


Methodology for estimation of intrinsic dimensions and state variables of microstructuresVeera Sundararaghavan ^{*}*Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA*Megna N. Shah[†] and Jeff P. Simmons[‡]*Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, Ohio 45433, USA* (Received 11 October 2022; revised 12 July 2023; accepted 16 August 2023; published 6 September 2023)

According to the manifold hypothesis, real data can be compressed to lie on a low-dimensional manifold. This paper explores the estimation of the dimensionality of this manifold with an interest in identifying independent degrees of freedom and possibly identifying state variables that would govern materials systems. The challenges identified that are specific to materials science are (i) accurate estimation of the number of dimensions of the data, (ii) coping with the intrinsic random and low-bit-depth nature of microstructure samples, and (iii) linking noncompressed domains such as processing to microstructure. Dimensionality estimates are made with the maximum-likelihood-estimation method with the Minkowski p -norms being used as a measure of the distance between microstructural images. It is found that, where dimensionality estimates are required to be accurate, it is necessary to use the Minkowski 1-norm (also known as the L_1 -norm or Manhattan distance). This effect is found to be due to image quantification and proofs are given regarding the distortion produced by quantization. It is also found that homogenization is an effective way of estimating the dimension of random microstructure image sets. An estimate of 40 dimensions for the fibers of a SiC/SiC fiber composite is obtained. It is also found that, with images generated from a sparse domain (surrogate to the process domain), it is possible to infer the nature of the process manifold from images alone.

DOI: [10.1103/PhysRevE.108.035001](https://doi.org/10.1103/PhysRevE.108.035001)**I. INTRODUCTION**

It is widely accepted that controlling the microstructure of a material will enable control of its properties. However, it is less clear which, or even how many, of the features of the microstructure represent its variability. Recently, Chen *et al.* [1] identified intrinsic dimensions in complex and chaotic dynamical systems, using only short videos of their behavior, and proposed that state variables of complex systems may be identified in this way. This suggests the tantalizing prospect of identification of a minimal set of microstructural state variables, equal in number to the intrinsic dimension, that would govern the material's behavior. This minimum number of features would encode all of the dimensions in the microstructure necessary to make design decisions, much like when the Wright brothers "invented the airplane" by discovering how to control *all* dimensions of rotation. Finding and controlling all dimensions of the microstructure could enable a completely new way of exploiting design spaces.

Recent advances in characterization techniques and computing have led to the generation and analysis of large data sets, enabling an improved understanding of microstructure. Much work has been done to quantify various aspects of the microstructure, such as particle size and shape distributions,

orientation distributions, and n -point statistics among others [2–5], enabling significant advancements in the understanding of processing structure properties. However, this has relied on domain experts manually identifying which features should be characterized. While advancing understanding, this approach still leaves uncertainty about whether all of the important variation in the structure was captured and quantified. Recently, deep learning generative methods have shown promise towards capturing the key variables accounting for most of the variation in image-based data sets [6,7] using latent space representations, which are low-dimensional representations of image data. Such models have been trained on microstructural data [8–11]. Deep learning methods often leverage the fact that although each microstructure image can be represented as a vector of size n , the actual dimensionality is expected to be much lower.

In formal terms, a data set containing points of dimensionality n is said to have intrinsic dimensionality (ID) equal to $\mu < n$ if every point lies entirely within an μ -dimensional manifold of \mathbb{R}^n . The methods of dimensionality estimation can be categorized as local and global approaches. Global methods for ID estimation rely on the spread of the entire data set, as exemplified by projection methods such as principal component analysis (PCA). Linear methods such as PCA and multidimensional scaling were explored for microstructural data in Refs. [12–14]. However, it is known that such methods tend to fail on nonlinear manifolds [15]. Other global approaches to dimension reduction such as Isomap and its variants treat nonlinear manifolds using geodesic distances

^{*}Corresponding author: veeras@umich.edu[†]megna.shah.1@us.af.mil[‡]jeff.simmons.3@afml.af.mil

[16] and have been used to reduce the dimensionality of microstructures [15,17]. Local approaches use the local geometry of the high-dimensional space to estimate the intrinsic dimension and tend to be more computationally efficient [18]. Levina and Bickel [19] developed such an estimate by choosing an optimal dimension in which the local neighborhood of points would be uniformly spaced. Pope *et al.* [20] applied this methodology to estimate the dimensionality of some well-known benchmark data sets such as the MNIST [21] and CIFAR [22] data sets and found that the information in those had a surprisingly low number of dimensions, i.e., degrees of freedom. Estimates ranged from 10 to 25 dimensions from the simplest to the most complex data set. Much of the prior work relies on human judgment as to the reasonableness of the dimensionality estimates and did not have any ground truths by which to evaluate such reasonableness. Consequently, assessing the validity of the methods becomes problematic.

Materials science has specific challenges that lead to considerations over and above the above approaches, which are heavily centered around natural images.

(a) In other fields, intrinsic dimensionality is applied to reduce the size of a search space and a factor of 2 is an acceptable error. When using the intrinsic dimensionality to actually identify the state variables, it is desirable to have a much more accurate estimate.

(b) In natural images, the feature of interest is usually local in the image: A human or animal face, an animal's profile, or a tree all have a place in the image where they can be found with high probability. With microstructures, the features of interest are nonlocal. Correlations, size distributions, and orientations have all been used in an attempt to circumvent this problem. This is normally treated with homogenization theory, where an estimator can be made arbitrarily precise by requiring the system size to be above a threshold value.

(c) In materials science, decisions are often made in three domains: (i) processing, (ii) microstructure, and (iii) property. These are linked to one another and used for control, inspection, and production, respectively. More to the point, they are all sparse, in that their ambient dimensionality is much higher than their intrinsic dimensionality.

This paper addresses these by (i) testing the methodology on a set of phantom image sets whose dimensionality is known by construction (we use the MLE method of Levina and Bickel [19] using different Minkowski p -norms), (ii) applying the methods to real (random) microstructures to show evidence of a homogenization limit, and (iii) doing some exploratory work in a generator domain, which maps onto an image domain. Each domain is sparse, but could be linked through a dense latent domain. With these approaches, it is found that there is a potential overestimate by a factor of 2 in the dimensionality in low-bit-depth images, particularly prevalent in materials sciences. A real SiC/SiC reinforced fiber composite matrix is used as an example, showing a homogenization in the dimensionality estimate. It is possible to infer some geometric features of generator domains for a synthetic so-called Swiss roll generator of images as well as for phase field data sets of sequences of images of grain evolution.

The layout of this paper is as follows. Section II describes the basic mathematical details of our approach. Section III

describes the maximum-likelihood-estimation (MLE) method used for dimensionality estimation in this paper. Section IV describes our methodology in this work. Section V presents the results of our investigation, followed by a more detailed discussion in Sec. VI, including the additional mathematical treatment necessary to obtain consistent estimations. The conclusions of the paper are summarized in Sec. VII.

II. MICROSTRUCTURES, MANIFOLDS, AND DISTANCES

In the above, a qualitative description of the concept of microstructures, manifolds, and their dimensionality was given. This section makes these concepts more quantitative. The main concepts to be developed here are (a) a more precise definition of microstructure and how this will be used in this paper, (b) manifolds, and (c) distance measures.

A. Microstructures as random variables

The concept of microstructure in material sciences is familiar and tends to go without definition. A practitioner can examine a series of micrographs and feel some confidence that they understand what the microstructure is. With the development of machine learning, it is important to define the microstructure in a more operational way. We start by asserting that a microstructure is a latent state of a materials system, from which only examples of images may be observed. This is analogous to the concept in probability of a random variable, which can have outcomes of observations that form a sample set on which the analysis is performed [23]. Many statements about the (latent) random variable may be proven, but the only thing that can actually be observed is the outcomes. In this analogy, the representation of the microstructure is analogous to the random variable and the observations made in a microscope with the outcomes of the experiment. This distinction allows us to directly utilize the concepts developed in sampling theory [24].

At this point, it is noted that practitioners use the word microstructure to represent two things: (i) the abstract description of the structural state of the material and (ii) the observed image of a specific area of the material. Practitioners who seek to change material properties tend to use microstructure to describe the first of these and microscopists the second. In everything that follows in this paper, when we use the word, we will italicize it if we mean the first of these and use roman font if we mean the second. So equiaxed grain structure is an example of a *microstructure*, while a picture of an equiaxed grain structure would be an example of a microstructure. For the mathematical description in this paper, we will use \mathcal{M} to describe the former and \mathbf{m} to represent the latter. Since our approach is data driven, this will require that we work on samples of microstructure. We will represent a set of such samples with a capital \mathbf{M} .

A final note on the distinction between *microstructure* and microstructure is available in machine learning, specifically, the variational autoencoder [25], where the latent space itself forms a representation of the *microstructure* and the images generated (or used for training) are examples of microstructures.

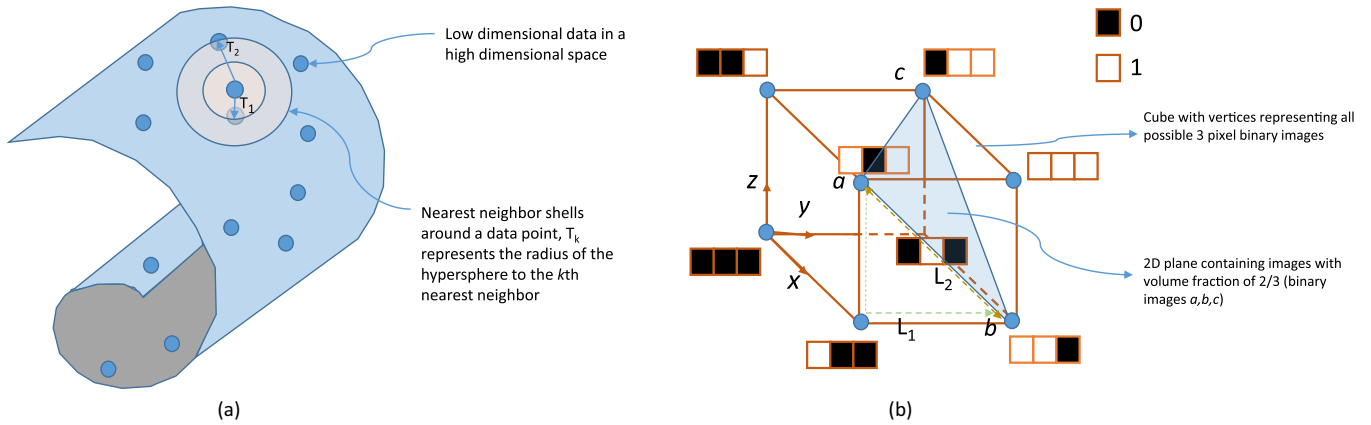


FIG. 1. (a) Swiss roll manifold containing image data represented as points. Nearest-neighbor shells around a data point are illustrated which can be used to estimate the intrinsic dimensionality. (b) Manifold representation of binary images with n pixels, which exist on vertices of a cube of dimension n . The space of three pixel images is shown with a 2D domain representing images (marked a , b , and c) with pixel values that sum to 2.

B. Manifolds

The concept of lower-dimensional manifolds is familiar to the machine learning community. Its familiarity in the materials community is inconsistent. This section gives a brief discussion of what a manifold is and why it is important.

With microstructure observations as images, each pixel can be viewed as a separate dimension in some very large dimensional space. As sensors improve, the pixel density increases, expanding this space further. Beyond the resolution of the beam and lens system of the machine though, this added dimensionality does not present any additional information. More importantly, elements in a microstructure image are generally correlated, so knowledge of one element can be used to infer things about another. This dependence between pixels reduces the size of the space needed to represent the microstructure image. So, for example, a 1280×720 photograph formally contains 921 600 dimensions. However, actual *microstructures*, we assert, require much less than this, shown for natural images in [20].

This is a well-known property of data and is known as the manifold hypothesis: The actual data lie on a lower-dimensional manifold that is embedded in the ambient space. Though not known by that word, most people are familiar with the manifold concept. We know that the surface of the earth is spherical. However, it looks like a two-dimensional plane. We view the surface of the earth as a manifold that has the property that, so long as the distance between two observations is not too great, the surface may be represented by a two-dimensional Euclidean space \mathbb{R}^2 , that is, the surface of the earth is a two-dimensional (2D) Euclidean space embedded in the three-dimensional ambient space in which we live. Figures 1(a) and 1(b) show other examples of manifolds in \mathbb{R}^3 . Figure 1(a) is known as the Swiss roll manifold. Essentially, this is a plane that contains all of the data, but has been “rolled up” into a spiral so that it exists in \mathbb{R}^3 , but the points themselves only occupy \mathbb{R}^2 . Figure 1(b) shows a manifold embedded in \mathbb{R}^3 that would be referred to in crystallography as a $(11\bar{1})$ plane, i.e., one that intercepts coordinate axes at $[1,0,0]$, $[0,1,0]$, and $[0, 0, -1]$. This can, for example,

represent a three-pixel image, where the intercepts of the axes correspond to images consisting of one black pixel and two white ones. The shaded area of the plane represents interpolations between these points.

In the same sense, the manifold hypothesis asserts that data lie in a lower-dimensional space that appears to be Euclidean for small displacements. The aim of this paper is to estimate the dimension of the manifolds that represent specific *microstructures* in material systems.

C. Minkowski distance measures

The discussion above of manifolds implicitly referenced the concept of distance by asserting that, so long as distances between two observations are not too great, etc. This work builds on the work by Levina and Bickel [19], which uses the Euclidean distance for dimensionality estimation. We generalize this approach by using the Minkowski distance measures. We had expected all the Minkowski distance measures to give the same answer for the intrinsic dimensionality. Indeed, for nonquantized images, we find this to be correct, but significant errors appear when too few gray levels are used to make the dimensionality estimate. This is explored and discussed extensively in Sec. VI. The concept of Minkowski distance measures is introduced here.

Formally, if distance between vectors x and y is denoted by $d(x, y)$, then the distance has the following properties: (1) $d(x, y) > 0$, (2) $d(x, y) = 0$ if and only if $x = y$, (3) $d(x, y) = d(y, x)$, and (4) $d(x, y) \leq d(x, z) + d(y, z)$ for any z . The Euclidean distance ($p = 2$) obeys all of these properties. An extension of this is the Minkowski distance

$$d(x, y) \triangleq \left(\sum_i |x_i - y_i|^p \right)^{1/p}, \tag{1}$$

where p is the Minkowski distance parameter. In this work, we extend the estimation to Minkowski distances.

Particular cases of the Minkowski distance family are the Manhattan distance ($p = 1$), or the L_1 -norm, and Euclidean distance ($p = 2$), or the L_2 -norm. Only metric p -norms are

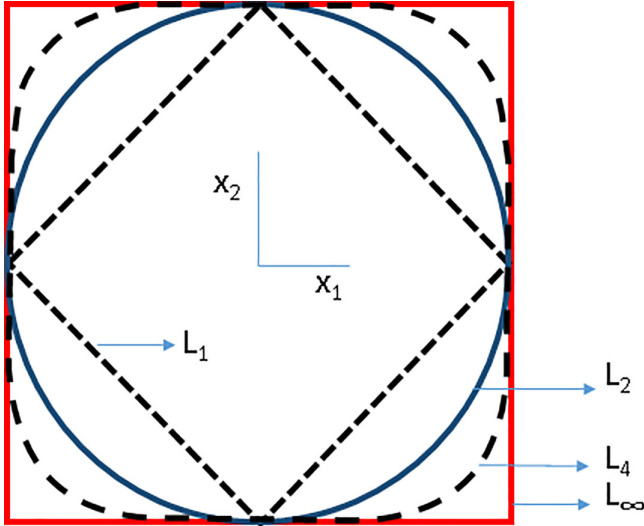


FIG. 2. Minkowski circle geometric representation of a 2D circle for $p = 1, 2, 4,$ and ∞ .

considered here ($p \geq 1$) since triangle inequality (property 4 above) is violated for $p < 1$. A geometric representation of a 2D circle for $p = 1, 2, 4,$ and ∞ is shown in Fig. 2(a), where the surface describes all points equidistant from the origin under the respective p -norm.

III. DIMENSIONALITY ESTIMATION

There are a number of methods in use for estimating the dimensionality of data, including PCA, isometric feature mapping, local linear embedding, multidimensional scaling, and correlation-based methods [26]. Among the most popular is the maximum-likelihood-estimation approach [19], which is the one used here. In this section we describe the approach. The maximum-likelihood-estimation approach is built on the nearest-neighbor (NN) method, published by Pettis *et al.* [27], which is a geometric estimator of the intrinsic dimensionality of the manifold on which the data lie. The assumptions behind this approach are that (i) the samples are independent and identically distributed from some distribution, (ii) in a space of proper dimension the samples will be uniformly distributed locally, (iii) the mapping between the latent space and the ambient space is continuous, and (iv) the distance between two points in the ambient space is the same as that in the latent space locally. The importance of the data being uniformly distributed is that there would be no bias towards one region over another in the space. In practice, it is not always possible to populate an entire space this way, so a local uniformity assumption is used, that the points are uniformly spread in some neighborhood of each point. Representative samples may be drawn randomly in this case. The unique point process that will ensure such a uniform distribution is the Poisson process [28]. The intuitive meaning of continuous is that neighboring points in the latent space correspond to neighboring points in the ambient space, a notion more rigorously defined in topology [29].

There is one subtle complication that arises because data are generally not on a linear manifold but may be on one that

is curved and twisted. The distance between two points on a curved manifold would be measured as its geodesic distance, whereas in the ambient space it would be measured as a Euclidean distance or similar. Since differentiable manifolds are approximately Euclidean for small distances, this amounts to a requirement that the distance between points be made small, and thus the fourth assumption is also made locally. With these assumptions, the dimensionality estimate of a data set may be made, knowing only a distance between the points.

A. Nearest-neighbor method

The NN method aims to estimate the intrinsic dimensionality using the number of nearest neighbors of each data point [19]. The data are modeled as having been produced by a set of independent and identically distributed samples from a uniform probability density in some low-dimensional latent space. Here we use the denotation that μ is the true dimensionality, n is the dimension of the ambient space, and k is the number of neighbors of each data point (when measured by a Minkowski distance, for example). Note that, since the intrinsic dimension is lower than the ambient dimension, $\mu \leq n$. The latent space is in \mathbb{R}^μ .

Let the set $\mathbf{M} \triangleq \{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_s\} \subset \mathbb{R}^n$ be the microstructure samples. Under these assumptions, the average number of data points \bar{k} that fall into a hypersphere in \mathbb{R}^μ around a point \mathbf{m}_i will be proportional to the volume of the hypersphere

$$\bar{k}_i = f(\mathbf{m}_i) \mathcal{V}_2(\mu), \quad (2)$$

where the proportionality constant $f(\mathbf{m}_i)$ defines the uniform probability density defining number of points per unit volume about point i in \mathbb{R}^μ . Here $\mathcal{V}_2(\mu)$ refers to the volume of the hypersphere of dimensionality μ that has an expected number \bar{k} of nearest neighbors with distances represented using an L_2 -norm. The volume of the hypersphere is given by the particular choice of the distance measure. The volume is given by the formula

$$\mathcal{V}_2(\mu) = V_2(\mu) [T_k]^\mu, \quad (3)$$

where $V_2(\mu)$ is the volume of a hypersphere of unit radius in \mathbb{R}^μ and T_k is the distance from a fixed point \mathbf{m}_i to its k th nearest neighbor in the ambient space. By the locally isometric assumption, this is the same as the distance that would be measured in the latent space. Equation (3) is presented in the Euclidean norm. Later, we show that this relation holds for all p -norms, with only $V_2(\mu)$ being a function of p .

Using Eqs. (2) and (3), the expected number of points within a distance T_k from a point \mathbf{m}_i can be written as

$$\bar{k}_i = c_{2,i} [T_k]^\mu, \quad (4)$$

where $c_{2,i} = f(\mathbf{m}_i) V_2(\mu)$ is a constant, with the subscript 2 indicating the use of Euclidean distance. Equation (4) can be solved for μ as a function of the average number of points within T_k of point i . This could be used to estimate μ , but this requires us to fix a radius about point i and count the number of data points lying within that range. When the data are sparse, this can easily result in erroneous estimates because of the sparsity of the data. In practice, it makes more sense to fix the number of points about point i on which this estimate is

made, that is, we are changing the independent variable from T_k to k , so now T_k is dependent on point i ; T_{ki} is the distance of the k th nearest neighbor of point i .

The relationship in Eq. (4) can be used to estimate the dimension by linear regression of $\ln T_k$ on $\ln k$ over a suitable range of k (e.g., from $k = k_{a_i}$ to $k = k_{b_i}$, where k_{a_i} and k_{b_i} are the a th and b th nearest neighbors of point i) for every point i . The intrinsic dimension is obtained as the slope

$$\mu_i = \frac{\ln k_{b_i} - \ln k_{a_i}}{\ln T_{k_{b_i}} - \ln T_{k_{a_i}}} = \ln \left(\frac{k_{b_i}}{k_{a_i}} \right) \left(\ln \frac{T_{k_{b_i}}}{T_{k_{a_i}}} \right)^{-1}. \quad (5)$$

B. Maximum-likelihood-estimation method

Equation (5) can be used to estimate the dimensionality of the data, but makes that estimate, pair by pair, of the data. The MLE method uses multiple points to make a single, more precise estimate. The data are modeled as having come from a Poisson process and the parameters of this process, μ and λ , are estimated from a maximum-likelihood algorithm [30].

In the nearest-neighbor method, the data are modeled as having been produced by a set of independent and identically distributed samples from a uniform probability density in some low-dimensional latent space. This is true if and only if the data are produced by a Poisson process [24], that is, the points are placed with the μ -dimensional Poisson process

$$P(k - 1) = \frac{\bar{k}^{k-1}}{\Gamma(k)} e^{-\bar{k}}, \quad (6)$$

where \bar{k} is the expected value of the Poisson process.

From the preceding section, \bar{k}_i can be written as

$$\bar{k}_i = f(\mathbf{m}_i) V_2(\mu) = f(\mathbf{m}_i) V_2(\mu) [T_k]^\mu. \quad (7)$$

By definition, the rate of the Poisson process is

$$\lambda_i = \frac{d\bar{k}}{dV_2(\mu)} = f(\mathbf{m}_i). \quad (8)$$

Levina and Bickel’s approach uses a one-dimensional equivalent of the Poisson rate with respect to the radius of the hypervolume. Changing variables to take the derivative with respect to $T_k(\mu)$, we obtain

$$\lambda_i^*(T_k) = \frac{d\bar{k}_i}{dT_k} = \frac{d\bar{k}_i}{dV_2(\mu)} \frac{dV_2(\mu)}{dT_k}. \quad (9)$$

Substituting Eq. (3), we obtain

$$\frac{dV_2(\mu)}{dT_k} = V_2(\mu) \mu T_k^{\mu-1}, \quad (10)$$

$$\lambda_i^*(T_k) = f(\mathbf{m}_i) V_2(\mu) \mu T_k^{\mu-1}, \quad (11)$$

where $\lambda^*(T_k)$ corresponds to Levina and Bickel’s $\lambda(t)$ in Eq. (3) [19], which represents a one-dimensional equivalent of the Poisson rate. The log-likelihood of the 1D Poisson process can be written as [24]

$$L(\mu_R(\mathbf{m}_i), \theta_i) = \int_0^R \ln(\lambda^*) dN(r) - \int_0^R \lambda^* dr, \quad (12)$$

where $\mu_R(\mathbf{m}_i)$ is the estimate of the dimensionality estimate about point \mathbf{m}_i , using points within a radius R from that point, $\theta_i = \ln f(\mathbf{m}_i)$, and $N(r)$ is the number of points within

a distance r from point \mathbf{m}_i . Specifically, L is parametrized by μ , the dimensionality we wish to recover, and $f(\mathbf{m}_i)$, the unknown probability density in the neighborhood of point \mathbf{m}_i , which is a nuisance parameter that we may need to evaluate to do the problem but otherwise do not need.

Maximizing the likelihood with respect to μ and θ about each data point gives the optimal values of these parameters. A detailed derivation of this is given in Appendix C and yields, for μ ,

$$\mu_{k_a}(\mathbf{m}_i) = (k_a - 1) \left[\sum_{j=1}^{k_a-1} \ln \left(\frac{T_{k_a}}{T_j} \right) \right]^{-1}, \quad (13)$$

where k_a is a choice of radius in terms of nearest neighbors made by the user. Comparing the ID estimators, i.e., Eqs. (5) and (13), both formulas use the inverse logarithm of the ratio of distances.

C. Minkowski generalized distance estimates

The results of the preceding section are specialized for the Euclidean distance between data points. There is nothing in the analysis that requires this, only that the distance be measured with a distance metric. Since we do not have ground truth values for empirical data, we wish to have multiple ways of evaluating the dimensionality and use the consistency between those as a measure of our confidence of the answer. For this purpose, we use the Minkowski p -norms, each of which can provide a separate estimate of the dimensionality.

The preceding section reiterated the development of the MLE estimate for the Euclidean norm. Inspection of all of the equations indicates that the only place where dimensionality enters the problem is the constant V_2 , which multiplies T_k^μ to give the volume of a ball of radius T_k about a point [see Eq. (12)]. This is a simple constant, representing a hypersphere of unit radius, measured with the Euclidean norm. So the change to Minkowski norms amounts to changing this constant only.

To extend the notation to include the parameter p , we use either subscripts, such as V_p as an extension of V_2 for the Euclidean norm, or parentheses, such as $T_k(p)$ as an extension of T_k in the preceding section. Equation (3) shows that, for the Euclidean norm, the volume of a hypersphere of radius T_k varies as T_k^μ , with a proportionality constant we reported as V_2 . This property is general: The volume of a hypersphere with any of the p -norms is also a homogeneous function of degree μ , but the proportionality constant changes and is dependent on p . In this section we show that this is the only consideration in the equations for making p -norm estimates of the dimensionality.

For the Euclidean distance, the volume of a unit hypersphere is $\frac{\pi^{\mu/2}}{\Gamma(\mu/2+1)}$ [19], where $\Gamma(x)$ is Euler’s Γ function equal to $(x - 1)!$ if x is an integer and a smooth interpolation between integer points if it is not. The general formula for the volume of a unit hypersphere with the p -norm is [31]

$$V_p(\mu) = \frac{2^\mu [\Gamma(1/p + 1)]^\mu}{\Gamma(\mu/p + 1)} \quad (14)$$

and, by the assertion above, the volume of a hypersphere in \mathbb{R}^μ is

$$V_p(\mu) = V_p(\mu)[T_k(p)]^\mu. \quad (15)$$

Similar to Eq. (4), the expected number of points within a distance T_k from a point \mathbf{m}_i can be generalized based on a p -norm distance as

$$\bar{k}_i = c_{p,i}[T_k(p)]^\mu, \quad (16)$$

where $c_{p,i} = f(\mathbf{m}_i)V_p(\mu)$ is a constant with the subscript p now indicating the use of the Minkowski p -norm distance. The equations of dimensionality estimation (5) and (13) remain unchanged except T_k now represents the distance measured using the Minkowski p -norm. The MLE estimate is given by (see Appendix C for the derivation)

$$\mu_{k_a}(\mathbf{m}_i, p) = (k_a - 1) \left[\sum_{j=1}^{k_a-1} \ln \left(\frac{T_{k_a}(p)}{T_j(p)} \right) \right]^{-1}. \quad (17)$$

Note that this is a direct consequence of the assumption that the data could be modeled as having originated from a Poisson point process. Accordingly, the data points will be distributed evenly in space [24], so the expected number of points occurring within a region will be proportional to the hypervolume of that region. In this case, this is the hypervolume within our μ -dimensional space in which the data may be obtained by uniform sampling.

More specifically, upon examination, we should find that the expected number of data points occurring within a thin shell of distance r_p of our target image scales as the volume within that shell, from the properties of a Poisson point process. For an accurate estimate of $\hat{\mu}$, that uniform density assumption would hold and the expected number of points would scale as $r_p^{\hat{\mu}}$ [see Eq. (14)]. On the other hand, for an underestimate of ($\hat{\mu} < \mu$) the expected number of points in the model would be an underestimate of that observed in the data and the density of empirical data points would have to increase with r_p in order for all to be accounted for. The data would have a tendency to accumulate on the surface of a hypersphere of radius of r_p in $\mathbb{R}^{\hat{\mu}}$. Were an overestimate of $\hat{\mu}$ obtained, the data would appear to concentrate near the origin. It is only at the ‘‘correct’’ value of $\hat{\mu}$ that the data would uniformly spread in the space.

IV. METHODS

The MLE estimates of dimensionality were made on several data sets. The method was proved out with phantom data sets, that is, synthetic data sets for which we knew the ground truth by construction. It was then applied to two materials data sets: one from a phase field simulation and the other from fibers segmented from a continuously reinforced fiber composite. The details of this are explained below.

A. MLE

Dimensionality estimation was performed with Eq. (17) for a range of p values on each of the data sets. It was expected that the dimensionality estimates would be consistent over p . This proved to be incorrect, so exploratory investigations

were undertaken to produce a reliable best practice approach to providing the best estimates of dimensionality.

In Eq. (17), every data point $\mathbf{m} \in \mathbf{M}$ gives a dimension estimate for every neighbor count k , yielding $k \times |D|$ dimension estimates, where $|D|$ is the number of images in the data set. The intrinsic dimension was estimated as the mean of the values in the dimension estimate matrix and the full histogram typically reported for inspection.

Equation (17) has two free parameters, i.e., the minimum and maximum k values, to use for the estimate. Too small a range may lead to underestimation of the dimensionality due to counting statistics or local variations, while too large a k range may lead to overestimation due to global curvature and jumps across layers in a manifold (consider, e.g., a Swiss roll). In practice, it was possible to find a range of k values in which the dimensionality estimation was fairly insensitive to the exact range and the estimates made in these range constituted the high peak in the histogram of estimates. It should be appreciated that the range of k also depends implicitly on the size of the data set, since large- k values are generally inappropriate in small data sets due to severe sparsity of points. The actual ranges of k values chosen for each data set are listed in Table II.

One further complication was encountered. Because of the quantization in the center position of the features in some of the data sets, it often happened that duplicate distances occurred during sampling. If these positions were sampled from \mathbb{R}^2 , instead of on a quantized grid in \mathbb{Z}^2 , almost always the distance between sample points would be unique. However, forming a pixel grid made it a common occurrence that some near neighbors had the exact same distance and Eq. (17) would have a 0 in the denominator. We adopted one of two strategies in this case: (i) simply set the minimum k value high enough that this problem was avoided and (ii) construction of the phantom image at ten times the resolution and rescaling with a bicubic interpolation. In practice, each approach worked equally well, so we did not attempt to document exactly which was used.

B. Data sets

The method was proved with data sets that were constructed having a known ground truth dimensionality and then applied to materials data for which the ground truth was not known but for which some sanity checks could be made to show that our results were reasonable. The ground truth data sets were constructed from synthetic images containing a single particle, for which the ground truth dimensionality could be listed from the independent variables pertaining to locations of particles and their size and shape. Methods were developed for reliably reproducing the ground truth dimensionality on these data sets before the more complex materials data sets were attempted.

Materials data sets were used in which there were more than single particles. We examined the dimensionality of an evolving grain system, using phase field simulations, and of a real SiC/SiC continuously reinforced composite. A dimensionality of 1 was expected from the phase field simulations because each image was closest to the neighbors on either side in a temporal sequence.

The composite structure was not as straightforward. For these data, we expected that the number of dimensions would not be able to exceed (the number of particles) \times (the dimensionality of the space where each particle center could be placed) \times (the number of geometric degrees of freedom of the particles). On further reflection, the actual dimensionality was expected to be lower than this number because of the correlations induced by the requirement that they not overlap and larger than this because of the unknown true number of shape and orientation parameters on the particles. So we estimated the upper bound dimensionality by constructing a surrogate data set with the same numbers of fiber cross sections but of constant orientation that were placed according to a Poisson point process [24]. This was expected to produce larger than the true dimensionality, but not strictly an upper bound because not all geometric degrees of freedom were used.

1. Ground truth: Single-particle tests

Phantom data sets (DSs) were created for simplified, single-particle images, for which a ground truth dimensionality was known. The DS Circles consisted of circles of constant radius that were placed at random horizontal (x) and vertical (y) positions in the image. Here random means sampled from a uniform random variate. Since the radius was fixed, this data set had two degrees of freedom, by construction. The estimate of the dimensionality of this data set should be 2. In DS Rectangles, we produced, for example, images of a single square in the matrix. In this case, there were three degrees of freedom: x , y , and the width a . We also relaxed the square constraint and allowed the aspect ratio to vary, giving a rectangle. This gave us a known dimensionality of 4. Other tests were conducted with this data set as well.

The DS Rectangles and DS Circles were sampled directly from Euclidean manifolds of known dimensionality. Real data points are generally not acquired from a latent manifold but from a low-dimensional manifold embedded in a higher-dimensional space. For example, strain components during processing would theoretically occupy a six-dimensional space, since the strain matrix has six independent components. However, a real process is not expected to be able to sample every point in some subset of \mathbb{R}^6 . More likely, some manifold within that space would be all that could be practically sampled. For this reason, we produced a data set DS SwissRoll, with a latent space with the ubiquitous Swiss roll data set [32]. Here we sampled from a plane rolled up into a spiral.

These data sets are described in greater detail in the following section. In all of these, we define the following spaces: \mathcal{G} , the generator space for sampling; \mathcal{L} , the dense latent space in which the data points are assumed to be uniformly distributed; and \mathcal{A} , the ambient space of image pixels \mathbf{M} set of samples used for dimensionality estimation. The following data sets are used for single-particle tests.

(i) *DS Circles : randomly centered circles of a constant radius in a matrix.* Here we used $\mathcal{G} = \mathcal{L} \in \mathbb{R}^2$ to generate points that were embedded in \mathcal{A} with an embedding function that mapped these points into 256×256 images with circles of a fixed radius, centered at the x and y coordinates of each point in \mathbf{M} . We constructed four data sets with $r \in$

TABLE I. Test cases used in Fig. 4. Here η represents a random quantity and c represents a fixed one. Subscripts on η mean that the random variables are independent. Those on fixed quantities mean they may have different values but are still fixed. Column 6 (Dimensionality) is simply a counting of the number of independent η values in each case. In the header, x and y refer to the (x, y) centers of the objects and r_x and r_y are the dimensions of the objects in the x and y directions, respectively

Case	x	y	r_x	r_y	Dimensionality
Rectangles					
A	η_1	η_2	η_3	η_4	4
B	η_1	η_2	c_1	c_2	2
C	η_1	c	η_2	η_3	3
D	η	c_1	c_2	c_3	1
E	c_1	c_2	η_1	η_2	2
Squares					
F	η_1	η_2	η_3	η_3	3
G	η_1	η_2	c	c	2
H	η_1	c	η_2	η_2	2
I	η	c_1	c_2	c_2	1
J	c_1	c_2	η	η	1

{24, 36, 48, 58} pixels. These are referred to as DS Circles A, DS Circles B, DS Circles C, and DS Circles D, respectively. In order to avoid contact with the boundaries of the images by the circles, \mathbf{M} was sampled from the set

$$\mathbf{M} \in \{(\eta_1, \eta_2) \in \mathcal{L} | r < \eta_1 < 256 - r, r < \eta_2 < 256 - r\}. \quad (18)$$

(ii) *DS Rectangles: rectangles and squares in a matrix.* Ten different synthetic data sets were tested containing rectangular and square shapes in a matrix following different size and positional constraints which dictate the intrinsic dimensionality. The ground truth dimensionality could be calculated from the number of independent variables, which are shown in Table I.

Images of size 128×128 were constructed from each point in a Euclidean space \mathbb{R}^μ , where μ was taken from column 6 of Table I. More formally, we have a generator space \mathcal{G} in \mathbb{R}^μ , from which we sampled points and embedded them in an ambient space \mathcal{A} , which in this case was $\mathbb{R}^{128 \times 128}$. We constructed a set of samples

$$\mathbf{M} = \{(\eta_1, \eta_2, \dots, \eta_\mu)_1, (\eta_1, \eta_2, \dots, \eta_\mu)_2, (\eta_1, \eta_2, \dots, \eta_\mu)_3, \dots, (\eta_1, \eta_2, \dots, \eta_\mu)_n\} \in \mathcal{G}. \quad (19)$$

Like DS Circles, \mathcal{G} was taken to be the same as \mathcal{L} .

The samples drawn from \mathcal{G} were then embedded in the ambient space with

$$f : \mathcal{G} \rightarrow \mathcal{A}, \quad (20)$$

where f maps each point of \mathbf{M} to a 128×128 image. These images have pixel values of 1 for a rectangle (as described in Table I) and 0 for the background. Examples of these images are shown in Table I.

By construction, the ground truth dimensionality of these image data sets is given in column 6 (Dimensionality) in

Table I. The task of this work is to recover this as μ from the set of images.

(iii) *DS SwissRoll: a Swiss roll.* A data set was prepared, in which the generator space \mathcal{G} was a larger space in which the latent space was embedded. This is the case we would anticipate, where processing determines a structure but the processing variables would not form a dense space. For this purpose we defined a latent space

$$\mathcal{L} = \{(\eta_1, \eta_2) \in \mathbb{R}^2 | \eta_1, \eta_2 \in [0, 1]\}. \quad (21)$$

This was deformed into a Swiss roll, where the spiral was viewed edge on in the x - z plane and the axis along the y axis, as shown in Fig. 5(a). The 2D space was expanded into a space of size $\frac{3\pi}{2} \times 30$ with the transformation

$$t \triangleq \frac{3\pi}{2}(1 + 2\eta_1), \quad (22)$$

where t is physically the distance traveled from a reference point around the axis of the Swiss roll. This was mapped onto the x - z plane with the transformations

$$\begin{aligned} x &= t \cos t + c_1, \\ z &= t \sin t + c_3, \end{aligned} \quad (23)$$

where c_1 and c_3 were offset values placing the center of the Swiss roll in the x - z plane. In addition, η_2 was stretched and translated according to

$$y = 30\eta_2 + c_2, \quad (24)$$

where c_2 was an offset of the zero point of η_2 to the zero point of y . In this, the parameters c_1 , c_2 , and c_3 were assigned the values of 62, 50, and 20, respectively. The reason for these extra parameters is to allow for large enough objects in the images that aliasing effects are minimized.

Once embedded in this space, $(x, y, z) \in \mathbb{R}^3$, the coordinates were defined as the x center and y center of a circle of radius z . This description was embedded into $\mathcal{A} \in \mathbb{Z}^{128 \times 128}$, where each coordinate represented pixels of an image, with 0 being assigned to the background and 1 to the interior and boundary of the circle [see Fig. 5(a)]. The η_1 and η_2 were sampled from a uniform random variate and given the transformations above to construct DS SwissRoll with a ground truth dimensionality of 2.

2. Microstructure tests

We performed MLE estimates of dimensionality in two *microstructure* examples: phase field simulations of grain growth and of the fibers in a SiC/SiC fiber composite. For this purpose, we prepared two additional data sets. The DS PhaseField was produced by phase field simulations and represented a synthetic data set over which we had control through its parameters. The DS SiC/SiC was a real microstructure consisting of the commercial S200 SiC/SiC fiber composite. The data were taken from the Globus archive of Sherman *et al.* [33] of the data used in their publication [34]. This consisted of a set of segmented ellipses of the fiber cross sections and was reported as (x, y, r_1, r_2, α) values, where x and y were the centers of the fibers, r_1 and r_2 were the principle axes of the ellipses, and α was the orientation angle of the ellipses with the horizontal. Note that to go along with the real fiber

microstructure data set, a synthetic fiber microstructure data set was also created to explore the difference between real and synthetic microstructures.

These microstructure data sets are described in more detail in the following.

(i) *DS PhaseField: a phase field.* This DS contains results from a phase field simulation of grain growth based on the Allen-Cahn equation following the numerical formulation of Fan and Chen [35]. The data are in the form of binary images (128×128) containing grain boundaries at different time steps of a single simulation. Since all the model parameters are fixed at the start of the simulation and images are only a function of time, the intrinsic dimensionality of all images from a single simulation run is expected to be one. The data set contains results from three different simulation runs. Ten order parameters were used and kinetic parameters in the simulation were all taken to be one, including the relaxation coefficient, coefficients in the free-energy density, and the gradient energy coefficients, and the simulation was run for 1000 time steps with a time step of 1.

(ii) *DS SiC/SiC: a S200 SiC/SiC fiber composite.* Here we present microstructures that represent fibers in a matrix. We use both simulated and real (segmented) data sets.

(a) *DS SiC/SiC A: experimental data containing elliptical fibers.* We hypothesized that the dimensionality of a random structure was a homogenous property and that the estimates of the dimensionality would stabilize as larger and larger window sizes were used. Data were sampled from the published S200 fiber data set [33] containing 30 slices of composite tows from a 3D composite. In that study, the fiber center and radii were reported in microns. Circular windows were placed randomly in this data set and the fibers whose centers were contained in the interior of the window were selected. These were placed in a square image of size twice the window radius plus $30 \mu\text{m}$ on each side such that the fibers did not intersect the edges of the image [as shown in Fig. 7(a)]. To study the effect of window size on dimensionality estimate, window radii of 40–240 μm were chosen in increments of 20 μm . All images were resized to 128×128 pixels.

(b) *DS SiC/SiC B: synthetic circles.* Since real data do not have a ground truth, we resorted to idealized estimates of the dimensionality of this data set with a surrogate data set consisting of circles that were placed according to a Poisson point process [24]. The mean number of fibers per unit area was estimated from the real data set of Sherman *et al.* and was used as λ to generate the number of circles, according to a Poisson distribution. The circle radii were generated according to a Poisson distribution with mean 12 μm but including a maximum cutoff at $1.4 \times 12 \mu\text{m}$, mirroring the maximum cutoff observed in real data. The centers of the circles were generated according to a nonoverlapping Poisson point process on a window of changing size in microns. These $\{x, y, r\}$ values were then embedded in an ambient space $\mathcal{A} \in \mathbb{Z}^{1000 \times 1000}$, where the interior and boundaries of the circles were assigned a value of 1 and the background a value of 0. From here the images were downsampled to the final ambient space of $\mathcal{A} \in \mathbb{Z}^{128 \times 128}$, with bicubic interpolation, resulting in 32-bit float images. Because the window radius in microns changed (to match the data in data set DS SiC/SiC A) but the image size in pixels stayed the same, the pixel size increased with increasing window size. A

TABLE II. Data and simulation parameters.

Data set index	Size (unique images)	Sets	k range
Circles A	984	2 (binary, Gaussian)	10–40
Circles B	983	2 (binary, Gaussian)	10–40
Circles C	982	2 (binary, Gaussian)	10–40
Circles D	969	2 (binary, Gaussian)	10–40
Rectangles A	1998	1	10–40
Rectangles B	1598	1	10–40
Rectangles C	1869	1	10–40
Rectangles D	73	1	6–12
Rectangles E	284	1	6–12
Squares F	1960	1	10–40
Squares G	1593	1	10–40
Squares H	880	1	10–40
Squares I	65	1	6–12
Squares J	17	1	6–12
SwissRoll	840	1	10–40
PhaseField	181–216	3 (three runs)	6–12
SiC/SiC experiment	913–1000	11 (window sizes)	10–40

15- μm buffer was included for each window radius to ensure particles did not touch the edge. At each window size, 1000 images were generated for a dimensionality estimate.

estimate; we use the L_1 -norm for the distance measure for MLE dimensionality estimates. These results are presented in the next section and extensively discussed in Sec. VI.

V. RESULTS

All the data sets and simulation parameters, i.e., the sizes of the data sets (unique images) and the ranges of k values used in the MLE estimator, are listed in Table II. To give a quick preview of the results, this study produced the unexpected result that, for binary images, the MLE estimate of the intrinsic dimension varied linearly with p , the Minkowski parameter. This is shown in Fig. 3(a). For this reason, we expanded the original intended scope of the work to investigate the effects of image quantization and Minkowski parameter on the MLE

A. Single-particle studies

1. DS Circles: Circles of constant radius

Figure 3(a) shows the variation of the estimate of the intrinsic dimension with the choice of Minkowski parameter for binary image DS Circles D. In the case of the binary image, this resulted in our estimates being a function of p . The intrinsic dimension is, by construction, 2, corresponding to the (x, y) coordinates of the center of the circle. This effect was less pronounced as the number of quantization levels in the image increased.

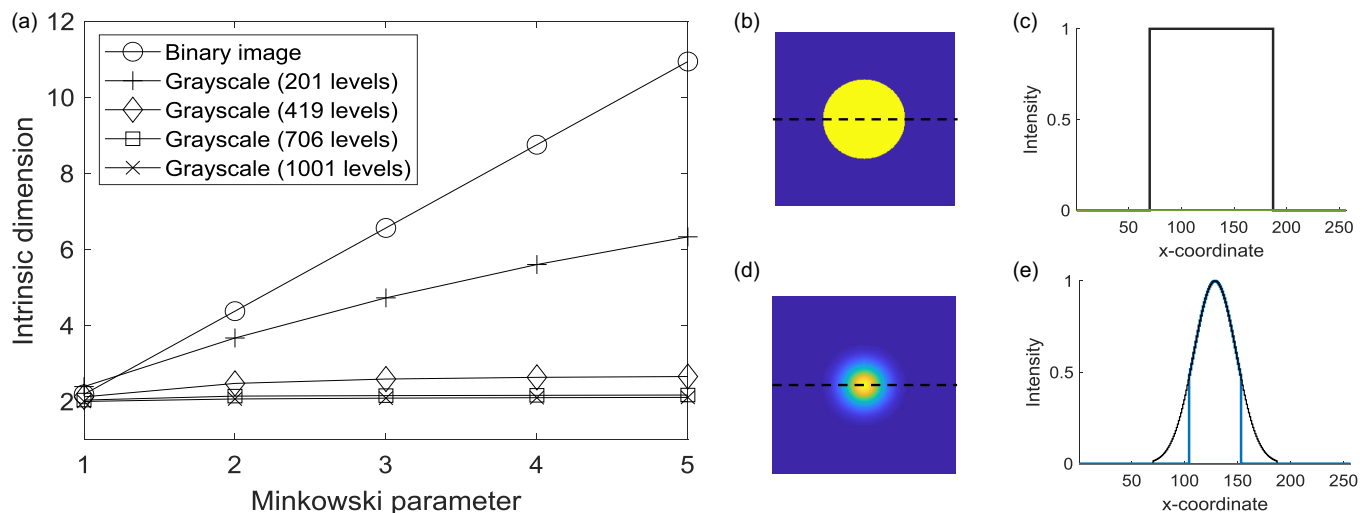


FIG. 3. Variation of intrinsic dimension with the choice of Minkowski parameter for a circle with varying position, represented as binary and grayscale images. The ground truth dimensionality is 2. (a) As the number of grayscale levels increases, the dimensionality estimates for higher- p values converge toward the ground truth value. (b) Microstructure and (c) intensity profile for binary quantization cases. (d) Microstructure and (e) intensity profile for nonbinary quantization cases. In (e) intensities outside of the vertical lines are clipped to zero to allow for different-size particles (see the text for details).

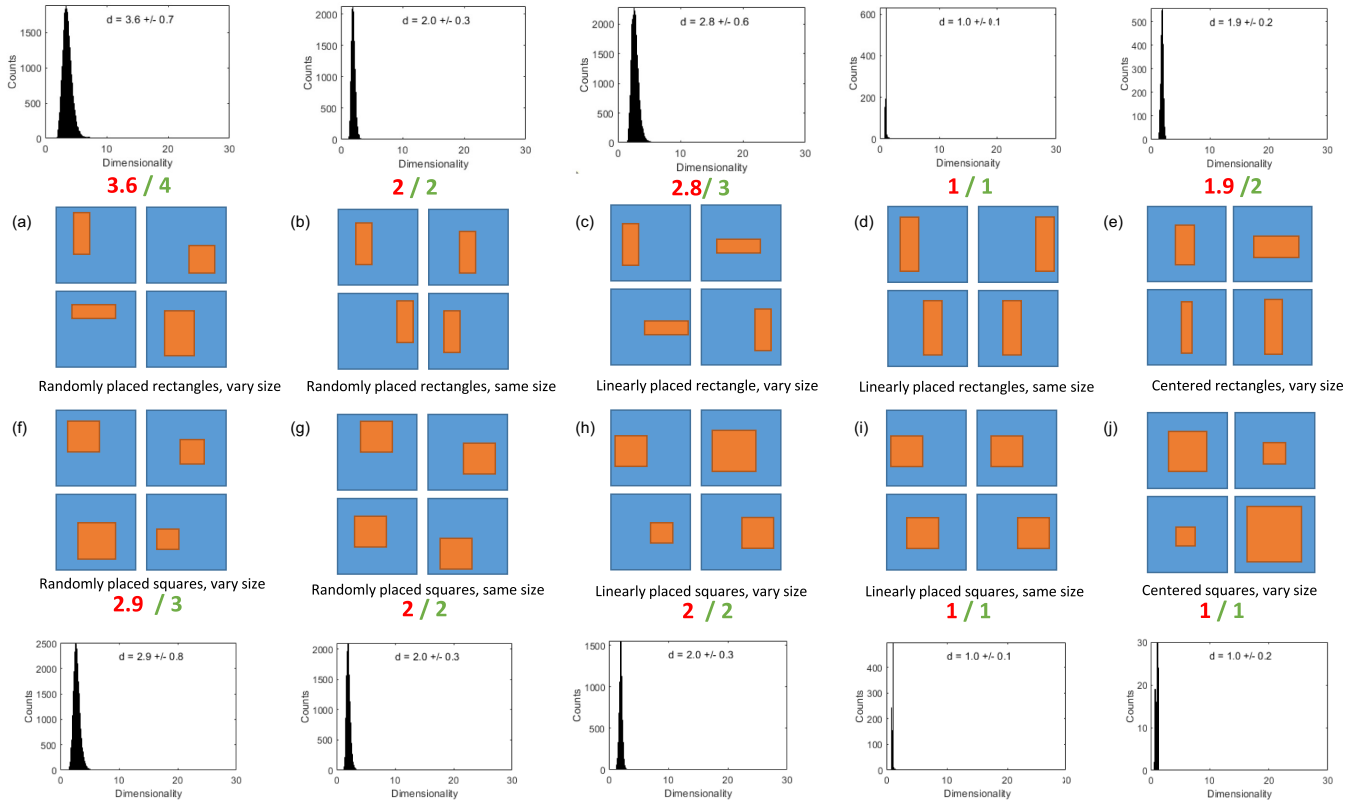


FIG. 4. Variation of MLE-based intrinsic dimension for a variety of synthetic data sets using the L_1 -norm. The predicted dimension for various data points is shown as a histogram. The numbers in red indicate mean predictions for each case. The numbers in green are the expected dimensionality. In all these cases, while the L_1 -norm provides excellent ID estimates, the use of the L_2 -norm provides exactly twice the obtained mean ID estimate, consistent with the discussion in Sec. VI A.

The images in DSs 2 A–2 D consisted of circles (unit intensity) and background (zero intensity). In order to investigate the effects of quantization, the circles were replaced with Gaussian point spread functions with a standard deviation of 20 pixels, centered at the same positions. To capture the effects of different radii, the intensities were clipped to zero intensity outside the radii of 24, 36, 48, and 58 for DSs 2 A, 2 B, 2 C, and 2 D, respectively. This gave the number of unique grayscale levels for these data sets of 201, 419, 706, and 1001, respectively.

As the number of grayscale levels increased, the intrinsic dimension obtained from the use of higher Minkowski p -norm distance measures converged toward the ground truth value. This point is explored more fully in Sec. VI.

2. DS Rectangles: Rectangles and squares

Using the results of Sec. V A 1, the dimensionality estimates of the DS for rectangles and squares were made with the L_1 -norm. Selected test cases were run, according to Table I. The results are shown in Fig. 4. In this figure, a histogram of values in the dimension estimates from each data point is also plotted and the standard deviation of the histogram is reported in addition to the mean. As seen from these results, the MLE approach with the L_1 -norm gives a sound estimate of the intrinsic dimension in all cases.

3. DS SwissRoll: Swiss roll data set

The dimensionality of the image data in DS SwissRoll is expected to be 2, given the images were sampled using data from the Swiss roll. Both MLE and NN methods furnish the correct dimensionality of 2 with the L_1 -norm as seen in Fig. 5. However, the dimensionality predicted using the L_2 -norm is 4 and subsequent norms show linear scaling of the dimensionality with the p value in the p -norm. The MLE and NN estimates of dimensionality are close at $p = 1$ and the difference gets amplified with the increase in Minkowski parameter.

4. DS PhaseField: Phase field data

Snapshots from phase field simulations [Fig. 6(a)] in time form a highly statistically dependent sequences of images. This violates one of the assumptions of the MLE approach, namely, that the data are independent and identically distributed. Rather than giving the dimensionality of the *microstructures*, it gives the dimensionality of one trajectory through \mathcal{L} of a given process. The intrinsic dimensionality is expected to be 1, the time dimension, which is confirmed by the results of the MLE algorithm with the L_1 -norm in Fig. 6(b). However, the dimensionality again shows linear scaling with the p -norm, with both the mean and the standard deviation of the dimensionality doubling when using the L_2 -norm.

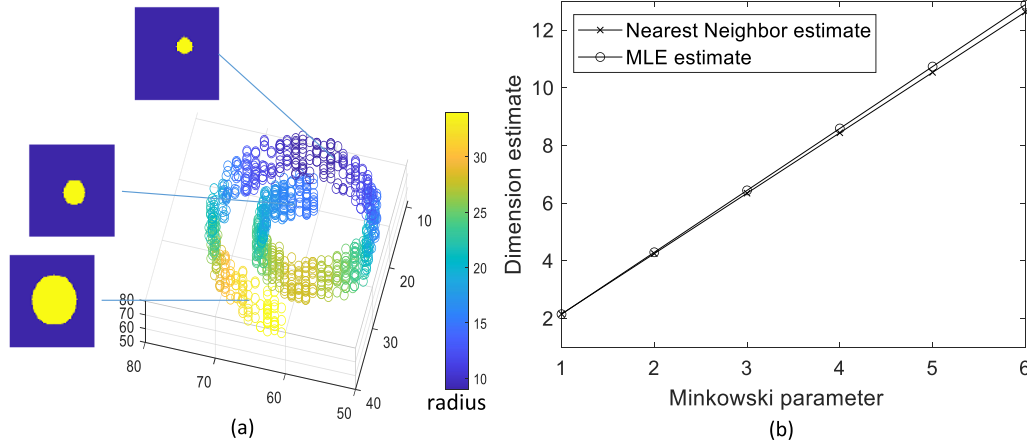


FIG. 5. (a) Swiss roll data set containing images of circles whose centers and radii are sampled from the coordinates of a Swiss roll. (b) Comparison of predicted dimensionality versus the Minkowski parameter for both MLE and NN estimates. In both cases, the linear scaling is evident. The L_1 -norm identifies the correct dimensionality of 2.

5. DS SiC/SiC: Dimensionality of a SiC/SiC composite

In this study we investigated the intrinsic dimension of an experimental microstructure containing elliptical SiC fibers as a function of window size [Fig. 7(a)]. The window size represents the length scale at which the microstructure is observed and analyzed. We extracted fibers from within randomly placed circular windows that contained all fibers whose centers fell within them. As the window size increased, the number of fibers also increased, leading to a higher intrinsic dimension as shown in Fig. 7(b). However, we found that the intrinsic dimension converged to a constant value of about 40 when the window size reached 140 μm .

VI. DISCUSSION

The significant findings of this work may be summarized as follows.

(i) While it was possible to make rational estimates of all data sets, there was an inconsistency in the results obtained

from different Minkowski norms. We attribute this to the effects of image quantization: For unquantized analog images, the estimates are consistent, but for any level of quantization, inconsistencies appear. We explain this below, including proofs of our assertions.

(ii) The dimensionality of the real microstructure images in this study homogenized, that is, it was possible to identify a representative volume element, where the dimensionality estimate became insensitive to the actual images used. So long as the images were collected within a sufficiently large window, the dimensionality estimates were consistent.

(iii) For the dual sparse domain example of points selected from a (sparse) Swiss roll embedded in \mathbb{R}^3 and used to construct images in the (sparse) image domain of $\mathbb{Z}^{128 \times 128}$, it was possible to recover the geometry of the Swiss roll domain from the images. This suggests the possibility of recovering the processing domain from a set of example microstructures. Specifically, we recovered the manifold of the original sampled data from the images, which suggests that we could

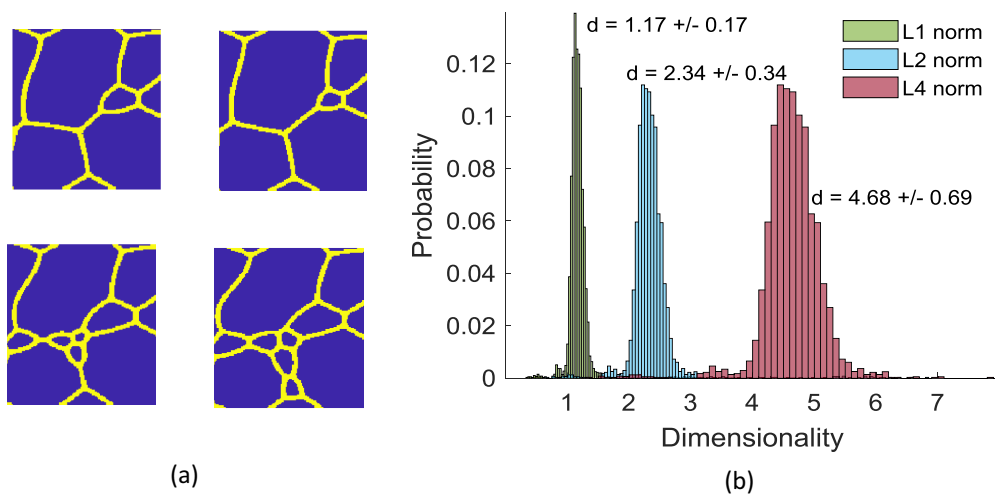


FIG. 6. (a) Images from a single phase field simulation. (b) Histograms of dimensionality from MLE estimation. The L_1 -norm computes the ground truth dimensionality of 1 (mean d is 1.17), corresponding to the variable time, and the estimated dimensionality increases proportionally to the p -norm used (the 2-norm gives $d = 2.34$ and the 4-norm gives $d = 4.68$).

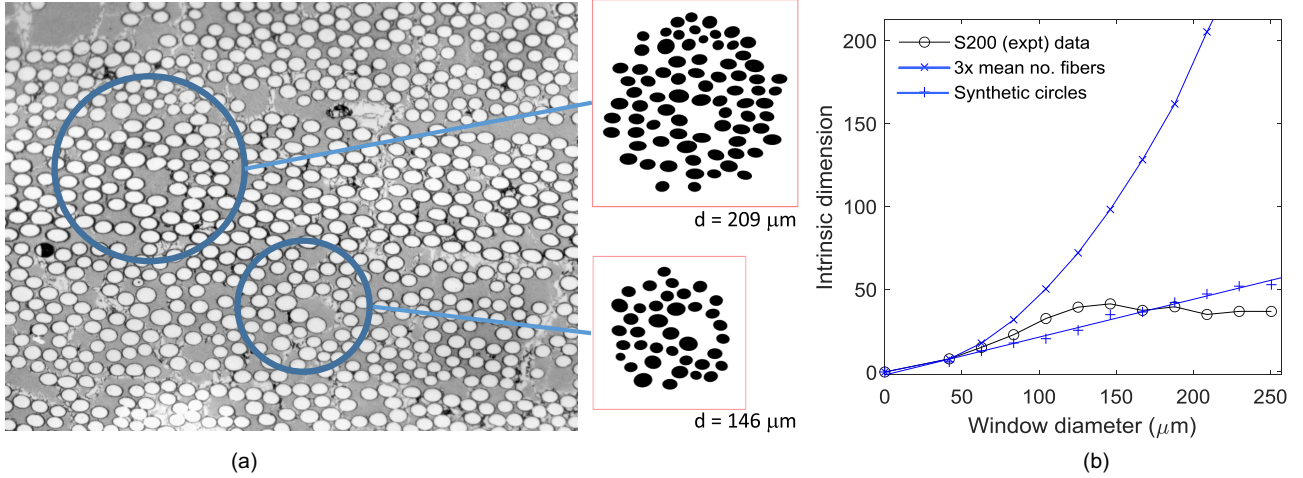


FIG. 7. (a) Intrinsic dimension of the microstructure as a function of window size compared with (b) the expected dimension based on the mean number of ellipses.

recover the processing conditions of an out-of-specification environmental event, for example, where the *microstructure* was altered by this event. These results are preliminary, of course, but they do suggest a unique application of these techniques in materials science.

These three points are discussed in greater detail in what follows.

A. Effects of quantization on dimensionality estimation

We summarize the findings from the examples as follows.

(i) Figure 3 shows that where there is minimal effect of image quantization, i.e., q is large, the dimensionality estimate is independent of the Minkowski parameter p .

(ii) In the other extreme, where the quantization $q = 2$, which leads to a binary image, the dimensionality estimate is a function of p , specifically, it is a linear function as seen in Figs. 5 and 6. The estimate with the L_2 -norm is exactly twice that when using the L_1 -norm.

(iii) Figure 3 also shows that when considering two Minkowski norms p and q ($p < q$), the dimensionality estimate $\mu_p < \mu_q$ for any quantization. However, the dimensionality estimated with the L_1 -norm appears to be independent of quantization level.

We explain these findings through four theorems.

Theorem 1. In the limit that the intensities are not quantized ($q \rightarrow \infty$), the dimensionality estimate is independent of the choice of p . By examining the development of the MLE and nearest-neighbor approaches, we show that the only effect of different Minkowski dimensions is a multiplicative constant on the distance between nearest neighbors. Since the dimensionality estimate involves ratios of distances, this completely eliminates any difference between dimensionality estimates.

Theorem 2. For binary images, the dimensionality estimate is proportional to p . By examining the Minkowski distance formula, for a binary image, intensities are some multiple of either 0 or 1. Both of these values are not affected by the operations of computing the Minkowski norm.

Theorem 3. The dimensionality estimate is a nonstrictly increasing function of p for all levels of quantization., that

is, the L_1 -norm always produces the lowest dimensionality estimate over all p -norms of quantized images. This follows directly from fundamental properties of Minkowski distances. The effect of this on estimates of dimensionality is a ratio of distances made with different Minkowski p -norms.

Theorem 4. For finite q , the dimensionality estimate assumes its lowest value for the L_1 -norm. This follows directly from the properties of Minkowski norms applied to the estimation procedures used here.

The consequence of this is that, for infinite precision in quantization ($q \rightarrow \infty$), the dimensionality estimate is consistent over all Minkowski parameters. For all other cases, the estimate using the L_1 -norm gives the smallest number of dimensions. For the most extreme case, the binary image, the dimensionality estimate scales directly with p . We have not shown that the L_1 -norm estimates were independent of the level of quantization, but the experimental results shown in Fig. 9 indicate that the distortion is less than for the other norms.

Proofs of theorems

Theorem 1. For analog data, the nearest-neighbor estimate of intrinsic dimension $\hat{\mu}(p)$ is independent of the choice of the Minkowski distance parameter p .

The key to this result is that the Poisson point density varies with a $\text{const} \times r_p^\mu$, where the constant is the only dependence on p of the estimate. This constant cancels when the dimensionality estimate is made. The proof is as follows.

Proof. For a Poisson point process, there is a stochastic number of particles in a unit region, distributed according to the Poisson distribution. In this case, the Poisson rate is the product of the hypervolume of the region and the local probability density per unit hypervolume, that is,

$$\lambda = f(\mathbf{m})V_p(\mu)r_p^\mu, \quad (25)$$

where $f(\mathbf{m})$ is the local probability density about point \mathbf{m} and $V_p(\mu)$ is the hypervolume of a unit sphere measured with the Minkowski p -norm. Their product is a constant for a particular \mathbf{m} , so we will represent this as c_p .

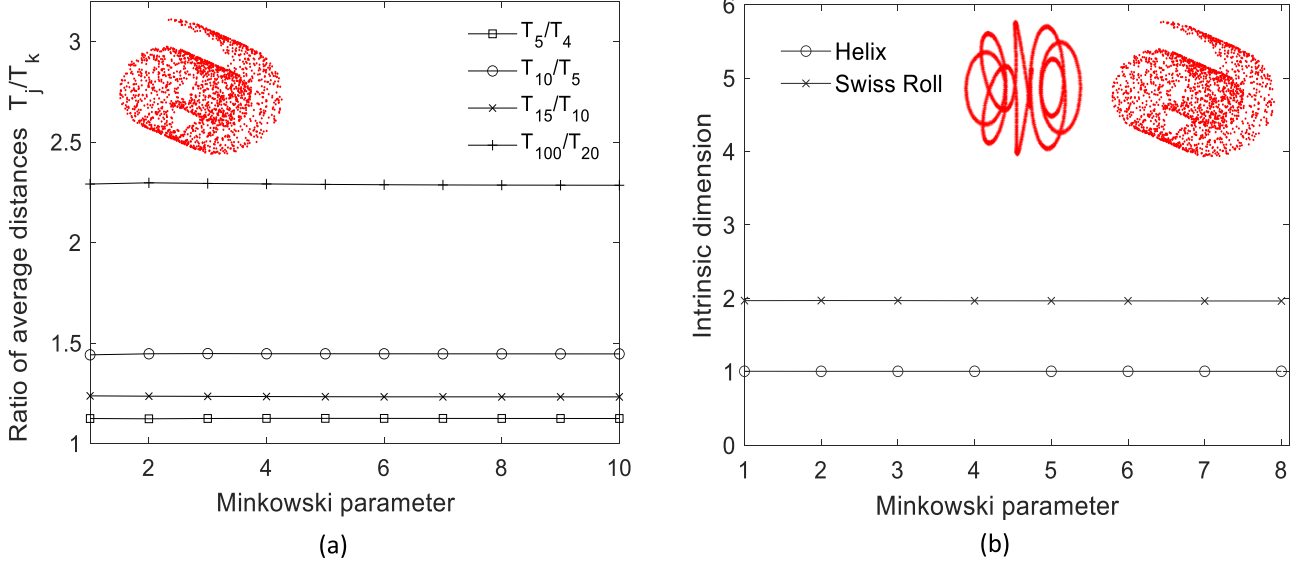


FIG. 8. (a) Estimation of the ratio of average distances to the j th and k th nearest-neighbor shells for the Swiss roll data set as a function of the Minkowski parameter. (b) MLE-based intrinsic dimension for different Minkowski parameters: helix, $\mu = 1$ and Swiss roll, $\mu = 2$. Note that for these data, which are not discrete, the intrinsic dimensionality estimate is independent of the Minkowski parameter.

Equation (25) says that the Poisson rate scales as r_p^μ , with c_p the proportionality constant,

$$\lambda = c_p r_p^\mu, \quad (26)$$

which is the Poisson rate of points within a ball of radius r_p , measured according to the Minkowski p -norm. Defining r_p to be the distance between \mathbf{m} and its k th nearest neighbor, there will be $k - 1$ points within the (open) ball of radius r_p . Thus, the probability of finding these $k - 1$ points within the ball is

$$P(k - 1) = \frac{(c_p r_p^\mu)^{k-1}}{(k - 1)!} \exp(-c_p r_p^\mu). \quad (27)$$

Equation (27) gives the (stochastic) number of points at a given radius. What is needed is a way of estimating the radius to the k th point, so this equation is inverted [27],

$$\begin{aligned} F_k(r_p) &= \left(\frac{(c_p r_p^\mu)^{k-1}}{(k - 1)!} \exp(-c_p r_p^\mu) \right) \frac{d}{dr_p} c_p r_p^\mu \\ &= \left(\frac{(c_p r_p^\mu)^{k-1}}{(k - 1)!} \exp(-c_p r_p^\mu) \right) c_p \mu r_p^{\mu-1}, \end{aligned} \quad (28)$$

where $F_k(r_p)$ is the probability of observing a ball of radius r_p containing $k - 1$ points. From this expression, the expectation of a distance r_p from \mathbf{m} to its k th neighbor can be found as

$$\begin{aligned} \bar{T}_k(p) &= \int_0^\infty \rho F_k(\rho) d\rho \\ &= \frac{c_p^k \mu}{(k - 1)!} \int_0^\infty \rho^{\mu k} e^{-c_p \rho^\mu} d\rho. \end{aligned} \quad (29)$$

Recalling that $\Gamma(k) = (k - 1)!$, this equation simplifies to (see Appendix B)

$$\bar{T}_k(p) = [c_p^{-1/\mu}] \left[\frac{\Gamma(k + \frac{1}{\mu})}{\Gamma(k)} \right]. \quad (30)$$

The leading term $c_p^{-1/\mu}$, which is a function of the Minkowski parameter p , is independent of k . This implies that the ratio of expected distances for different values of k will be independent of the Minkowski parameter. This can be verified for the case of a Swiss roll in Fig. 8(a), which shows that the estimates of the ratio of average distances to the j th and k th nearest-neighbor shells $\frac{T_j(p)}{T_k(p)}$ for different Minkowski parameters is constant.

Equations (5) and (17) are the estimators for the dimensionality for the nearest-neighbor and the MLE approaches, respectively. Both involve only the ratios of distances. In the above, we showed that the only effect of p in the Minkowski p -norm is a multiplicative constant, which cancels in both cases. ■

Remark 1. Figure 8(b) shows the variation of computed intrinsic dimension by this approach against the choice of Minkowski parameter for a helix and a Swiss roll data set. The correct intrinsic dimension is found for all Minkowski parameters tested: 1 for the helix and 2 for the Swiss roll.

Theorem 2. For binary images, the intrinsic dimensionality estimates scale with the value of p in the p -norm, that is, $\hat{\mu}_{p,2} = p \hat{\mu}_{1,2}$.

Proof. Consider the distance between any two images in the Minkowski distance family for the case of binary images (two-level quantized). Here one has discrete samples where images occupy vertices of a cube as shown in Fig. 1(b).

When \mathbf{m}_v is quantized, it assumes discrete values that are an integer multiple of a basic value Δm . We can define a variable $\zeta^q \in \mathbb{Z}^q$, where q is the number of quantization levels in the representation of the image, so the elements of ζ^q are in

$\{0, 1, 2, \dots, q-1\}$,

$$\mathbf{m}_v \triangleq \Delta m \zeta_v^q, \quad (31)$$

where v denotes the v th image. In this case, $q = 2$, since it is a binary image. Since the absolute difference between any two binarized pixels is either zero or one ($|\zeta_{v,i}^2 - \zeta_{w,i}^2| = 0$ or $|\zeta_{v,i}^2 - \zeta_{w,i}^2| = 1$),

$$|\zeta_{v,i}^2 - \zeta_{w,i}^2|^p = |\zeta_{v,i}^2 - \zeta_{w,i}^2| \forall i \in \{1, 2, 3, \dots, n\}, \quad (32)$$

where i denotes the coordinates of ζ^2 .

The Minkowski p -norm distance between any two binary image instances \mathbf{m}_v and \mathbf{m}_w can be written as

$$\begin{aligned} d_p(\mathbf{m}_v, \mathbf{m}_w) &= \left(\sum_{i=1}^n |\mathbf{m}_{v,i} - \mathbf{m}_{w,i}|^p \right)^{1/p} \\ &= (\Delta m)^{1/p} \left(\sum_{i=1}^n |\zeta_{v,i}^2 - \zeta_{w,i}^2|^p \right)^{1/p} \\ &= \left(\Delta m \sum_{i=1}^n |\zeta_{v,i}^2 - \zeta_{w,i}^2| \right)^{1/p} \\ &= \left(\sum_{i=1}^n |\mathbf{m}_{v,i} - \mathbf{m}_{w,i}| \right)^{1/p} \\ &= [d_1(\mathbf{m}_v, \mathbf{m}_w)]^{1/p}. \end{aligned} \quad (33)$$

In terms of the distance estimates, this leads to $T_k(p) = [T_k(1)]^{1/p}$, using the nearest-neighbor estimate in Eq. (5). The intrinsic dimension estimates are then a linear function of p ,

$$\begin{aligned} \hat{\rho}_{p,2} &= \ln \left(\frac{k_b}{k_a} \right) \left(\ln \frac{T_{k_b}(p)}{T_{k_a}(p)} \right)^{-1} \\ &= \ln \left(\frac{k_b}{k_a} \right) \left(\ln \frac{T_{k_b}(1)^{1/p}}{T_{k_a}(1)^{1/p}} \right)^{-1} \\ &= \ln \left(\frac{k_b}{k_a} \right) \left(\frac{1}{p} \ln \frac{T_{k_b}(1)}{T_{k_a}(1)} \right)^{-1} \\ &= p \hat{\rho}_{1,2}, \end{aligned} \quad (34)$$

where the superscript 2 indicates that this estimate is made with $q = 2$, one-bit precision.

The same behavior is also obtained with the MLE equation (17) because it uses similar distance ratios. Thus, an estimate of the intrinsic dimension of $p \hat{\rho}_{1,2}$ would be obtained for the Minkowski p -norm measure when using binary (one-bit quantized) images. ■

Remark 2. Note also that the condition described in its extreme by Theorem 2 is usually inescapable. Even if care were taken to use a large value of q to quantize the image, most of the levels would not be used because of the intrinsic sparsity of the data. With materials images, the intensities tend to cluster about only a few intensities that represent, for example, phases, compositions, and crystal orientations. The end result will inevitably be a quantized image for which we recommend the L_1 -norm as the metric for estimating the intrinsic dimension.

Remark 3. Theorem 2 is the opposite extreme from Theorem 1. Whereas Theorem 1 pertains to the infinite quantization limit, Theorem 2 pertains to the one-bit binary image extreme. The results are general, irrespective of the data. For q values between these extremes, the results are in general dependent on the data. Still some general conclusions may be drawn. The following two theorems describe the behavior of the estimates between these two extremes.

Theorem 3. The dimensionality estimate with the L_1 -norm is lower than that using any other Minkowski norm for q -level quantized images with pixels represented as positive integers.

Proof. We do this in two steps: We prove that (i) the L_1 -norm is greater than or equal to any other Minkowski p -norm and (ii) there is a proportionality constant between the estimate of nearest-neighbor dimension made for the L_1 -norm and any other p -norm. The generalization of the p proportionality constant in Eq. (34) is a constant $\gamma_q > 1$ and is a function of q , the quantization level.

In the proof of Theorem 2, we showed that the scaling of the k -nearest-neighbor radius goes as $T_k(p) = [T_k(1)]^{1/p}$ for binary images [Eq. (33)]. This theorem generalizes this to any level of quantization.

To generalize this idea, one can first show that $T_k(p) \leq T_k(1)$ for q -level quantized images. This can be proved using repeated applications of the triangle inequality ($\|f + g\|_p \leq \|f\|_p + \|g\|_p$), which applies for metric p -norms ($p \geq 1$) [36]:

$$\begin{aligned} \|d\|_p &= \left(\sum_{i=0}^n |d_i|^p \right)^{1/p} \\ &= \left(\sum_{i=0}^{n-1} |d_i|^p + |d_n|^p \right)^{1/p} \\ &\leq \left(\sum_{i=0}^{n-1} |d_i|^p \right)^{1/p} + |d_n|^p \\ &= \left(\sum_{i=0}^{n-1} |d_i|^p \right)^{1/p} + |d_n| \\ &= \left(\sum_{i=0}^{n-1} |d_i|^p \right)^{1/p} + \|d_n\|_1. \end{aligned} \quad (35)$$

By recursion, all of the terms of the sum can be separated out, using the triangle inequality to yield

$$\begin{aligned} \|d\|_p &\leq \sum_{i=0}^n \|d_n\|_1 \\ &= \|d\|_1, \end{aligned} \quad (36)$$

which proves that the L_1 -norm gives a value greater than or equal to any other metric p -norm.

For part (ii) of the proof, recall that the distances $T_k(1)$ and $T_k(p)$ are just the L_1 - and L_p -norms, respectively, of the distance to the k th neighbor of the target point. Using the result just proved, the following is true:

$$T_k(1) \leq T_k(p). \quad (37)$$

Since logarithms are increasing functions, it is also true that

$$\ln[T_k(1)] \leq \ln[T_k(p)]. \quad (38)$$

Under the condition $T_k(1) > 1$, the logarithms will always be positive and

$$1 \leq \frac{\ln[T_k(p)]}{\ln[T_k(1)]} \triangleq \gamma_{1,p;q}, \quad 1 < T_k(t). \quad (39)$$

Rearranging Eq. (39), we obtain $T_k(1) = [T_k(p)]^{1/\gamma_{1,p;q}}$. Using this relation in the nearest-neighbor estimate in Eq. (5), it is shown that

$$\hat{\mu}_{1;q} = \gamma_{1,p;q} \hat{\mu}_{p;q}, \quad (40)$$

which is guaranteed if the estimate is made from neighbor distances $T_k(1) > 1$. This implies that $\hat{\mu}_{p;q} > \hat{\mu}_{1;q}$. Thus, the L_1 -norm provides the lowest nearest-neighbor estimate of intrinsic dimensionality from among the p -norms. ■

Remark 4. Theorem 3 is a generalization of Theorem 2, for $1 < q < \infty$.

Theorem 4. Let $\hat{\mu}_{p;q}$ and $\hat{\mu}_{t;q}$ be the p -norm and t -norm estimates, respectively, of q -level quantized images, where the pixels are represented as positive integers. If $p \leq t$, then, for all q , $\hat{\mu}_{p;q} \leq \hat{\mu}_{t;q}$.

Proof. We do this in two parts. First, we show that the t -norm is less than or equal to the p -norm for $t > p$, in general. Then we show that the relationship between the dimensionality estimates using the t -norm are greater than those using the p -norm, in general.

Consider a p -normalized vector $d^* \triangleq d/\Lambda$, where $\Lambda \triangleq \|d\|_p$. We have $\|d^*\|_p = 1$. With this, the t -norm of d^* may be written

$$\|d^*\|_t = \left(\sum |d_i^*|^t \right)^{1/t}, \quad (41)$$

which is not expected to be 1 in the general case. For each i , $|d_i^*| \leq 1$, so if $p \leq t$, we have

$$\begin{aligned} |d_i^*|^t &= [|d_i^*|^{t-p}] |d_i^*|^p \\ &\leq |d_i^*|^p. \end{aligned} \quad (42)$$

Raising the t -norm of Eq. (41) to the power of t and using Eq. (42) gives

$$\begin{aligned} \|d^*\|_t^t &= \sum |d_i^*|^t \\ &\leq \sum |d_i^*|^p. \end{aligned} \quad (43)$$

However, this is just 1, since the p -norm of d^* is 1, by construction. This implies that $\|d^*\|_t \leq \|d^*\|_p$ for $p \leq t$, establishing the inequality for all vectors of unit p -norm. This can be extended to any vector by multiplying by Λ to recover the original vector d :

$$\Lambda \|d^*\|_t = \|\Lambda d^*\|_t = \|d\|_t. \quad (44)$$

Applying the analogous transformation to the p -norm establishes that $\|d\|_t \leq \|d\|_p$ if $p < t$.

For part (ii) of the proof, recall that the distances $T_k(t)$ and $T_k(p)$ are just the t -norm and p -norm, respectively, of the distance to the k th neighbor of the target point. Using the result just proved, the following is true:

$$T_k(t) \leq T_k(p), \quad t \leq p. \quad (45)$$

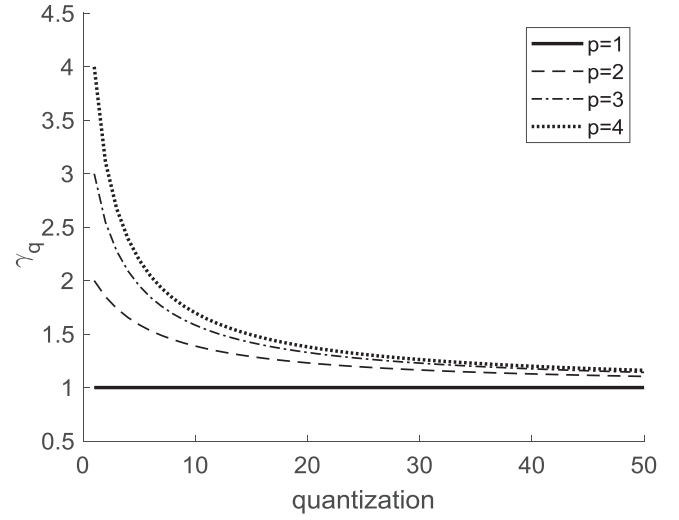


FIG. 9. Illustration of features of γ_q as a function of quantization q . Sampled images of size 256×256 contain pixels represented with integers uniformly sampled in the range $(0, 2^q - 1)$. For binary images, γ_q scales with the p -norm. On the other hand, the estimates converge at high levels of quantization. In addition, the higher the Minkowski norm parameter, the higher the dimensionality estimate.

Since logarithms are increasing functions, it is also true that

$$\ln[T_k(t)] \leq \ln[T_k(p)], \quad t \leq p. \quad (46)$$

Under the condition that $T_k(t) > 1$, the logarithms will always be positive and

$$1 \leq \frac{\ln[T_k(p)]}{\ln[T_k(t)]} \triangleq \gamma_{t,p;q}, \quad t \leq p, \quad 1 < T_k(t). \quad (47)$$

Rearranging Eq. (47), we obtain $T_k(t) = [T_k(p)]^{1/\gamma_{t,p;q}}$. Using this relation in the nearest-neighbor estimate in Eq. (5), it is shown that

$$\hat{\mu}_{t;q} = \gamma_{t,p;q} \hat{\mu}_{p;q}, \quad p \leq t, \quad (48)$$

which is guaranteed if the estimate is made from neighbor distances $T_k(t) > 1$. This implies that $\hat{\mu}_{t;q} \geq \hat{\mu}_{p;q}$ when $t \geq p$, that is, the p -norm estimate is an increasing function of p . This is not strictly increasing: At infinite quantization, they are all equal. ■

Remark 5. Theorem 4 is a generalization of Theorem 3 to any pair of Minkowski distance measures.

Remark 6. The constraint in the above proofs is that $T_k(1) > 1$ and this is realized when pixel values are represented in the form of integers, where q -level quantization indicates pixels represented in the range $(0, q - 1)$. Most common is an eight-bit representation in the range $(0, 255)$. In such a case, the minimal L_1 distance between two different images is one, in the case where one pixel intensity is changed by one and the images are otherwise identical. At higher neighbor distances (e.g., the second nearest neighbor), it is guaranteed that $T_k(1) > 1$. Note that in the implementation of the MLE (17), the lower limit of the k neighborhood has to be necessarily greater than one [19]. As numerical verification, Fig. 9(a) shows the value of γ_q as a function of number of bits for sampled images of size 256×256 . The figure shows the features of γ_q as a function of quantization q , indicating that γ_q

scales with the p -norm for binary images, $\gamma_q > 1$ for all cases, and γ_q trends towards one in the limit of infinite quantization. In addition, following Theorem 4, the higher the Minkowski norm parameter, the higher the dimensionality estimate.

Remark 7. In Ref. [1] it was found that the intrinsic dimension estimate based on raw images was curiously higher than the dimension estimates from the same images coded to a smaller latent vector using an autoencoder. This was specifically for cases where the images were quantized (e.g., binary images) and not the case for images represented in grayscale. For example, in the case of the circular motion example, which is highly quantized, the intrinsic dimension was 3.67 compared to the ground truth of 2, and 2.19 estimated using the latent vectors. The dimension estimates should be the same if the images are losslessly compressed. The reason for the increase in dimensionality, we propose, is given by the proofs in Theorems 1 and 4. Specifically, they used an L_2 -norm for the quantized images leading to close to double the estimate provided by the autoencoder. The autoencoder generates real numbers in the latent vector and hence the estimate from the L_2 -norm will be close to the true dimensionality according to Theorem 4. In the next section, we further explore the topology of the space generated by an autoencoder.

B. Composite microstructure dimensionality

Figure 7 shows that the microstructure has a self-similar or fractal property and that 40 variables are sufficient to describe its features. The intrinsic dimension is much smaller than the expected dimension based on the number of fibers, which would be five times the mean number of ellipses (x and y positions of centroid, major, and minor axes and orientation angle). A comparison of the results with three times the number of fibers, which is the upper bound for circular fibers, is shown in Fig. 7(b).

The reason for the low dimensionality is that the fiber packing is correlated as the window size increases, so some variables become redundant or less informative. To demonstrate this, we compared the dimensionality of experimental images with the synthetic data set DS SiC/SiC B with random placement of fibers in Fig. 7(b). Circular fibers were used in the synthetic data set, which initially had lower dimensionality compared to elliptical fibers, which additionally had differences in the minor axis and orientation. However, the intrinsic dimensionality continued to increase with the window size linearly, unlike the experimental data set.

One implication of this result is that it can help to determine the optimal size and centroid placement of a microstructural representative volume element (RVE), which is a sample of a heterogeneous material that can be used to predict its effective properties using computational models. A common criterion for choosing an RVE size is that it should be large enough to capture the statistical variability of properties or spatial correlation of the microstructural features but small enough to reduce the computational cost and complexity. Based on our intrinsic dimension analysis, we can suggest that an RVE size of 140 μm or larger would be suitable for representing the fibers in a tow for this data set.

C. Retrieving state variables using an autoencoder

While the MLE algorithm recovers the intrinsic dimensionality, it is of interest to identify the geometry of the latent space and to correlate the dimensions to microstructural features. A variety of applications can benefit from such analysis, including identification of novel processing paths and inverse design of microstructures for a given property as shown in Ref. [9]. To generate a proof of concept, a synthetic DS SwissRoll was used where the generator space is known. Our objective was to check if the generator space can be retrieved solely from the image data.

The state variables were identified using an autoencoder architecture. An autoencoder [37] is a multilayer neural network that learns the identity function such that the output $\tilde{\mathbf{x}}$ approximates the input \mathbf{x} . In the architecture, the hidden layers have fewer nodes than the input dimension and act as a bottleneck. In the first few layers, the autoencoder compresses the input to a compressed (latent space) representation in a process called encoding. At its simplest, a single hidden layer operates on the input $\mathbf{x} \in \mathbb{R}^n$ and generates an encoding $\mathbf{y}_1 \in \mathbb{R}^j$, $j < n$, such that

$$\mathbf{y}_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \quad (49)$$

where \mathbf{W}_1 represents the $j \times n$ weight matrix and \mathbf{b}_1 is the $j \times 1$ bias vector for the first layer. The function σ is typically a nonlinear activation function and a logistic sigmoid function is used in this work. Later layers reconstruct the output from this latent space representation in a process called decoding. An example is another layer that maps the latent vector \mathbf{y}_1 in the previous step to output $\tilde{\mathbf{x}} \in \mathbb{R}^n$ such that

$$\tilde{\mathbf{x}} = \sigma(\mathbf{W}_2\mathbf{y}_1 + \mathbf{b}_2), \quad (50)$$

where \mathbf{W}_2 represents the $n \times j$ weight matrix and \mathbf{b}_2 is the $n \times 1$ bias vector of the second layer. Multiple layers can be used to develop a deep network. The parameters in W and b are found by minimizing the cost $\frac{1}{2}(\|\mathbf{x} - \tilde{\mathbf{x}}\|_2)^2$ by training via backpropagation.

In this work a stacked autoencoder configuration comprised of five fully connected layers in total as shown in Fig. 10(c) was employed. The first autoencoder comprised of three layers (input, bottleneck, and output) was trained to reduce the dimensions to a bottleneck of 100 first. This was followed by a second autoencoder that used the 100-dimensional feature from the first autoencoder as input and reduced it to the intrinsic dimension identified by the MLE algorithm or the generator dimension. The two autoencoders were sequentially trained first, followed by retraining a combined five-layer autoencoder. We used a mean-square error (MSE) loss function and an Adam optimizer with a learning rate of 0.001. We trained the first two networks for 400 epochs and the stacked autoencoder for 1000 epochs. The network achieved a final MSE of 0.01 on the test set, indicating a good reconstruction performance.

Figure 10 shows a schematic of the approach using DS SwissRoll A [circles sampled from a Swiss roll, Fig. 10(b)]. The stacked autoencoder reads the images into a network with a bottleneck equal to the generator space dimension (equal to 3). The decoded images from the last layer are shown in Fig. 10(d). The three variables corresponding to each image

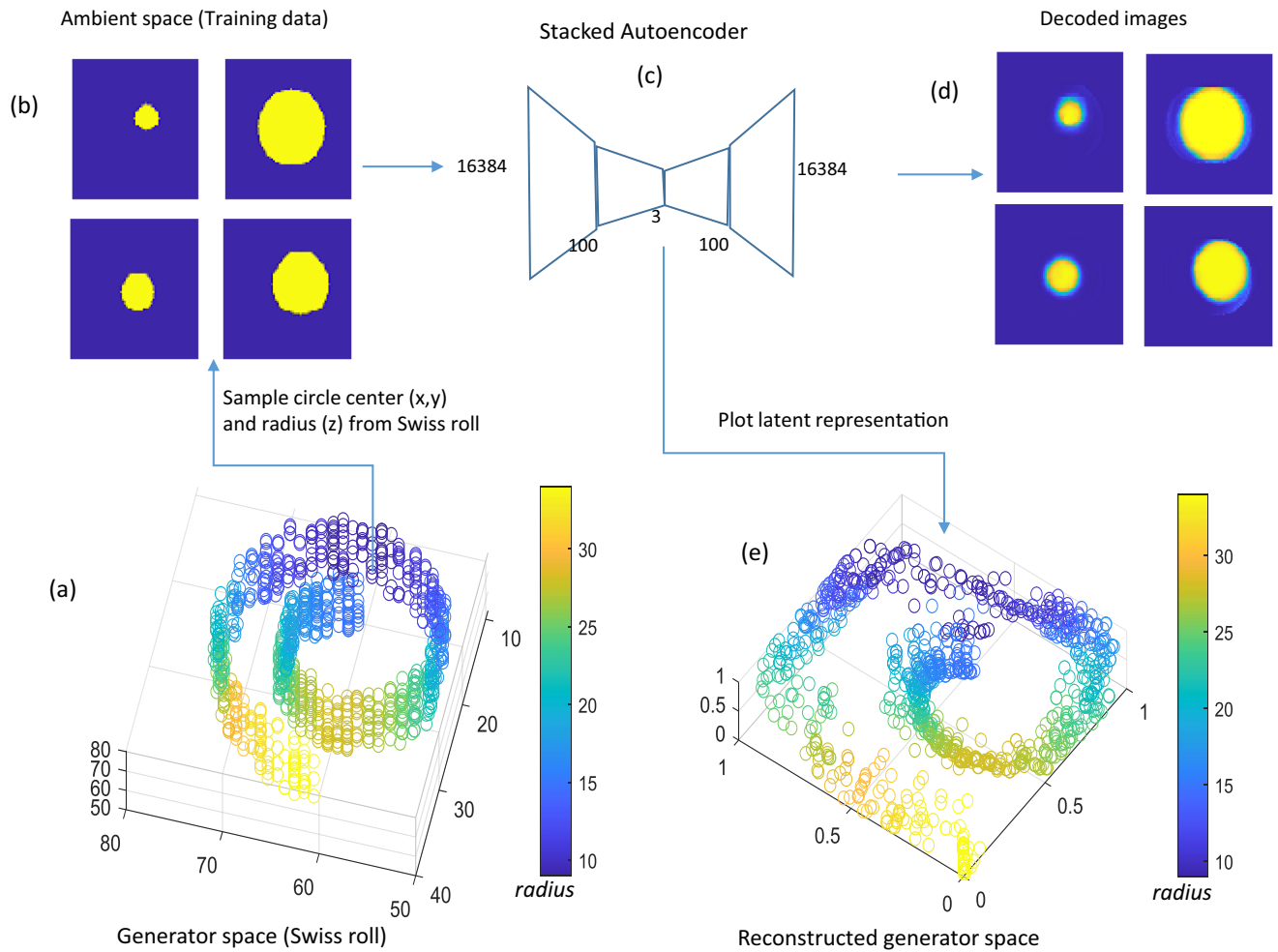


FIG. 10. Use of autoencoders for building a generator space shown using a synthetic data set. (a) The Swiss roll is the generator space: Each point has a (x_1, x_2, x_3) coordinate equal to (x, y, R) in an image where R is the circle radius and (x, y) is the center of the circle in the matrix representation (row, column) of the image. (b) Images from the database. (c) The stacked autoencoder reads in the images into a network with a bottleneck equal to the generator space dimension (equal to 3). (d) Decoded images from the last layer of the network. (e) The 1000 images are reduced to three variables in the bottleneck. These variables (when plotted) show the generator space.

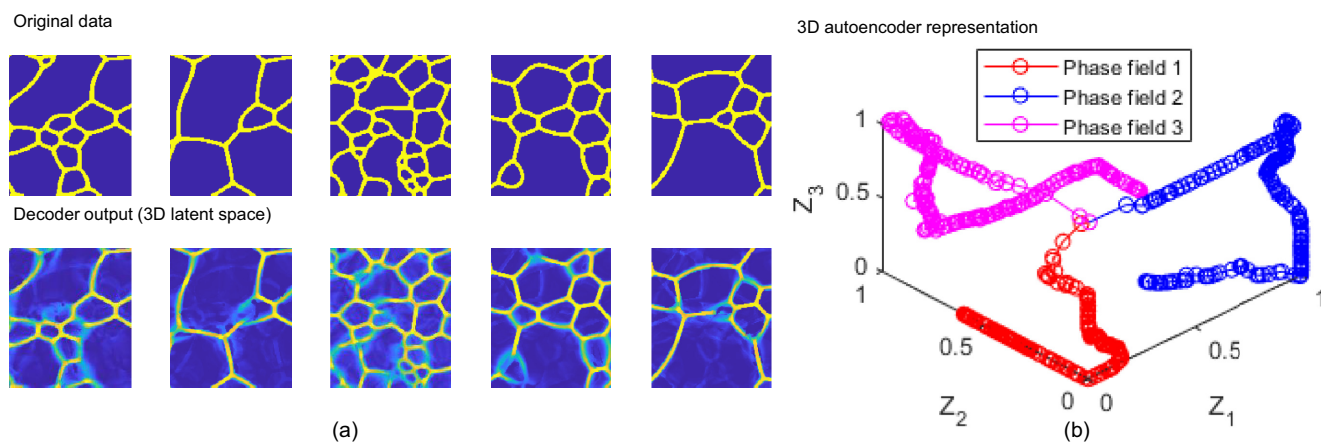


FIG. 11. (a) Decoder outputs for combined data set containing all three phase field trajectories. (b) A 3D representation from the autoencoder showing all three trajectories, with the first trajectory going towards the origin, the second to the right, and the third to the left of the plot.

are plotted in Fig. 10(e), which shows that the generated topology is similar to the actual generator space in Fig. 10(a). The points shown are colored according to the circle radius in the images. Note that the autoencoder, by default, restricts the range of values to between 0 and 1. However, the topology of the space is generally well reconstructed demonstrating a proof of concept that variables that define the microstructural state can be identified using stacked autoencoders.

Since the dimensionality of microstructure data sets and its topology can be estimated, the present approach can be used to understand structure-property relationships directly in the microstructure space. As an example, Fig. 11 shows a 3D space generated by the autoencoder for the DS Phase-Field (phase field data) containing three different phase field simulations. Initial images were generated via small random perturbations to the order parameters leading to different evolution pathways for the microstructure. In all three simulations, the initial images are close and are represented by the central points in the latent space. The trajectories from the three simulations emerge in different directions from the initial point, resulting in different microstructures. This latent space is an illustration of the microstructural latent space for a grain coarsening process. This falls short of estimating the true dimensionality of the space, for which a large number of such phase field trajectories is required. Nevertheless, this example is used to illustrate how the framework can be used to visualize a multitude of complex processes within a single low-dimensional manifold. The use of nonlinear manifolds (as opposed to linear PCA) as demonstrated here is expected to significantly improve the state of the art in the future.

VII. CONCLUSION

A methodology for reliably estimating the intrinsic dimensionality of random media was developed. The method resolves the ambiguity in results for images with low bit depth when using state-of-the-art techniques that employ the Euclidean (L_2) norm. The following are the main contributions of this work.

(i) Maximum-likelihood estimates of dimensionality can undergo a significant distortion because of image quantization. For infinite level quantization (analog), all Minkowski p -norms yield the same value for dimensionality. When the image is quantized, the MLE estimate increases with the Minkowski p parameter. We provided proofs of this result in Sec. VI.

(ii) The L_1 -norm produced the most consistent estimate. This was the consequence of mathematical properties of the MLE estimate for different levels of quantization.

(iii) For microstructure images, where the object of interest is not localizable, it was possible to estimate the dimensionality of the random images because the estimate homogenized. Specifically, for a SiC/SiC fiber composite, it was found that the dimensionality estimate homogenized, in this case, at 40 dimensions.

(iv) We conjectured that the processing-microstructure-property paradigm is actually a linking of three sparse domains, each describing the material. We showed that it was possible to infer characteristics of a sparse generator domain from high-dimensional images produced by the generator.

This was shown for a simulated data set and one constructed with phase field modeling.

ACKNOWLEDGMENTS

This research was supported in part by the Air Force Research Laboratory Materials and Manufacturing Directorate, through the Air Force Office of Scientific Research Summer Faculty Fellowship Program, Contracts No. FA8750-15-3-6003 and No. FA9550-15-0001. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. This work used software and services provided by Globus.

APPENDIX A: EFFECTS OF PARTICLE INTERSECTIONS WITH BOUNDARIES

In the data sets used in this paper, we avoided the intersection of the particle shape with the image boundary to get an unbiased estimation of the dimensionality. In general, image boundaries can play a role in biasing the intrinsic dimension. Consider the case of a single circular shape placed in a matrix. If the shape intersects the boundary, only a part of the shape is seen and an intrinsic dimensionality of 3 as judged from the data set assumes that the shape always remains a circle. To test this, we have plotted results from two data sets, one with and one without boundary intersections. We consider sufficiently high sample size (3000 images) and image size (128×128) for each case.

The histogram of estimated dimension per data point is plotted in Fig. 12(a), showing that the algorithm is able to predict the correct mean dimension for both cases. The key difference is that the histogram is broader and a higher standard deviation is obtained when the circles intersect the boundary. Another case is shown in Fig. 12(b), where the circle is linearly placed along the centerline. Here two distinct peaks are seen in the histogram where the circles are freely placed, with the first peak at a lower intrinsic dimension. The mean dimension is again correctly estimated for both cases with a higher standard deviation for the case where circles intersect the image boundaries. A case where the circles are periodic is also shown here, where the circles wrap around on the opposite side when they intersect the boundary. While this case shows a single sharp peak as in the case where boundaries are avoided, the standard deviation is higher than that case.

APPENDIX B: ESTIMATE OF MINKOWSKI p -NORM DISTANCE TO THE k th NEIGHBOR

Equation (28) simplifies to

$$\begin{aligned} F_k(r_p) &= \left(\frac{(c_p r_p^\mu)^{k-1}}{\Gamma(k)} e^{-c_p r_p^\mu} \right) c_p \mu r_p^{\mu-1} \\ &= \left[c_p^k \mu \left(\frac{1}{\Gamma(k)} \right) \right] r_p^{\mu k-1} e^{-c_p r_p^\mu}. \end{aligned}$$

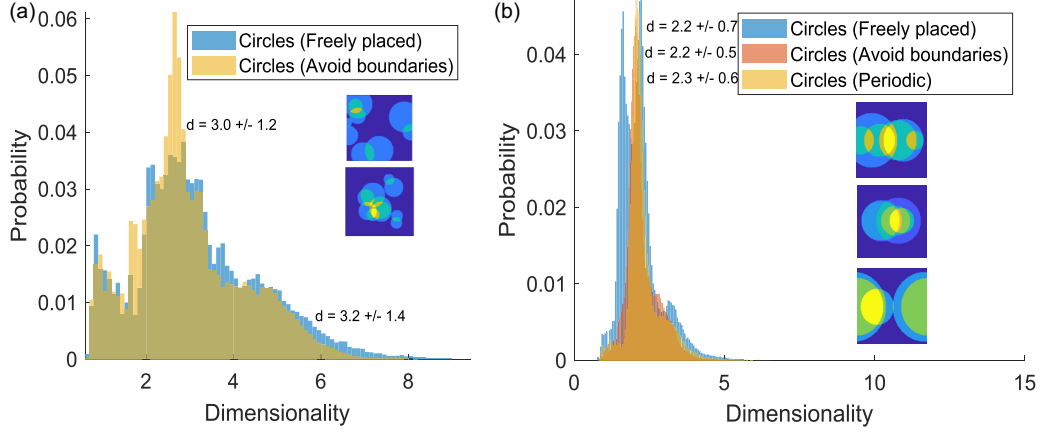


FIG. 12. Histograms of estimated dimension per data point. (a) Two cases are considered: A circle of varying radius is freely placed in one case ($d = 3.2$) and avoids boundary ($d = 3$) in another case. (b) Here the circle is linearly placed along the centerline. A case with periodicity is included. The insets show a superposition of a few different images in each data set.

Inserting this into Eq. (29), we obtain

$$\begin{aligned}\bar{T}_k(p) &= \int_0^\infty r_p F_k(r_p) dr_p \\ &= \left[c_p^k \mu \left(\frac{1}{\Gamma(k)} \right) \right] \int_0^\infty r_p^{\mu k} e^{-c_p r_p^\mu} dr_p.\end{aligned}$$

Changing the integration variable with $\xi \triangleq c_p r_p^\mu$ ($r_p = \xi^{1/\mu} c_p^{-1/\mu}$), we obtain

$$\begin{aligned}\bar{T}_k(p) &= \left[c_p^k \mu \left(\frac{1}{\Gamma(k)} \right) \right] \int_0^\infty (\xi^{1/\mu} c_p^{-1/\mu})^{\mu k} e^{-\xi} d(\xi^{1/\mu} c_p^{-1/\mu}) \\ &= c_p^{-1/\mu} \left(\frac{1}{\Gamma(k)} \right) \int_0^\infty \xi^{k+1/\mu-1} e^{-\xi} d\xi.\end{aligned}$$

The integral is, by definition, $\Gamma(k + 1/\mu)$ [38], which we obtain by substituting

$$\bar{T}_k(p) = c_p^{-1/\mu} \frac{\Gamma(k + \frac{1}{\mu})}{\Gamma(k)}.$$

APPENDIX C: MLE DERIVATION WITH THE MINKOWSKI DISTANCE MEASURE

The Poisson rate can be written in terms of the radius r based on a Minkowski p -norm as

$$\lambda^*(r) = f(\mathbf{m}) V_p(\mu) \mu r^{\mu-1}. \quad (\text{C1})$$

The log-likelihood of the 1D Poisson process can be written as

$$L(\mu, \theta) = \int_0^R \ln(\lambda^*) dN(r) - \int_0^R (\lambda^*) dr. \quad (\text{C2})$$

Using the conversion $\ln f(\mathbf{m}_i) = \theta$, we get $\lambda^*(r) = e^\theta V_p(\mu) \mu r^{\mu-1}$. We can plug this into L as follows:

$$L(\mu, \theta) = \int_0^R [\theta + \ln(V_p \mu r^{\mu-1})] dN(r) - \int_0^R e^\theta V_p \mu r^{\mu-1} dr. \quad (\text{C3})$$

Maximizing likelihood with respect to θ using $\frac{\partial L}{\partial \theta} = 0$, the second term independent of θ vanishes, θ terms comes out

of the integral, and the remaining integrals can be trivially evaluated as

$$\frac{\partial L}{\partial \theta} = N - e^\theta V_p(\mu) R^\mu = 0, \quad (\text{C4})$$

$$N = e^\theta V_p(\mu) R^\mu. \quad (\text{C5})$$

Maximizing likelihood with respect to μ , we get

$$\frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu} \int_0^R \ln(e^\theta V_p \mu r^{\mu-1}) dN - \frac{\partial}{\partial \mu} \int_0^R e^\theta V_p \mu r^{\mu-1} dr. \quad (\text{C6})$$

Evaluating the first term only, we obtain

$$\begin{aligned}\frac{\partial}{\partial \mu} \int_0^R \ln(e^\theta V_p \mu r^{\mu-1}) dN \\ = \frac{\partial}{\partial \mu} \left(\ln(e^\theta V_p \mu) N + (\mu - 1) \int_0^R \ln(r) dN \right) \quad (\text{C7})\end{aligned}$$

$$= \frac{\partial}{\partial \mu} \left(\theta + \ln(V_p \mu) N + (\mu - 1) \int_0^R \ln(r) dN \right). \quad (\text{C8})$$

The first term vanishes and the second and third terms can be differentiated as follows, using $\frac{\partial V_p}{\partial \mu} = V_p'$:

$$\frac{\partial}{\partial \mu} \int_0^R \ln(e^\theta V_p \mu r^{\mu-1}) dN = \left(\frac{V_p'}{V_p} + \frac{1}{\mu} \right) N + \int_0^R \ln(r) dN. \quad (\text{C9})$$

The second term is given by

$$\begin{aligned}\frac{\partial}{\partial \mu} \int_0^R e^\theta V_p \mu r^{\mu-1} dr \\ = \frac{\partial}{\partial \mu} e^\theta V_p \mu \int_0^R r^{\mu-1} dr \\ = e^\theta \frac{\partial}{\partial \mu} (V_p R^\mu) \quad (\text{C10})\end{aligned}$$

$$= e^\theta V_p' R^\mu + e^\theta V_p(\mu) R^\mu \ln(R). \quad (\text{C11})$$

Using $N = e^\theta V_p(\mu)R^\mu$, we get

$$\frac{\partial}{\partial \mu} \int_0^R e^\theta V_p \mu r^{\mu-1} dr = \frac{V'_p}{V_p} N + N \ln(R). \quad (\text{C12})$$

Combining the first and second terms and putting that back in $\frac{\partial L}{\partial \mu}$, we get

$$\frac{\partial L}{\partial \mu} = \left(\frac{V'_p}{V_p} + \frac{1}{\mu} \right) N + \int_0^R \ln(r) dN - \frac{V'_p}{V_p} N - N \ln(R). \quad (\text{C13})$$

Simplifying

$$\frac{\partial L}{\partial \mu} = \frac{N}{\mu} + \int_0^R \ln(r) dN - N \ln(R) = 0 \quad (\text{C14})$$

and using a numerical approximation for the integral, we obtain

$$\int_0^R \ln(r) dN = \sum_{k=1}^N \ln T_k. \quad (\text{C15})$$

This comes from the fact that numerical increments in the number of neighbors N is one and the integral is equivalent to a summation over the logarithm of distances to the k th

neighbor. This leads to

$$\begin{aligned} \frac{N}{\mu} &= - \left(\sum_{k=1}^N \ln T_k \right) + N \ln R \\ &= \sum_{k=1}^N (-\ln T_k + \ln R) \\ &= \sum_{k=1}^N \ln \frac{R}{T_k}. \end{aligned} \quad (\text{C16})$$

Rearranging the terms

$$\mu = N \left(\sum_{k=1}^N \ln \frac{R}{T_k} \right)^{-1} \quad (\text{C17})$$

and fixing the number of neighbors k rather than the radius of the sphere R results in the MLE equation being identical to that of Levina and Bickel except that T_k is the Minkowski p -norm:

$$\mu = (k-1) \left[\sum_{j=1}^{k-1} \ln \left(\frac{T_k}{T_j} \right) \right]^{-1}. \quad (\text{C18})$$

-
- [1] B. Chen, K. Huang, S. Raghupathi, I. Chandratreya, Q. Du, and H. Lipson, Automated discovery of fundamental variables hidden in experimental data, *Nat. Comput. Sci.* **2**, 433 (2022).
- [2] S. Torquato, *Random Heterogeneous Materials: Microstructure and Macroscopic Properties* (Springer, New York, 2002).
- [3] J. Ohser and F. Mücklich, *Statistical Analysis of Microstructures in Materials Science* (Wiley, New York, 2001).
- [4] R. Bostanabad, Y. Zhang, X. Li, T. Kearney, L. C. Brinson, D. W. Apley, W. K. Liu, and W. Chen, Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques, *Prog. Mater. Sci.* **95**, 1 (2018).
- [5] V. Sundararaghavan and N. Zabarar, Classification and reconstruction of three-dimensional microstructures using support vector machines, *Comput. Mater. Sci.* **32**, 223 (2005).
- [6] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).
- [7] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, *arXiv:1312.6114*.
- [8] N. Lubbers, T. Lookman, and K. Barros, Inferring low-dimensional microstructure representations using convolutional neural networks, *Phys. Rev. E* **96**, 052111 (2017).
- [9] S. Sundar and V. Sundararaghavan, Database development and exploration of process-microstructure relationships using variational autoencoders, *Mater. Today Commun.* **25**, 101201 (2020).
- [10] R. Bostanabad, Reconstruction of 3D microstructures from 2D images via transfer learning, *Comput.-Aided Design* **128**, 102906 (2020).
- [11] D. Fokina, E. Muravleva, G. Ovchinnikov, and I. Oseledets, Microstructure synthesis using style-based generative adversarial networks, *Phys. Rev. E* **101**, 043308 (2020).
- [12] S. Thakre, V. Harshith, and A. K. Kanjarla, Intrinsic dimensionality of microstructure data, *Integrat. Mater. Manuf. Innov.* **10**, 44 (2021).
- [13] V. Sundararaghavan and N. Zabarar, A dynamic material library for the representation of single-phase polyhedral microstructures, *Acta Mater.* **52**, 4111 (2004).
- [14] V. Sundararaghavan and N. Zabarar, Linear analysis of texture property relationships using process-based representations of Rodrigues space, *Acta Mater.* **55**, 1573 (2007).
- [15] B. Ganapathysubramanian and N. Zabarar, A non-linear dimension reduction methodology for generating data-driven stochastic input models, *J. Comput. Phys.* **227**, 6612 (2008).
- [16] J. B. Tenenbaum, V. de Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* **290**, 2319 (2000).
- [17] Z. Li, B. Wen, and N. Zabarar, Computing mechanical response variability of polycrystalline microstructures through dimensionality reduction techniques, *Comput. Mater. Sci.* **49**, 568 (2010).
- [18] J. Costa and A. Hero, Manifold learning with geodesic minimal spanning trees, *arXiv:cs/0307038*.
- [19] E. Levina and P. Bickel, in *Advances in Neural Information Processing Systems 17, Vancouver, 2004*, edited by L. Saul, Y. Weiss, and L. Bottou (MIT Press, Cambridge, 2004).
- [20] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein, *Proceedings of the Ninth International Conference on Learning Representations* (OpenReview.net, 2021).
- [21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* **86**, 2278 (1998).

- [22] A. Krizhevsky, Learning multiple layers of features from tiny images, University of Toronto Report No. TR-2009, 2009 (unpublished).
- [23] S. R. Niezgoda, Stochastic representation of microstructure via higher-order statistics: Theory and application, Ph.D. thesis, Drexel University, 2010.
- [24] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space* (Springer Science + Business Media, New York, 2012).
- [25] D. P. Kingma and M. Welling, An introduction to variational autoencoders, *Found. Trends Mach. Learn.* **12**, 307 (2019).
- [26] F. Camastra and A. Staiano, Intrinsic dimension estimation: Advances and open problems, *Inf. Sci.* **328**, 26 (2016).
- [27] K. W. Pettis, T. A. Bailey, A. K. Jain, and R. C. Dubes, An intrinsic dimensionality estimator from near-neighbor information, *IEEE T. Pattern Anal.* **1**, 25 (1979).
- [28] A. Baddeley, I. Bárány, and R. Schneider, in *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, edited by W. Weil, Lecture Notes in Mathematics Vol. 1892 (Springer, Berlin, 2007), pp. 1–75.
- [29] R. Brown, *Topology and Groupoids* (McGraw-Hill, New York, 2006).
- [30] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann. Math. Stat.* **9**, 60 (1938).
- [31] X. Wang, Volumes of generalized unit balls, *Math. Mag.* **78**, 390 (2005).
- [32] A. Schumacher, Swiss rolls dataset, https://github.com/ajschumacher/gadsdc1/blob/master/dataset_research/john_k_swissrolls.md (2014).
- [33] S. Sherman, J. Simmons, and C. Przybyla, S200 SiC/SiC seral section data and fiber chirality codes with tutorial, <https://doi.org/10.18126/M2135G> (2019).
- [34] S. Sherman, J. Simmons, and C. Przybyla, Mesoscale characterization of continuous fiber reinforced composites through machine learning: Fiber chirality, *Acta Mater.* **181**, 447 (2019).
- [35] D. Fan and L.-Q. Chen, Computer simulation of grain growth using a continuum field model, *Acta Mater.* **45**, 611 (1997).
- [36] H. P. Mulholland, On generalizations of Minkowski's inequality in the form of a triangle inequality, *Proc. London Math. Soc.* **s2-51**, 294 (1949).
- [37] P. Baldi and K. Hornik, Neural networks and principal component analysis: Learning from examples without local minima, *Neural Netw.* **2**, 53 (1989).
- [38] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products* (Academic, New York, 2014).