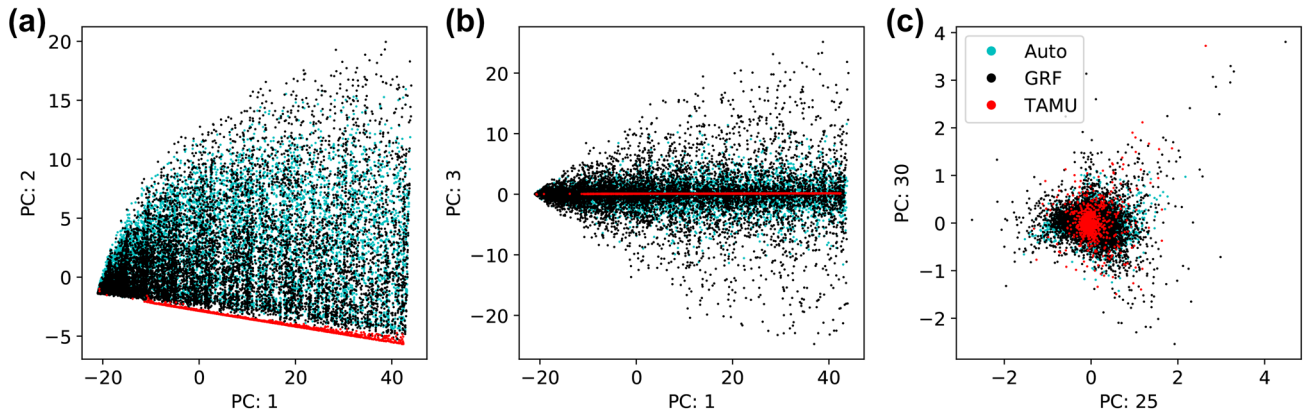

Fig. 8 Comparison of the original autocorrelation dataset distribution (cyan) with the samples from the Gaussian Random Field (black) and the TAMU spinodal dataset (red). Each subfigure displays PC subspaces from a standard 2-point statistics and PCA analysis performed on the entire MICRO2D dataset. The original autocorrelation dataset and the spinodal dataset are projected in after the fact and were not included in the generation of the PC basis

[15] Other microstructures, like grain boundary structures, could be generated by the local diffusion model [64, 76].

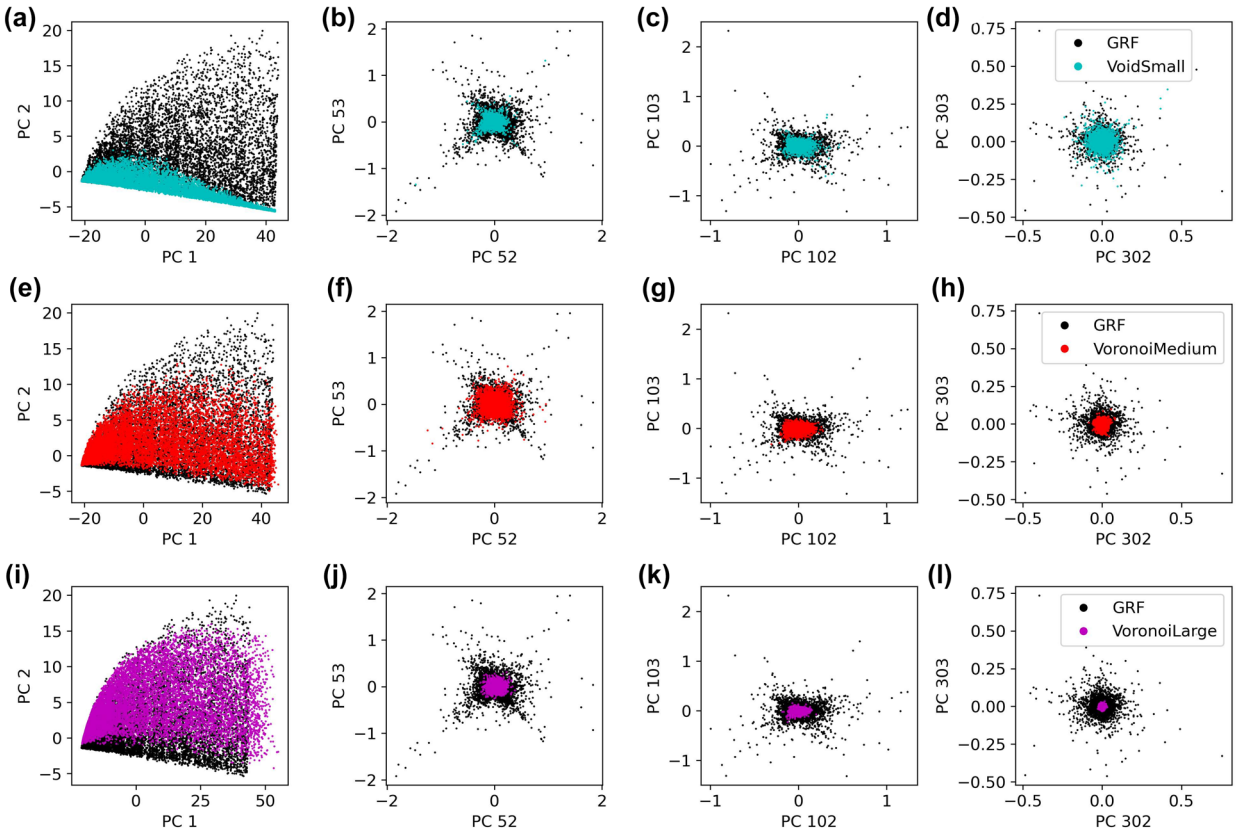

Fig. 9 Comparison of coverage of the distribution of GRF class samples (black) with VoidSmall (cyan), VoronoiMedium (red), and VoronoiLarge (magenta). Each figure displays PC subspaces from a standard 2-point statistics and PCA analysis performed on the entire generated dataset

90, 91] of the entire MICRO2D dataset. The representational efficiency of the PCA dimensionality reduction provides an initial quantification of the achieved diversity. 123 PC basis vectors are necessary to achieve 99.9% explained variance on the MICRO2D dataset. This is a significant increase in comparison with other available 2-phase microstructure datasets. For example, Marshall and Kalidindi [23] report requiring 10. Similarly, the PYMKS dataset requires 2 [108, 126]. Researchers at TAMU recently generated a large, open-source spinodal decomposition dataset [73] for process modeling. This dataset requires 85 PC scores to capture the same amount of information for its 2-phase microstructures. In addition to requiring numerous basis vectors to achieve a high explained variance, MICRO2D's variance is more evenly spread between the components than previous studies. For example, Fig. 7 contrasts the saturation of explained variance against number of PC basis vectors between MICRO2D and the TAMU spinodal dataset. The MICRO2D dataset displays slow saturation, indicating that each basis vector is well-sampled. In contrast, the spinodal dataset displays a sharp saturation which is consistent with previous efforts [23, 88].

Figure 8 contrasts the subset of microstructures in MICRO2D generated by the GRF with the autocorrelation dataset used to seed MICRO2D and the spinodal dataset.[16] Each point represents the entire autocorrelation map for a single microstructure. The PC basis was generated using the entire MICRO2D dataset. The autocorrelation dataset and spinodal dataset[17] are projected into the PC basis afterwards and are not included in training. The GRF dataset displays good distributional agreement with the original autocorrelation dataset. In fact, the GRF's statistical scatter (extensively documented previously [63, 64]) causes the GRF distribution to cover a greater volume of the autocorrelation space than the original autocorrelation dataset, Fig. 8b. Unsurprisingly, this is most noticeable in the lower PC scores. The

[16] The TAMU microstructures are rescaled down to 256×256 for comparison.

[17] The average relative L_2 reconstruction error of the projection is 0.0071 ± 0.0077 for the spinodal dataset. This is comparable with the reconstruction error of MICRO2D, Appendix B. Therefore, the dataset is well represented by the basis. Additionally, including the spinodal dataset in training the PC basis did not change the structure of the latent space.

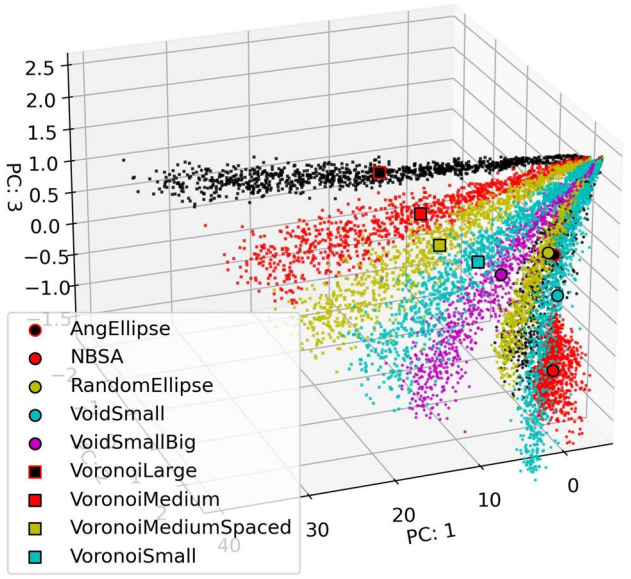

Fig. 10 Visual summary of the 3-point statistics coverage achieved by the generated dataset. All nine non-GRF classes are compared. The distribution of the GRF class engulfs the remaining classes

GRF is prone to larger scatter in these subspaces because they are more sensitive to larger features (see Appendix B) [63, 64]. In comparison, the spinodal dataset is restricted to a smaller subset of the autocorrelation space. This further supports the previous conclusion; the MICRO2D dataset achieves a wide sampling of the autocorrelation space. Before continuing, we emphasize that, in our opinion, the spinodal dataset remains an extremely useful microstructure dataset and a important contribution to dataset curation: it is still very structurally diverse, it contains a large number of 3-phase microstructures which were excluded from this analysis, and, most importantly, it contains processing information necessary for developing and benchmarking Process-Structure linkages [92].

The remaining classes also display significant diversity. Like the TAMU dataset, their coverage is only a subset of the full GRF distribution. This occurs because of the coupling that exists between neighborhood distributions and 2-point statistics [64]. Figure 9 contrasts the coverage of three example classes—VoidSmall, VoronoiMedium, and VoronoiLarge—against the GRF. The size of the features in the neighborhood distribution dictate each class's coverage with respect to the GRF class distribution. For example, VoronoiLarge—whose neighborhood features are nearly one fifth of the RVE domain—displays strong agreement with the GRF class for the lower eigenvectors, Fig. 9i. Again, these are the eigenvectors which capture large feature differences. For higher index eigenvectors, the VoronoiLarge distribution becomes very tight, Fig. 9l, because the selected neighborhood precludes

Table 2 Mechanical (young's modulus) and thermal (conductivity) properties for each phase

Mechanical (GPa)		Thermal (W/mK)	
Phase 1 (black)	Phase 2 (white)	Phase 1 (black)	Phase 2 (white)
10	1	10	1
100	1	100	1
1000	1	1000	1
1	10	1	10
1	100	1	100
1	1000	1	1000

the introduction of small features. In contrast, VoidSmall displays the opposite behavior. The placed circles, e.g., Fig. 6e, breaks the homogeniety of large features. As a result, the spread on the lower eigenvectors is smaller than the GRF. The spread grows to eventually match the GRF for the higher eigenvectors, Fig. 9d. VoronoiMedium, unsurprisingly, displays an intermediate behavior, maximizing its agreement for intermediate indexes, Fig. 9f. It is important to emphasize that the small and medium feature neighborhoods still display large microstructural features due to hierarchical clustering. However, these features are not composed of a continuous phase as they are in the GRF; instead they are a tight collection of individual neighborhoods, Fig. 6. Therefore, they do not display the same coverage in the lower PC basis vectors because the larger aggregate cluster features are not smooth. This analysis also emphasizes that the initial PC variance metric is only a rudimentary metric. Clearly, the large index PC basis vectors still contain important microstructural information and the true dimensionality of the MICRO2D dataset is much larger than just the initial 123 dimensions.

Importantly, the dataset is not just limited to second order diversity. Because we vary the neighborhood distribution during generation, the microstructure's higher order statistics vary as well. We were unable to identify a subspace of the autocorrelations using PCA in which the 10 classes could be distinguished [130]. Instead, we can see the importance of the classes and the present higher order variations in microstructure statistics by performing a standard MKS analysis using 3-point statistics[18]. Figure 10 contrasts the 3-point statistics of each of the non-GRF classes. The GRF class is excluded because it covers the entire space achieved by the remaining classes. The remaining classes are clearly separated in the 3-point statistics space. Note that the 3-point statistics space

[18] We use an analysis congruent to the analysis reported in Robertson et al. [64]. Only the subset of 3-point statistics in which the first shift is equal to 3 are considered.

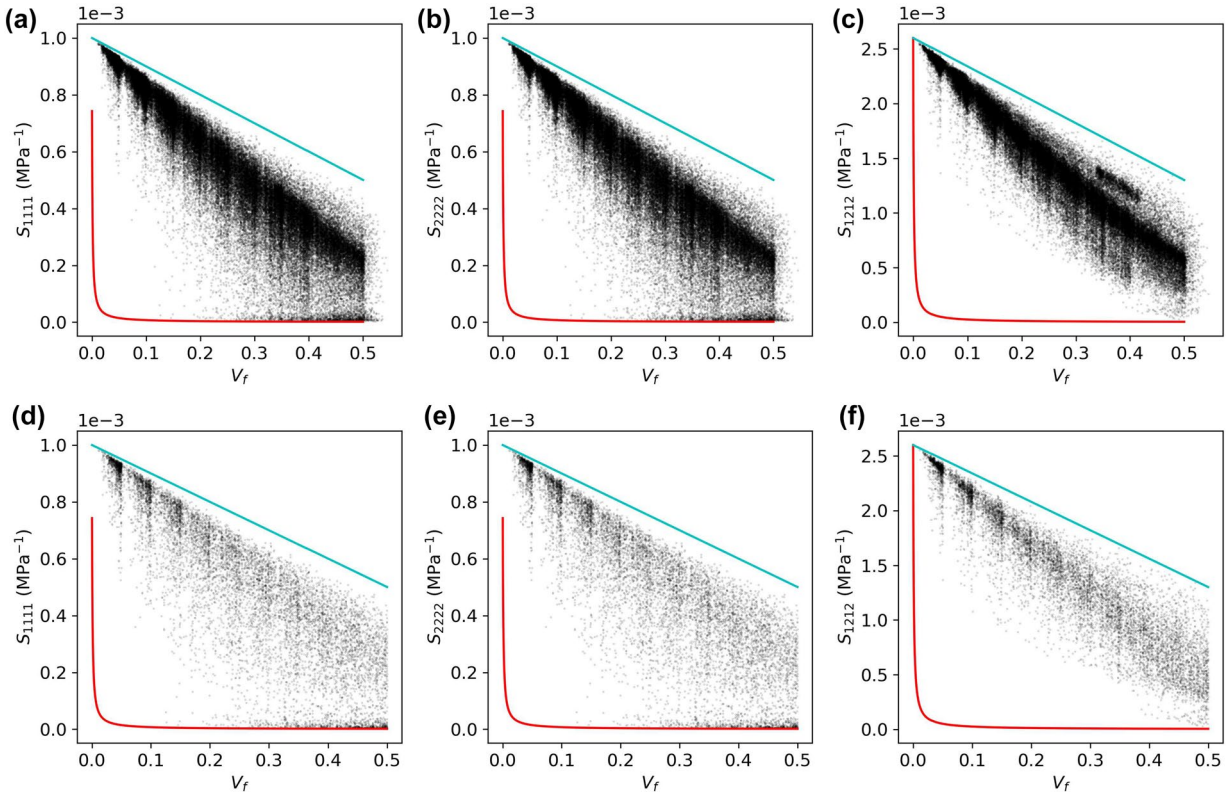

Fig. 11 Visual summary of the distribution of MICRO2D's homogenized properties against microstructure volume fraction. The Hill-Paul theoretical bounds are included for context. The computed upper bound is demarcated using a cyan line and the lower bound is demarcated using a red line. Subfigures **a**–**c** summarize the property distribution for the complete dataset. Subfigures **d**–**f** summarize the property distribution for just the GRF class

displays clearly interpretable structure—similar classes are arranged next to each other and the feature size recognizably decreases with increasing PC 2 (i.e., shifting left to right in the image). The various generated classes provide a systematic method to study the impact of higher order statistics.

In total, the generated dataset displays significant diversity with respect to 2-point statistics, neighborhood distributions and higher order n-point statistics. Appendix C displays randomly sampled examples from each class.

Diversity in Property Space

Beyond providing the microstructures themselves, we also include a collection of 42 homogenized property values for each microstructure in the MICRO2D dataset. Specifically, these include homogenized orthotropic thermal and mechanical (elastic) coefficients.[19] The simulations were performed using a standard periodic finite-element-based homogenization scheme [21, 90, 131, 132]. In both cases, both phases in the microstructure were treated isotropically. For each microstructure, 6 combinations of both constituent elastic and thermal property assignments were used. The Poisson's ratio was kept constant in both phases (equal to 0.3).

For the mechanical properties, three simulations were performed for each constituent property set using different plane-stress boundary conditions: uniaxial tension (X-direction), uniaxial tension (Y-direction) and pure shear. These boundary conditions were selected to evaluate the homogenized values of E_x, E_y, G_{xy}, ν_{xy} and ν_{yx} for each microstructure. All mechanical simulations were performed with an applied stress of 1 MPa. In a similar manner, two thermal simulations were performed with an applied heat flux of 1 W/m^2 in the X-direction and Y-direction, respectively. These simulations produced estimates of the homogenized thermal conductivities, k_x and k_y. Including the six different sets of constituent properties summarized in Table 2, a total of 30 simulations were performed for each microstructure. We included different contrast ratios in our work because previous research has demonstrated that producing surrogate ML models for the homogenized properties becomes increasingly difficult as the contrast ratio between the phases increases [21, 133, 134].

[19] Additionally, we computed localized elastic strain fields that are not included in the dataset due to the extreme memory cost. Interested readers should contact the authors.

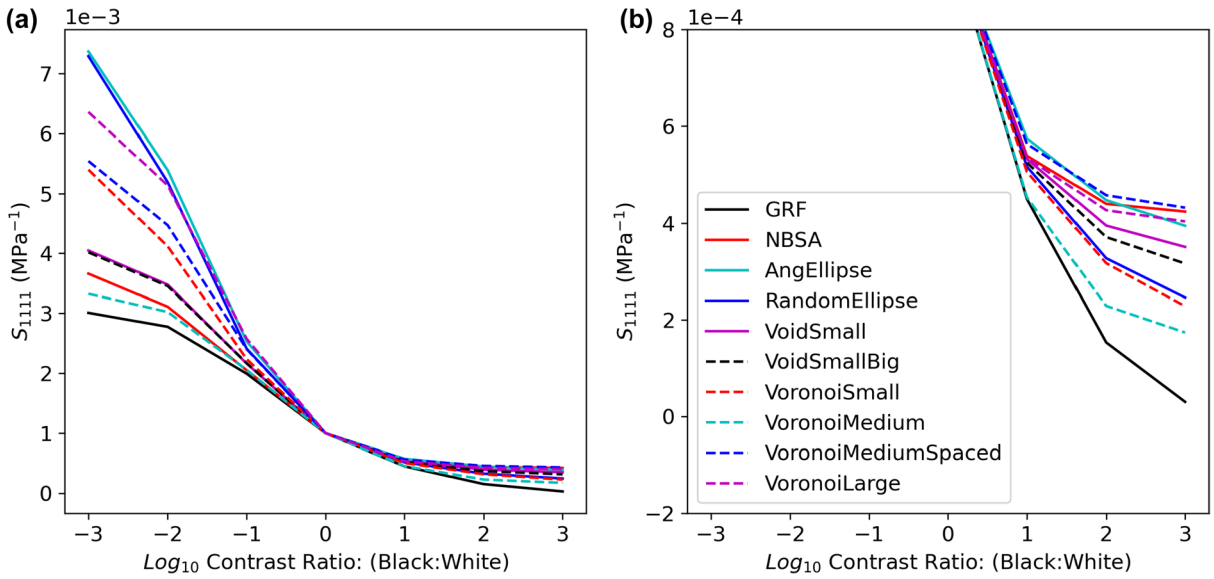

Fig. 12 Trend of S_{1111} against increasing contrast ratio between the two phases. The far right contrast ratio, highlighted in (**b**), is the 1000:1 contrast ratio displayed in Fig. 11

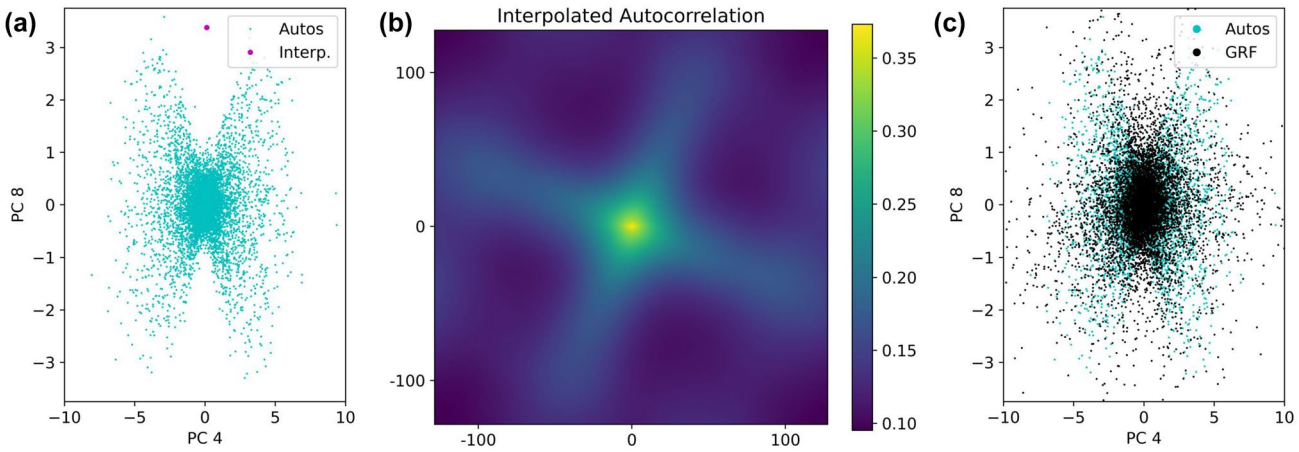

Fig. 13 Illustration of the incomplete coverage produced by the adopted autocorrelation parameterization. **a** PC subspace projection of the original autocorrelation dataset. **b** A interpolated autocorrelation contained in the unpopulated convex full of the dataset. **c** A projection of the autocorrelations of the GRF-class microstructures onto the subspace from (**a**)

Figure 11 exemplifies the distribution of computed mechanical properties for the extreme contrast case 1000 : 1 contrast ratio (black:white in, for example, Fig. 5). The first row summarizes the property distribution for the entire dataset while the second row summarizes the distribution for just the GRF class. In addition, the first-order Hill-Paul bounds [135, 136] are reported in order to contextualize the achieved values. The extracted properties are presented as compliance coefficients to facilitate comparison against these theoretical bounds. Reflecting on the first row, MICRO2D achieves good coverage of the displayed property space. This statement remained true for every property we checked. We emphasize that the framework was not designed to produce a coverage of any specific material property space. Instead, the framework—and the dataset—were designed to diversely cover the space of 2-point statistics. The diverse coverage of a wide variety of material properties is only achieved *indirectly*. In this context, the achieved covered emphasizes an extremely valuable element of the dataset. The MICRO2D dataset is likely to automatically display diversity with respect to many materials properties—even those not considered here—because of the important role of 2-point

statistics in estimating the homogenized properties using statistical materials theory [60, 86, 88, 93, 96, 100, 101].

Contrasting the first and second rows of Fig. 11 illuminates a similar trend as the one observed in the microstructure space. The GRF class displays the widest variety of homogenized properties. This reflects the flexibility of the GRF to achieve a variety of long range spatial patterns as well as local feature sizes. In contrast, the remaining classes cover subsets of the property space.

The incorporation of the neighborhood distributions is also important; variation in the neighborhood distribution can significantly shift the homogenized properties that characterize a material. This is even true if the generating 2-point statistics are held constant. For example, Fig. 12 contrasts the homogenized S_{1111} of the ten artificial microstructures from Fig. 6. The right most contrast ratio—Fig. 12b—corresponds to the contrast ratio displayed in Fig. 11. Contrasting the variation, it is clear that adjusting the neighborhood distribution, even when the generating 2-point statistics are maintained constant, causes a half order of magnitude change in the displayed property—nearly half the width of the theoretical bounds. This indicates that specific properties can likely be achieved by either varying the spatial statistics or the neighborhood distribution.

The achieved variation in the property space reflects the microstructural diversity of MICRO2D and highlights the importance of varying the 2-point statistics and the neighborhood distributions. In Fig. 11d–f, we see that variation in the 2-point statistics, even for a constant neighborhood distribution and volume fraction, results in extremely significant shifts in homogenized properties.[20] Similarly, variation in the neighborhood distribution, keeping the volume fraction and spatial statistics constant, also leads to large shifts, Fig. 12. By varying both, MICRO2D provides a holistic environment for studying complex Structure-Property linkages.

Discussion

The proposed framework represents a tremendous advance toward developing a systematic method for exploring the microstructure space. Using the framework, we were able to generate an open-source dataset displaying tremendous variability in its constituent 2-phase microstructures as well as in its homogenized material properties. However, we see this work as one initial effort in the ongoing drive toward data curation in Materials Informatics.

In this paper, we argue that diverse datasets should be curated with respect to a statistical measure—we used autocorrelations. The sheer magnitude and complex constraints of the space of autocorrelations make it extremely challenging to cover this space and quantitatively decide whether a diverse coverage has been achieved. Of course, this problem is not unique to the space of autocorrelations; we would expect this problem to be further exacerbated for full 2-point statistics needed in multiphase (with more than two phases) and/or polycrystalline material systems [88] as well as higher-order statistics (e.g., 3-point statistics). As an initial test, there are some simple measures that can be used to grade the coverage. For example, Niezgoda et al. [71] demonstrate that the space of autocorrelations is convex. Our earlier analysis showed that the initial autocorrelation dataset (recall that this dataset was used to generate MICRO2D) is convex in most projections. However, this is not always the case. Figure 13a displays a PC subspace projection of the autocorrelation dataset, which clearly shows a non-convex "butterfly" structure. This is a clear indication that the initial autocorrelation coverage is incomplete. In fact, this intermediate gap is a result of limitations in the proposed parameter generation scheme. We can explore the gap by interpolating between autocorrelations on its left and right. While the left and right butterfly wings contain autocorrelations with single central features, the gap contains autocorrelations with multiple, orthogonal central peaks, e.g., Fig. 13b. It was largely unexplored because our generation scheme only included single central peaks. However, we emphasize that in practice this identified gap is very small; in fact, during generation, it is largely filled by the random scatter in the GRF, Fig. 13c. Even though this specific example was largely inconsequential, it emphasizes an important point. It is extremely unlikely that the coverage is complete (the gaps in the property space confirm this). This is especially true because we are *passively* producing a coverage by using expert knowledge to guide generation. Active strategies (e.g., active space filling [139, 140] and output driven coverage [108, 141]), in which generation is systematically optimized, are likely necessary to fill these unknown gaps. We emphasize that the development of the frameworks necessary to do this and whether or not it is even necessary are significant open problems given the high dimensionality and complex constraints of these spaces.

Conclusions

In this work, we present a computational framework for directly curating statistically diverse big 2-phase microstructure datasets. In addition, we introduce an open-source large, statistically diverse microstructure dataset intended to enable future microstructure informatics efforts. The core

[20] In practice, such second order variability arises in many important material classes and is important to study to achieve desirable properties (e.g., rafting in nickel superalloy [137, 138]).

theoretical contribution of the framework is a passive strategy for generating an efficient coverage of the space of autocorrelations. The strategy has two components. The first is a flexible, parametric approximation of an arbitrary autocorrelation. The approximation is presented in algorithmic form in Eqs. (5)–(10). Importantly, this parameterization provide a highly intuitive mechanism for systematically constructing valid, synthetic autocorrelations. The second element is a MaxPro-based thinning procedure that identifies an efficient, space-filling coverage from a candidate set of autocorrelations produced using this parameterization. The framework performs dataset curation by combining this sampling procedure and a set of identified neighborhood distributions with established statistically conditioned generative models [63, 64].

To demonstrate the proposed framework, we synthesize a second-order diverse, large, 2-phase microstructure dataset, MICRO2D. We analyze the dataset's 2- and 3-point statistics to quantify the achieved diversity. In addition to the microstructures, we provide several important homogenized properties for each microstructure that could be used as regression targets for future learning problems. In addition to microstructural diversity, MICRO2D encloses a large range of theoretically-possible homogenized property values. This achievement demonstrates the clear value in using n-point statistics as a statistical measure for constructing the dataset. Because of their relationship to statistical materials theory [60, 61], diversity with respect to the n-point statistics indirectly guarantees diversity with respect to arbitrary material properties. This allows us to avoid the painful process of targeting properties individually when creating the dataset and instead focus exclusively on the microstructural domain. In theory, this connection means we expect MICRO2D to display diversity even with respect to unconsidered properties.

Finally, we end with a discussion of the dataset generation problem. We emphasize, again, that this study represents just an initial step in generating second-order statistically-diverse datasets. Many important research paths remain. First, for practical reasons, the generated dataset contains microstructures limited to a constant resolution and a 256×256 pixel discretization. In addition, the LGD generation strategy incorporates two important lengthscales of features: long range spatial patterns and local neighborhood distributions. Arbitrary real experimental microstructures can contain much more complex hierarchies of features which require larger discretizations to capture. We need more advanced generative models and dataset curation strategies to consider these complex systems. Second, a clear definition of "sufficient diversity" in the autocorrelation space remains a challenging open question due to the extreme dimensionality of this space. Further, for practical engineering purposes, it would be extremely useful to extend this to systems with more complex local states, such as polycrystalline microstructures. It remains unclear how to extend the proposed autocorrelation parameterization to more complex states. The generated dataset is also only well-sampled in a second-order sense, despite some higher-order diversity being introduced via the neighborhood distributions. Extension to higher-order statistics would incorporate structures with nonlinear long range patterns (e.g., copolymers [65, 66, 76]).

Appendix A: Real Space Mixtures

The mixture method adopted in this paper differs slightly from the approach adopted historically in Spectral Mixture-based Gaussian Process Regression modeling [97, 98]. Specifically, instead of constructing the kernel function via a mixture model approximation to a probability density in the frequency space (e.g., a Symmetric Gaussian mixture model [97]), we approximate the kernel using a real-space symmetric Gaussian mixture model and, subsequently, enforce the spectral requirements of the kernel function via two linear projections. This approach takes the following path. The approximate kernel function, $\hat{k}(\boldsymbol{\tau})$, is constructed via a mixture of symmetric Gaussians.

$$\hat{k}(\boldsymbol{\tau}) = \sum_{i=1}^{M} \frac{\alpha_i}{2} \left[\phi(\boldsymbol{\tau}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \phi(\boldsymbol{\tau}; -\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right] \tag{11}$$

$$\phi(\boldsymbol{\tau}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(2^k \pi^k |\boldsymbol{\Sigma}| \right)^{-1/2} \exp\left(-0.5 (\boldsymbol{\tau} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\tau} - \boldsymbol{\mu}) \right) \tag{12}$$

Here, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariances of each mixture. $|\cdot|$ is the determinant operator. The mixture weights, α_i, are selected to add to unity.

An approximate kernel structure of this form produces the following expression when transformed into frequency space [142].

$$\mathcal{F}(\hat{k}(\boldsymbol{\tau}))(\boldsymbol{\xi}) \propto \sum_{i=1}^{M} \left[\exp\left(-i \boldsymbol{\mu}_i^T \boldsymbol{\xi} \right) + \exp\left(i \boldsymbol{\mu}_i^T \boldsymbol{\xi} \right) \right] \exp\left(-\boldsymbol{\xi}^T \boldsymbol{\Sigma}_i \boldsymbol{\xi} \right) \tag{13}$$

$$\propto \sum_{i=1}^{M} \cos\left(\boldsymbol{\mu}_i^T \boldsymbol{\xi} \right) \exp\left(-\boldsymbol{\xi}^T \boldsymbol{\Sigma}_i \boldsymbol{\xi} \right) \tag{14}$$

Here, superscripts refer to exponentiation not indexing. Note that this produces the inverse of the kernel structure proposed by Wilson and Adams [97]—with the cosine fluctuations in the frequency space instead of the real space. This kernel structure meets only one of the minimum requirements outlined in Sects. 2.2 and 2.1: it is real valued. The presence of the cosine fluctuations introduces negative

values in the spectrum. These fluctuations are removed by zeroing the negative values [92].

$$\mathcal{F}(k(\boldsymbol{\tau}))(\boldsymbol{\xi}) = \max\left(\mathcal{F}(\hat{k}(\boldsymbol{\tau}))(\boldsymbol{\xi}), \epsilon\right) \tag{15}$$

Here, ϵ is a near zero, positive value added for computational stability of the subsequent steps. Finally, the generated kernel function is produced by applying the inverse cosine transform, $\mathcal{C}^{-1}[\cdot]$. This operation returns the real space equivalent of the kernel without introducing spurious imaginary components in either the real or Fourier space representation of the kernel. In practice, we discretely sample the approximate covariance kernel, $\hat{k}(\boldsymbol{\tau})$, in real space to a discrete covariance kernel, $\hat{k}_r$, and, subsequently, apply the two identified projections discretely. This procedure produces the following set of expressions.

$$\hat{k}(\boldsymbol{\tau}) = \sum_{i=1}^{M} \frac{\alpha_i}{2}\left[\phi(\boldsymbol{\tau};\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \phi(\boldsymbol{\tau}; -\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right] \tag{16}$$

$$\hat{k}_r = \hat{k}(\boldsymbol{\tau}_r) \tag{17}$$

$$k_r = \mathcal{C}^{-1}[\max\left(\mathcal{F}[\hat{k}_r]_t, \epsilon\right)]_r \tag{18}$$

Here, $\boldsymbol{\tau}_r$ is the value of $\boldsymbol{\tau}$ at the center of pixel r. The autocorrelation is derived from the kernel function via addition of the mean squared [63].

$$f_r^{\beta\beta} = k_r + (v_f^{\beta})^2 \tag{19}$$

As noted in the main body of the paper, we used this alternative structure instead of the traditional method for two important reasons. Most importantly, the traditional Fourier-space mixture model produces spatially compact real-space kernels. This means that it cannot easily produce kernels with multiple modes. Mathematically, this is clear in the original real-space expression provided by Wilson and Adams [97] (here, for simplicity, we reproduce the 1D single mixture expression).

$$k(\tau) = \exp(-2\pi^2\tau^2\sigma^2)\cos(2\pi\tau\mu) \tag{20}$$

Clearly, the dominant exponential term has zero mean. As a result, this type of kernel cannot easily reproduce the important longer range peaks (for example, secondary peaks in layered composites that statistically represent the repetition of the layering [63]) that are present in 2-point statistics maps [90, 94, 123]. In contrast, our real-space formulation can directly construct these secondary peaks via the direct placement of the means of the individual symmetric mixtures. The second reason is practical and an extension of the first: in real space, we can use our expert knowledge of 2-point statistics [21, 72, 88, 90, 92, 93, 123] to guide the placement of the symmetric mixtures into common regions. The unfamiliar nature of the Fourier representation makes it challenging to embed domain knowledge into the construction of the kernel function (and, by extension, the autocorrelation). Of course, this second reason would be irrelevant if one was using an optimization-based placement strategy for the mixtures instead of an expert driven one.

Appendix B: PCA Truncation for MaxPro Filtering

We used PCA to perform distance preserving dimensionality reduction of the initial candidate autocorrelation set. This facilitated the framework's spacefilling filtering operation by significantly decreasing the computational expense of the Min–Max optimization central to the MaxPro algorithm [113]. Importantly, the extracted latent space must be a good approximation of the original 2-point statistics. Therefore, it must have sufficient representational capacity to recreate the original autocorrelations' salient features with high fidelity. Recent work by Generale et al. observed that PCA is relatively inefficient for generative tasks like this one [21]. As a result, we expect the number of necessary principal components to be quite high. We selected the truncation level for the number of principal components by tracking the reconstruction error (the relative L_2 error—i.e., the L_2 distance between a proposed autocorrelation and the reconstruction from the principal component basis normalized by the L_2

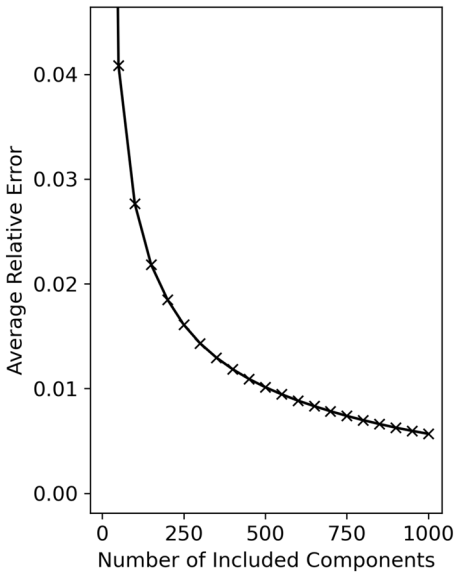

Fig. 14 Trend of the average reconstruction error (the relative L_2 error) as a function of the number of retained principal components

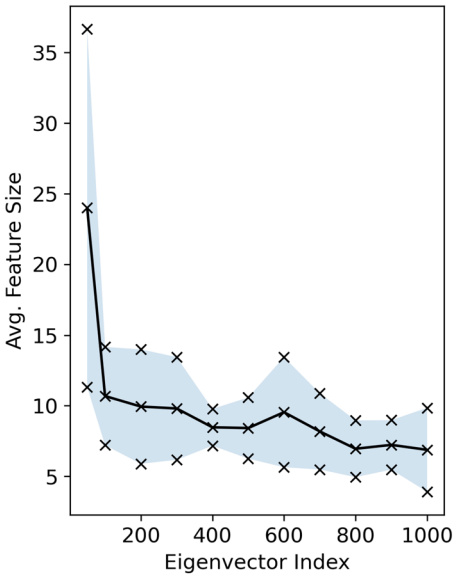

Fig. 15 Trend of the expected feature size as a function of the principal component index

magnitude of the original autocorrelation) of the original candidate autocorrelation dataset. Figure 14 summarizes the relative error. We selected a truncation level of 750, achieving an average reconstruction error of approximately 0.75%.

Additionally, we also explored the average feature size of stochastic microstructure functions differentiated by each eigenvector. Here, our aim was to ensure that the selected principal component basis was sensitive to every feature lengthscale in the initial candidate autocorrelation set. This is important to check because PCA is known to filter out short lengthscale (i.e., high frequency) features [112, 143–145]. For several eigenvectors, we identified a set of diverse autocorrelations with respect to the selected eigenvector. We did so by performing spacefilling in just the subspace defined by that eigenvector using the described MaxPro procedure. Then, we used Berryman's method [63, 146, 147] to estimate the feature size of microstructures corresponding to the identified autocorrelations. Figure 15 depicts the trend. Importantly, the decay largely stabilizes well before the 750th eigenvector. Therefore, we expect this selected cutoff to create a subspace which is sufficiently sensitive to all salient lengthscales in the candidate autocorrelation set. We also note that the trend displays largely monotonic decay—i.e., lower index eigenvectors correspond to larger features.

Appendix C: Examples from MICRO2D

Here, we simply display a collection of randomly selected examples from each class in the MICRO2D dataset: GRF—Fig. 16, NBSA—Fig. 17, AngEllipse—Fig. 18, RandomEllipse—Fig. 19, VoidSmall—Fig. 20, VoidSmallBig—Fig. 21, VoronoiLarge—Fig. 22, VoronoiMedium—Fig. 23, VoronoiMediumSpaced—Fig. 24, and VoronoiSmall—Fig. 25.

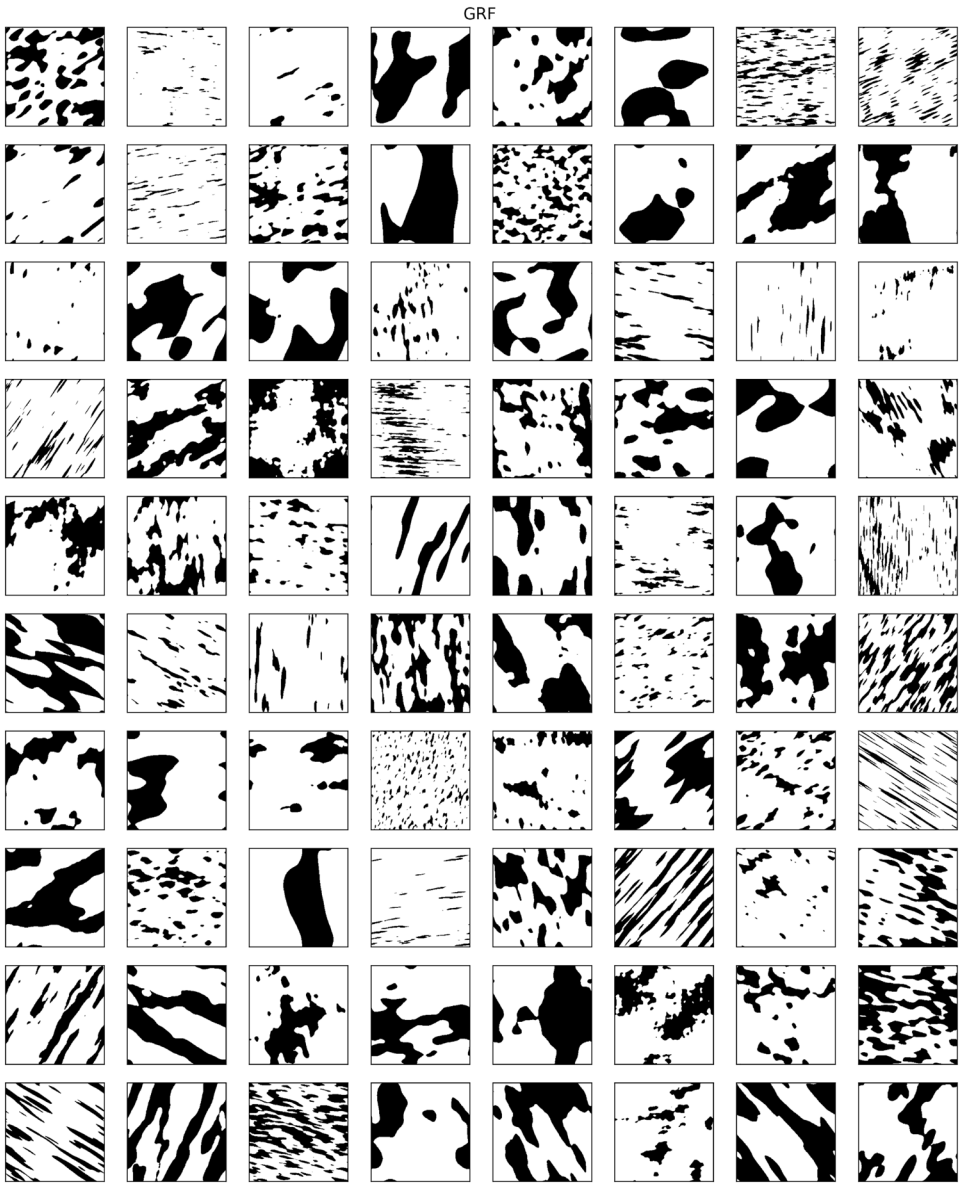

Fig. 16 Eighty randomly selected microstructures corresponding to the GRF class

Fig. 17 Eighty randomly selected microstructures corresponding to the NBSA class

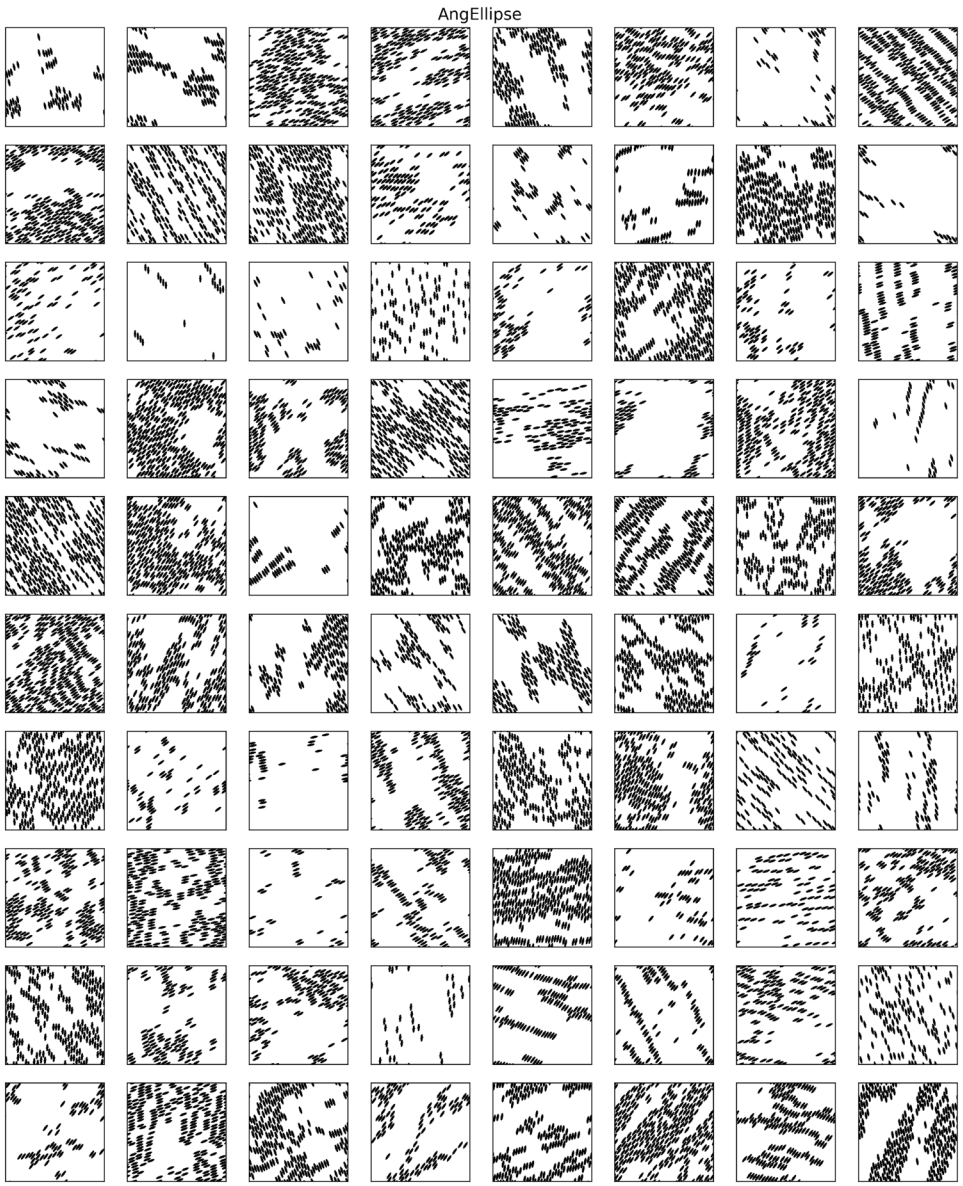

Fig. 18 Eighty randomly selected microstructures corresponding to the AngEllipse class

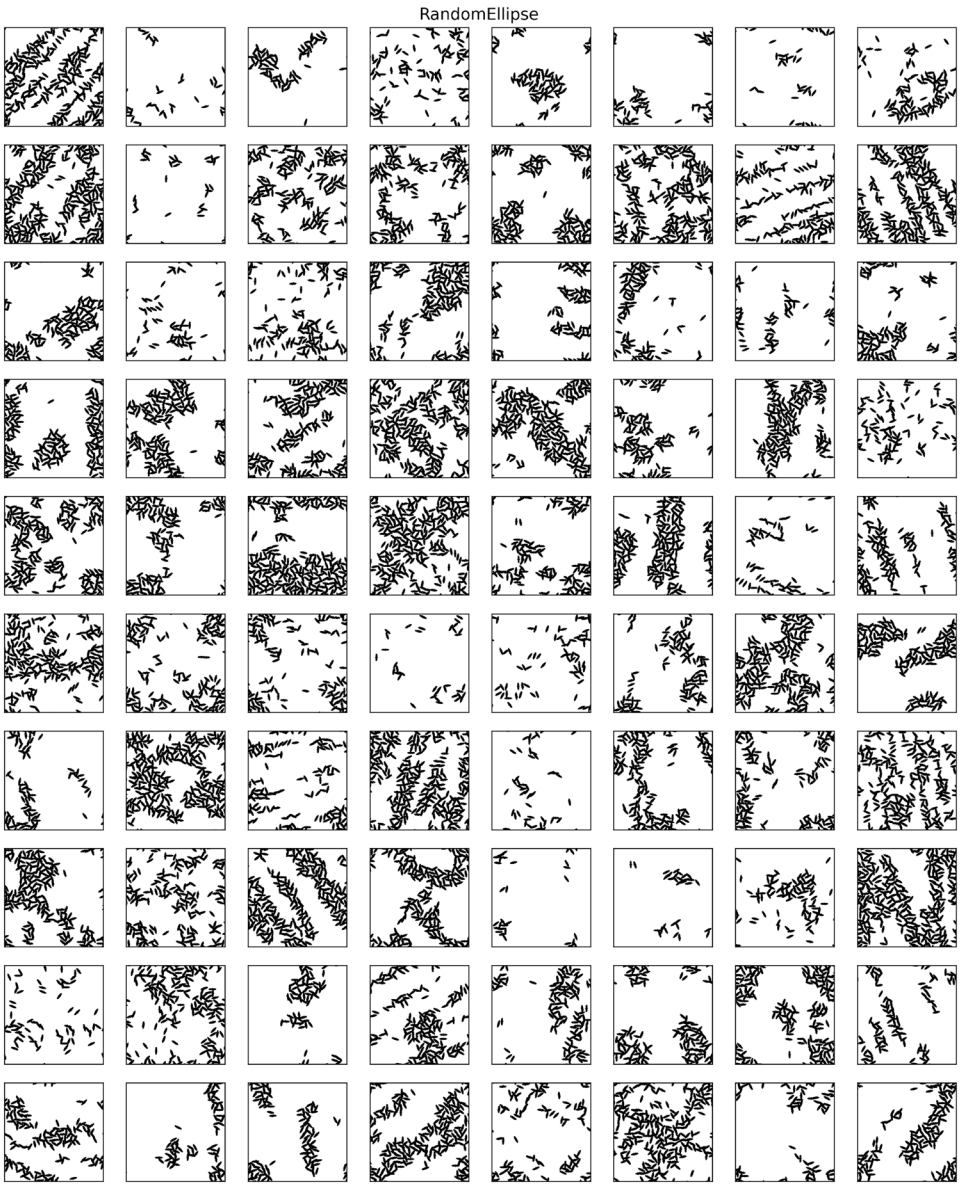

Fig. 19 Eighty randomly selected microstructures corresponding to the RandomEllipse class

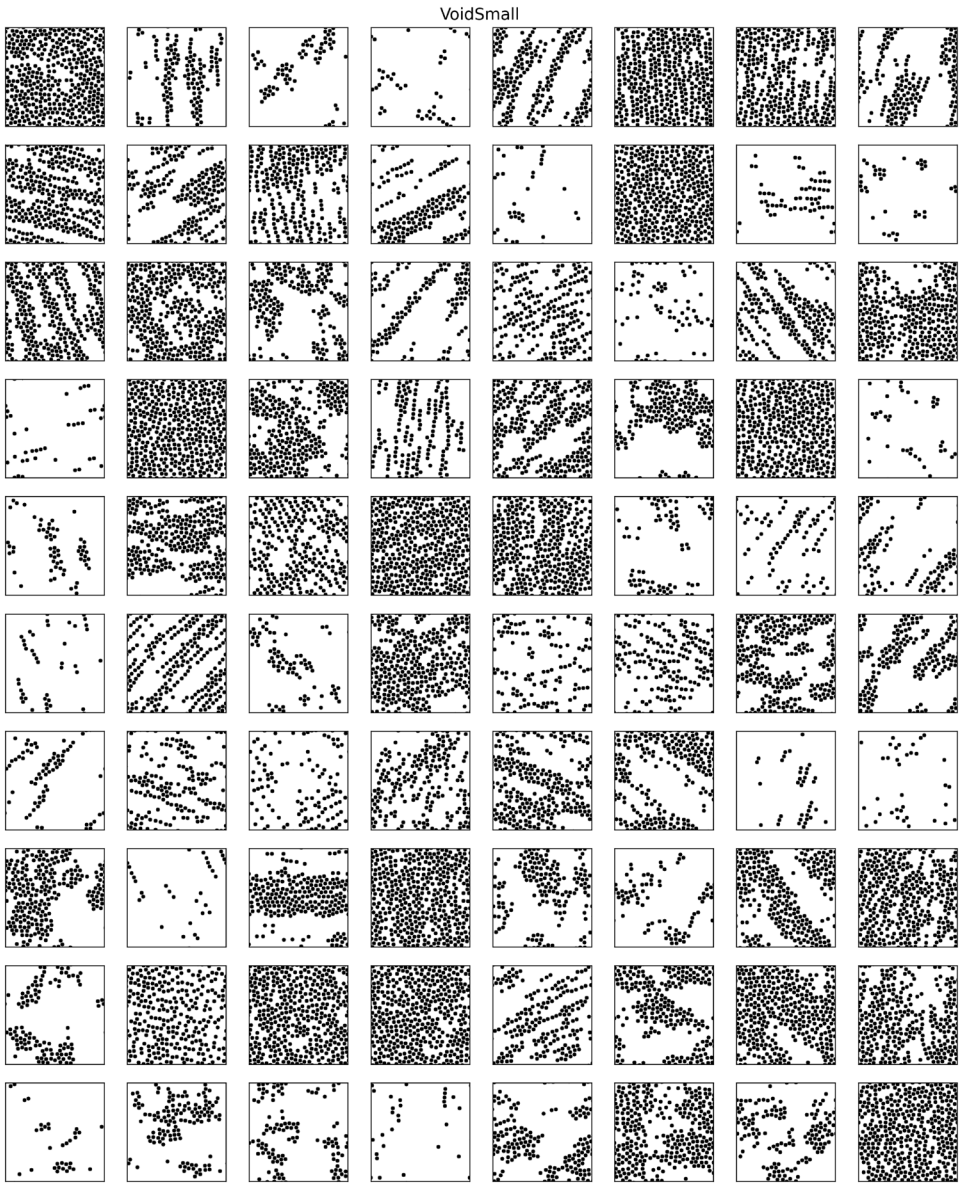

Fig. 20 Eighty randomly selected microstructures corresponding to the VoidSmall class

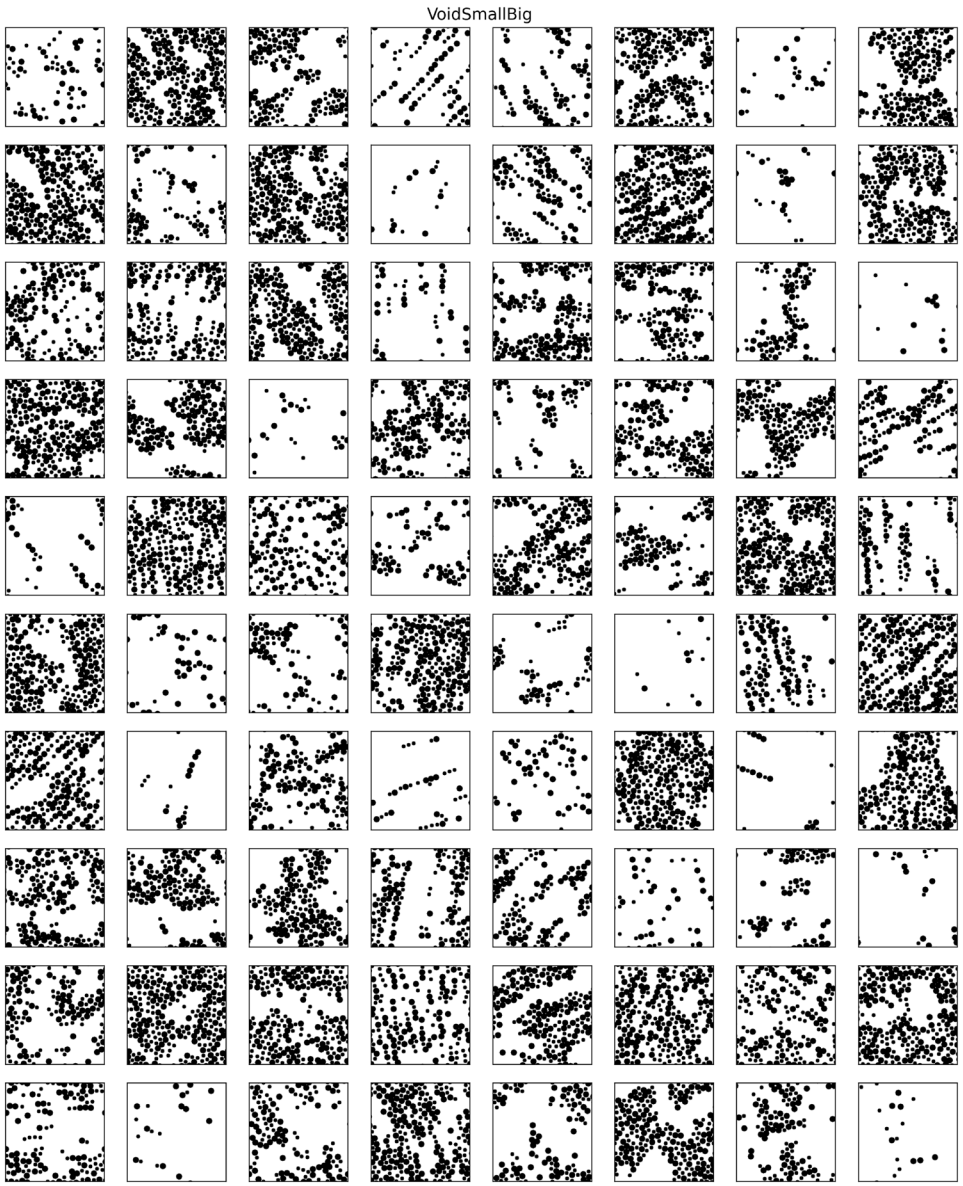

Fig. 21 Eighty randomly selected microstructures corresponding to the VoidSmallBig class

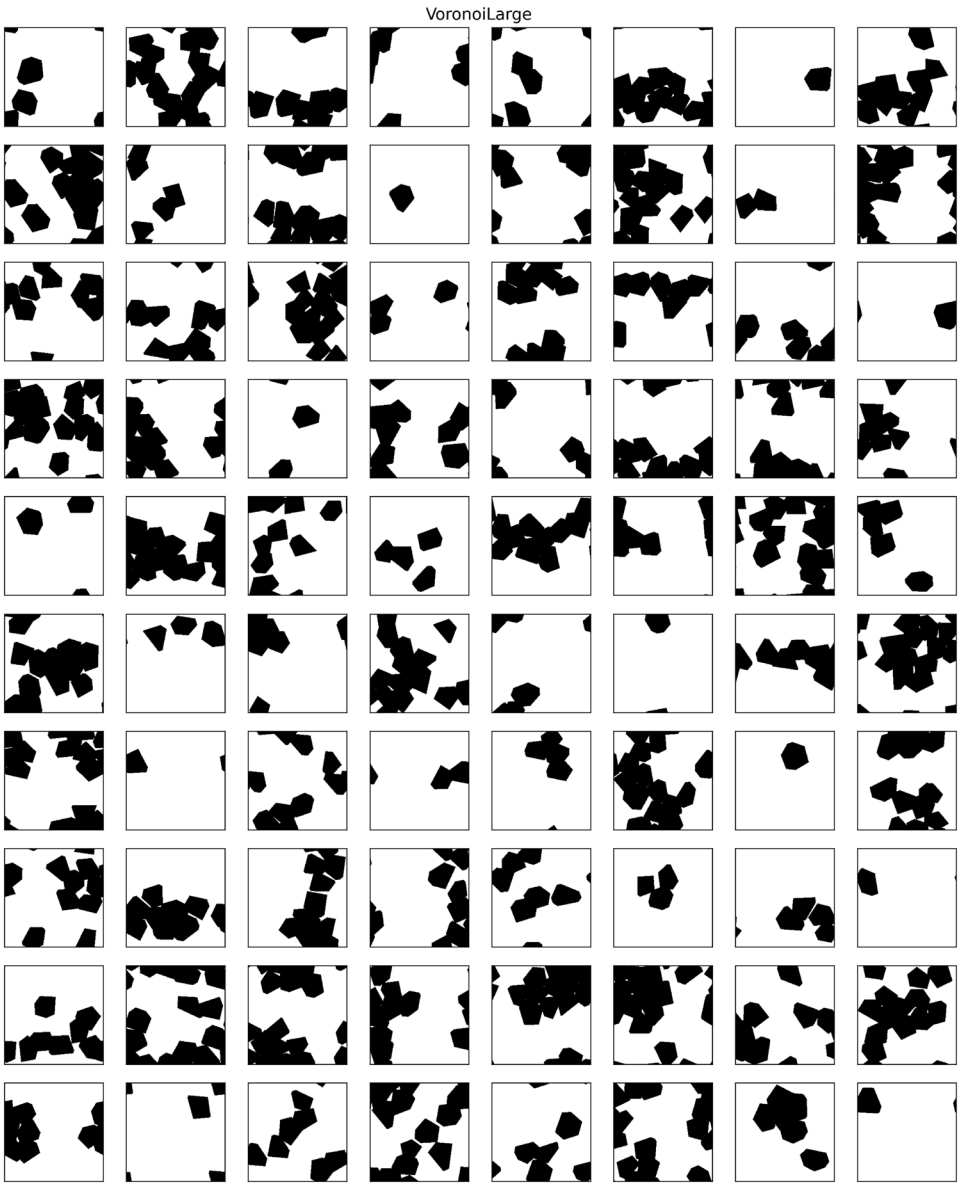

Fig. 22 Eighty randomly selected microstructures corresponding to the VoronoiLarge class

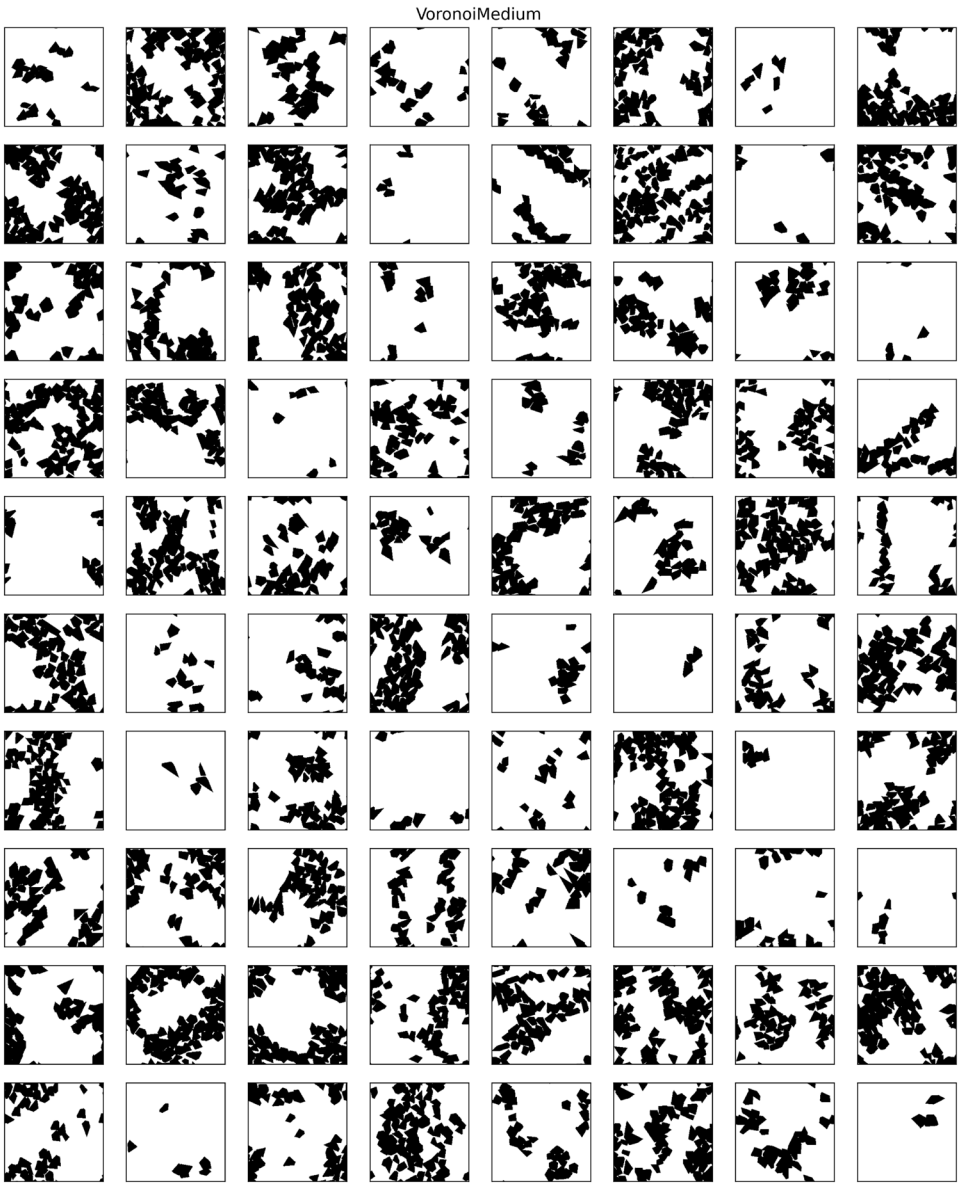

Fig. 23 Eighty randomly selected microstructures corresponding to the VoronoiMedium class

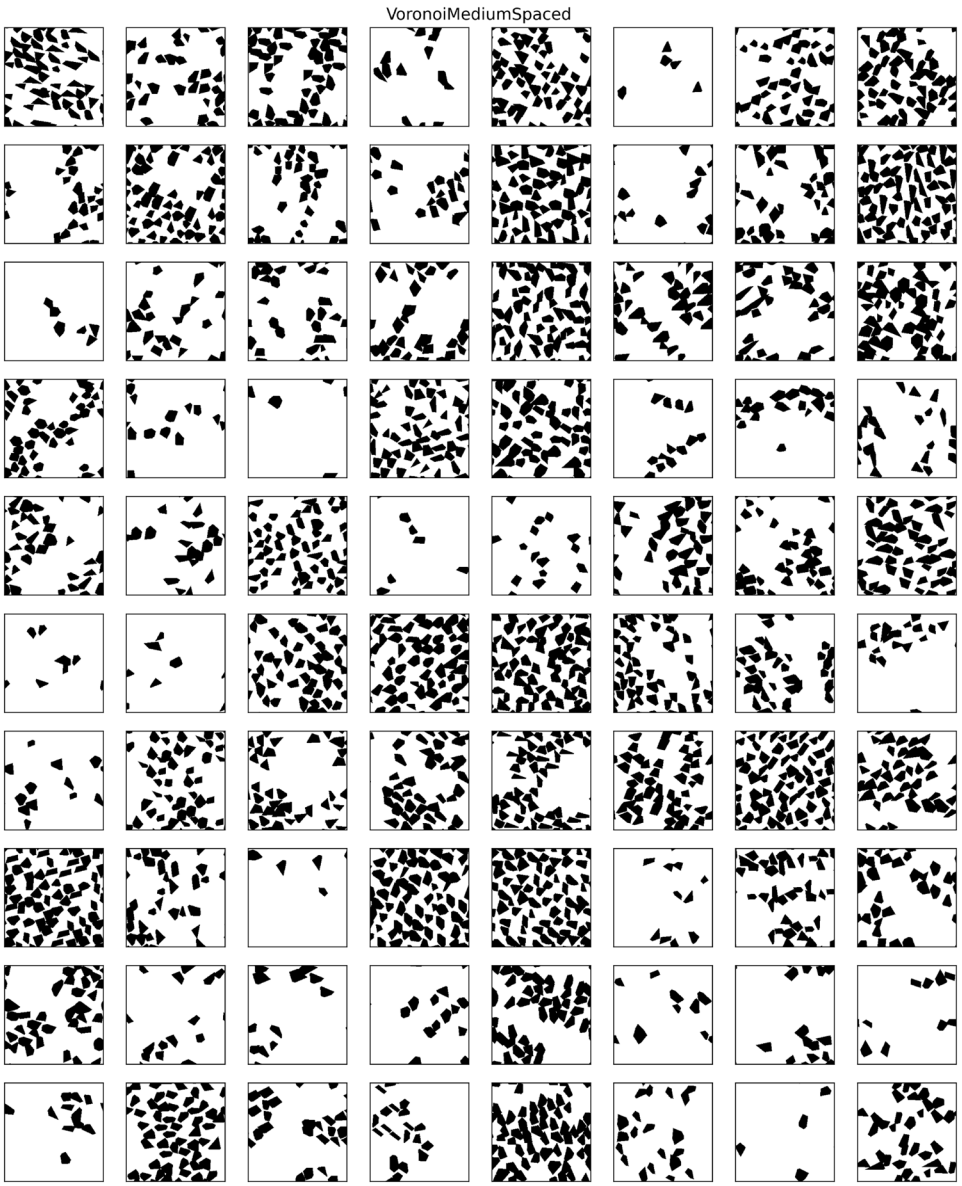

Fig. 24 Eighty randomly selected microstructures corresponding to the VoronoiMediumSpaced class

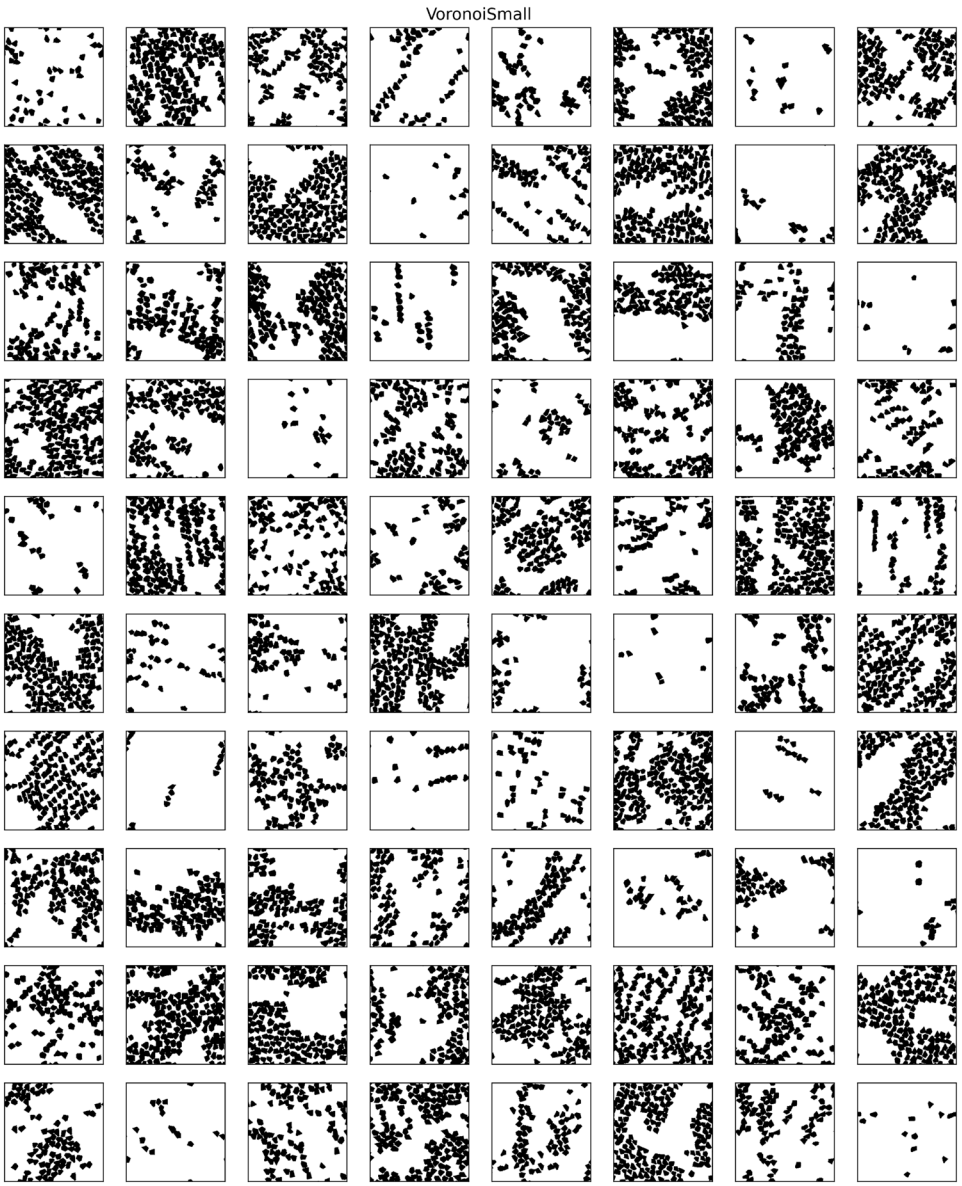

Fig. 25 Eighty randomly selected microstructures corresponding to the VoronoiSmall class

Acknowledgements A.E. Robertson and S.R. Kalidindi thank the National Science Foundation for their support under NSF 2027105. A.P. Generale acknowledges Pratt & Whitney and the Alfred P. Sloan Foundation. C. Kelly acknowledges NSF 2027105, NSF Graduate Research Fellowship DGE-1650044, and ONR N00014-18-1-2879. M. Buzzy acknowledges support from NSF DMREF 2119640. Additionally, A.E. Robertson would like to acknowledge the continued support of the Jack Kent Cooke Foundation. A.P. Generale would like to acknowledge the continued support of the Alfred P. Sloan Foundation.

Code Availability The MICRO2D dataset and the code used in this paper will be freely provided upon publication at https://arobertson38.github.io/MICRO2D.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260. https://doi.org/10.1126/science.aaa8415
2. Vaswani A, Shazeer N, Parmar N, Uskoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. Attention is all you need, NeurIPS
3. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y Generative adversarial networks, NeurIPS
4. Chen N, Zhang Y, Zen H, Weiss R, Norouzi M, Chan W (2009) Wavegrad: estimating gradients for waveform generation, https://doi.org/10.48550/arxiv.2009.00713
5. Mahdavifar S, Ghorbani AA (2019) Application of deep learning to cybersecurity: a survey. Neurocomputing 347:149–176. https://doi.org/10.1016/j.neucom.2019.02.056
6. Cai L, Gao J, Zhao D (2020) A review of the application of deep learning in medical image classification and segmentation. Ann Translat Med 8:713. https://doi.org/10.21037/atm.2020.02.44
7. Jiang W (2021) Applications of deep learning in stock market prediction: recent progress. Expert Syst Appl 184:115537. https://doi.org/10.1016/j.eswa.2021.115537
8. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805
9. Song Y, Sohl-Dickstein J, Kigma DP, Kumar A, Ermon S, Poole B (2021) Score-based generative modeling through stochastic differential equations. In: International congress for learning representation, pp 1–36
10. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. NeurIPS
11. Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet D. Video diffusion models, https://doi.org/10.48550/arxiv.2204.03458
12. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks, In: Pereira F, Burges C, Bottou L, Weinberger K (eds), Advances in neural information processing systems, vol. 25, Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
13. Hamilton W, Ying Z, Leskovec J (2017) Inductive representation learning on large graphs, In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds), Advances in neural information processing systems, vol. 30, Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf
14. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K, Klingner J, Shah A, Johnson M, Liu X, Kaiser L, Gouws S, Kato Y, Kudo T, Kazawa H, Stevens K, Kurian G, Patil N, Wang W, Young C, Smith J, Riesa J, Rudnick A, Vinyals O, Corrado G, Hughes M, Dean J (2016) Google's neural machine translation system: bridging the gap between human and machine translation. arXiv:1609.08144
15. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronnberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl S, Ballard A, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Peterson S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior A, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with alphafold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
16. Anand N, Achim T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models, https://doi.org/10.48550/arxiv.2205.15019
17. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Ganesan S, Goodsell DS, Ghosh S, Green RK, Guranovi V, Guzenko D, Hudson BP, Lawson CL, Liang Y, Lowe R, Namkoong H, Peisach E, Persikova I, Randle C, Rose A, Rose Y, Sali A, Segura J, Sekharan M, Shao C, Tao Y-P, Voigt M, Westbrook JD, Young JY, Zardecki C, Zhuravleva M (2020) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res 49(D1):D437–D451. https://doi.org/10.1093/nar/gkaa1038
18. Fersht A (2021) Alphafold: a personal perspective on the impact of machine learning. J Mol Biol 433(20):167088. https://doi.org/10.1016/j.jmb.2021.167088
19. Zheng S, He J, Liu C, Shi Y, Lu Z, Feng W, Ju F, Wang J, Zhu J, Min Y, Zhang H, Tang S, Hao H, Jin P, Chen C, Noé F, Liu H, Liu T-Y (2023) Towards predicting equilibrium distributions for molecular systems with deep learning. arxiv:2306.05445
20. Materials genome initiative for global competitiveness
21. Generale A, Robertson A, Kelly C, Kalidindi S. Inverse stochastic microstructure design, SSRN: preprint https://doi.org/10.2139/ssrn.4590691
22. Gao Y, Liu Y. Relibaility-based topology optimization with stochastic heterogeneous microstructure properties. Mater Des. https://doi.org/10.1016/j.matdes.2021.109713
23. Marshall A, Kalidindi S (2021) Autonomous development of a machine-learning model for the plastic response of two-phase composites from micromechanical finite element models. JOM 73:2085–2095. https://doi.org/10.1007/s11837-021-04696-w
24. Kalidindi S, Binci M, Fullwood D, Adams B (2006) Elastic properties closures using second-order homogenization theories: case studies in composites of two isotropic constituents. Acta Mater 54:3117–3126. https://doi.org/10.1016/j.actamat.2006.03.005
25. Hasan M, Mao Y, Tavazza F, Choudhary A, Agrawal A, Acar P. Data-driven multi-scale modeling and optimization for elastic properties of cubic microstructures. Integr Mater Manuf Innov. https://doi.org/10.1007/s40192-022-00258-3
26. Acar P, Sundararaghavan V (2019) Stochastic design optimization of microstructural features using linear programming for robust design. AIAA J 57:448–455

27. Xiong Y, Duong P, Wang D, Park S-I, Ge Q, Raghavan N, Rosen D (2019) Data-driven design space exploration and exploitation for design for additive manufacturing. J Mech Des 141:101101. https://doi.org/10.1115/1.4043587
28. Morris C, Bekker L, Haberman M, Seepersad C (2018) Design exploration of reliably manufacturable materials and structures with applications to negative stiffness metamaterials and microstereolithography. J Mech Des 140:111415. https://doi.org/10.1115/1.4041251
29. Pei Z, Rozman KA, Dogan ÖN, Wen Y, Gao N, Holm EA, Hawk JA, Alman DE, Gao MC (2021) Machine-learning microstructure for inverse material design. Adv Sci 8:2101207. https://doi.org/10.1002/advs.202101207
30. Fung V, Zhang J, Hu G, Ganesh P, Sumpter BG (2021) Inverse design of two-dimensional materials with invertible neural networks. npj Comput Mater 7:200. https://doi.org/10.1038/s41524-021-00670-x
31. Abram M, Burghardt K, Steeg GV, Galstyan A, Dingreville R. Inferring topological transitions in pattern forming processes with self supervised learning, NPJ: Comput Mater 8. https://doi.org/10.1038/s41524-022-00889-2
32. Diehl M, Groeber M, Haase C, Molodov D, Roters F, Raabe D (2017) Identifying structure-property relationships through dream. 3d representative volume elements and damask crystal plasticity simulations: An integrated computational materials engineering approach. JOM 69:848–855. https://doi.org/10.1007/s11837-017-2303-0
33. Muir C, Swaminathan B, Almansour A, Sevener K, Smith C, Presby M, Kiser J, Pollock T, Daly S. Damage mechanism identification in composites via machine learning and acoustic emission, NPJ: Comput Mater 7. https://doi.org/10.1038/s41524-021-00565-x
34. Hashemi S, Kalidindi SR (2023) Gaussian process autoregression models for the evolution of polycrystalline microstructures subjected to arbitrary stretching tensors. Int J Plast 162:103532. https://doi.org/10.1016/j.ijplas.2023.103532
35. Yabansu YC, Steinmetz P, Hötzer J, Kalidindi SR, Nestler B (2017) Extraction of reduced-order process-structure linkages from phase-field simulations. Acta Mater 124:182–194. https://doi.org/10.1016/j.actamat.2016.10.071
36. Dornheim J, Morand L, Zeitvogel S, Iraki T, Link N, Helm D. Deep reinforcement learning methods for structure-guided processing path optimization. J Intell Manuf 33. https://doi.org/10.1007/s10845-021-01805-z
37. Vlassis NN, Sun W (2023) Denoising diffusion algorithm for inverse design of microstructures with fine-tuned nonlinear material properties. Comput Methods Appl Mech Eng 413:116126. https://doi.org/10.1016/j.cma.2023.116126
38. Jain A, Ong S, Hautier G, Chen W, Richards W, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson K (2013) Commentary: The materials project: a materials genome approach to accelerating materials innovation. APL Mater 1:011002. https://doi.org/10.1063/1.4812323
39. Groeber M, Jackson M (2014) Dream.3d: a digital representation environment for the analysis of microstructure in 3d, Integrating Materials and Manufacturing. Innovation 3:56–72. https://doi.org/10.1186/2193-9772-3-5
40. Groeber M, Ghosh S, Uchic M, Dimiduk D (2008) A framework for automated analysis and simulation of 3d polycrystalline microstructures. part 2: synthetic microstructure generation. Acta Mater 56:1274–1287. https://doi.org/10.1016/j.actamat.2007.11.040
41. Pilchak AL, Shank J, Tucker JC, Srivatsa S, Fagin PN, Semiatin SL(2016) A dataset for the development, verification, and validation of microstructure-sensitive process models for near-alpha titanium alloys. Integr Mater Manuf Innov, 1–18 https://doi.org/10.1186/s40192-016-0056-1
42. DeCost BL, Holm EA (2016) A large dataset of synthetic SEM images of powder materials and their ground truth 3d structures. Data Brief 9:727–731. https://doi.org/10.1016/j.dib.2016.10.011
43. Kalidindi S, Khosravani A, Yucel B, Shanker A, Blekh A (2019) Data infrastructure elements in support of accelerated materials innovation: ELA, PyMKS, and MATIN. Integr Mater Manuf Innov 8:441–454
44. Hart KA, Rimoli JJ (2020) Microstructpy: a statistical microstructure mesh generator in python. SoftwareX 12:100595. https://doi.org/10.1016/j.softx.2020.100595
45. Song K, Yan Y (2013) A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Appl Surf Sci 285P:858–864. https://doi.org/10.1016/j.apsusc.2013.09.002
46. DeCost BL, Hecht M, Francis T, Webler BA, Picard YN, Holm E (2017) Uhcsdb: ultra high carbon steel micrograph database. Integr Mater Manuf Innov 6:197–205. https://doi.org/10.1007/s40192-017-0097-0
47. Barber Z, Leake J, Clyne T. The doitpoms project: micrograph library. https://www.doitpoms.ac.uk/miclib/index.php
48. Saal J, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials design and discovery with high-throughput denisty functional theory: the open quantum materials database. JOM 65:1501–1509. https://doi.org/10.1007/s11837-013-0755-4
49. Choudhary K, Garrity KF, Reid ACE, DeCost B, Biacchi AJ, Walker ARH, Trautt Z, Hattrick-Simpers J, Kusne AG, Centrone A, Davydov A, Jiang J, Pachter R, Cheon G, Reed E, Agrawal A, Qian X, Sharma V, Zhuang H, Kalinin SV, Sumpter BG, Pilania G, Acar P, Mandal S, Haule K, Vanderbilt D, Rabe K, Tavazza F, The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. npj Comput Mater 6. https://doi.org/10.1038/s41524-020-00440-1
50. Tanifuji M, Matsuda A, Yoshikawa H (2019) Materials data platform: a fair system for data-driven materials science, In: 2019 8th International congress on advanced applied informatics (IIAI-AAI), pp 1021–1022. https://doi.org/10.1109/IIAI-AAI.2019.00206
51. Ma R, Luo T (2020) PI1M: a benchmark database for polymer informatics. J Chem Inf Model 60(10):4684–4690. https://doi.org/10.1021/acs.jcim.0c00726
52. Borysov S, Geilhufe R, Balatsky A. Organic materials database: an open-access online database for data mining. PLoS ONE 12. https://doi.org/10.1371/journal.pone.0171501
53. Kench S, Squires I, Dahari A Microlib: A library of 3d microstructures generated from 2d micrographs using slicegan. Sci Data 9. https://doi.org/10.1038/s41597-022-01744-1
54. Bargmann S, Klusemann B, Markmann J, Schnabel J, Schneider K, Soyarslan C, Wilmers J (2018) Generation of 3d representative volume elements for heterogeneous materials: a review. Prog Mater Sci 96:322–384. https://doi.org/10.1016/j.pmatsci.2018.02.003
55. Mosser L, Dubrule O, Blunt M (2018) Stochastic reconstruction of oolitic limestone by generative adversarial networks. Transp Porous Med 125:81–103. https://doi.org/10.1007/s11242-018-1039-9
56. Kench S, Cooper S (2021) Generating three-dimensional structures from a two-dimensional slice with generative adversarial network-based dimensionality expansion. Nature Mach Intell 3:299–305. https://doi.org/10.1038/s42256-021-00322-1
57. Fokina D, Muravleva E, Ovchinnikov G, Oseledets I (2020) Microstructure synthesis using style-based generative adversarial networks. Phys Rev E 101:043308. https://doi.org/10.1103/PhysRevE.101.043308

58. Noguchi S, Inoue J (2021) Stochastic characterization and reconstruction of material microstructures for establishment of process-structure-property linkage using the deep generative model. Phys Rev E 104:025302. https://doi.org/10.1103/PhysRevE.104.025302
59. Fullwood D, Niezgoda S, Adams B, Kalidindi S (2010) Microstructure sensitive design for performance optimization. Prog Mater Sci 55:477–562. https://doi.org/10.1016/j.pmatsci.2009.08.002
60. Torquato S (2002) Random heterogeneous materials. Springer, New York
61. Adams B, Kalidindi S, Fullwood D (2013) Microstructure sensitive design for performance optimization. Butterworth-Heinemann, Waltham
62. Gao Y, Jiao Y, Liu Y (2021) Ultra-efficient reconstruction of 3d microstructure and distribution of properties of random heterogeneous materials containing multiple phases. Acta Mater 204:116526. https://doi.org/10.1016/j.actamat.2020.116526
63. Robertson A, Kalidindi S (2022) Efficient generation of n-field microstructures from 2-point statistics using multi-output gaussian random fields. Acta Mater 232:117927. https://doi.org/10.1016/j.actamat.2022.117927
64. Robertson AE, Kelly C, Buzzy M, Kalidindi SR (2023) Local-global decompositions for conditional microstructure generation. Acta Mater 253:118966. https://doi.org/10.1016/j.actamat.2023.118966
65. Seibert P, Ambati M, Rabloff A, Kastner M (2021) Reconstructing random heterogeneous media through differentiable optimization. Comput Mater Sci 196:110455. https://doi.org/10.1016/j.commatsci.2021.110455
66. Seibert P, Rabloff A, Ambati M, Kastner M (2022) Descriptor-based reconstruction of three-dimensional microstructures through gradient-based optimization. Acta Mater 227:117667. https://doi.org/10.1016/j.actamat.2022.117667
67. Seibert P, Husert M, Wollner M, Kalina K, Kastner M. Fast reconstruction of microstructures with ellipsoidal inclusions using analytic descriptors, https://doi.org/10.48550/arxiv.2306.08316
68. Falco S, Jiang J, Cola FD, Petrinic N (2017) Generation of 3d polycrystalline microstructures with a conditioned Laguerre–Voronoi tessellation technique. Comput Mater Sci 136:20–28. https://doi.org/10.1016/j.commatsci.2017.04.018
69. Prasad M, Vajragupta N, Hartmaier A (2019) Kanapy: a python package for generating complex synthetic polycrystalline microstructures. J Open Source Softw 4:1732. https://doi.org/10.21105/joss.01732
70. Mandal S, Lao J, Donegan S, Rollett A (2018) Generation of statistically representative synthetic three-dimensional microstructures. Scripta Mater 146:128–132. https://doi.org/10.1016/j.scriptamat.2017.11.034
71. Niezgoda S, Fullwood D, Kalidindi S (2008) Delineation of the space of 2-point correlations in a composite material system. Acta Mater 56:5285–5292. https://doi.org/10.1016/j.actamat.2008.07.005
72. de Oca Zapiain DM, Stewart J, Dingreville R (2021) Accelerating phase field based microstructure evolution predictions via surrogate models trained by machine learning methods. NPJ Comput Mater 3:1–11. https://doi.org/10.1038/s41524-020-00471-8
73. Attari V, Honarmandi P, Duong T, Sauceda DJ, Allaire D, Arroyave R (2020) Uncertainty propagation in a multiscale calphad-reinforced elastochemical phase-field model. Acta Mater 183:452–470. https://doi.org/10.1016/j.actamat.2019.11.031
74. Hsu T, Epting WK, Kim H, Abernathy HW, Hackett GA, Rollett AD, Salvador PA, Holm EA (2021) Microstructure generation via generative adversarial network for heterogeneous, topoligically complex 3d materials. JOM 73:90–102. https://doi.org/10.1007/s11837-020-04484-y
75. NIMS, Nims materials database. https://mits.nims.go.jp/en/
76. Lee K, Yun G Microstructure reconstruction using diffusion-based generative models
77. Lin H, Brown LP, Long AC (2011) Modelling and simulating textile structures using texgen, In: Advances in textile engineering, vol. 331 of advanced materials research, pp 44–47. https://doi.org/10.4028/www.scientific.net/AMR.331.44
78. Krishnamoorthi S, Bandyopadhyay R, Sangid MD (2023) A microstructure-based fatigue model for additively manufactured ti-6al-4v, including the role of prior β boundaries. Int J Plast 163:103569. https://doi.org/10.1016/j.ijplas.2023.103569
79. Du P, Zebrowski A, Zola J, Ganapathysubramanian B, Wodo O. Microstructure design using graphs. Comput Mater 4. https://doi.org/10.1038/s41524-018-0108-5
80. Dureth C, Seibert P, Rucker D, Handford S, Kastner M, Gude M. Conditional diffusion-based microstructure reconstruction
81. Jung J, Yoon JI, Park HK, Jo H, Kim HS (2020) Microstructure design using machine learning generated low dimensional and continuous design space. Materialia 11:100690. https://doi.org/10.1016/j.mtla.2020.100690
82. Tang J, Geng X, Li D, Shi Y, Tong J, Xiao H, Peng F (2021) Machine learned-based microstructure prediction during laser sintering of alumina. Sci Rep 11:10724. https://doi.org/10.1038/s41598-021-89816-x
83. Iyer A, Dey B, Dasgupta A, Chen W. A conditional generative model for predicting material microstructures from processing methods
84. Kanit T, Forest S, Galliet I, Mounoury V, Jeulin D (2003) Determination of the size of the representative volume element for random composites: statistical and numerical approach. Int J Solids Struct 40(13):3647–3679. https://doi.org/10.1016/S0020-7683(03)00143-4
85. Kim Y, Jung J, Park H, Kim H (2023) Importance of microstructural features in bimodal structure-property linkage. Met Mater Int 29:53–58. https://doi.org/10.1007/s12540-022-01200-0
86. Paulson N, Priddy M, McDowell D, Kalidindi S (2019) Reduced-order microstructure-sensitive protocols to rank-order the transition fatigue resistance of polycrystalline microstructures. Int J Fatigue 119:1. https://doi.org/10.1016/j.ijfatigue.2018.09.011
87. Latypov M, Toth L, Kalidindi S (2019) Materials knowledge system for nonlinear composites. Comput Methods Appl Mech Eng 346:180. https://doi.org/10.1016/j.cma.2018.11.034
88. Paulson N, Priddy M, McDowell D, Kalidindi S (2017) Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics. Acta Mater 129:428. https://doi.org/10.1016/j.actamat.2017.03.009
89. Kaundinya PR, Choudhary K, Kalidindi SR. Machine learning approaches for feature engineering of the crystal structure: application to the prediction of the formation energy of cubic compounds, https://doi.org/10.48550/arXiv.2105.11319
90. Generale A, Kalidindi S (2021) Reduced-order models for microstructure-sensitive effective thermal conductivity of woven ceramic matrix composites with residual porosity. Compos Struct 274:114399. https://doi.org/10.1016/j.compstruct.2021.114399
91. Fast T, Wodo O, Ganapathysubramanian B, Kalidindi S (2016) Microstructure taxonomy based on spatial correlations: application to microstructure coarsening. Acta Mater 108:176. https://doi.org/10.1016/j.actamat.2016.01.046
92. Harrington G, Kelly C, Attari V, Arroyave R, Kalidindi S (2022) Application of a chained-ann for learning the process-structure mapping in $mg_2si_xsn_{1-x}$ spinodal decomposition.

Integr Mater Manuf Innov 11:433–449. https://doi.org/10.1007/s40192-022-00274-3
93. Barry MC, Gissinger JR, Chandross M, Wise KE, Kalidindi SR, Kumar S (2023) Voxelized atomic structure framework for materials design and discovery. Comput Mater Sci 230:112431. https://doi.org/10.1016/j.commatsci.2023.112431
94. Yabansu YC, Iskakov A, Kapustina A, Rajagopalan S, Kalidindi S. Application of gaussian process regression models for capturing the evolution of microstructure statistics in aging of nickel-based superalloys. Acta Mater 178
95. Altschuh P, Yabansu YC, Hötzer J, Selzer M, Nestler B, Kalidindi SR (2017) Data science approaches for microstructure quantification and feature identification in porous membranes. J Membr Sci 540:88–97. https://doi.org/10.1016/j.memsci.2017.06.020
96. Latypov M, Kalidindi S (2017) Data-driven reduced order models for effective yield strength and partitioning of strain in multiphase materials. J Comput Phys 346:242–261. https://doi.org/10.1016/j.jcp.2017.06.013
97. Wilson A, Adams R (2013) Gaussian process kernels for pattern discovery and extrapolation, In: Proceedings of the 30th international conference on machine learning, vol 28 of proceedings of machine learning research, PMLR, pp 1067–1075
98. Lazaro-Gredilla M, Quinonero-Candela J, Rasmussen C, Figueiras-Vidal A (2010) Sparse spectrum gaussian process regression. J Mach Learn Res, 1865–1881
99. Soutis C (2005) Fibre reinforced composites in aircraft construction. Prog Aerosp Sci 41:143–151. https://doi.org/10.1016/j.paerosci.2005.02.004
100. Brown Jr WF (1955) Solid mixture permittivities. J Chem Phys 23:1514–1517
101. Kroner E (1977) Bounds for effective elastic moduli of disordered materials. J Mech Phys Solids 25:137–155
102. Safdari M, Baniassadi M, Garmestani H, Al-Haik M (2012) A modified strong-constrast expansion for estimating the effective thermal conductivity of multiphase heterogeneous materials. J Appl Phys 112:114318
103. Torquato S (1997) Effective stiffness tensor of composite media: 1. Exact series expansions. J Mech Phys Solids 45:1421–1448
104. Torquato S (1998) Effective stiffness tensor of composite media: 2. Applications to isotropic dispersions. J Mech Phys Solids 46:1411–1440
105. Fullwood D, Adams B, Kalidindi S (2008) A strong contrast homogenization formulation for multi-phase anistropic materials. J Mech Phys Solids 56:2287–2297
106. Hashemi S, Kalidindi S (2021) A machine learning framework for the temporal evolution of microstructure during static recrystallization of polycrystalline materials simulated by cellular automaton. Comput Mater Sci 188:110132. https://doi.org/10.1016/j.commatsci.2020.110132
107. Fullwood D, Adams B, Kalidindi S (2007) Generalized pareto front methods applied to second-order material property closures. Comput Mater Sci 38:788–799. https://doi.org/10.1016/j.commatsci.2006.05.016
108. Mann A, Kalidindi S (2022) Development of a robust cnn model for capturing microstructure-property linkages and building property closures supporting material design. Front Mater 9:851085. https://doi.org/10.3389/fmats.2022.851085
109. Rossin J, Leser P, Pusch K, Frey C, Vogel S, Saville A, Torbet C, Clarke A, Daly S, Pollock T (2022) Single crystal elastic constants of additively manufactured components determined by resonant ultrasound spectroscopy. Mater Charact 192:112244. https://doi.org/10.1016/j.matchar.2022.112244
110. Kroner E (1972) Statistical continuum mechanics. Springer, New York
111. Niezgoda S, Yabansu Y, Kalidindi S (2011) Understanding and visualizing microstructure and microstructure variance as a stochastic process. Acta Mater 59:6387–6400. https://doi.org/10.1016/j.actamat.2011.06.051
112. Shlens J (2020) A tutorial of principal component analysis. Accessed 28 Nov 2020. https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf
113. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11(1):137–148. https://doi.org/10.1080/00401706.1969.10490666
114. Mak S, Joseph V (2018) Minimax and minimax projection designs using clustering. J Comput Graph Stat 27:166–178. https://doi.org/10.1080/10618600.2017.1302881
115. Huang C, Joseph V, Ray D (2021) Constrained minimum energy designs. Stat Comput 31:80. https://doi.org/10.1007/s11222-021-10054-2
116. Fullwood D, Niezgoda S, Kalidindi S (2008) Microstructure reconstruction from 2-point statistics using phase recovery algorithms. Acta Mater 56:942–948. https://doi.org/10.1016/j.actamat.2007.10.044
117. Jiao Y, Stillinger F, Torquato S (2007) Modeling heterogeneous materials via two-point correlation functions: basic principles. Phys Rev E 76:031110. https://doi.org/10.1103/PhysRevE.76.031110
118. Jiao Y, Stillinger F, Torquato S (2009) A superior descriptor of random textures and its predictive capacity. PNAS 106:17634–17639. https://doi.org/10.1073/pnas.0905919106
119. Niezgoda SR, Turner DM, Fullwood DT, Kalidindi SR (2010) Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics. Acta Mater 58(13):4432–4445. https://doi.org/10.1016/j.actamat.2010.04.041
120. Helton J, Davis F (2003) Latin hypercube sampling and propogation of uncertainty in analyses of complex systems. Reliab Eng Syst Saf, 23–69. https://doi.org/10.1016/S0951-8320(03)00058-9
121. Swayer S (2023) Wishart distributions and inverse-wishart sampling. Accessed 4 Oct 2023. https://www.math.wustl.edu/~sawyer/hmhandouts/Wishart.pdf
122. Odell PL, Feiveson AH (1966) A numerical procedure to generate a sample covariance matrix. J Am Stat Assoc 61(313):199–203. https://doi.org/10.1080/01621459.1966.10502018
123. Cecen A (2017) Calculation, utilization, and inference of spatial statistics in practical spatio-temporal data. Georgia Tech Library, Atlanta
124. Cecen A, Yucel B, Kalidindi S (2021) A generalized and modular framework for digital generation of composite microstructures. J Compos Sci 5:1–20. https://doi.org/10.3390/jcs5080211
125. Brough D, Wheeler D, Kalidindi S (2017) Materials knowledge systems in python: a data science framework for accelerated development of hierarchical materials. Integr Mater Manuf Innov 6:36–53. https://doi.org/10.1007/s40192-017-0089-0
126. Kelly C, Kalidindi S (2021) Recurrent localization networks applied to the Lippmann–Schwinger equation. Comput Mater Sci 192:110356. https://doi.org/10.1016/j.commatsci.2021.110356
127. You H, Zhang Q, Ross C, Lee C, Yu Y (2022) Learning deep implicit fourier neural operators (ifnos) with applications to heterogeneous material modeling. Comput Methods Appl Mech Eng 398:115296. https://doi.org/10.1016/j.cma.2022.115296
128. Chun S, Roy S, Nguyen Y, Choi J, Udaykumar H, Baek S (2020) Deep learning for synthetic microstructure generation in a materials-by-design framework for heterogeneous energetic materials. Sci Rep 10:13307. https://doi.org/10.1038/s41598-020-70149-0
129. Ostoja-Starzewski M, Kale S, Karimi P, Malyarenko A, Raghavan B, Ranganathan S, Zhang J (2016) Chapter two-scaling to RVE in random media, vol 49 of Advances in Applied Mechanics, pp 111–211. https://doi.org/10.1016/bs.aams.2016.07.001

130. Zerhouni O, Brisard S, Danas K. Quantifying the effects of two-point correlations on the effective elasticity of specific classes of random porous materials with and without connectivity. Int J Eng Sci. https://doi.org/10.1016/j.ijengsci.2021.103520
131. Li S (1999) On the unit cell for micromechanical analysis of fibre-reinforced composites. Proc R Soc A 455:815–838. https://doi.org/10.1098/rspa.1999.0336
132. Li S (2001) General unit cells for micromechanical analyses of unidirectional composites. Compos A Appl Sci Manuf 32(6):815–826. https://doi.org/10.1016/S1359-835X(00)00182-2
133. Landi G, Niezgoda N, Kalidindi S (2010) Multi-scale modeling of elastic propoerties of three-dimensional voxel-based microstructure datasets using novel DFT-based knowledge systems. Acta Mater 58:2716–2725. https://doi.org/10.1016/j.actamat.2010.01.007
134. Fast T, Kalidindi SR (2011) Formulation and calibration of higher-order elastic localization relationships using the MKS approach. Acta Mater 59:4595–4605. https://doi.org/10.1016/j.actamat.2011.04.005
135. Proust G, Kalidindi S (2006) Procedures for construction of anisotropic elastic-plastic property closures for face-centered cubic polycrystals using first-order bounding relations. J Mech Phys Solids 54:1744–1762. https://doi.org/10.1016/j.jmps.2006.01.010
136. Hill R (1963) Elastic properties of reinforced solids: some theoretical principles. J Mech Phys Solids 11:357–372
137. Yang M, Zhang J, Wei H, Zhao Y, Gui W, Su H, Jin T, Liu L. Study of γ' rafting under different stress states: a phase field simulation considering viscoplasticity. J Alloys Compounds. https://doi.org/10.1016/j.jallcom.2018.07.317
138. Blesgen T, Chenchiah I. Cahn–Hilliard equations incorporating elasticity: analysis and comparison to experiments. Philos Trans R Soc. https://doi.org/10.1098/rsta.2012.0342
139. Chen W, Fuge M (2017) Beyond the known: detecting novel feasible domains over unbounded design space. J Mech Des 139:111405. https://doi.org/10.1115/1.4037306
140. Chen W, Fuge M (2019) Synthesizing designs with interpart dependencies using hierarchical generative adversarial networks. J Mech Des 141:111403. https://doi.org/10.1115/1.4044076
141. Wang S, Generale AP, Kalidindi SR, Joseph VR (2023) Sequential designs for filling output spaces. Technometrics, 1–12 https://doi.org/10.1080/00401706.2023.2231042
142. Ahrendt P (2023) The multivariate gaussian probability. Accessed 4 Oct 2023. https://d1wqtxts1xzle7.cloudfront.net/49874923/The_Multivariate_Gaussian_Probability_Di20161026-27105-77g7a0-libre.pdf?1477466954= &response-content-disposition=inline%3B+filename%3DThe_multivariate_gaussian_probability_di.pdf &Expires=1696429097 &Signature=EbY-smInGeeMVvC0qsTaERE9jTZTSJF8NC9MZl0fOkqTiBgWVcmYqZ~u-8vaYnjyuJyCgV-40kYMMHThOOAhgEGQ8~2dzZG~TV7Rn69mTy1I1ieWafwrsatRpsj3CB6KIbhRn6Y2MgwENUL0RVxnycgT2uiSJiAAoucqbOw5cxBO9H2OrgzgT2SywfSb2hxmr~GLayEwsCWUA~QRgm4AYcbK-YwWebZcZ6RkMOCMotDks-aCd66kbFpBz8bdM3avpmNpYJRWn9jxUFhDhJOnhz0OFdidp~fN96dS-J7~hSJDeK4dGDBE03b5sUd4Px7YrFf4jCCD6KOn1ldefSJR9w__ &Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA
143. Chawla M (2011) PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: a survey and comparison. Appl Soft Comput 11(2):2216–2226. https://doi.org/10.1016/j.asoc.2010.08.001
144. Hastie T, Tibshirani R, Friedman J (2016) The elements of statistical learning. Springer, New York
145. Vetterli M, Kovacevic J, Goyal V (2014) Foundations of signal processing. Cambridge University Press, Cambridge
146. Berryman J (1987) Relationship between specific surface area and spatial correlation functions for anistropic porous media. J Math Phys 28:244–245
147. Blair S, Berge P, Berryman J (1996) Using two-point correlation functions to characterize microgeometry and estimate permeabilities of sandstone and porous glass. J Geophys Res 101:20359–20375. https://doi.org/10.1029/96JB00879