Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 359 (2020) 112657
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

A nonlinear data-driven reduced order model for computational homogenization with physics/pattern-guided sampling

Satyaki Bhattacharjee, Karel Matouš*

Department of Aerospace and Mechanical Engineering, Center for Shock Wave-processing of Advanced Reactive Materials, University of Notre Dame, Notre Dame, IN, 46556, USA

Available online 14 October 2019

Abstract

Developing an accurate nonlinear reduced order model from simulation data has been an outstanding research topic for many years. For many physical systems, data collection is very expensive and the optimal data distribution is not known in advance. Thus, maximizing the information gain remains a grand challenge. In a recent paper, Bhattacharjee and Matouš (2016) proposed a manifold-based nonlinear reduced order model for multiscale problems in mechanics of materials. Expanding this work here, we develop a novel sampling strategy based on the physics/pattern-guided data distribution. Our adaptive sampling strategy relies on enrichment of sub-manifolds based on the principal stretches and rotational sensitivity analysis. This novel sampling strategy substantially decreases the number of snapshots needed for accurate reduced order model construction (i.e., $\sim 5\times$ reduction of snapshots over Bhattacharjee and Matouš (2016)). Moreover, we build the nonlinear manifold using the displacement rather than deformation gradient data. We provide rigorous verification and error assessment. Finally, we demonstrate both localization and homogenization of the multiscale solution on a large particulate composite unit cell.

Keywords: Computational homogenization; Nonlinear manifold; Reduced order model; Machine learning; Parallel computing; Big data

1. Introduction

Over time, a large number of scientific and engineering fields have come to rely on computer simulations to handle complex real world phenomena with large amounts of data. Such phenomena appear in Very Large Scale Integration (VLSI) design [1], chemical engineering [2], and modeling of turbulence [3], as well as in mechanics of materials [4]. These complex electrical/chemical/mechanical systems with complex multi-physics interactions require detailed numerical analysis tools (i.e., Direct Numerical Modeling/Simulation (DNM/DNS)) to resolve them.

Among these applications is the modeling of complex materials (e.g., biological materials). A very popular and effective method when material complexity involves multiple length scales is Computational Homogenization (CH) [5–8]. This includes the so-called FE^2 implementation where both scales are solved together concurrently [9,10]. However, the complexity is increased when deformation reaches the high strain regime. It is worth commenting that although problem definitions differ in different fields, these can broadly be categorized as nonlinear optimization problems. CH poses extra complexity as it involves multiple length scales along with costly

* Corresponding author.
E-mail address: kmatous@nd.edu (K. Matouš).

https://doi.org/10.1016/j.cma.2019.112657
0045-7825/

functional evaluations to solve the inherent optimization/variational problem [9–11]. Unfortunately, computational costs associated with solving these complex problems limit the use of these high-fidelity simulations, and an efficient means of acceleration would promote their industrial use.

To circumvent computational challenges, many data-driven Reduced Order Models (ROMs) have gained attention in the computational physics/mechanics community. These include proper orthogonal/generalized decomposition [12,13], kernel methods [14], manifold methods [15,16] and many others [17–19]. Consequently, in many fields there is a great interest in tools and techniques that facilitate the construction of such regression models, while minimizing the computational costs and maximizing model accuracy. For example, an efficient surrogate model using dimension reduction and the Gaussian process has been proposed for a single scale two-dimensional system [20]. Moreover, the applicability of Neural Networks (NNs) [21,22], kernel methods [14], and other ROM techniques [23–25] has become indispensable to the computational physics/mechanics community as these methods drastically reduce the computational overhead. However, most of the aforementioned techniques are applicable only when dealing with fairly low-dimensional problems.

In the context of CH, the literature mainly involves Proper Orthogonal Decomposition (POD) [24], discrete material maps [26,27], NN techniques [22], and recently wavelet-based model reduction [25]. In our previous work [16], we proposed a manifold-based technique that can perform both homogenization and localization of the material response in the 3D finite strain setting with a realistic Representative Unit Cell (RUC). It is important to mention that all these methodologies involve pre-computation of solutions to construct a digital database using highly expensive solvers. Although utilizing modern high-performance computers speeds up the process of generating data [9,28,29], the database construction is still a bottleneck. Furthermore, ROMs demand extra care since the complexity increases with the number of simulations. Therefore, data sampling is necessary in order to eliminate redundant simulations.

Because the data collection is computationally expensive and the optimal data distribution is not known in advance, data sampling is a grand challenge to maximize the information gain [30]. Although there are several statistical techniques available to address this issue [30–33], none of them are universal. These methods are categorized as classical Response Surface Model (RSM) schemes and space-filling methods [30]. Generally, minimization of bias is more important in ROMs for better predictability [34,35]. Therefore, the uniformity of the points in the design space is the main objective. This uniformity is achieved either by maximizing the minimum distances among input points or by minimizing correlation measures among the sample data [36]. Although both criteria are important and intuitively linked, achieving them together is almost impossible [36].

Practical implementation of space filling techniques include orthogonal array testing [37], Latin hypercube design [32], minmax/maxmin design [33], entropy based design [38], and others. These techniques also have considerably intuitive appeal. The orthogonal arrays produce uniform samples, but can generate particular forms of point replications. On the other hand, the Latin hypercube does not replicate points, but can lack uniformity. Ultimately, all mentioned methodologies have no control over the output space. It is true that in the limit, uniformity of the input space induces uniformity of the output space, but that requires a large number of sample points. This large sample point requirement increases exponentially with the dimension of the input space [35]. Unfortunately, the generation of such a rich dataset is practically impossible. For engineering implementation, we have to rely on very few simulations that eventually maximize the information and lead to an enriched ROM. However, in this particular case and when extremely high-dimensionality is involved, a uniform input space does not guarantee a uniform output space because of the curse of dimensionality. When the process of generating data is expensive as in the case of CH, sequential sampling is more efficient [30,39]. Moreover, adaptive sequential sampling is superior to one without adaptation [40].

In this work, we propose a novel sampling strategy in the context of CH. In particular, we employ the Manifold-based Nonlinear ROM (MNROM) developed in [16]. The MNROM operates directly on the finite element data and can be used to substantially accelerate fully-coupled FE^2 simulations [9,29]. The key components of MNROM are nonlinear manifold learning, such as Isomap (i.e., projection to a Euclidean space), which unfolds the solution manifold, a kernel based inverse map, and a NN. The efficacy of Isomap depends on how the limited simulation points can represent the manifold [41]. Since CH is a highly nonlinear process, a uniform distribution of points in the input space may lead to many pockets of extremely sparse regions on the manifold. On the other hand, some regions become unnecessarily dense. This fails to capture the global information of the manifold. Consequently, Isomap fails as it is a global learning technique and produces enormous noise [42]. Unfortunately, all other maps associated with MNROM are noise sensitive as well and that leads to a poorly performing ROM.

Therefore, we propose a deterministic strategy to efficiently sample in the input space such that the sparse regions of the high dimensional output manifold are minimized with respect to the specific number of simulations. This novel sampling method is completely pattern/physics-driven. In this method, the input and output spaces are coupled. We explore the data distribution on the manifold and find physical parameters which guide the pattern. We enrich the complete manifold by enriching the sub-manifolds. Some sub-manifolds require a small number of data points. On the other hand, some sub-manifolds require more points depending on the volume of the sub-manifold. From the learned pattern, we can easily determine the number of necessary points (i.e., the simulation snapshots). This method distributes the snapshots more evenly in the output space. This deterministic sampling technique is sequential and adaptive in nature. However, it does not imply the creation of a uniform input space.

We also address the issue of reducing the expense of costly Isomap as the enrichment technique proposed here is sequential in nature. Every time we enrich the manifold, we need to perform a highly expensive algorithm (i.e., Isomap) to reduce dimensionality. To mitigate this particular problem, we propose to use a greedy algorithm instead of the full-scale Isomap to project any new point into the reduced space. This greedy algorithm is extremely effective if the current dataset is dense enough. Our native physics-guided sampling strategy supports this aspect of the model reduction as well and lend itself well to FE^2 acceleration.

This paper is organized as follows. In Section 2, we briefly summarize the theory of computational homogenization for heterogeneous materials and the framework of the MNROM. In Section 3, we describe a new physics-guided sampling methodology. In Sections 4 and 5, we present numerical examples to show different aspects of this physics-guided technique and present both the homogenization and localization of the multiscale response of a RUC with 95 stiff inclusions. Finally, in Section 6, we summarize the work and discuss the issues and scope for further improvement.

2. Scale coupling using dimension reduction

In this section, we briefly summarize the theory of computational homogenization, which is described in detail in [5–11]. We also summarize the reduced order scheme for CH, which is based on the nonlinear manifold and machine learning introduced in [16]. This abbreviated theoretical section is for the completeness of the presentation. More details can be found in [16].

2.1. Computational homogenization

Let a body $\Omega_0 \subset \mathbb{R}^3$, consisting of material points $\boldsymbol{X} \in \Omega_0$, represent a macro-continuum (see Fig. 1), where the position vector ${}^0\boldsymbol{x}$ and the deformation gradient ${}^0\boldsymbol{F}$ are defined as

$$\begin{aligned} {}^0\boldsymbol{x}(\boldsymbol{X}) &= \boldsymbol{X} + {}^0\boldsymbol{u}(\boldsymbol{X}) && \forall \boldsymbol{X} \in \Omega_0, \\ {}^0\boldsymbol{F}(\boldsymbol{X}) &= \nabla_{\boldsymbol{X}}\, {}^0\boldsymbol{x} = \boldsymbol{1} + \nabla_{\boldsymbol{X}}\, {}^0\boldsymbol{u} && \forall \boldsymbol{X} \in \Omega_0. \end{aligned} \tag{1}$$

Next, a microstructure (RUC) $\Theta_0 \subset \mathbb{R}^3$ consisting of microscale points $\boldsymbol{Y} \in \Theta_0$ is locally attached to each macroscale point $\boldsymbol{X} \in \Omega_0$ (see Fig. 1). In the first-order homogenization theory, the microscale position vector $\boldsymbol{x}$ and the deformation gradient $\boldsymbol{F}$ are assumed to be functions of both the macro and micro variables as

$$\begin{aligned} \boldsymbol{x}(\boldsymbol{X}, \boldsymbol{Y}) &= {}^0\boldsymbol{F}(\boldsymbol{X})\boldsymbol{Y} + \boldsymbol{u}(\boldsymbol{Y}) && \forall \boldsymbol{Y} \in \Theta_0, \\ \boldsymbol{F}(\boldsymbol{X}, \boldsymbol{Y}) &= {}^0\boldsymbol{F}(\boldsymbol{X}) + \nabla_{\boldsymbol{Y}}\boldsymbol{u}(\boldsymbol{Y}) && \forall \boldsymbol{Y} \in \Theta_0. \end{aligned} \tag{2}$$

In short, CH solves two nested nonlinear Boundary Value Problems (BVP). Neglecting inertial forces, the macroscale BVP is given by

$$\begin{aligned} \nabla_{\boldsymbol{X}} \cdot {}^0\boldsymbol{P} + {}^0\boldsymbol{b} &= \boldsymbol{0} && \text{in } \Omega_0, \\ {}^0\boldsymbol{P} \cdot {}^0\boldsymbol{N} &= \bar{\boldsymbol{t}} && \text{on } \partial\Omega_0^t, \\ {}^0\boldsymbol{u} &= \bar{\boldsymbol{u}} && \text{on } \partial\Omega_0^u, \end{aligned} \tag{3}$$

where ${}^0\boldsymbol{P} = \partial\, {}^0W / \partial\, {}^0\boldsymbol{F}$ is the macroscopic first Piola–Kirchhoff stress tensor, ${}^0W$ is the hyperelastic macroscopic strain energy density function, ${}^0\boldsymbol{b}$ is the macroscale body force, and ${}^0\boldsymbol{N}$ is the unit normal to the surface $\partial\Omega_0^t$. After neglecting body forces and without prescribed tractions, the microscale equilibrium equation reads

$$\nabla_{\boldsymbol{Y}} \cdot \left({}^1\boldsymbol{P}\right) = \boldsymbol{0} \quad \text{in } \Theta_0, \tag{4}$$

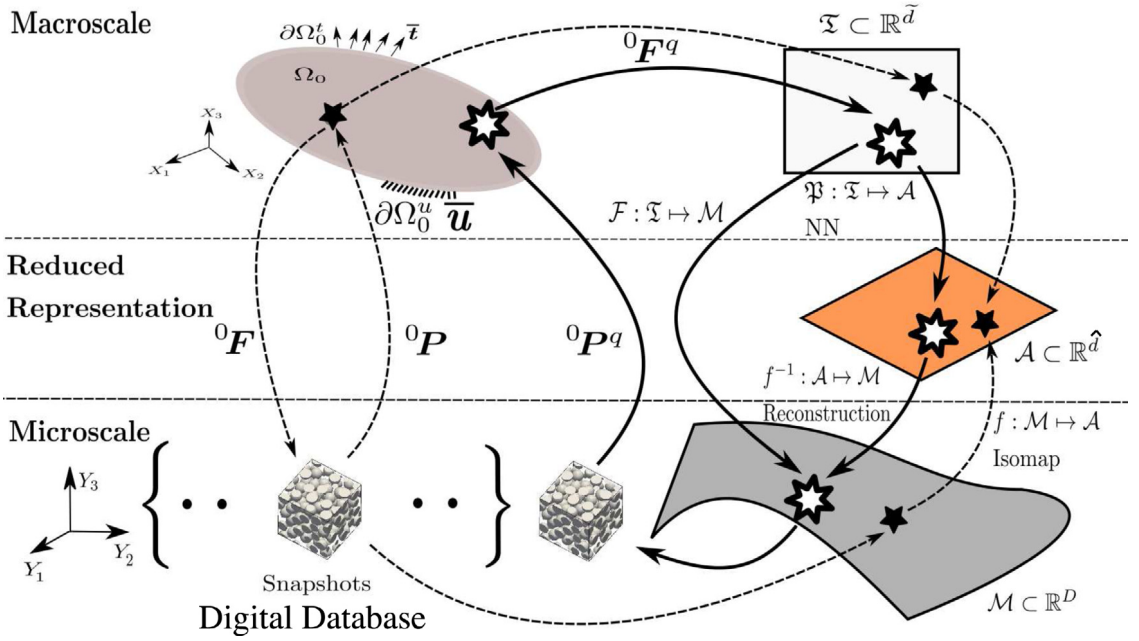

Fig. 1. Schematic of CH and MNROM: Dashed lines represent solution paths for snapshots, and solid lines show solution paths for reduced analysis of query points (i.e., points not in the database). The digital database (i.e., snapshots) is constructed in the offline stage. The data flow represented by solid lines also shows utilization of MNROM in the FE^2 setting.

where $^1\boldsymbol{P} = \partial\, ^1W/\partial \boldsymbol{F}$ is the microscopic first Piola–Kirchhoff stress tensor. The microscopic free energy density function, ^1W, is needed to close the constitutive equation across the scales. Moreover, the microscopic equilibrium equation (4) requires boundary conditions. In CH, the behavior of a material point at the macroscale is linked to the microscale through the Hill–Mandel stationarity condition [5]:

$$^0W(^0\boldsymbol{u}) = \inf_{\boldsymbol{u}} \frac{1}{|\Theta_0|} \int_{\Theta_0} {}^1W(^0\boldsymbol{F} + \nabla_Y \boldsymbol{u})\, \mathrm{d}\Theta. \tag{5}$$

The above equation implicitly suggests that the fine-scale deformation only influences the coarse scale behavior through its volume average over the RUC. This entails that the boundary conditions prescribed on the RUC must satisfy certain restrictions [11]. Periodic boundary conditions are a natural choice and are used in this work [8].

2.2. Computational homogenization from a manifold perspective

In this section, the macro and microscale coupling, (see Fig. 1), is summarized from a nonlinear manifold point of view (MNROM) as discussed in [16]. Since the entire quality of any data driven modeling rests on the ensemble nature of the data, the grand challenge lies in how to construct a representative manifold, $\mathcal{M}$, from a limited number (N) of data points. This is an emphasis of this work.

In this work, we propose a novel physics-driven adaptive technique to construct the solution manifold by sampling only non-redundant simulations. In contrast to our original work [16], where the deformation gradient was used, here a nonlinear manifold is constructed using the microscale displacement field which is also referred to as a snapshot. This substantially reduces the dimensionality of the problem (i.e., $\boldsymbol{F}$ is a second order tensor (3×3) and $\boldsymbol{u}$ is a vector (3×1)). We note that the gradient operator is a linear function of the nodal displacements, and thus, both approaches are closely related. By vectorizing the snapshots (i.e., the RUC microscale simulations), we construct a high-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ (the gray region in Fig. 1) containing the set $\mathfrak{Z} = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N\}$. The point $\boldsymbol{\xi}^i$ on the manifold $\mathcal{M}$ is given by the ordered set of the microscale displacements at each node (discretized by the Finite Element Method (FEM), for example), $\boldsymbol{\xi}^i = \{\boldsymbol{u}^1, \ldots, \boldsymbol{u}^{N_n}\} \in \mathcal{M}$, where N_n is the number of nodes in the RUC. This set governs the dimension of the embedding/ambient space of the manifold $D = 3 \times N_n$. The proximity

of the points on the manifold is described by the distance metric $\mathcal{D}$ (in terms of the discrete L_2-norm), suggesting that the topological space $(\mathcal{M}, \mathcal{D})$ is a metric space, which is bounded, dense, complete and compact [16,43].

The manifold points are associated with the corresponding macroscale loading conditions, $\mathfrak{T} = \{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^N\} \subset \mathbb{R}^{\widetilde{d}}$, where $\boldsymbol{\eta}^i$ is a vector of loading/controlling parameters and $\widetilde{d}$ is the dimension. In this work, we use 3 principal stretches and 3 principal directions as loading/controlling parameters (i.e., $\widetilde{d} = 6$, see Section 3). Although mathematically there is a map $\mathcal{F} : \mathfrak{T} \mapsto \mathcal{M}$ (the elliptic Eq. (4)) that defines the manifold $\mathcal{M}$ parametrically in terms of the input space $\mathfrak{T}$, it is not a trivial task to construct this map analytically or even by deep learning [44], especially when the map is extremely nonlinear and involves very high dimensionality. For well-posed problems [45], if the parameter space $\mathfrak{T}$ is compact, the manifold $\mathcal{M}$ is also compact and inherits the topological dimension of the parameter space. We construct it as a composition of two maps f^{-1} and $\mathfrak{P}$ ($\mathcal{F} = f^{-1} \circ \mathfrak{P}$).

Thus, MNROM consists of three individual components: (i) dimension reduction (map f), (ii) linkage of the parameter space to the reduced space (map $\mathfrak{P}$), and (iii) reconstruction of the microscale solution (map f^{-1}) [16].

For dimension reduction, we apply Isomap [42] as we idealize the data on the nonlinear manifold. This idealization is consistent since the displacement field is nonlinear over the macroscopic loading space. Therefore, our data reduction problem statement is as follows: given a set of N-unordered points (i.e., snapshots) belonging to a manifold $\mathcal{M}$ embedded in a high-dimensional space $\mathbb{R}^D$, find a low-dimensional region $\mathcal{A} \subset \mathbb{R}^{\widehat{d}}$ that is isometric to $\mathcal{M}$, where $\widehat{d} \ll D$. Isomap globally unfolds the manifold and returns a Euclidean space, $\mathcal{A} = \{\boldsymbol{\zeta}^1, \ldots, \boldsymbol{\zeta}^N\} \subset \mathbb{R}^{\widehat{d}}$, where $\boldsymbol{\zeta}^i$ is a vector of reduced/latent variables and $\widehat{d}$ is the intrinsic dimension of the manifold. Because the reduced space $\mathcal{A}$ is isometric to $\mathcal{M}$, the geodesic distances on manifold $\mathcal{M}$ are approximately preserved as the Euclidean distances on $\mathcal{A}$. This allows us to map vectors $\boldsymbol{\xi}$ (i.e., N vectors of dimension D) in $\mathcal{M}$ to vectors $\boldsymbol{\zeta}$ (i.e., N vectors of dimension $\widehat{d}$) in $\mathcal{A}$ more accurately than in other methods (e.g., POD).

Next, we construct the relationship between the loading space $\mathfrak{T}$ (i.e., a space containing a set of loading conditions) and the reduced space $\mathcal{A}$ (i.e., a space containing a set of reduced/latent variables) by a NN. We refer to this map as $\mathfrak{P}$.

Finally, we complete the correspondence between the manifold and the reduced space by constructing an appropriate inverse map, f^{-1}, by nonparametric regression. Note that the inverse map is exact only when the number of data points $N \to \infty$. Thus in what follows, f^{-1} represents the approximate inverse and is constructed from a finite set of points.

This concludes the three individual maps/components of MNROM: (i) the dimension reduction where vectors $\boldsymbol{\xi} \in \mathcal{M} \subset \mathbb{R}^D$ are mapped to vectors $\boldsymbol{\zeta} \in \mathcal{A} \subset \mathbb{R}^{\widehat{d}}$ by Isomap, *(ii)* the link between generating parameters $\boldsymbol{\eta} \in \mathfrak{T} \subset \mathbb{R}^{\widetilde{d}}$ and the low-dimensional space $\boldsymbol{\zeta} \in \mathcal{A} \subset \mathbb{R}^{\widehat{d}}$ by NN, and *(iii)* the reconstruction of the solution, which is the inverse process to dimension reduction. Fig. 1 shows the macroscopic level with the space of generating parameters $\mathfrak{T}$, the reduced order level with the space $\mathcal{A}$ of latent variables as well as the microscopic level with simulation data (i.e., the manifold $\mathcal{M}$). Moreover, the information flow is highlighted. Discussion about possible non-uniqueness of the deformation gradient and about incompressible solids can be found in [16].

2.2.1. Isomap and kernel isomap

As mentioned in the previous section, we establish the map f using Isomap, which assumes that $\mathcal{M} = f^{-1}(\mathcal{A})$ is globally isometric to $\mathcal{A}$ [42]. In the limit, $N \to \infty$, the data representation is a smooth manifold $\mathcal{M}$ without holes, as the parametric space $\mathfrak{T}$ is without singularity (i.e., the loading conditions are finite). Since f is globally isometric and $\mathcal{M}$ is compact and without holes, the Euclidean space $\mathcal{A}$ is convex. For the smooth manifold $\mathcal{M}$, we can define the geodesic distance, $D_M(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j)$, between all pair of points $\boldsymbol{\xi}^i$ and $\boldsymbol{\xi}^j$ where $i, j = 1 \ldots N$ [46]:

$$D_M(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) = \inf_{\gamma}\{\text{length}(\gamma) : \gamma : [0,\ 1] \mapsto \mathcal{M} \text{ and } \gamma(0) = \boldsymbol{\xi}^i, \gamma(1) = \boldsymbol{\xi}^j\}. \tag{6}$$

Note that the length of the arcs in the equation above is defined using the distance metric $\mathcal{D}$.

Isomap exploits the concept of the distance kernel trick with the geodesic kernel $\mathfrak{K}(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) = -D_M^2(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j)$ that is conditionally positive definite (CPD) and returns the parameter space $\mathcal{A}$ [47]. The manifold is represented by the graph, G, constructed from the points, $\boldsymbol{\xi}^i$, on the manifold $\mathcal{M}$ using the k-rule, where k nearest neighbors are connected through the graph edges. What points are neighbors in the graph G construction is determined based on the pair-wise distances $\mathcal{D}$ (i.e., the discrete L_2 measure). Next, we construct the neighborhoods for each $\boldsymbol{\xi}^i$. There are several methods to construct the k-neighborhood (e.g., local linear interpolation [43]). In this work, we use both

kernel Isomap with the reproducing kernel map construction as in our previous work [16] and regular Isomap with the k-neighborhood local linear interpolation.

Isomap involves the computation of an affinity matrix, $\overline{D}_{ij} = D_G^2(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j)$, from the dataset $\mathfrak{Z}$. Here, $D_G(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j)$ is the approximate geodesic computed from the graph, G [48]. Then, the matrix $\overline{D}$ is transformed into a positive semi-definite (PSD) double-centered matrix, $\widetilde{D}$, as follows

$$\widetilde{D} = -\frac{1}{2} H \overline{D} H, \tag{7}$$

where $H = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ is a centering matrix, I is the $N \times N$ identity matrix, and $\mathbf{1}$ is a N-vector with all elements equal 1. Note, $\widetilde{D}$ is a PSD Gram matrix [47].

To construct the reduced space $\mathcal{A}$ (i.e., a space containing a set of reduced/latent variables, $\boldsymbol{\zeta}$), Isomap utilizes Merecer's theorem [49]: let $\mathcal{M} \subset \mathbb{R}^D$ be closed with a strictly positive Borel measure on $\mathcal{M}$, and $\mathfrak{K}$ be a continuous function on $\mathcal{M} \times \mathcal{M}$ which is positive definite (PD) and square integrable, then,

$$\mathfrak{K}(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) = \sum_{k=1}^{\infty} \sigma_k \boldsymbol{v}^k(\boldsymbol{\xi}^i) \cdot \boldsymbol{v}^k(\boldsymbol{\xi}^j), \tag{8}$$

where σ_k is the kth eigenvalue and $\boldsymbol{v}^k$ is the corresponding eigenvector of $\widetilde{D}$.

We note that the series converges absolutely for each pair $(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) \in \mathcal{M} \times \mathcal{M}$ and uniformly on each compact subset of $\mathcal{M}$. Mercer's theorem still holds pointwise if $\mathcal{M}$ is a finite set $\mathfrak{Z}$. Exploiting Eq. (8), the embedding is constructed by

$$\zeta_k^i = \sqrt{\sigma_k}\, v_i^k, \tag{9}$$

Here, ζ_k^i is the kth component of $\boldsymbol{\zeta}^i$ out of $\widehat{d}$ components (i.e., $k = 1, \ldots, \widehat{d}$), and v_i^k is the ith component of $\boldsymbol{v}^k$ out of N components (i.e., $i = 1, \ldots, N$). Truncation in the decaying spectrum to compute $\widehat{d}$ is done by applying the Beardwood–Halton–Hammersley (BHH) theorem [50]. Once $\widehat{d}$ is computed, embedding can be constructed by taking the first $\widehat{d}$ eigenvalues and eigenvectors. This will create the low-dimensional space $\mathcal{A}$. It can be shown that convergence takes place asymptotically for an isometric map.

One of the major components of Isomap is the computation of approximate geodesic distances in terms of all-pair shortest paths. We use Dijkstra's algorithm [51] to compute the shortest path. The approximate geodesic distances computed over the graph G obey the following theorem: let $\mathcal{M}$ be a compact manifold, embedded in $\mathbb{R}^D$, isometrically equivalent to a convex domain $\mathcal{A} \subset \mathbb{R}^{\widehat{d}}$ and let some parameters ω_1, ω_2 and μ be given in $(0,\ 1)$ (i.e., $0 < \omega_1, \omega_2 < 1$ and $0 < \mu < 1$). Neglecting the boundary effects, we can assert with probability at least $1 - \mu$ that the following inequalities

$$(1 - \omega_1) D_M(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) \leq D_G(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) \leq (1 + \omega_2) D_M(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j), \tag{10}$$

hold on $\mathcal{M}$ for all $\boldsymbol{\xi}^i, \boldsymbol{\xi}^j$. Eq. (10) guarantees the asymptotic convergence of the computed approximate geodesics in limiting conditions ($N \to \infty$). See [43,48] for more details.

The inequality (10) indicates the possibility of noise associated with the computed geodesic distances. This implies that the Gram matrix, $\widetilde{D}$, may not be PSD and violates the fundamentals of the kernel operations. Although Isomap replaces the negative eigenvalues with zeros, which is mathematically sound, this can underestimate the dimension. To circumvent this situation, a kernel version of the Isomap algorithm has been proposed in [52]. The kernel version employs the analytical framework of the additive constant problem and transforms the kernel into a PD one [53]. Although topological instability can be an issue for Isomap [54], we can circumvent it by carefully constructing a finite number of disjoint sub-manifolds and analyze those separately [16].

2.2.2. Greedy algorithm

Isomap is an expensive algorithm, since we need to compute all the shortest paths (approximate geodesics) and carry out the eigenvalue analysis. However, once the manifold reaches a certain density, we would like to enrich the manifold and the corresponding reduced space further very quickly. This enrichment should allow adding new snapshots to the manifold without recomputing all existing geodesics (i.e., $\mathcal{O}(N^2)$ algorithm). Therefore, to avoid

the computational cost, we apply a greedy algorithm to project any new snapshot, $\boldsymbol{\xi}^{new}$, added to the manifold by projecting it to the current reduced space (in terms of the existing eigenvalues and eigenvectors) as follows [41]:

$$\zeta_k(\boldsymbol{\xi}^{new}) = \frac{1}{2\sqrt{\sigma_k}} \sum \boldsymbol{v}_k^i (\mathbb{E}[\widetilde{D}^2(\boldsymbol{\xi}^{new}, \boldsymbol{\xi}^i)] - \widetilde{D}^2(\boldsymbol{\xi}^{new}, \boldsymbol{\xi}^i)). \tag{11}$$

Here, we only need to compute the geodesic distances for the newly added point, $\boldsymbol{\xi}^{new}$, and $\mathbb{E}$ is the average operator. Note, the value of the geodesic distances can change drastically for a manifold that is not well populated. This implies that the accuracy of this greedy technique depends on the current density of the manifold data, and we will show this in Section 4.1. For more details on greedy algorithms see [55].

2.2.3. Neural network

The map $\mathfrak{P}$ is essentially the relationship between the macroscale loading parameter space $\mathfrak{T}$ and the reduced space $\mathcal{A}$ (see Fig. 1). We note that the connection between the reduced space, $\mathcal{A}$, and the space of physical macroscale parameters, $\mathfrak{T}$, is not straightforward for nonlinear problems. Because feed-forward neural network is a universal approximator [56,57], in this work, the map between the reduced space, $\mathcal{A}$, and the macroscale loading parameter space, $\mathfrak{T}$, is established through use of a NN by minimizing the mean-square error between the reduced space, $\mathcal{A}$, and the approximate output of the NN [56,58]. The NN architecture is discussed in [16].

One of the very common problems with NN is over-fitting [58,59]. To circumvent this issue, we have used Bayesian regularization [58,60,61]. Another important issue with NN is the possibility of getting stuck at the local minima. In this work, we use an ensemble NN to limit this problem [62].

2.2.4. Reconstruction of the microscale solution

The map f is globally isometric [42], which leads to a bijection since the geodesic distance on $\mathcal{M}$ is equal to the Euclidean distance on $\mathcal{A}$ and the distance between two distinct points cannot be zero. Additionally, the entire framework assumes that the embedding, f^{-1}, is a diffeomorphism. Although the isometric map can be constructed explicitly, in our work the map f is without an explicit form and does not have an explicit inverse. Moreover, the computed Euclidean distance does not exactly correspond to the geodesic distance over manifold $\mathcal{M}$ [16]. In order to predict the microstructural response from the reduced dimensional parameters, we numerically construct an approximation to the inverse map through regression [16,43,63,64]. In this work, we exploit the concept of Reproducing Kernel Hilbert Space ($RKHS$) to construct the inverse map, f^{-1} [65,66]. For every $RKHS$, there is a corresponding PD kernel. However, most radial functions do not correspond to a PD kernel [67]. The generalized version of $RKHS$ is Reproducing Kernel Krein Space ($RKKS$), where the associated kernel is CPD. Note, that the geodesic kernel, $\mathcal{K}(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) = -D_M^2(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j)$, is a CPD one [47]. This allows us to represent each component as [68,69]

$$\xi_j^q = \sum_{i=1}^{N} \theta_j^i \mathcal{K}(\boldsymbol{\xi}^i, \boldsymbol{\xi}^q). \tag{12}$$

Here, ξ_j^q is the jth component of the solution vector ($\boldsymbol{\xi}^q$) corresponding to its reduced space representation $\boldsymbol{\zeta}^q$. The reproducing kernel, $\mathcal{K}(\boldsymbol{\xi}^i, \boldsymbol{\xi}^q) = -D_M^2(\boldsymbol{\xi}^i, \boldsymbol{\xi}^j) = -\|\boldsymbol{\zeta}^i - \boldsymbol{\zeta}^q\|_{L_2}$, is a scale-free radial basis function (RBF). Note that here, we represent the L_2 norm in the discrete sense. The coefficients θ_j^i are determined using the training dataset ($\boldsymbol{\zeta}^i, \boldsymbol{\xi}^i$) as follows

$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_D^1 \\ \vdots & \ddots & \vdots \\ \theta_1^N & \cdots & \theta_D^N \end{bmatrix} = \begin{bmatrix} \mathfrak{K}(\boldsymbol{\zeta}^1, \boldsymbol{\zeta}^1) & \cdots & \mathfrak{K}(\boldsymbol{\zeta}^1, \boldsymbol{\zeta}^N) \\ \vdots & \ddots & \vdots \\ \mathfrak{K}(\boldsymbol{\zeta}^N, \boldsymbol{\zeta}^1) & \cdots & \mathfrak{K}(\boldsymbol{\zeta}^N, \boldsymbol{\zeta}^N) \end{bmatrix}^{-1} \begin{bmatrix} \xi_1^1 & \cdots & \xi_D^1 \\ \vdots & \ddots & \vdots \\ \xi_1^N & \cdots & \xi_D^N \end{bmatrix}. \tag{13}$$

This regression training requires one simple $N \times N$ matrix inversion followed by $N \times D$ matrix vector product operations and can be performed quickly even for large D. However, it is worth mentioning that the solution in $RKKS$ may not represent the global minima of the cost function [69]. Moreover, a Bayesian framework [70] is very expensive in the case of very high-dimensional regression, hence we avoid it in this work. The successful implementation of this technique for the reconstruction of an extremely high dimensional manifold can be found in [16].

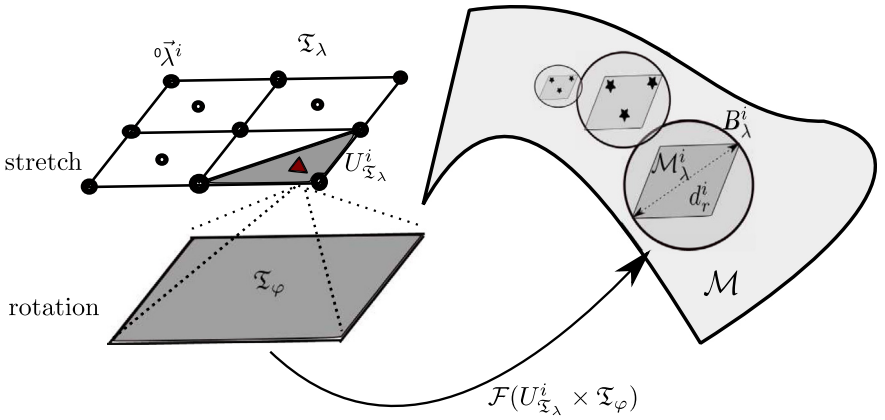

Fig. 2. Concept of the physics-based construction of the manifold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Physics-guided manifold construction

As discussed in the previous section (see Section 2.2), the importance of the representativeness of the manifold $\mathcal{M}$ is crucial. Therefore, we explore the snapshot patterns on the manifold $\mathcal{M}$. This novel sampling strategy substantially reduces the number of snapshots (i.e., the RUC microscale simulations by the FEM) needed for accurate MNROM representation. Moreover, it can be used in several similar model reduction schemes. In this work, we limit ourselves to statistically isotropic hyperelastic solids, and decompose the microscopic free energy density function in Eq. (5) into its deviatoric and volumetric parts, ${}^1W = {}^1W_{\mathrm{C}}(\widehat{\boldsymbol{C}}) + {}^1W_{\mathrm{J}}(J)$. Here, $J = \det(\boldsymbol{F})$ is the Jacobian of the deformation gradient and $\widehat{\boldsymbol{C}} = J^{-2/3}\boldsymbol{C}$ is the deviatoric right Cauchy–Green deformation tensor, where $\boldsymbol{C} = \boldsymbol{F}^T\boldsymbol{F}$.

First, we would like to highlight the parameter space, $\mathfrak{T}$, which stores the macroscopic deformation gradients ${}^0\boldsymbol{F}$. Using the polar decomposition, ${}^0\boldsymbol{F} = {}^0\boldsymbol{R}\,{}^0\boldsymbol{U}$, we obtain

$$ {}^0\boldsymbol{U} = {}^0\lambda_1({}^0\boldsymbol{e}^1 \otimes {}^0\boldsymbol{e}^1) + {}^0\lambda_2({}^0\boldsymbol{e}^2 \otimes {}^0\boldsymbol{e}^2) + {}^0\lambda_3({}^0\boldsymbol{e}^3 \otimes {}^0\boldsymbol{e}^3), \tag{14} $$

where ${}^0\lambda_I \in (0, \infty)$ are principal stretches and ${}^0\boldsymbol{e}^I$ are the orthogonal principal directions generated by three angles φ^I (the axis-angle representation), with $I = 1, 2, 3$. Axis is characterized by the angles $\varphi^1 \in [0, 2\pi)$, $\varphi^2 \in [0, \frac{\pi}{2})$, and the rotation about the axis $\varphi^3 \in [0, \frac{\pi}{2})$. Note that we assume an isotropic material, which makes the rotational space smaller. We construct the parameter space $\mathfrak{T}$ as a Cartesian product of two spaces: (i) the stretch space $\mathfrak{T}_\lambda$ and (ii) the rotational space $\mathfrak{T}_\varphi$ as depicted in Fig. 2. In our numerical implementation, stretches ${}^0\vec{\lambda}$ are from a closed interval, and thus, the parameter space, $\mathfrak{T}$, is a compact subset of Euclidean space. The objective of this analysis is to better understand the distribution pattern for the rotational parameters.

3.1. Rotational sensitivity

A concept of the rotational sensitivity analysis is illustrated in Fig. 3. In this analysis, the stretch space is a cube with the line of volumetric deformations (the dotted line in Fig. 3) as its main diagonal. All the points on this line manifest pure volumetric deformation (${}^0\lambda_1 = {}^0\lambda_2 = {}^0\lambda_3 = {}^0\lambda$), because ${}^0\boldsymbol{U} = {}^0\lambda\boldsymbol{1}$. From the physical perspective regardless of the constitutive model, the pure volumetric deformations are invariant of rotation. Therefore, if the stretch vector is near the volumetric deformation line, it will show less rotational dependency. Moreover, the volumetric part of the energy density, ${}^1W_{\mathrm{J}}(J)$, is typically more dominant in the high stretch regime (i.e., $J \ll 1$ or $J \gg 1$) [71]. When performing the rotational sensitivity analysis, we first cover the space $\mathfrak{T}_\lambda$ with the finite number (^{i}) of open sets $U^i_{\mathfrak{T}_\lambda}$ (see Fig. 2). Thus, the parameter space $\mathfrak{T}$ can be represented as

$$ \mathfrak{T} = (\cup U^i_{\mathfrak{T}_\lambda}) \times \mathfrak{T}_\varphi = \cup(U^i_{\mathfrak{T}_\lambda} \times \mathfrak{T}_\varphi). \tag{15} $$

Since the manifold $\mathcal{M}$ is compact, the sub-manifolds, $\mathcal{M}^i_\lambda$, can be defined as

$$ \mathcal{M}^i_\lambda = \mathcal{F}(U^i_{\mathfrak{T}_\lambda} \times \mathfrak{T}_\varphi), \tag{16} $$

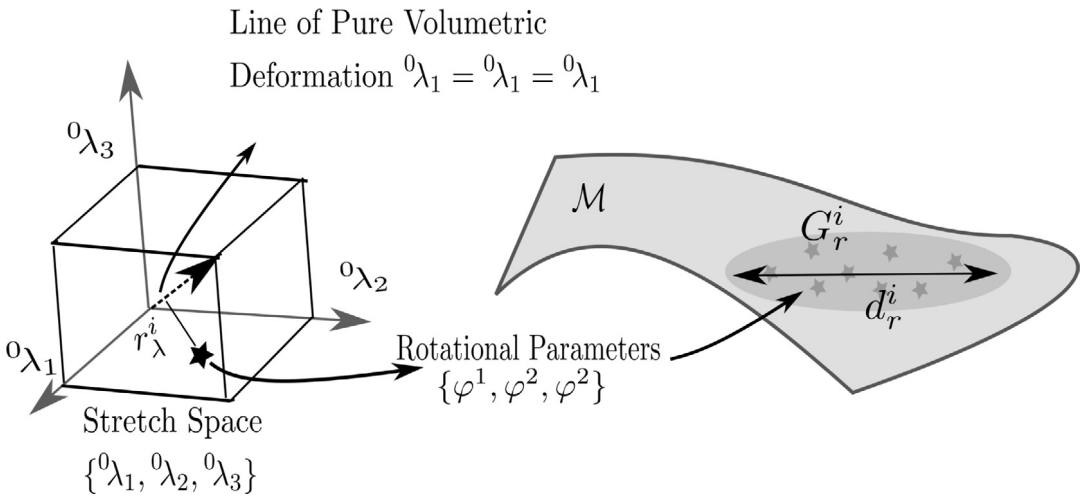

Fig. 3. Concept of the rotational sensitivity analysis.

where $\mathcal{M} = \bigcup \mathcal{M}^i_\lambda$. Note that these sub-manifolds also have the same dimension, $\widehat{d}$, because of the manifold $\mathcal{M}$ smoothness. In the discrete setting, sub-manifolds are constructed as a subset of points with all rotations for a specific stretch vector ${}^0\vec{\boldsymbol{\lambda}}^i$ (see Fig. 2). Therefore, the discrete space $\mathcal{M}^i_\lambda$ reads

$$\mathcal{M}^i_\lambda = \mathcal{F}({}^0\vec{\lambda}^i \times \mathfrak{T}_\varphi), \tag{17}$$

for the finite number of stretch points. Note that the number of generating parameters for $\mathcal{M}^i_\lambda$ is 3, which implies $\mathcal{M}^i_\lambda$ is a 3-manifold. Here we construct the subgraph G^i_r, which graphically represents the sub-manifold, $\mathcal{M}^i_\lambda$ (see Fig. 3). The graph diameter, d^i_r, represents the diameter of the minimum covering ball, $\mathcal{B}^i_\lambda$, for the corresponding sub-manifold $\mathcal{M}^i_\lambda$ (see Fig. 2). The graph diameter is computed simply as the longest shortest path (i.e., the longest graph geodesic) in the graph [72].

Physically, it is expected that the rotational dependency will increase with the perpendicular distance, r^i_λ, from any point on the line of volumetric deformation (see Fig. 3). It is also important to note that the effect is not the same for all volumetric deformations. As the volumetric strain energy increases, the graph diameter should be a function of the volumetric deformation (i.e., the third invariant of the Green–Lagrange strain tensor, $I_3 = \det(\boldsymbol{C}) = J^2$ or its square root) and the distance, r^i_λ, of the stretch vector from the corresponding pure volumetric point. In our implementation, we neglect this coupling.

Rotational sensitivity analysis allows for the optimal distribution of rotational points for the corresponding number of stretches. Lower diameter suggests less rotational dependency and higher diameter suggests higher rotational dependency of that particular point in the stretch space. From this relationship, we can decide the number of rotational points required for a particular point in the stretch space or equivalently construct the representative sub-manifold $\mathcal{M}^i_\lambda$. Now by utilizing the concept of rotational sensitivity analysis, we propose an adaptive data enrichment technique to represent the manifold by enriching the sub-manifolds and eliminating possible redundant points.

3.2. Adaptive snapshot redistribution

In the first step, we start with a few simulations, N_0. After determining all diameters, d^i_r, we learn as a function of the distance, r^i_λ. This function allows us to adaptively add extra points in the stretch space $\mathfrak{T}_\lambda$. To do so, we compute the average number of points, $\overline{n}_g$, in the ϵ neighborhood over all points in the entire dataset as follows

$$\overline{n}_g = \frac{1}{N}\sum_{i=1}^{N} n_{g_i}, \tag{18}$$

where n_{g_i} is the number of points within ϵ neighborhood centered on any point $\boldsymbol{\xi}^i$ on the manifold, $\mathcal{M}$. Next, we compute $\overline{n}_g^i$ for all data points associated with each closed domain, $U^i_{\mathfrak{T}_\lambda}$, for the same ϵ neighborhood. If $\overline{n}_g^i$ is less than $\max(\overline{n}_g, k)$, we add an extra stretch point to the center of $U^i_{\mathfrak{T}_\lambda}$ (shown as a filled triangle in the 2D representation in Fig. 2). This adaptive point insertion refines the stretch space (i.e., the initial uniform stretch cubical grid) without duplicating data as was possible in our previous blind sampling strategy [16]. Here we use local linear interpolation [43], where k is the k-nearest neighbor of the graph representation of the manifold, $\mathcal{M}$, and $\overline{n}_g$ measures the average data density. This strategy helps to add data points in the sparse regions of the manifold by increasing the average local data density, $\overline{n}_g^i$ to more than k and $\overline{n}_g$. In this work, we construct $U^i_{\mathfrak{T}_\lambda}$ using tetrahedral refinement. Note any extra points (N_e) will be added to the manifold, $\mathcal{M}$, by enriching the sub-manifolds, $\mathcal{M}^i_\lambda$. The number of required enrichment points is proportional to the volume of the sub-manifolds, $V_{\mathcal{M}^i_\lambda}$. To accomplish this purpose, first we compute the volume of these sub-manifolds. However, it is difficult to compute $V_{\mathcal{M}^i_\lambda}$ exactly from a few points. Therefore, we assume that the volume $V_{\mathcal{M}^i_\lambda}$ is proportional to the volume of its minimum covering ball [73], $\mathcal{B}^i_\lambda$, which yields

$$V_{\mathcal{B}^i_\lambda} = \frac{\pi^{\widehat{d}/2}}{\Gamma(\widehat{d}/2+1)}(d^i_r/2)^{\widehat{d}}. \tag{19}$$

Here, Γ is the gamma function and π is a mathematical constant. The number of sub-manifolds, N_λ, is the same as number of the points in the stretch space. The number of enrichment points N^i_φ in the rotational space, $\mathfrak{T}_\varphi$, for the corresponding stretch point is calculated as follows

$$N^i_\varphi = \left\lceil \frac{V^i_{\mathcal{M}_\lambda}}{\sum_{i=1}^{N_\lambda} V^i_{\mathcal{M}_\lambda}} N_e \right\rceil = \left\lceil \frac{V^i_{\mathcal{B}_\lambda}}{\sum_{i=1}^{N_\lambda} V^i_{\mathcal{B}_\lambda}} N_e \right\rceil. \tag{20}$$

Note that N^i_φ takes the next integer value as computed from Eq. (20). The tacit assumption behind the above algorithm is that the manifold $\mathcal{M}$ is the disjoint union of N_λ sub-manifolds ($\mathcal{M} = \bigsqcup \mathcal{M}^i_\lambda$). To verify this assumption, we compute the degree of overlap $\mathcal{P}_O$ as follows

$$\mathcal{P}_O = \left(\frac{V_I}{V_\mathcal{M}} - 1\right) \times 100\,[\%], \tag{21}$$

where V_I is the volume of the data associated with the interface of the closed domains, $U^i_{\mathfrak{T}_\lambda}$, with all attached rotations. Note that the interface regions are counted twice during volume calculations. We also note that once the manifold is adaptively built, the greedy algorithm proposed in Section 2.2.2 can be used to add points without complex geodesic computation and we will show this in Section 4.1.

4. Numerical examples

Numerical examples are performed on a randomly configured (statistically isotropic [74,75]) particulate hyperelastic material. We create the unit cell consisting of monodisperse spheres with radius R = 0.5 μm by using the packing algorithm *Rocpack* [76]. The size of the unit cell is selected based on statistical analysis, details of which are provided in [16]. The unit cell with a side length of $l_{RUC} = 10R = 5$ μm contains 95 particles and is identical to the unit cell in our previous work [16] for ease of comparison.

Fig. 4 shows the geometry of the cell used in this work. After a FEM convergence study, the RUC is discretized with 96,252 nodes and 486,051 elements. This discretization results in 250,035 nonlinear degrees of freedom to solve the micro-problem (see Eq. (4)). This sets the size of the dimension of the embedding/ambient space of the manifold, $\mathcal{M}$, and gives $D = 288,756$. As a constitutive model in the RUC, we use an isotropic hyperelastic strain–energy density function similar to the Mooney–Rivlin solid,

$$\begin{aligned} {}^1W_{\mathrm{C}}(\widehat{\boldsymbol{C}}) &= \mu_{10}\left[\mathrm{tr}(\widehat{\boldsymbol{C}}) - 3\right] + \frac{\mu_{01}}{2}\left[(\mathrm{tr}\widehat{\boldsymbol{C}})^2 - \widehat{\boldsymbol{C}} : \widehat{\boldsymbol{C}} - 6\right], \\ {}^1W_{\mathrm{J}}(J) &= \frac{\kappa}{2}\left[\exp(J-1) - \ln(J) - 1\right]. \end{aligned} \tag{22}$$

Here, μ_{10} and μ_{01} are the shear moduli and κ is the bulk modulus. The associated material properties of the microscale constituents are provided in Table 1. We note that more nonlinear models (e.g., visco-plasticity) will make the problem more complex. In this case, a new development will be needed to reduce both the nonlinear partial

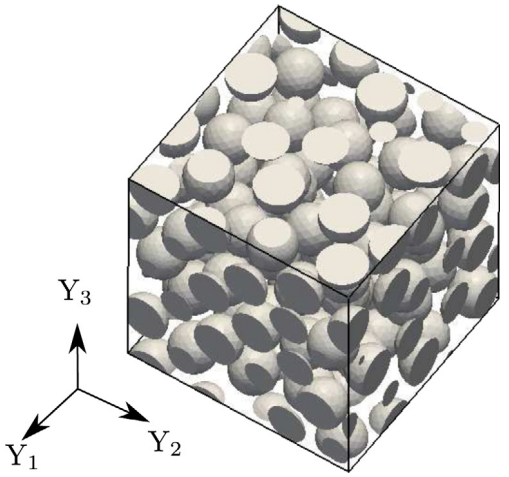

Fig. 4. RUC with the side length of $l_{RUC} = 10\text{R} = 5$ μm. The cell contains 95 hard spherical particles with 1 μm diameter. The unit cell is taken from our previous work [16] for ease of comparison.

Table 1
Mooney–Rivlin material properties.

Material	E [MPa]	ν	κ [MPa]	G [MPa]	μ_{01} [MPa]	μ_{10} [MPa]
Particle	5e3	0.25	3.33e3	2e3	5e2	5e2
Matrix	50	0.35	55.56	18.52	4.63	4.63

differential equation (PDE) and the nonlinear material model (i.e., nonlinear ordinary differential equations (ODEs)). The snapshots are generated using the high-performance $PGFem3D$ solver[1] that performs FEM operations in parallel with a high degree of efficiency [4,29]. Such parallelism is important for both the database construction as well as for fast re-localization of a new solution vector.

4.1. Physics-guided learning and manifold enrichment

We are interested in developing a microstructural ROM (i.e., MNROM for Eq. (4)) in the context of CH (see Fig. 1). The RUC is loaded by an arbitrary macroscale deformation gradient, ${}^0\boldsymbol{F}$, which consists of the principal stretches and directions (see Section 3, Eq. (14)). For our analysis, we take 10% maximum principal stretch in each direction (i.e., 30% maximum stretch) and all possible principal rotations as the loading envelope. This leads to a large design space of solutions of the nonlinear PDE without assumptions about the particular loading (e.g., shear only or tension only). This broad loading envelope is important for industrial design where loading conditions vary from point to point at the macroscale. We note that CH is a highly nonlinear process (i.e., depending on the free energy function, ${}^1W$). Moreover, two different loading conditions in $\mathfrak{T}$ (i.e., different prescribed principal stretches and rotations in Eq. (14)) can produce similar ${}^0\boldsymbol{U}$ and lead to similar microscale solutions or snapshots. Similarity of the microscale data can lead to misrepresentation of the manifold, which is referred to as topological instability [54]. Note that we do not consider loss of ellipticity and bifurcation in this work. As discussed in our previous work [16], capturing limit and bifurcation points requires a special numerical treatment and MNROM would have to be constructed and analyzed carefully for each bifurcation branch. However, morphological imperfections are common when working with random materials and will guide the solution path. Moreover, we use displacement driven CH where the macro-deformation is uniquely prescribed.

To avoid any potential topological instability, we divide our parameter space to four modes based on the loading conditions. Here, mode 1 is purely tensile and mode 4 is purely compressive. On the other hand, modes 2 and 3 are mixed ones as described in [16]. Next, we construct each of these manifolds separately for our reduced order

[1] https://github.com/C-SWARM.

Table 2
Description of the macroscale loading modes. Recall that N is the number of FEM snapshots used to construct the manifold $\mathcal{M}$.

Loading case	Description	Step 0	Step 1	Step 2
Mode 1	${}^0\lambda_1 \geq 1,\ {}^0\lambda_2 \geq 1,\ {}^0\lambda_3 \geq 1$	405	530	730
Mode 2	${}^0\lambda_1 \leq 1,\ {}^0\lambda_2 \geq 1,\ {}^0\lambda_3 \geq 1$	405	1133	2770
Mode 3	${}^0\lambda_1 \leq 1,\ {}^0\lambda_2 \leq 1,\ {}^0\lambda_3 \geq 1$	405	1060	2471
Mode 4	${}^0\lambda_1 \leq 1,\ {}^0\lambda_2 \leq 1,\ {}^0\lambda_3 \leq 1$	405	524	790
Total		$N = 1620$	$N = 3247$	$N = 6761$

model. In doing so, we implement our novel physics-guided sampling technique for the performance enhancement of the ROM. This technique is deterministic and sequential in nature.

To begin with our sequentially adaptive sampling technique, initially (step 0), we take a uniform grid ($3 \times 3 \times 3$) to create the stretch space, $\mathfrak{T}_\lambda$. This leads to 27 points. Next, to construct the rotational space, $\mathfrak{T}_\varphi$, associated with each stretch point, we take 15 randomly generated (from the uniform distribution) well-spaced points. To create these well spaced points, we first generate 10,000 random points inside the rotational space and call this set $\mathfrak{T}^*_\varphi$. Next, we pick any point from $\mathfrak{T}^*_\varphi$ to get the first point of $\mathfrak{T}_\varphi$. Then, we use the concept of Hausdorff distance and pick the next point from $\mathfrak{T}^*_\varphi$, which produces the farthest distance from $\mathfrak{T}_\varphi$. We continue this process for specified times (i.e., 14 here) to generate 15 points to construct the rotational space. Since the parameter space is represented as a product of the stretch and rotational spaces ($\mathfrak{T} = \mathfrak{T}_\lambda \times \mathfrak{T}_\varphi$), the total number of initial snapshots (i.e., FEM simulations of RUC using Eq. (4)) for step 0 is 405 (see Table 2).

After the initial snapshot set is defined, we employ both regular and kernel Isomap to estimate the dimension of the manifold. Both Isomap representations unfold the manifold into 6-dimensional Euclidean space (i.e., $\widehat{d} = 6$), which is the dimension of the parameter space, $\mathfrak{T}$ (i.e., $\widetilde{d} = 6$). Note we construct our reduced space by regular Isomap and use local linear interpolation with k-nearest neighbors. We have used $k = 21$ based on comparison with kernel Isomap and numerical tests, as well as the errors we encountered while maintaining the algorithm's ability to make accurate predictions.

Although the number of enrichment steps is not limited in this study, we only show the results for two steps (step 1 and step 2). We have selected 2 steps to show the convergence of the algorithm, but only the first step is needed to accurately distribute data as will be shown in Section 5. Also, we can enrich the manifold $\mathcal{M}$ with any number of points in each step. Here, we double the total number of snapshots in every step to understand the convergence properties of the enrichment (see Table 2). The maximum number of simulations/snapshots is $N = 6761$ (i.e., step 2), but we will show later that such manifold density is not needed and step 1 enrichment with $N = 3247$ is sufficient. In each step, we compute the volume of each manifold corresponding to each mode (See Eq. (19)) and proportionally distribute the added points among these four manifolds. After determining the number of extra points in each mode, we first enrich the stretch space $\mathfrak{T}_\lambda$ by exploiting a density analysis as described in Fig. 2 (red marker). Then, we employ a rotational sensitivity analysis and decide the number of rotational points for each stretch vector, which is proportional to the volume of the sub-manifolds, $\mathcal{M}^i_\lambda$ (see Eq. (20)).

For the first enrichment step (step 1), we conduct a density analysis on step 0 dataset as described in Section 3.2. We compute $\overline{n}_g$, from Eq. (18) on the entire dataset for a particular ϵ-neighborhood. We have considered $\epsilon = (1/4)\, d_G$ for all modes. Here, d_G is the diameter of the complete graph corresponding to the whole manifold, $\mathcal{M}$. Next, we consider each closed domain, $U^i_{\mathfrak{T}_\lambda}$, in the stretch space (see Fig. 2). We construct the ith stretch space grid by tetrahedral elements covering $U^i_{\mathfrak{T}_\lambda}$. We compute $\overline{n}^i_g$ by taking the data associated with the ith element (see the gray triangle in the stretch space in Fig. 2) with all rotations. If $\overline{n}^i_g$ is less than $\max(\overline{n}_g, k)$, we add an extra stretch point in the center of that element (see added red data point in the stretch space in Fig. 2). This ensures that after the enrichment step, the sparse regions in the manifold have at least k neighboring points (in an average sense, in the ϵ-neighborhood). This completes the stretch space enrichment for step 1.

Next, we explore the pattern using rotational sensitivity analysis to distribute the rotational points in the newly constructed stretch space. To distribute the rotational points, we separate out the sub-manifolds, $\mathcal{M}^i_\lambda$. We approximate the sub-manifolds in terms of the subgraphs, G^i_r. All subgraphs initially contain 15 rotational points (step 0). Next, we compute the diameter, d^i_r, for each subgraph which is the longest shortest path (i.e., the longest graph geodesic) in the graph [72]. Although the diameter is a function of both r^i_λ and I^i_3, we notice the effect of I^i_3

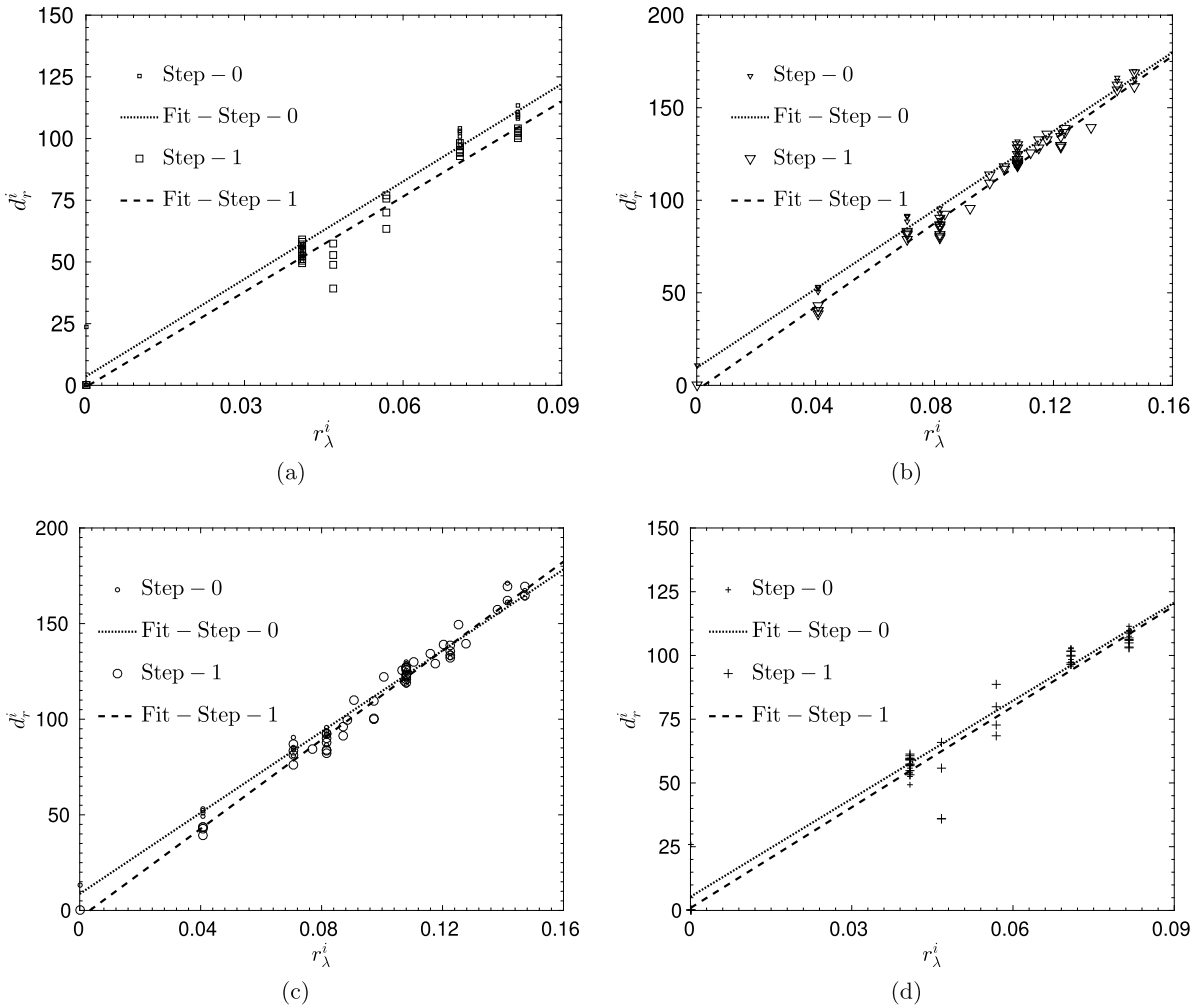

Fig. 5. Rotational sensitivity analysis with step 0 and step 1 data. (a) Mode 1. (b) Mode 2. (c) Mode 3. (d) Mode 4.

is not very significant. Therefore, we plot the graph diameters d_r^i as a function of r_λ^i in Fig. 5 for all four modes. Because the graph diameters are clustered around a line, we learn diameter as the linear least-square fit. The small scatter of the diameters of the sub-manifolds is due to the effect of the third invariant, I_3, and noise associated with the geodesic computations. Using this linear relationship obtained from data in Fig. 5, we estimate the diameter for any new stretch point and compute the volume by Eq. (19) with $\widehat{d} = 3$ (see the discussion in Section 3.1). Next we distribute new points among the sub-manifolds $\mathcal{M}_\lambda^i$ by Eq. (20). This completes the first enrichment process (step 1).

For the second enrichment step (step 2), we use the current dataset (step 1) with total 3247 snapshots (distributed among the modes, see Table 2) and repeat the process for further enrichment. Note that the second enrichment step is performed to assess the convergence of the algorithm only. As listed in Table 2, mode 2 and mode 3 take most of the extra points. On the other hand, mode 1 and mode 4 enrichment is not that significant. Note that mode 2 and mode 3 are mixed modes and more deviatoric in nature, which is consistent with the rotational sensitivity analysis (see Section 3.1). It is also indicative that the manifolds corresponding to mode 2 and mode 3 are more nonlinear than mode 1 and mode 4, respectively. To summarize our novel physics-guided enrichment strategy, we provide a complete algorithm that entails all steps discussed above (see Algorithm 1).

To understand the quality of the enrichment scheme for two steps, in Fig. 6, we present the distribution of the number of points, n_{gi}, within ϵ-neighborhood (density) with all steps for all four modes. In Fig. 6, we can

Algorithm 1 Adaptive Physics-guided MNROM

1: Start with a number of uncorrelated snapshots, N_0.
2: Compute all-pair geodesic/shortest paths (see Eq. (6)) and carry out a density distribution analysis (see Eq. (18)) to predict the number of stretch points required and their locations. The current number of stretch points is N_λ.
3: Carry out a rotational sensitivity analysis (see Section 3.1) and learn the diameter of the sub-manifolds, d_r^i, as a function of the parameter, r_λ^i.
4: From the rotational sensitivity analysis, determine the diameter of the sub-manifold corresponding to any stretch point (see Section 3.1). Next, compute the volume of those sub-manifolds by Eq. (19). Then, determine the number of the rotational points corresponding to each stretch point, given the total number of extra points by following Eq. (20).
5: Any further extra snapshot will be added to the database based on the pattern of the current enrichment state.

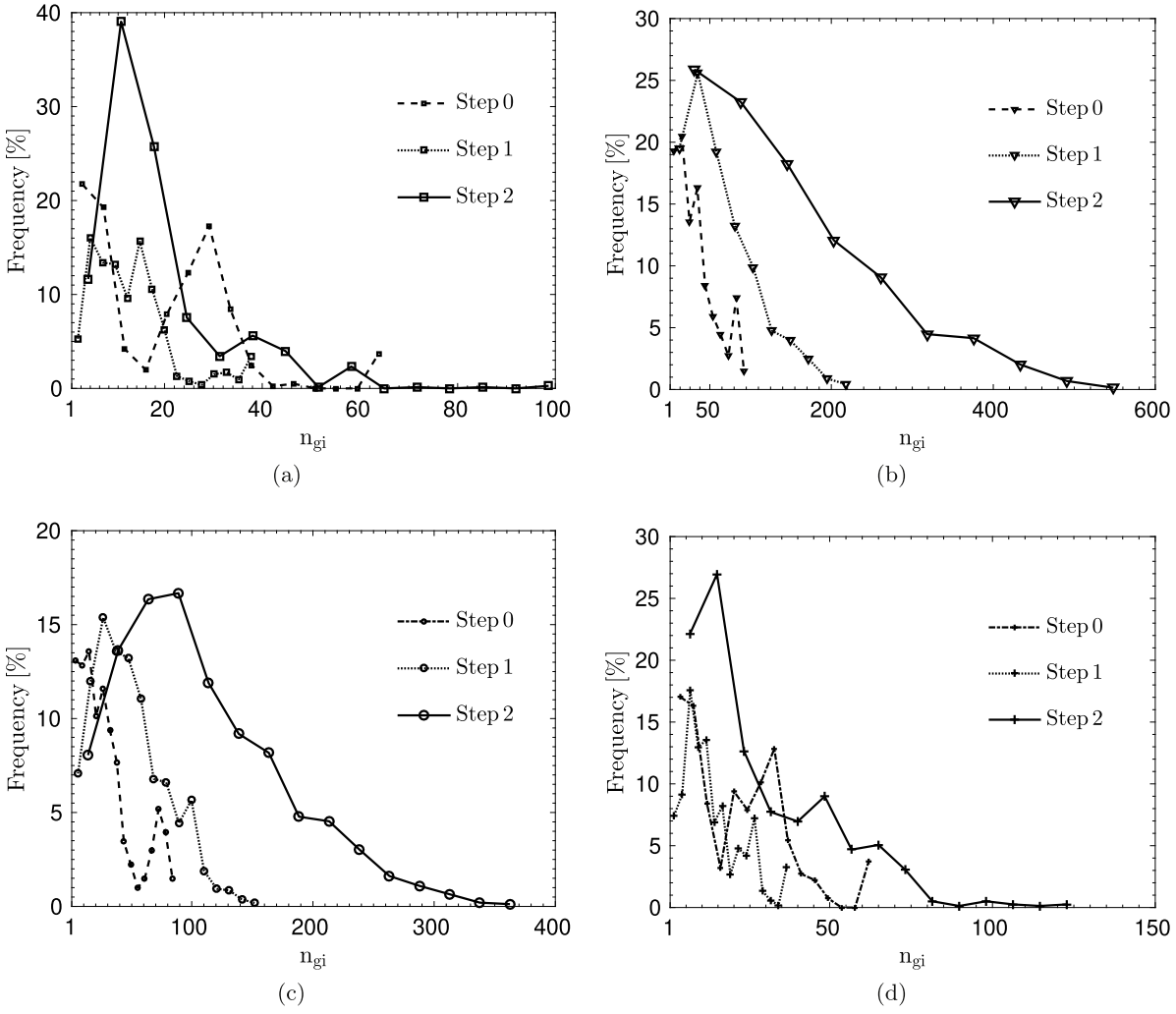

Fig. 6. Distribution of n_{gi} for all steps with all modes. The size of the markers indicates the enrichment step. (a) Mode 1. (b) Mode 2. (c) Mode 3. (d) Mode 4.

see that the distributions become smoother with this enrichment scheme. Moreover, the distributions have distinct tails, which indicates smaller data redundancy. Since the sparsity of each neighborhood decreases in the process,

Table 3
Percentage of the overlap, $\mathcal{P}_O$, for each mode and in each step.

	Step 0	Step 1	Step 2
Mode 1	3.65	3.99	4.49
Mode 2	2.49	2.66	2.66
Mode 3	2.52	2.62	2.72
Mode 4	3.67	3.88	4.64

in effect, it creates a higher possibility of capturing the local features/scales of the manifold. This is especially important for mixed modes 2 and 3, which manifest more nonlinearity. In the tensile mode (Mode 1), we see that most neighborhoods have around 10 points, and while the largest neighborhoods have around 100 points their frequency is very low. On the other hand, the more nonlinear mixed mode (Mode 2) requires around 50 points in the most frequent neighborhoods and the maximum neighborhood density approaches 600 points. This shows the importance of heterogeneous data distribution to capture the features of the full manifold.

In each step, we also compute the percentage of overlap, $\mathcal{P}_O$, for each mode (see Table 3). The percentage of overlap estimates the degree of redundancy of the data. Note the percentage of overlap remains almost unchanged with manifold enrichment. This suggests that the adaptive enrichment technique helps to distribute data on the manifold more evenly.

For our refinement strategy, we use Isomap at every enrichment step (i.e., we rebuild the geodesic distances for all points between enrichment steps). However, as mentioned previously, the Isomap algorithm is expensive. To avoid Isomap, a greedy algorithm is proposed in Section 2.2.2, where newly added points are projected into the existing reduced space $\mathcal{A}$. Therefore, when a new FEM unit cell simulation is performed, we compute the geodesic distance only for that one point and then project it to the reduced space using Eq. (11) without rebuilding the complete geodesic paths. This substantially improves the performance of the algorithm when new snapshots come online as in the FE^2 setting. To verify the greedy algorithm, we first take the step 2 dataset and construct the reduced space, $\mathcal{A}_2$, by using the regular Isomap. The subscript 2 indicates that the reduced space involves all the data in step 2. Similarly, we have constructed $\mathcal{A}_0$ and $\mathcal{A}_1$ with step 0 and step 1 datasets, respectively. Next, we project the extra snapshots in the step 2 dataset, which are not in the step 0 dataset by using Eq. (11) and denote this reduced space as $\mathcal{A}_2(0)$. In the same way, we construct $\mathcal{A}_2(1)$, where all extra snapshots are projected using the eigenvalues and eigenvectors computed from the step 1 data. To compare $\mathcal{A}_2$ with $\mathcal{A}_2(0)$ and $\mathcal{A}_2(1)$, we first vectorize these matrices into $\boldsymbol{\zeta}_2$, $\boldsymbol{\zeta}_2(0)$, and $\boldsymbol{\zeta}_2(1)$, respectively. Note the reduced space $\mathcal{A}$ is a $6 \times N$ matrix, $\widehat{d} = 6$. However, in the continuum sense ($N \to \infty$) these vectors can be viewed as the finite set of points constructed from the scalar field, $\widehat{\zeta}$. To verify the greedy algorithm, we plot the distribution of the scalar fields $\widehat{\zeta}_2$, $\widehat{\zeta}_2(0)$, and $\widehat{\zeta}_2(1)$ in Fig. 7. Fig. 7 shows that the enrichment points on the manifold have been projected to the reduced space satisfactorily as $\widehat{\zeta}_2(1)$ (projected) converges to $\widehat{\zeta}_2$ (directly computed). This indicates that the step 1 dataset is enriched enough and that any future snapshot can be projected to the reduced space, $\mathcal{A}_1$, by avoiding expensive Isomap computation of all-pair shortest paths and eigenvalue analysis.

Finally, we investigate the effect of the enrichment on the convergence of the physical fields, such as the deformation gradient. The local error in the physical fields (localization) is computed as follows

$$\mathcal{E}_r = \frac{\|\boldsymbol{F}_r^{ROM} - \boldsymbol{F}_r^{FEM}\|_{\mathcal{F}}}{\|\boldsymbol{F}_r^{FEM}\|_{\mathcal{F}}} \times 100\,[\%]. \tag{23}$$

Here, we use the Frobenius norm ($\|\bullet\|_{\mathcal{F}}$). The subscript r represents a material phase (particle or matrix). In order to quantify the cumulative error, the following volume averaged (homogenization) error, ${}^0\mathcal{E}_r$, reads

$${}^0\mathcal{E}_r = \frac{\frac{1}{\Theta_0}\int_{\Theta_r} \mathcal{E}_r \mathrm{d}\Theta}{\|{}^0\boldsymbol{F}_r\|_{\mathcal{F}}} \times 100\,[\%], \tag{24}$$

where ${}^0\boldsymbol{F}_r = \frac{1}{\Theta_0}\int_{\Theta_r} \boldsymbol{F}\, \mathrm{d}\Theta$.

For the error analysis, we first establish all the maps associated with the MNROM framework for step 0, step 1 and step 2 datasets. Next, we compute the microscopic deformations for 200 query points (i.e., the cells/loading

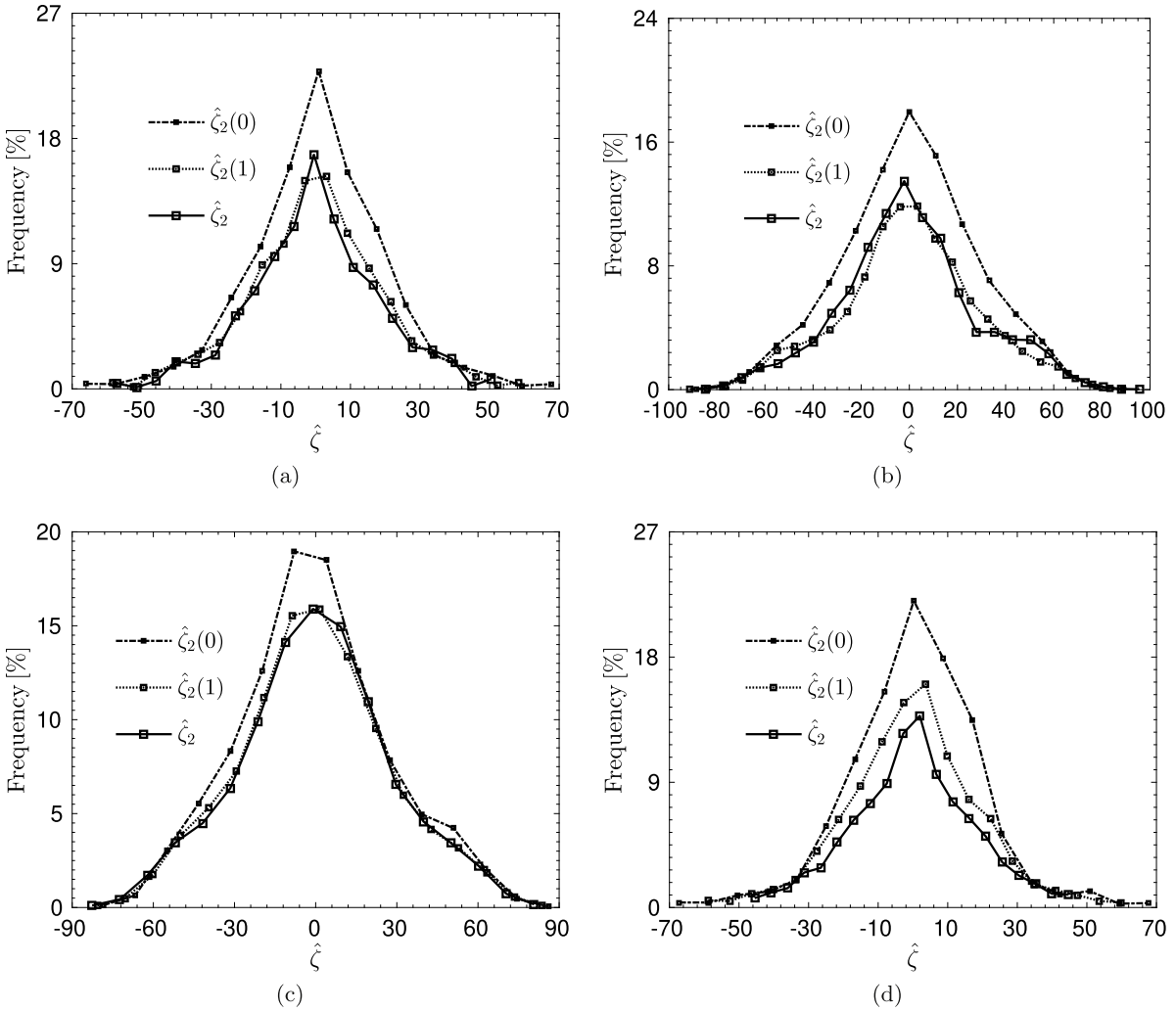

Fig. 7. Efficacy of the greedy algorithm: enrichment points in the step 2 dataset are projected to the step 0 and step 1 reduced spaces by Eq. (11). (a) Mode 1. (b) Mode 2. (c) Mode 3. (d) Mode 4.

conditions not in the database) in each mode. To ensure that the query points are well spaced and uncorrelated with the existing dataset, we first generate 10,000 samples from a uniform distribution in the macroscopic loading space and refer to this set as $\mathfrak{T}^*$. Then we pick the query point $\boldsymbol{\eta}^q$ from $\mathfrak{T}^*$ which produces the maximum Euclidean distance from the existing loading set, $\mathfrak{T}$. After picking $\boldsymbol{\eta}^q$ we remove this point from $\mathfrak{T}^*$ and add it to $\mathfrak{T}$. We iterate this process 200 times. Once this process ends, we take out all these query points from $\mathfrak{T}$ and construct the set of query points, $\mathfrak{T}^q$. After computing the displacement field for all the query points by the MNROM framework, we can determine the deformation gradient locally and compute the volume averaged error by Eq. (24) in comparison with the deformation gradient computed from the FEM framework (i.e., the numerical model of Eq. (4)). Next, we plot the mean and maximum errors in Fig. 8. We notice that in step 1 the maximum error has decreased to below 2 %. However, from step 1 to step 2 error decay has saturated.

Considering the enrichment result presented in this section, we proceed in further studies with the dataset from step 1 (i.e., 3247 snapshots) since this manifold is well populated and covers all loading conditions, ${}^0\boldsymbol{F}$, to approximate solutions of Eq. (4) within 10% principal stretch. This constitutes a large reduction in the number of required samples relative to our previous work [16], where we used 16128 simulations to construct the manifold.

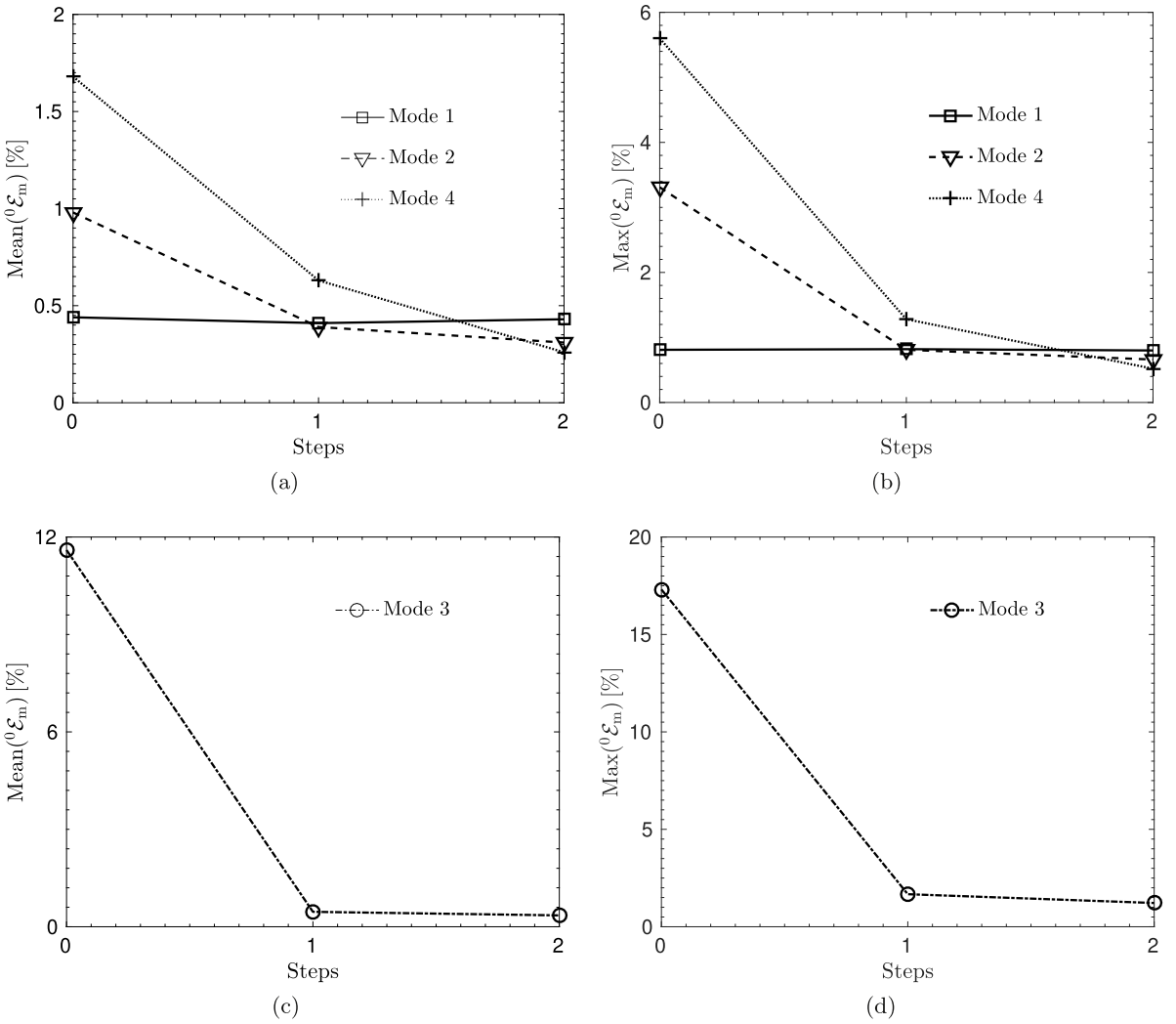

Fig. 8. Convergence of the adaptive data enrichment scheme in terms of the microscopic deformation gradient field. (a) Matrix-mean for Modes 1, 2 and 4. (b) Matrix-max for Modes 1, 2 and 4. (c) Matrix-mean for Mode 3. (d) Matrix-max for Mode 3.

5. Verification of the complete MNROM

In this section, we provide the verification of each map associated with MNROM individually with the step 1 dataset (i.e., we use only 3247 snapshots for the full manifold construction). Subsequently, we also present the quality of the solution predicted for out-of-data macroscopic loading condition (query points). The overall quality of MNROM is the cumulative quality of the individual maps. Note that all snapshots are computed in parallel using the high-performance FEM solver *PGFem3D*. This allows fast snapshot construction even for $\mathcal{O}(1000)$ samples needed to create the database.

5.1. Isomap, f

We use Isomap as a dimension reduction technique. We see in Fig. 9 that regular Isomap converges to dimension $\widehat{d} = 6$ for all four loading modes. As mentioned earlier, we have verified this result with kernel Isomap to confirm the intrinsic dimensionality of the manifold $\mathcal{M}$. Once more we would like to mention that the manifold $\mathcal{M}$ is a parametric manifold with 6 macroscopic loading parameters. This implies that Isomap has unfolded the manifold

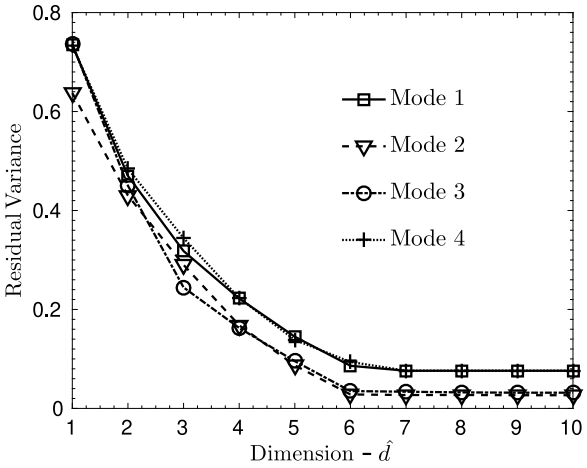

Fig. 9. Isomap dimensionality, $\widehat{d}$.

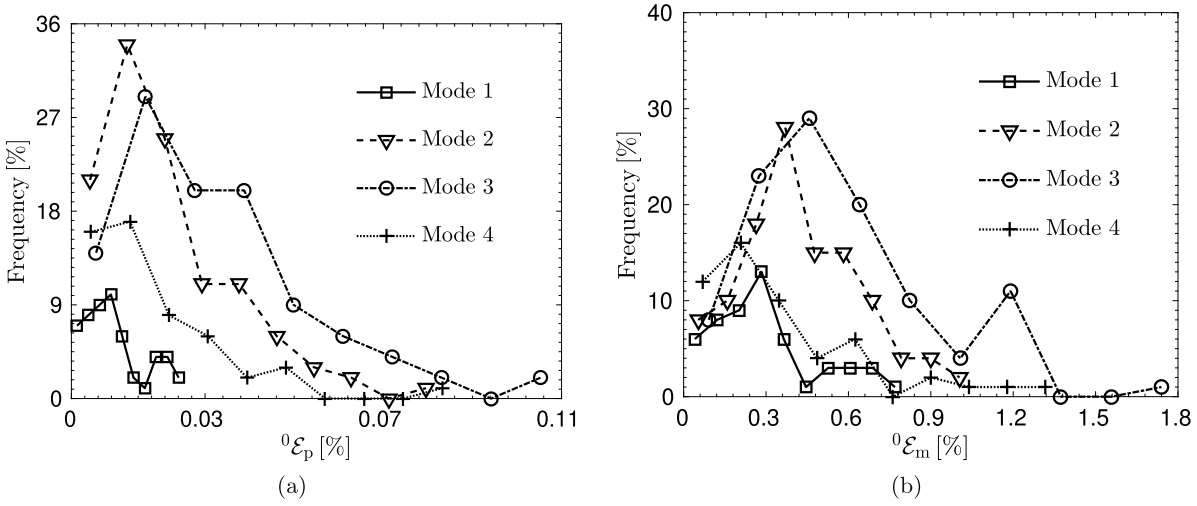

Fig. 10. Distribution of the volume averaged error (homogenization) given by Eq. (24) for randomly selected 10% of points from database (one-out cases). (a) Particle. (b) Matrix.

correctly and has returned a reduced space, $\mathcal{A}$, where the Euclidean distances are approximately equal to the geodesic distances on the manifold. Because we construct the map, $\mathcal{F}$ ($\mathcal{F} : \mathfrak{T} \mapsto \mathcal{M}$) as a composition of the inverse map (f^{-1}) and the parameter linkage map ($\mathfrak{P}$), for the next step we verify these maps individually.

5.2. *Inverse map,* f^{-1}

We check the accuracy of the inverse map, $f^{-1} : \mathcal{A} \mapsto \mathcal{M}$, by a standard leave-one-out experiment on 10% randomly selected (from the uniform distribution) points from each loading mode. For each of these points, we compare the inverse map solution with the FEM solution in terms of the volume-averaged error for the matrix and particles (see Eq. (24)) of the deformation gradient field, $\boldsymbol{F}$. The distribution of the volume averaged error is presented in Fig. 10. The particle error is less than 0.11% (see Fig. 10(a)) and the matrix error is bounded by 1.8% (see Fig. 10(b)). To check the quality of the localized MNROM solution, we show the distribution of $\|\boldsymbol{F}\|_{\mathcal{F}}$ for the particle in Fig. 11(b) and for the matrix in Fig. 11(a). Comparison with the FEM solution yields the highest volume averaged error in the matrix (i.e., $^0\mathcal{E}_m = 1.789\%$). We can see in Fig. 11(a) that the inverse map can predict the localized field remarkably well.

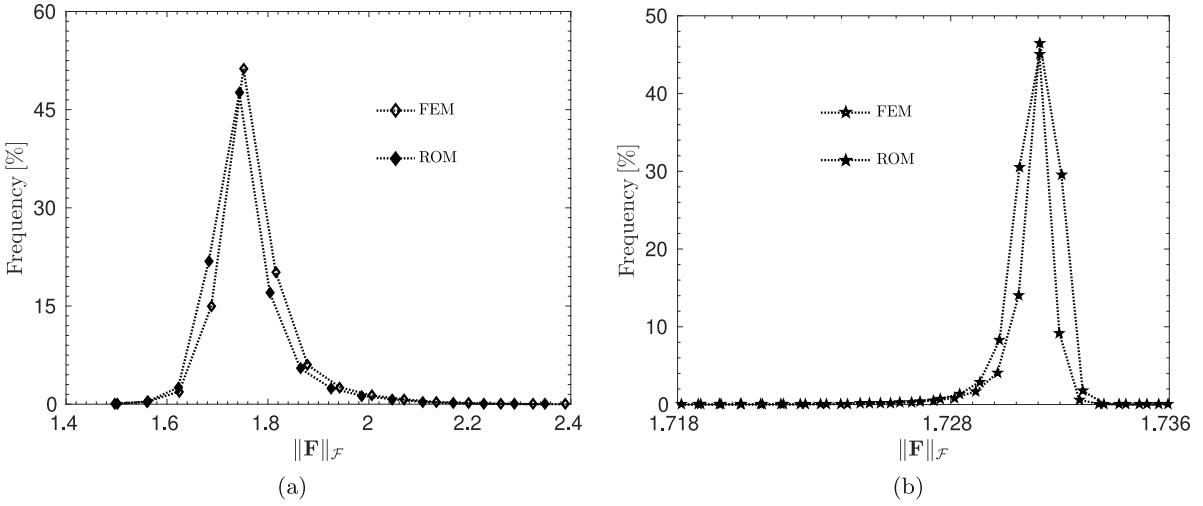

Fig. 11. Distribution of $\|\boldsymbol{F}\|_{\mathcal{F}}$ from FEM and MNROM over RUC in Fig. 4. (a) Over matrix. (b) Over particle.

5.3. NN map, $\mathfrak{P}$

To establish a link between the macroscopic loading space and the reduced space, we build the map $\mathfrak{P}$ using the neural network. As in our previous work, we use the NN implementation in MATLAB [58]. The NN architecture, target-function/error, training and other parameters are identical to those from our work in [16]. As elucidated in Section 2.2.3, NN is very prone to getting trapped in local minima. To circumvent this issue, we use ensemble NN. We train NN 10 times for each mode and take the direct average of the output. Also, to overcome over-fitting, we have employed a Bayesian regularization method [60,61]. Then, we reconstruct the reduced space as the NN output for the entire dataset. This reconstructed reduced space, $\overline{\mathcal{A}}$, deviates from the reduced space, $\mathcal{A}$. The physical interpretation of this error is the drift of the predicted MNROM solution from the FEM solution along the manifold. Note that the vector $\boldsymbol{\zeta}^i \in \mathcal{A}$ contains the displacement data. It is worth mentioning that the effect of this error on the local micro-field is extremely small.

In Fig. 12, we plot the distribution of the error, which is calculated as the Euclidean distance between the same index vectors $\boldsymbol{\zeta}^i$ and $\overline{\boldsymbol{\zeta}}^i$ from the two sets $\mathcal{A}$ and $\overline{\mathcal{A}}$, respectively. In order to understand the error, we provide the target range for vectors $\boldsymbol{\zeta}^i$ that is $[-150, 150]$ for Mode 1 and Mode 4, and $[-220, 220]$ for Mode 2 and Mode 3. We can see that the NN error has been reduced drastically from step 0 to step 1 (see Fig. 12). However, error reduction is not noticeable from step 1 to step 2. The average error for all modes is around 3 unit with step 1 data, which is $\mathcal{O}(10^{-3})$ in $[-1,\ 1]$ range when normalized. This suggests that the step 1 data is enriched enough. The data shift for the highest frequency is only 3 units on average, and the maximum shift is ~ 15 units. We note that in [16], we observed the shift for the highest frequency well above 5 units with the maximum shift at ~ 20 units. This shows that the NN map has been improved significantly over the one in [16] with this novel physics-guided sampling technique.

5.4. Verification of complete MNROM framework

After verification of all maps, the MNROM framework (see Fig. 1) can be used to compute the macroscale (homogenization) as well as the microscale (localization) fields for any given loading conditions, ${}^0\boldsymbol{F}$, of the unit cell that are bounded by 10% principal stretch in each direction. Next, we predict the microscale deformation fields for 200 query points (see the process of generating query points in Section 4.1) for each of the four modes (i.e., 800 total query points).

We compute the microscale deformation gradient and compare it to the deformation gradient from FEM simulations. First, we compute the volume averaged error for 200 query points considering all four modes by Eq. (24). The volume averaged error is shown in Figs. 13(a) and 13(b) for the particles and matrix, respectively.

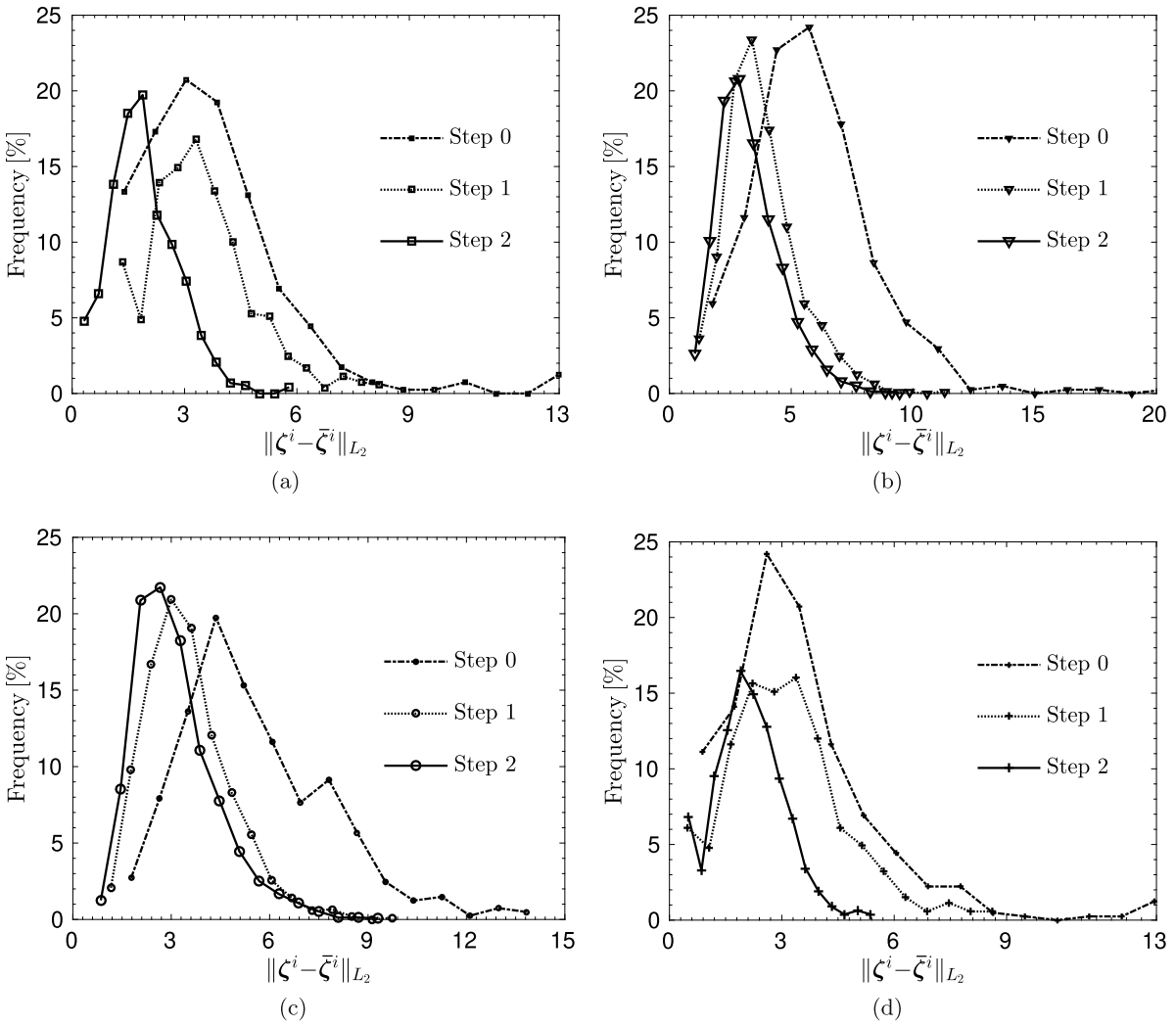

Fig. 12. Distribution of the neural network output error compared to the reduced space over the entire dataset for different enrichment steps. The size of the markers indicates the enrichment step. (a) Mode 1. (b) Mode 2. (c) Mode 3. (d) Mode 4.

The minimum, maximum, mean and median of the volume averaged error over all query points for all modes are listed in Table 4. Fig. 13 shows that the error in the matrix is higher than that in the particles. In Fig. 13(b), we note that the maximum error goes to 1.67%, but less than 10% of query points have a volume-averaged error above 1% considering all modes. The query points associated with higher error are either close to the loading envelope or in the regions where the manifold data density is still low. Next, we will analyze the quality of MNROM for the homogenization and localization of different fields of interest.

5.5. Homogenization of material response

In this section, we investigate the capability of MNROM to predict the homogenized response for 800 query points (200 in each mode) with the step 1 dataset. In doing so, we first compute the homogenized strain energy density (see Eq. (5)), which is the volume average of the local strain energy density. In Fig. 14, we plot the deviatoric (Fig. 14(a)) and the volumetric (Fig. 14(b)) part of the strain energy density. Looking at both the deviatoric and volumetric strain energy, we can observe a very high correlation between the FEM and MNROM predictions.

Table 4
Quality of the MNROM for all modes computed from Eq. (24) for 200 query points in each mode.

Loading cases		Min. [%]	Max.[%]	Mean [%]	Median [%]
Mode 1	Particle	0.0038	0.0284	0.0127	0.0124
	Matrix	0.1080	0.8224	0.4093	0.4013
Mode 2	Particle	0.0069	0.0541	0.0239	0.0214
	Matrix	0.1009	0.8058	0.3866	0.3810
Mode 3	Particle	0.0101	0.1083	0.0330	0.0295
	Matrix	0.1168	1.6605	0.4612	0.3796
Mode 4	Particle	0.0100	0.0722	0.0344	0.0334
	Matrix	0.1557	1.2821	0.6322	0.6335

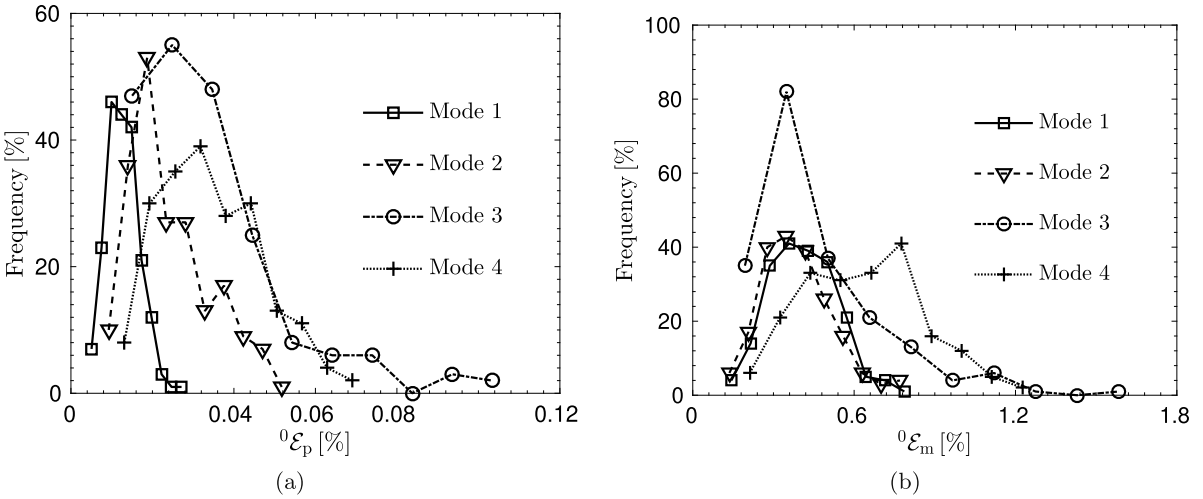

Fig. 13. Distribution of the volume averaged error (homogenization) given by Eq. (24) of the 200 randomly selected query points in each mode. (a) Particle. (b) Matrix.

Moreover, we compute the error associated with the homogenized potential energy:

$$\mathcal{E}_{0_W} = \frac{\left|{}^0W^{ROM} - {}^0W^{FEM}\right|}{{}^0W^{FEM}} \times 100\,[\%]. \tag{25}$$

Here, ${}^0W^{ROM}$ and ${}^0W^{FEM}$ are the homogenized strain energies computed using the MNROM framework and FEM respectively. Because mode 1 and mode 4 are purely tensile and compressive, and also more volumetric in nature (see Fig. 5), these are the modes contributing to the moderate to high volumetric strain energy in Fig. 14(b). On the other hand, mode 2 and mode 3 are more deviatoric and contribute more to the deviatoric energy in Fig. 14(a). In the entire analysis with 800 query points, 94.12% of the points have less than 10% error as computed from Eq. (25). Only a few points (around 5%) show more than 10% error. Moreover, the greedy procedure introduced in Section 2.2.2 can be used to enrich the neighborhoods with large errors and thus improve the overall accuracy further. This is a large improvement over our prior work, where only 75% of the cases had error below 20% [16]. This indicates the strength of the physics-guided manifold construction. Furthermore, we would like to learn the functional form of the homogenized potential energy, ${}^0W({}^0\boldsymbol{C})$. We do it separately for the volumetric part, ${}^0W_J({}^0J)$, and the deviatoric part, ${}^0W_C({}^0\widehat{\boldsymbol{C}})$. We could use a machine learning technique for this calibration, however, here we simply take the functional form given in Eq. (22) and learn it as a least square fit. Fig. 15 depicts the model calibration using the volumetric (Fig. 15(b)) and deviatoric (Fig. 15(a)) functions for both the FEM and MNROM data. The FEM and MNROM parameterizations show an excellent agreement for both the volumetric and deviatoric

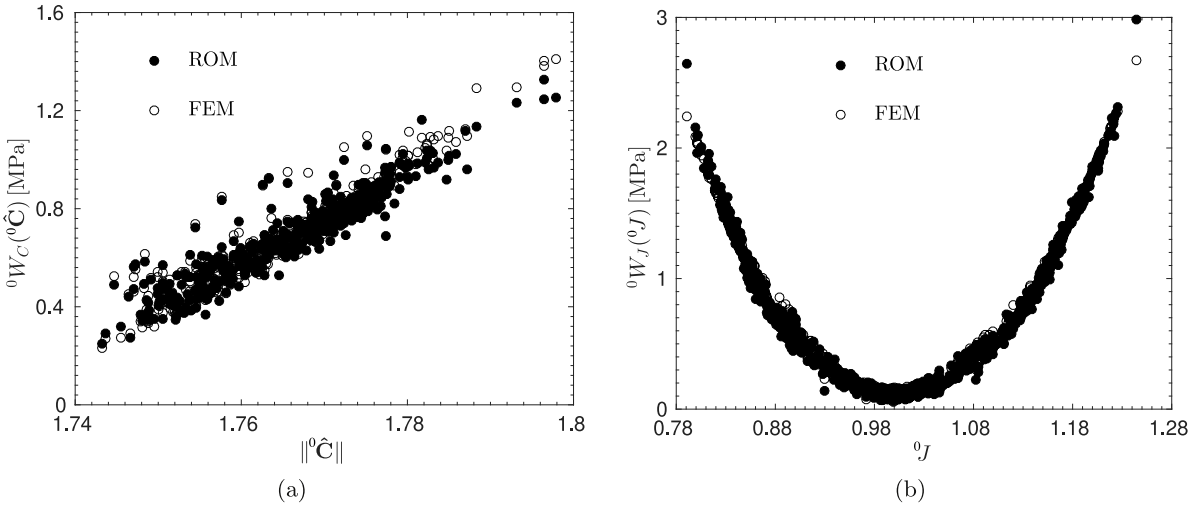

Fig. 14. Macroscale strain energy density function MNROM versus FEM. (a) The deviatoric potential, $^0W_C(^0\widehat{\boldsymbol{C}})$. (b) The volumetric potential, $^0W_J(^0J)$.

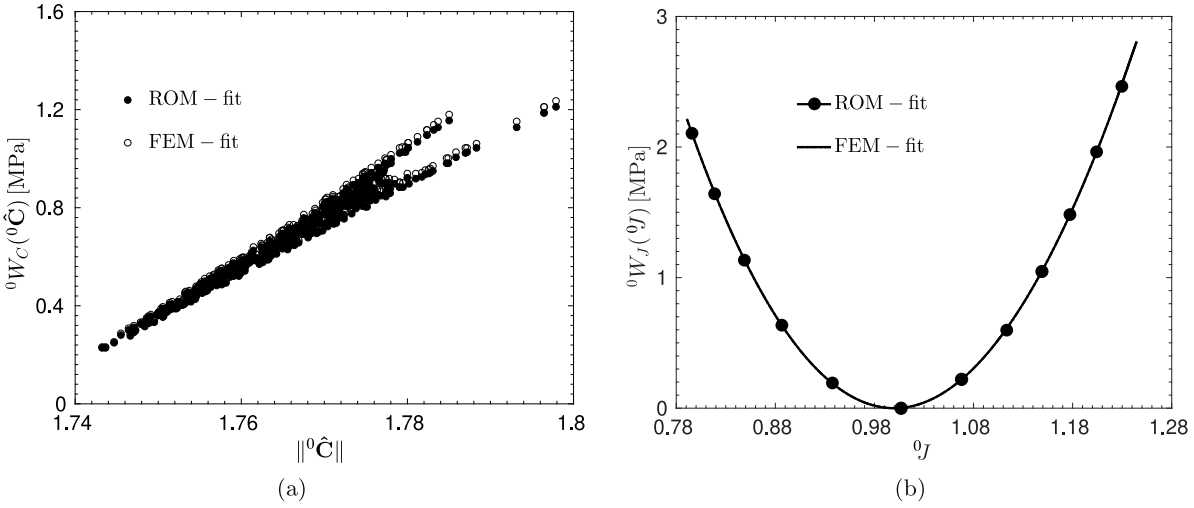

Fig. 15. Least-square fit of the macroscale strain energy density function from MNROM versus FEM (a) The deviatoric potential, $^0W_C(^0\widehat{\boldsymbol{C}})$. (b) The volumetric potential, $^0W_J(^0J)$.

Table 5
Homogenized material properties of the RUC in Fig. 4 determined from the least-square fit.

Source	κ [MPa]	μ_{10} [MPa]	μ_{01} [MPa]
MNROM	96.37	11.01	11.01
FEM	96.36	11.74	11.74

potentials. Finally, the homogenized material constants (i.e., the coefficients of the least-square fit) for both FEM and MNROM models are presented in Table 5.

Although we can derive the homogenized stress after we learn the functional form of the homogenized potential, here we would also like to check the quality of the homogenized stress computed from the microscale field using MNROM. We compare this stress with the FEM solution. In Fig. 16(a), we plot the Frobenius norm of

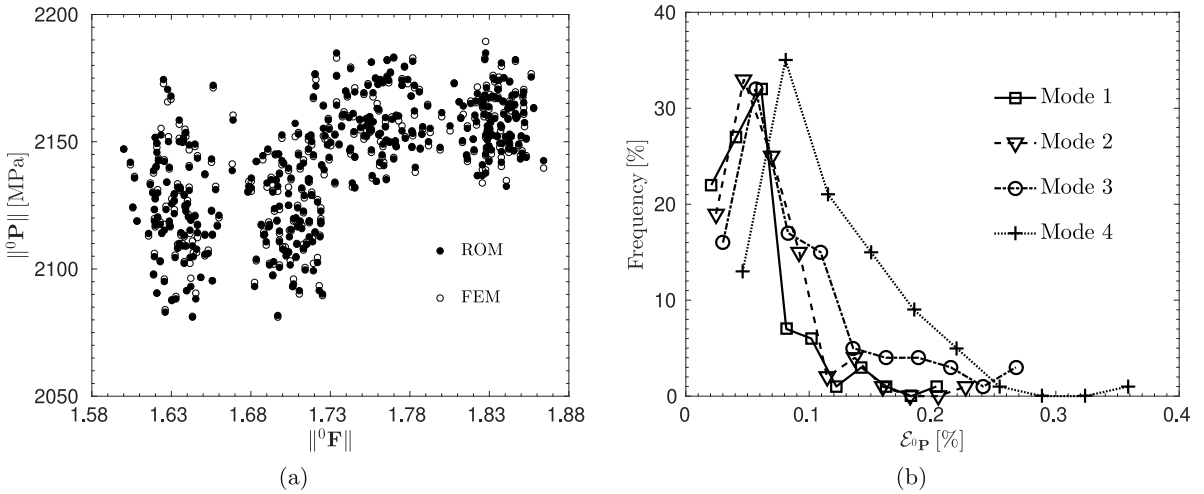

Fig. 16. Macroscale first Piola stress tensor, ${}^0\boldsymbol{P}$, MNROM versus FEM. (a) Frobenius norm of ${}^0\boldsymbol{P}$ against Frobenius norm of ${}^0\boldsymbol{F}$. (b) Distribution of the homogenization error given by Eq. (26).

the macroscopic first Piola stress tensor, ${}^0\boldsymbol{P} = \frac{1}{|\Theta_0|}\int_{\Theta_0} {}^1\boldsymbol{P}\,\mathrm{d}\Theta$, with the corresponding Frobenius norm of the macroscopic deformation gradient, ${}^0\boldsymbol{F}$, for all four modes. Note, ${}^0\boldsymbol{F}$ is the energy conjugate to ${}^0\boldsymbol{P}$. We can see that the points corresponding to $\|{}^0\boldsymbol{P}\|$ computed by MNROM are very close to the points which are calculated using the FEM. To close the comparison, we compute the error as follows:

$$\mathcal{E}_{0_{\boldsymbol{P}}} = \frac{\|{}^0\boldsymbol{P}^{ROM} - {}^0\boldsymbol{P}^{FEM}\|}{\|{}^0\boldsymbol{P}^{FEM}\|} \times 100\,[\%]. \tag{26}$$

The error plot is shown in Fig. 16(b), where the maximum error is $\mathcal{E}_{0_{\boldsymbol{P}}} = 0.38\%$. This is again a large improvement over the results presented in [16], where only 67.25% of all query points had less than 1% error with the maximum error approaching 20%.

5.6. Localization of material response

In this section, we explore the ability of our adaptive data enrichment scheme to improve MNROM in terms of the local (element-wise) engineering fields derived from the microscopic deformation. Note that localization of the material response is performed using the highly parallel FEM solver *PGFem3D*. When a new loading condition (i.e., $\boldsymbol{\eta}$) is selected, we use NN and find the reduced vector $\boldsymbol{\zeta}$. Next, we perform the reconstruction of the solution and obtain the solution vector $\boldsymbol{\xi}$ by applying the map f^{-1}. This solution vector is loaded into the parallel FEM solver and traditional FEM operations are performed to compute, strains, stresses, etc. Thus, the re-localization of the solution is obtained quickly and consistently with the FE^2 setting. Moreover, if the MNROM solution is not sufficient, we can execute the nonlinear solve with $\boldsymbol{\xi}$ as the initial guess and produce DNM with a high degree of acceleration.

In this study, we consider the deformation gradient in the matrix $\boldsymbol{F}_m$ and the Almansi strain, $\boldsymbol{e}$, defined as

$$\boldsymbol{e} = \frac{1}{2}(\mathbf{1} - \boldsymbol{F}^{-1}\boldsymbol{F}^{-T}). \tag{27}$$

Moreover, we define $\|\boldsymbol{e}\|_{\mathcal{F}}$ as the measure of the microscopic effective Almansi strain and use it for the localization assessment. To compare FEM and MNROM solutions, we quantify the deformation gradient error using $\mathcal{E}_m$ (i.e., see Eq. (23)) at each local (microscopic) point $\boldsymbol{Y}$ in the microstructure (i.e., each finite element in RUC). Moreover, we define an effective strain error as follows:

$$\mathcal{E}_e = \frac{|\|\boldsymbol{e}^{ROM}\|_{\mathcal{F}} - \|\boldsymbol{e}^{FEM}\|_{\mathcal{F}}|}{\|\boldsymbol{e}^{FEM}\|_{\mathcal{F}}} \times 100\,[\%]. \tag{28}$$

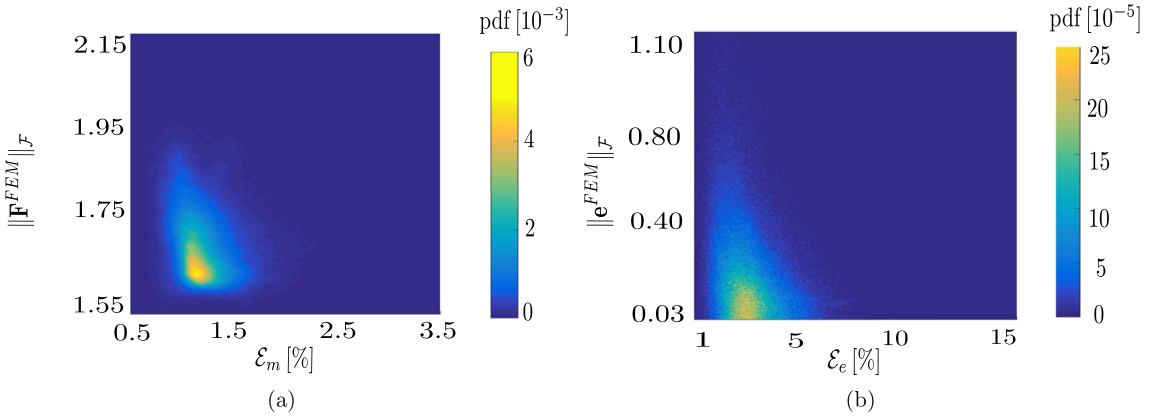

Fig. 17. Joint probability distribution (computed in discrete element-wise sense) of the local fields (in matrix phase). (a) $\mathcal{E}_m$ and $\|\boldsymbol{F}^{FEM}\|_{\mathcal{F}}$. (b) $\mathcal{E}_e$ and $\|\boldsymbol{e}^{FEM}\|_{\mathcal{F}}$.

Because mode 3 is relatively more complex in nature, we chose a query point from mode 3 to judge the quality of the localization process, which corresponds to a very high strain. We plot two joint probability distributions: Fig. 17(a) shows the joint probability distribution of $\mathcal{E}_m$ and $\|\boldsymbol{F}^{FEM}\|_{\mathcal{F}}$, and Fig. 17(b) displays the joint probability distribution of $\mathcal{E}_e$ and $\|\boldsymbol{e}^{FEM}\|_{\mathcal{F}}$. Because the deformation of the particles is negligible, we focus on the matrix. The joint probability distribution is constructed from a cloud of points representing individual finite elements within the matrix phase (285,380 elements in the matrix).

Fig. 17(a) shows that most finite elements in the matrix carry a mean localization error of 0.99% for this loading case. Although the range of deformations is [1.55, 2.15], the deformation is concentrated between $\|\boldsymbol{F}^{FEM}\|_{\mathcal{F}} = 1.5$ and $\|\boldsymbol{F}^{FEM}\|_{\mathcal{F}} = 1.65$, which characterizes very high compressive strain (note that $\|\mathbf{1}\|_{\mathcal{F}} = \sqrt{3} = 1.732$). However, this particular loading case also induces very high tensile strain locally. Fig. 17(a) shows that the error $\mathcal{E}_m$ is concentrated between $\sim 0.5 - 1.5\%$. Fig. 17(b) reveals that the mean microscopic effective Almansi strain error calculated at every finite element in the matrix is 4.53%. Note that the regions of high strain are concentrated between errors of $3 - 5\%$. Moreover, 70.46% of the finite elements have errors below 5% (median error is 2.84%), and all of the finite elements in the matrix have errors below 15%. These results show excellent localization prediction of MNROM. Moreover, all higher error points are associated with negligible effective strain (see strain region $\|\boldsymbol{e}^{FEM}\|_{\mathcal{F}} = 0.03$ in Fig. 17(b)).

Finally, Fig. 18 provides a visual comparison of the microscopic effective Almansi strain in the matrix between the FEM (left: Fig. 18(a)) and MNROM (right: Fig. 18(b)) analysis. We can observe that MNROM is able to capture the overall strain distribution extremely well including localized features.

6. Conclusion

We have designed and implemented a novel physics-guided adaptive sequential sampling technique in the context of the manifold-based reduced order model, MNROM, for multiscale modeling of nonlinear hyperelastic materials. The MNROM relies on different nonlinear maps, established through Isomap for dimension reduction, kernel inverse/reconstruction map and a NN. This manifold-based ROM not only drastically speeds up the CH process (i.e., by avoiding additional large parallel finite element simulations in the FE^2 setting), but it also provides both homogenization and localization of the multiscale analysis for complex three-dimensional hyperelastic materials.

Although MNROM is a promising data-driven approach, it lacks an efficient means of data sampling. Thus, we have developed a novel physics-driven sampling strategy that mitigates this issue. Our approach is a stepwise data enrichment method, intended to minimize the sparse regions of the high-dimensional manifolds which effectively maximizes information gain in each step. The method couples input and output space and explores the pattern in the output space, which is directly guided by the inherent physics in the input space (i.e., the macroscale loading condition). We learn the pattern in each step based on the current dataset which determines how extra simulations

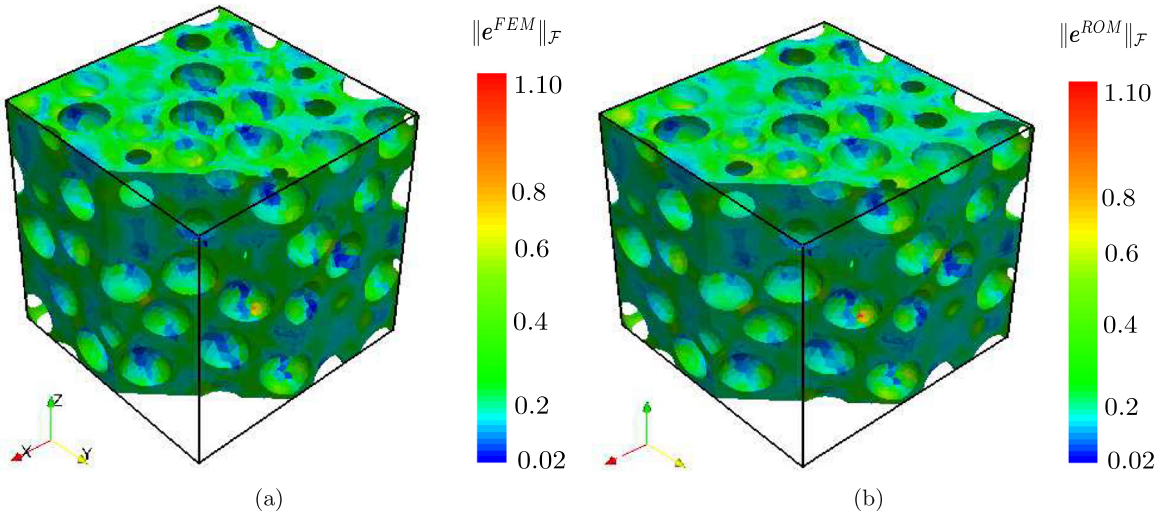

Fig. 18. Visualization of the microscopic effective Almansi strain ($\|\boldsymbol{e}\|_{\mathcal{F}}$) for $\boldsymbol{\eta}^q = \{{}^0\lambda_1 = 0.9284, {}^0\lambda_2 = 0.9058, {}^0\lambda_3 = 1.0839, \varphi^1 = 2.4217, \varphi^2 = 1.2802, \varphi^3 = 0.3165\}$. Note that the particles are rigid-like and hence are removed for visualization purpose. (a) FEM simulation. (b) MNROM analysis.

are decided and added to the database. In this work, we have used the same framework of MNROM, where the multiscale loading conditions are simulated in terms of the macroscopic principal stretches as well as the orthogonal principal directions, while the high-dimensional manifold has been constructed from the microscopic displacement fields. Moreover, we have shown the effectiveness of the proposed sampling technique through meticulous numerical simulations.

In the numerical example, we observed that the data distribution becomes a smooth stepwise data enrichment and accordingly, all the maps improved significantly. This physics-driven deterministic sampling strategy also appreciably reduced the number of required simulations by eliminating redundant data points. Thus, the computational complexity of MNROM is reduced enormously. Also, we have shown that with the proper enrichment the additional simulations can be added to the dataset seamlessly using a greedy algorithm. Moreover, we can observe that even the localization process produced extremely small error with a very small number of simulations. This sampling strategy is not only helpful for CH and FE^2 acceleration, but can potentially be applied to many other physical problems by identifying the influential parameters. The development of MNROM that considers also path-dependent material nonlinearity (e.g., visco-plasticity) is an important future direction.

Acknowledgments

This work was supported by the Department of Energy, National Nuclear Security Administration, under the reward No. DE-NA0002377 as part of the Predictive Science Academic Alliance Program II. We would like to thank two anonymous reviewers for providing useful suggestions and comments that improved the quality of the manuscript.

References

[1] B. Chazelle, L. Monier, A model of computation for VLSI with related complexity results, in: Proceedings of the Thirteenth Annual ACM Symposium on Theory of Computing, ACM, 1981, pp. 318–325.
[2] M.L. Klein, W. Shinoda, Large-scale molecular dynamics simulations of self-assembling systems, Science 321 (5890) (2008) 798–800.
[3] K.-S. Yang, J.H. Ferzigert, Large-eddy simulation of turbulent obstacle flow using a dynamic subgrid-scale model, AIAA J. 31 (8) (1993) 1406–1413.
[4] D. Yushu, S. Lee, K. Matouš, Sharp volumetric billboard based characterization and modeling of complex 3D Ni/Al high energy ball milled composites, Mech. Mater. 108 (2017) 93–106.
[5] R. Hill, On constitutive macro-variables for heterogeneous solids at finite strain, Proc. R. Soc. A 326 51 (1972) 245–271.
[6] M.G.D. Geers, V. Kouznetsova, W.A.M. Brekelmans, Multi-scale computational homogenization: Trends and challenges, J. Comput. Appl. Math. 234 (2010) 2175–2182.

[7] K. Matouš, M.G. Geers, V.G. Kouznetsova, A. Gillman, A review of predictive nonlinear theories for multiscale modeling of heterogeneous materials, J. Comput. Phys. 330 (2017) 192–220.
[8] M.G.D. Geers, V.G. Kouznetsova, K. Matouš, J. Yvonnet, Encyclopedia of computational mechanics, in: E. Stein, R. Borst, T.J. Hughes (Eds.), second ed., John Wiley & Sons, Inc., 2017, pp. 1–34, http://dx.doi.org/10.1002/9781119176817.ecm2107.
[9] M. Mosby, K. Matouš, Hierarchically parallel coupled finite strain multiscale solver for modeling heterogeneous layers, Internat. J. Numer. Methods Engrg. 102 (2015) 748–765.
[10] F. Feyel, J.-L. Chaboche, FE^2 Multiscale approach for modelling the elasto-viscoplastic behaviour of long fibre SiC/Ti composite materials, Comput. Methods Appl. Mech. Engrg. 183 (2000) 309–330.
[11] C. Miehe, Computational micro-to-macro transitions for discretized micro-structures of heterogeneous materials at finite strains based on the minimization of averaged incremental energy, Comput. Methods Appl. Mech. Engrg. 192 (2003) 559–591.
[12] F. Chinesta, P. Ladeveze, E. Cueto, A short review on model order reduction based on proper generalized decomposition, Arch. Comput. Methods Eng. 18 (4) (2011) 395.
[13] C.W. Rowley, Model reduction for fluids, using balanced proper orthogonal decomposition, Int. J. Bifurcation Chaos 15 (03) (2005) 997–1013.
[14] D. Wirtz, N. Karajan, B. Haasdonk, Model Order Reduction of Multiscale Models Using Kernel Methods, Preprint Series, SRC SimTech, University of Stuttgart, 2013, pp. 1–25.
[15] E. Lopez, D. Gonzalez, J.V. Aguado, E. Abisset-Chavanne, E. Cueto, C. Binetruy, F. Chinesta, A manifold learning approach for integrated computational materials engineering, Arch. Comput. Methods Eng. 25 (1) (2018) 59–68, http://dx.doi.org/10.1007/s11831-016-9172-5.
[16] S. Bhattacharjee, K. Matouš, A nonlinear manifold-based reduced order model for multiscale analysis of heterogeneous hyperelastic materials, J. Comput. Phys. 313 (2016) 635–653.
[17] A. Sawant, A. Acharya, Model reduction via parametrized locally invariant manifolds: Some examples, Comput. Methods Appl. Mech. Engrg. 195 (44) (2006) 6287–6311, http://dx.doi.org/10.1016/j.cma.2005.12.012.
[18] F. Fritzen, M. Leuschner, Reduced basis hybrid computational homogenization based on a mixed incremental formulation, Comput. Methods Appl. Mech. Engrg. 260 (2013) 143–154, http://dx.doi.org/10.1016/j.cma.2013.03.007.
[19] P. Acar, V. Sundararaghavan, Reduced-order modeling approach for materials design with a sequence of processes, AIAA J. 56 (12) (2018) 5041–5044, http://dx.doi.org/10.2514/1.J057221.
[20] W. Xing, A.A. Shah, P.B. Nair, Reduced dimensional Gaussian process emulators of parametrized partial differential equations based on isomap, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 471 (2174) (2015) 20140697.
[21] J. Ghaboussi, J. Garrett Jr., X. Wu, Knowledge-based modeling of material behavior with neural networks, J. Eng. Mech. 117 (1) (1991) 132–153.
[22] B. Le, J. Yvonnet, Q.-C. He, Computational homogenization of nonlinear elastic materials using neural networks, Internat. J. Numer. Methods Engrg. 104 (12) (2015) 1061–1084.
[23] J. Andrade, C. Avila, S. Hall, N. Lenoir, G. Viggiani, Multiscale modeling and characterization of granular matter: from grain kinematics to continuum mechanics, J. Mech. Phys. Solids 59 (2) (2011) 237–250.
[24] J. Yvonnet, E. Monteiro, Q.-C. He, Computational homogenization method and reduced database model for hyperelastic heterogeneous structures, Int. J. Multiscale Comput. Eng. 11 (3) (2013) 201–225.
[25] R.A. van Tuijl, C. Harnish, K. Matouš, J.J.C. Remmers, M.G.D. Geers, Wavelet based reduced order models for microstructural analyses, Comput. Mech. 63 (3) (2019) 535–554, http://dx.doi.org/10.1007/s00466-018-1608-3.
[26] I. Temizer, P. Wriggers, An adaptive method for homogenization in orthotropic nonlinear elasticity, Comput. Methods Appl. Mech. Engrg. 196 (2007) 3409–3423.
[27] I. Temizer, T. I. Zohdi, A numerical method for homogenization in non-linear elasticity, Comput. Mech. 40 (2007) 281–298.
[28] M. Mosby, K. Matouš, On mechanics and material length scales of failure in heterogeneous interfaces using a finite strain high performance solver, Modelling Simulation Mater. Sci. Eng. 23 (8) (2015) 085014.
[29] M. Mosby, K. Matouš, Computational homogenization at extreme scales, Extreme Mech. Lett. 6 (2016) 68–74, http://dx.doi.org/10.1016/j.eml.2015.12.009.
[30] V.C. Chen, K.-L. Tsui, R.R. Barton, M. Meckesheimer, A review on design, modeling and applications of computer experiments, IIE Trans. 38 (4) (2006) 273–291.
[31] G.E. Box, K. Wilson, On the experimental attainment of optimum conditions, in: Breakthroughs in Statistics, Springer, 1992, pp. 270–310.
[32] M.D. McKay, R.J. Beckman, W.J. Conover, Comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (2) (1979) 239–245.
[33] M.E. Johnson, L.M. Moore, D. Ylvisaker, Minimax and maximin distance designs, J. Statist. Plann. Inference 26 (2) (1990) 131–148.
[34] K. Palmer, K.-L. Tsui, A minimum bias latin hypercube design, IIE Trans. 33 (9) (2001) 793–808.
[35] J.L. Loeppky, J. Sacks, W.J. Welch, Choosing the sample size of a computer experiment: A practical guide, Technometrics 51 (4) (2009) 366–376.
[36] Y. Hung, Contributions to Computer Experiments and Binary Time Series, ProQuest, 2008.
[37] A.B. Owen, Orthogonal arrays for computer experiments, integration and visualization, Statist. Sinica (1992) 439–452.
[38] C. Currin, T. Mitchell, M. Morris, D. Ylvisaker, Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments, J. Amer. Statist. Assoc. 86 (416) (1991) 953–963.
[39] Y. Lin, F. Mistree, J. Allen, K.-L. Tsui, V. Chen, Sequential metamodeling in engineering design, in: 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, American Institute of Aeronautics and Astronautics, 2004, http://dx.doi.org/10.2514/6.2004-4304.

[40] R. Jin, W. Chen, A. Sudjianto, On sequential sampling for global metamodeling in engineering design, in: ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, 2002, pp. 539–548.
[41] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, M. Ouimet, Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering, Advances in neural information processing systems 16 (2004) 177–184.
[42] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2323–2326.
[43] B. Ganapathysubramanian, N. Zabaras, A non-linear dimension reduction methodology for generating data-driven stochastic input models, J. Comput. Phys. 227 (2008) 6612–6637.
[44] A.A. Shah, A survey of data-driven emulators for high dimensional spatio-temporal data, in: Proceedings of International Conference on Information Technology and Computer Science, 9788193137307, 2015.
[45] M. Renardy, R.C. Rogers, An Introduction to Partial Differential Equations, Vol. 13, Springer Science & Business Media, 2006.
[46] J.M. Lee, Riemannian Manifolds: An Introduction to Curvature, Vol. 176, Springer Science & Business Media, 2006.
[47] B. Schiilkopf, The kernel trick for distances, Adv. Neural Inf. Process. Syst. 13 (2001) 301–307.
[48] M. Bernstein, V.D. Silva, J. C. Langford, J. B. Tenenbaum, Graph Approximations to Geodesics on Embedded Manifolds, Technical report, Stanford University, Stanford, 2000.
[49] H.Q. Minh, P. Niyogi, Y. Yao, Mercer's theorem, feature maps, and smoothing, in: International Conference on Computational Learning Theory, Springer, 2006, pp. 154–168.
[50] J. Costa, A.O. Hero, et al., Geodesic entropic graphs for dimension and entropy estimation in manifold learning, IEEE Trans. Signal Process. 52 (8) (2004) 2210–2221.
[51] M. Sniedovich, Dijkstra's algorithm revisited: the dynamic programming connexion, Control Cybernet. 35 (3) (2006) 599.
[52] H. Choi, S. Choi, Kernel isomap on noisy manifold, in: Proceedings. the 4th International Conference on Development and Learning, 2005, IEEE, 2005, pp. 208–213.
[53] F. Cailliez, The analytical solution of the additive constant problem, Psychometrika 48 (2) (1983) 305–308.
[54] M. Balasubramanian, E.L. Schwartz, The isomap algorithm and topological stability, Science 295 (5552) (2002) 7.
[55] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, third ed., The MIT Press, 2009.
[56] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (1989) 359–366.
[57] B.G. Sumpter, C. Getino, D.W. Noid, Theory and applications of neural computing in chemical science, Annu. Rev. Phys. Chem. 45 (1994) 439–481.
[58] M.H. Beale, M.T. Hagan, H.B. Demuth, Neural Network Toolbox™, User's Guide, The MathWorks, Inc., Natick, MA, USA, 1992-2015.
[59] W.L. Buntine, A.S. Weigend, Baysian back-propagation, Complex Syst. 5 (1991) 603–643.
[60] F. Girosi, M. Jones, T. Poggio, Regularization theory and neural networks architectures, Neural Comput. 7 (2) (1995) 219–269.
[61] R. Hecht-Nielsen, Theory of the backpropagation neural network, in: Neural Networks, 1989. IJCNN., International Joint Conference on, IEEE, 1989, pp. 593–605.
[62] L.K. Hansen, P. Salamon, Neural network ensembles, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 993–1001.
[63] H. Wendland, Scattered Data Approximation, Cambridge University Press, 2004.
[64] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Science & Business Media, New York, Berlin, Paris, 2006.
[65] C. Preda, Regression models for functional data by reproducing kernel Hilbert spaces methods, J. Statist. Plann. Inference 137 (3) (2007) 829–840.
[66] H. Kadri, E. Duflos, P. Preux, S. Canu, M. Davy, et al., Nonlinear functional regression: a functional RKHS approach, in: AISTATS, vol. 10, 2010, pp. 111–125.
[67] C. Walder, O. Chapelle, Learning with transformation invariant kernels, in: Advances in Neural Information Processing Systems, 2007, pp. 1561–1568.
[68] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in: International Conference on Computational Learning Theory, Springer, 2001, pp. 416–426.
[69] C.S. Ong, X. Mary, S. Canu, A.J. Smola, Learning with non-positive kernels, in: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, 2004, p. 81.
[70] F. Liang, K. Mao, M. Liao, S. Mukherjee, M. West, Nonparametric Bayesian Kernel Models, Department of Statistical Science, Duke University, Discussion Paper, 2007, pp. 07–10.
[71] S. Doll, K. Schweizerhof, On the development of volumetric strain energy functions, J. Appl. Mech. 67 (1) (1999) 17–21, http://dx.doi.org/10.1115/1.321146.
[72] R.B. Ellis, J.L. Martin, C. Yan, Random geometric graph diameter in the unit ball, Algorithmica 47 (4) (2007) 421–438.
[73] X. Wang, Volumes of generalized unit balls, Math. Mag. 78 (5) (2005) 390–395, URL http://www.jstor.org/stable/30044198.
[74] A.T. H. Lee, K. Matouš, Three-dimensional reconstruction of statistically optimal unit cells of polydisperse particulate composites from microtomography, Phys. Rev. E 80 (2009) 061301–61312.
[75] S. Torquato, Random Heterogeneous Materials, Springer, New York, USA, 2002.
[76] D.S. Stafford, T.L. Jackson, Using level sets for creating virtual random packs of non-spherical convex shapes, J. Comput. Phys. 229 (2010) 3295–3315.