

A Riemannian stochastic representation for quantifying model uncertainties in molecular dynamics simulations

Hao Zhang, Johann Guilleminot*

Department of Civil and Environmental Engineering, Duke University, Durham, NC 27708, USA

Received 13 July 2022; received in revised form 2 October 2022; accepted 4 October 2022

Available online xxxx

Abstract

A Riemannian stochastic representation of model uncertainties in molecular dynamics is proposed. The approach relies on a reduced-order model, the projection basis of which is randomized on a subset of the Stiefel manifold characterized by a set of linear constraints defining, *e.g.*, Dirichlet boundary conditions in the physical space. We first show that these constraints are, indeed, preserved through Riemannian pushforward and pullback actions to, and from, the tangent space to the manifold at any admissible point. This fundamental property is subsequently exploited to derive a probabilistic model that leverages the multimodel nature of the atomistic setting. The proposed formulation offers several advantages, including a simple and interpretable low-dimensional parameterization, the ability to constraint the Fréchet mean on the manifold, and ease of implementation and propagation. The relevance of the proposed modeling framework is finally demonstrated on various applications including multiscale simulations on graphene-based systems.

© 2022 Elsevier B.V. All rights reserved.

Keywords: Model uncertainty; Molecular dynamics; Reduced-order modeling; Stiefel manifold; Uncertainty quantification

1. Introduction

Molecular dynamic (MD) simulations are widely employed to study microscopic processes and predict macroscopic thermodynamical properties in various science and engineering fields, including biophysics, computational chemistry, and materials science. The accuracy of MD simulations heavily depends on the interatomic potentials that are used to evaluate the force fields governing the interactions within the dynamical system. Such functions are usually designed and calibrated according to experimental data or first-principles calculations, leading to a myriad of models for the same atomic system.

For example, models for water–biomolecular interactions include 3-site potential models such as TIPS [1], SPC [2], TIP3P [3], SPC/E [4], as well as 4-site models such as BF [5], TIP4P [3], OPC [6], to name a few. The coarse-grained Optimized Potential for Liquid Simulation (OPLS) potential model and its variants [7–10] can be employed for MD simulations of polyethylene, which is one of the simplest polymer systems. In the case of graphene sheets (which are composed of carbon atoms only), many models were proposed in the literature, including AIREBO [11], BOP [12], REBO-2 [13], LCBOP [14], ReaxFF [15], and Tersoff-type potentials [16]. For

* Corresponding author.

E-mail address: johann.guilleminot@duke.edu (J. Guilleminot).

many systems of interest, and most critically for newly developed materials, the appropriate choice of interatomic potentials is rarely known in advance, and a detailed analysis must be conducted before MD simulations are performed, according to specific simulation targets and environment [17].

The selection and calibration of interatomic potentials can induce substantial uncertainties across scales. Several approaches have been pursued in the literature to propagate various types of uncertainties from fine to coarser atomistic scales (see, *e.g.*, [18–20]), as well as in MD-to-continuum coupling methods [21,22] where intermediate coarse-scale results are estimated through fine-scale MD simulations. There is also a very substantial number of papers that have reported on the impact of parametric uncertainties in a continuum multiscale setting; see, *e.g.*, [23–25], in a non-exhaustive manner. Restricting the discussion to uncertainty quantification for MD simulations, most of the works have focused on the propagation of parametric uncertainties in potential parameters [26–29], sampling-induced uncertainties [30], and uncertainties for models built using machine learning techniques [31].

In contrast, the consideration of model-form uncertainties has received relatively little attention to date. The issue of adaptive model selection under uncertainty, considering candidates that are obtained through different coarse-graining strategies for example, was addressed in [32,33] using Bayesian formalism. A framework enabling extrapolation from one model to another, using functional perturbations, was proposed in [34]. In the recent work [20], the probabilistic approach developed in [35] to capture model uncertainties through the construction of a probability measure for the projection basis of a reduced-order model was applied to identify, and investigate the effect of, model uncertainties in MD simulations. While the methodology was shown to perform satisfactorily in terms of variability representation, the parameterization to pullback from the tangent space to the manifold used in [20,35] does not allow for the mean of the fluctuations to be constrained. Furthermore, hyperparameter identification suffers from a curse of dimensionality unless assumptions about the structure of the statistical fluctuations are made—the number of hyperparameters scaling as $\mathcal{O}(n^2)$ for a reduced-order model of dimension n [35,36]. The aim of this work is to propose a new formulation that circumvents these limitations and fully leverages the multimodel nature of the atomistic setting, using *ad hoc* Riemannian projection and retraction operators. Note that this setting is intrinsically different from the one considered in [35], where only one model (about which statistical fluctuations are to be prescribed) is assumed to be available. The use of Riemannian operators is motivated by the preservation of linear constraints onto the tangent space, which enables the derivation of a particularly simple and easily implementable model, and allows the mean of the fluctuations to be constrained in the Fréchet sense.

The remainder of this paper is organized as follows. The stochastic modeling framework is introduced in Section 2. We present the reduced-order model, as well as strategies to parameterize the formulation on the tangent space to the Stiefel manifold. We subsequently define the probabilistic model and propose a strategy to constrain the Fréchet mean on the manifold. Section 3 is dedicated to various applications, including a toy example, related to sampling on the unit sphere, and multiscale MD-based predictions on a graphene system. Conclusion is finally presented in Section 4.

2. Riemannian stochastic representation of model uncertainties

2.1. Reduced-order modeling for molecular dynamics simulations

The evolution of the whole system, composed of N_a atoms in $\mathbb{R}^d$, is described by Newton's second law of motion

$$[M]\ddot{\mathbf{q}}(t) = \mathbf{f}(t; \mathbf{q}(t)), \quad (1)$$

where $[M] \in \mathbb{R}^{N \times N}$, $\mathbf{q} \in \mathbb{R}^N$, and $\mathbf{f} \in \mathbb{R}^N$ denotes the mass matrix, the position vector, and the force vector, respectively (with $N = d \times N_a$). Appropriate initial conditions are assumed and left unspecified throughout this section. In addition, we assume that $\mathbf{q}$ satisfies a set of linear constraints, written as

$$[B]^T \mathbf{q}(t) = [0_{N_{CD}, n}], \quad \forall t \geq 0, \quad (2)$$

where N_{CD} is the number of constrained degrees of freedom in the system and $[B] \in \mathbb{R}^{N \times N_{CD}}$ satisfies

$$[B]^T [B] = [I_{N_{CD}}]. \quad (3)$$

The above constraint can be used to specify homogeneous Dirichlet boundary condition, for example.

To fix ideas, assume that the system is defined for a *given* choice of interatomic potential (describing all types of interactions). A reduced-order model (ROM) can then be constructed by using a proper orthogonal decomposition (POD). To that end, we consider the state variable

$$\mathbf{z}(t) = \mathbf{q}(t) - \mathbf{q}(0), \quad \forall t \in [0, T], \quad (4)$$

which satisfies the equilibrium equation

$$[M]\ddot{\mathbf{z}}(t) = \tilde{\mathbf{f}}(t), \quad (5)$$

supplemented with appropriately modified boundary conditions, and $\tilde{\mathbf{f}}(t) = \mathbf{f}(\mathbf{z}(t) + \mathbf{q}(0), t)$.

Let $t_0 = 0 < t_1 < \dots < t_{N_t} = T$ be a discretization of the time interval $[0, T]$, where $t_j = j\Delta t$ and Δt is the time step. Let $\mathcal{J} = \{j_1, \dots, j_{N_s}\} \subset \{1, \dots, N_t\}$, with $1 \leq N_s \leq N_t$, not necessarily ordered and with distinct elements, and consider the sequence of snapshots $\{\boldsymbol{\mu}^{(k)}\}_{k=1}^{N_s}$ such that $\boldsymbol{\mu}^{(k)} = \mathbf{q}(t_{j_k})$. Let

$$[X] = [\underline{\boldsymbol{\mu}}^{(1)} \dots \underline{\boldsymbol{\mu}}^{(N_s)}], \quad (6)$$

where $\underline{\boldsymbol{\mu}}^{(k)} = \boldsymbol{\mu}^{(k)} - \mathbf{q}(0)$ for $1 \leq k \leq N_s$, and introduce the singular value decomposition

$$[X] = [U][S][V]^T, \quad (7)$$

where the sequence of singular values is nonincreasing. A reduced-order basis (ROB) $[\Phi]$ can be classically obtained by retaining the n first columns (which are referred to as POD modes) of $[U]$. The number n of modes can be determined through a convergence analysis enabling a tradeoff between dimensionality reduction and projection error.

Now consider the linear mapping

$$\mathbf{z}(t) = [\Phi]\mathbf{y}(t), \quad (8)$$

where $\mathbf{y}$ is the reduced variable with values in $\mathbb{R}^n$. The matrix $[\Phi]$ satisfies the orthogonality property

$$[\Phi]^T[\Phi] = [I_n], \quad (9)$$

as well as the boundary condition

$$[B]^T[\Phi] = [0_{N_{CD}, n}], \quad (10)$$

where $[B]^T$ defines a boundary condition operator and N_{CD} is the number of constrained degrees of freedom in the system. The Galerkin projection of Eq. (5) expressed in terms of atom displacements (with physical variable $\mathbf{z}$) reads as

$$[\mathcal{M}]\ddot{\mathbf{y}}(t) = \mathcal{F}(t), \quad (11)$$

where $[\mathcal{M}]$ and $\mathcal{F}$ are the projected mass matrix and reduced force vector, respectively:

$$[\mathcal{M}] = [\Phi]^T[M][\Phi], \quad \mathcal{F}(t) = [\Phi]^T\tilde{\mathbf{f}}(t). \quad (12)$$

Note at this stage that the previous reduced-order formulation is not introduced to accelerate simulations since most of the computational time is spent into assembling procedures. This numerical burden may be circumvented, in practice, by bypassing back-and-forth projections between the physical and reduced space, using, *e.g.*, machine-learning-based surrogates for forces in the reduced space. This aspect is out of the scope of the present work.

Since the reduced-order basis $[\Phi]$ satisfies the orthogonality property stated in Eq. (9), it is necessary to introduce the set of orthogonal matrices

$$St(N, n) = \{[Y] \in \mathbb{R}^{N \times n} \text{ such that } [Y]^T[Y] = [I_n]\}, \quad (13)$$

called the compact Stiefel manifold, where $\mathbb{R}^{N \times n}$ is the set of all $N \times n$ real matrices. Owing to the constraint given by Eq. (10), the matrix $[\Phi]$ then belongs to the subset $\mathbb{S}_{N,n} \subset St(N, n)$ defined as

$$\mathbb{S}_{N,n} = \{[Y] \in \mathbb{R}^{N \times n} \text{ such that } [Y]^T[Y] = [I_n], [B]^T[Y] = [0_{N_{CD}, n}]\}. \quad (14)$$

The set $\mathbb{S}_{N,n}$ constitutes the admissible space for $[\Phi]$, and can therefore be interpreted as the support of the probability measure for the stochastic counterpart of $[\Phi]$, denoted by $[\Phi]$. The main challenge then lies in the construction of a proper probabilistic model for $[\Phi]$.

Notation. In the following, we denote by $[W^{(1)}], \dots, [W^{(m)}]$ the reduced-order bases in $\mathbb{S}_{N,n}$ obtained by considering each interatomic potential separately, assuming here that m candidate models are available. The global reduced-order basis obtained by concatenating snapshots obtained for all considered potentials is denoted by $[W]$.

2.2. Construction of the stochastic reduced-order model

2.2.1. Problem statement

We consider the stochastic modeling of the random matrix $[\Phi]$, defined on a probability space $(\Theta, \mathcal{T}, P)$ and taking values in the subset $\mathbb{S}_{N,n}$ of the Stiefel manifold $St(N, n)$. We assume that the collection $\{[W^{(1)}], \dots, [W^{(m)}], [W]\}$ of reduced-order bases is given, and that $[W]$ belongs to the convex hull of $[W^{(1)}], \dots, [W^{(m)}]$. The tangent space of $St(N, n)$ at $[Y]$ is defined as

$$T_{[Y]}St(N, n) = \{[\Delta] \in \mathbb{R}^{N \times n} \mid [Y]^T[\Delta] + [\Delta]^T[Y] = [0_n]\} \subset \mathbb{R}^{N \times n}, \quad (15)$$

where $[Y] \in St(N, n)$ is called the base (or reference) point on the Stiefel manifold, and $[0_n]$ is the null matrix of size $n \times n$. The projection onto the tangent space to the Stiefel manifold at $[Y]$ (push-forward operation) is denoted by

$$P_{[Y]} : St(N, n) \rightarrow T_{[Y]}St(N, n), \quad (16)$$

while the retraction (pull-back operation) is denoted by

$$R_{[Y]} : T_{[Y]}St(N, n) \rightarrow St(N, n). \quad (17)$$

There are several ways to define such projection and retraction operators; see, *e.g.*, Chapter 4 in [37]. The retraction operator based on the polar decomposition (see Eq. (4.7), p. 59, in [37]), namely

$$R_{[Y]}([\Delta]) = ([Y] + [\Delta]) ([I_n] + [\Delta]^T[\Delta])^{-1/2}, \quad (18)$$

was used in [35], in particular (see [20] for an application in a molecular dynamics setting). In fact, using the parameterization

$$[\Delta] = [A] - [W][D] \quad (19)$$

on $T_{[W]}St(N, n)$, where $[A] \in \mathbb{R}^{N \times n}$ is arbitrary and $[D] = \text{Sym}([W]^T[A])$, it is seen that the pulled-back point

$$[\tilde{Y}] = R_{[Y]}(s[\Delta]) = ([W] + s[\Delta]) ([I_n] + s^2[\Delta]^T[\Delta])^{-1/2} \quad s \geq 0, \quad (20)$$

satisfies the Dirichlet boundary condition $[B]^T[\tilde{Y}] = [0_{N_{CD},n}]$ if $[A]$ also satisfies

$$[B]^T[A] = [0_{N_{CD},n}]. \quad (21)$$

The representation

$$[A] = ([I] - [B][B]^T)[U], \quad [U] \in \mathbb{R}^{N \times n} \text{ arbitrary}, \quad (22)$$

trivially satisfies Eq. (21) (see Eq. (3)) and was introduced in [35] to model uncertainties through the randomization of $[U]$. The main advantage of this approach is that the model ensures admissibility of samples by construction, since the stochastic reduced-order basis belongs to $\mathbb{S}$ almost surely. The complexity of (and nonlinearity in) the retraction operator, however, makes statistical inference intricate, since the mean of the stochastic model cannot be enforced for instance. In addition, the formulation introduced in [35] to model the stochastic version of $[U]$ introduces a curse of dimensionality in terms of hyperparameters, with a number of parameters that scales as $\mathcal{O}(n^2)$; see [36] for a discussion. In the following section, we propose a new representation that fully takes advantage of the *multimodel* molecular dynamics setting and in particular, of the dataset $\{[W^{(1)}], \dots, [W^{(m)}], [W]\}$.

2.2.2. Riemannian stochastic modeling

Let $[Y]$ be a reference point on the Stiefel manifold $St(N, n)$, and consider two points, denoted by $[\Delta]$ and $[\tilde{\Delta}]$, on the tangent space $T_{[Y]}St(N, n)$. The canonical inner product associated with the tangent space $T_{[Y]}St(N, n)$ is then given by

$$\langle [\Delta], [\tilde{\Delta}] \rangle_{[Y]} = \text{tr}([\Delta]^T([I_N] - \frac{1}{2}[Y][Y]^T)[\tilde{\Delta}]) \quad (23)$$

and induces the (canonical) metric $\|[\Delta]\|_{[Y]} = \langle [\Delta], [\Delta] \rangle_{[Y]}^{1/2}$. Note that $\|[\Delta]\|_{[Y]}$ is the length of tangent vector $[\Delta]$ on the tangent space at the base point $[Y]$ and corresponds to the arc length between $[\Delta]$ and $[Y]$ on the Stiefel manifold.

A Riemannian projection operator $P_{[Y]} : St(N, n) \ni [\tilde{Y}] \mapsto [\Delta] \in T_{[Y]}St(N, n)$ can be obtained as

$$[\Delta] = \log_{[Y]}^{St}([\tilde{Y}]), \quad (24)$$

where $\log_{[Y]}^{St}$ is the Riemannian Stiefel logarithm at point $[Y]$, defined such that $\|[\Delta]\|_{[Y]} = \langle [\Delta], [\Delta] \rangle_{[Y]}^{1/2}$ represents the geodesic distance between $[\tilde{Y}]$ and $[Y]$. The retraction operator $R_{[Y]} : T_{[Y]}St(N, n) \ni [\Delta] \mapsto [\tilde{Y}] \in St(N, n)$ is defined as

$$[\tilde{Y}] = \exp_{[Y]}^{St}([\Delta]), \quad (25)$$

where $\exp_{[Y]}^{St}$ is the Riemannian Stiefel exponential at $[Y] \in St(N, n)$; see Chapter 5 in [37] for a review.

No closed-form results exist for the computation of the Riemannian Stiefel logarithm, which must be evaluated numerically. An optimization-based approach was proposed in [38], while iterative algorithms based on matrix-algebraic representations geodesic can be found in [39] (see Algorithms 7 and 8 therein for the computation of the Stiefel exponential and logarithm, respectively); see also [40]. Note that there exists an empirical condition, given by $\|[Y] - [\tilde{Y}]\|_2 \leq 2$, that ensures that the Stiefel logarithm algorithm converges.

In this work, we rely on the algorithms proposed in [39,40] and utilize the iterative matrix construction to demonstrate important results related to the constraint given in Eq. (10). These results are presented in the form of propositions below. Note that $n \leq N/2$ and that for most dynamical systems of interest, the condition $n \ll N$ is met.

Proposition 1. *Let $[Y]$ and $[\tilde{Y}]$ be two points belonging to $\mathbb{S}_{N,n} \subset St(N, n)$. Then $[\Delta] = \log_{[Y]}^{St}([\tilde{Y}]) \in T_{[Y]}St(N, n)$, where the Riemannian Stiefel logarithm is defined through the matrix-algebraic representation proposed in [39], satisfies the linear constraint*

$$[B]^T[\Delta] = [0_{N_{CD} \times n}],$$

where $[B]$ is defined by Eq. (10).

Proof. Using the matrix-algebraic representation derived in [39] (see Algorithm 8 therein, as well as [40]), the Stiefel logarithm can be computed as

$$[\Delta] = \log_{[Y]}^{St}([\tilde{Y}]) = [Y][A_\tau] + [Q_L][B_\tau], \quad (26)$$

where $[Q_L] \in \mathbb{R}^{N \times n}$ stems from the compact (thin) QR decomposition

$$([I_N] - [Y][Y]^T)[\tilde{Y}] = [Q_L][N_L]. \quad (27)$$

The matrices $[A_k] \in \mathbb{R}^{n \times n}$ and $[B_k] \in \mathbb{R}^{n \times n}$ are associated with the sequence of matrices $\{[A_k], [B_k]\}_{k \geq 0}$ satisfying the system of nonlinear algebraic equations

$$\begin{bmatrix} [A_{k+1}] & -[B_{k+1}]^T \\ [B_{k+1}] & [C_{k+1}] \end{bmatrix} = \log \left\{ \exp \left\{ \begin{bmatrix} [A_k] & -[B_k]^T \\ [B_k] & [C_k] \end{bmatrix} \right\} \exp \left\{ \begin{bmatrix} [0_n] & [0_n] \\ [0_n] & -[C_k] \end{bmatrix} \right\} \right\}, \quad (28)$$

with

$$\begin{bmatrix} [A_0] & -[B_0]^T \\ [B_0] & [C_0] \end{bmatrix} = \log \left\{ \begin{bmatrix} [Y]^T[\tilde{Y}] & [X_0] \\ [N_L] & [Y_0] \end{bmatrix} \right\}. \quad (29)$$

The matrices $\{[X_0], [Y_0]\}$ are obtained by completion, and τ is the smallest integer such that $\|[C_\tau]\|_2 \leq \epsilon$, with ϵ a given threshold parameter. Assuming the invertibility of $[N_L] \in \mathbb{R}^{n \times n}$ (which follows when $\text{rank}([I_N] - [Y][Y]^T)[\tilde{Y}]) = n$), we have that

$$\begin{aligned} [B]^T[\Delta] &= [B]^T([Y][A_\tau] + [Q_L][B_\tau]), \\ &= [B]^T[Y][A_\tau] + [B]^T([I_N] - [Y][Y]^T)[\tilde{Y}][N_L]^{-1}[B_\tau], \\ &= [0_{N_{CD} \times n}], \end{aligned} \quad (30)$$

since $[B]^T[Y] = [B]^T[\tilde{Y}] = [0_{N_{CD} \times n}]$. $\square$

Proposition 2. Let $[Y] \in \mathbb{S}_{N,n} \subset St(N, n)$ and consider $[\Delta] \in T_{[Y]}St(N, n)$ satisfying $[B]^T[\Delta] = [0_{N_{CD} \times n}]$, where $[B]$ is defined by Eq. (10). Then $[\tilde{Y}] = \exp_{[Y]}^{St}([\Delta]) \in St(N, n)$ (Riemannian Stiefel exponential) satisfies the linear constraint

$$[B]^T[\tilde{Y}] = [0_{N_{CD} \times n}],$$

that is, $[\tilde{Y}] \in \mathbb{S}_{N,n} \subset St(N, n)$.

Proof. Following [41], the Riemannian exponential is evaluated as

$$St(N, n) \ni [\tilde{Y}] = \exp_{[Y]}^{St}([\Delta]) = [Y][M] + [Q_E][N_E], \quad (31)$$

where $[M], [N_E] \in \mathbb{R}^{n \times n}$ are defined as

$$\begin{bmatrix} [M] \\ [N_E] \end{bmatrix} = \exp \left\{ \begin{bmatrix} [Y]^T[\Delta] & -[R_E]^T \\ [R_E] & [0_n] \end{bmatrix} \right\} \begin{bmatrix} [I_n] \\ [0_n] \end{bmatrix}. \quad (32)$$

The matrices $[Q_E] \in \mathbb{R}^{N \times n}$ and $[R_E] \in \mathbb{R}^{n \times n}$ arise in the compact (thin) QR decomposition

$$([I_N] - [Y][Y]^T)[\Delta] = [Q_E][R_E]. \quad (33)$$

Assuming that $[R_E]$ is invertible (which follows when $\text{rank}([I_N] - [Y][Y]^T)[\Delta]) = n$), we deduce

$$\begin{aligned} [B]^T[\tilde{Y}] &= [B]^T([Y][M] + [Q_E][N_E]), \\ &= [B]^T[Y][M] + [B]^T([I_N] - [Y][Y]^T)[\Delta][R_E]^{-1}[N_E], \\ &= [0_{N_{CD} \times n}], \end{aligned} \quad (34)$$

as $[Y]$ and $[\Delta]$ satisfy $[B]^T[Y] = [B]^T[\Delta] = [0_{N_{CD} \times n}]$. $\square$

Propositions 1 and 2 imply that the satisfaction of the linear constraint is preserved through the pushforward and pullback actions defined by the Riemannian Stiefel logarithm and exponential. Applying these results to the proposed framework, we can now derive

Proposition 3. Let $\{[\Delta^{(i)}]\}_{i=1}^m$ be the projections of the reduced-order bases $\{[W^{(i)}] \in \mathbb{S}_{N,n}\}_{i=1}^m$ onto the tangent space $T_{[W]}St(N, n)$ at $[W] \in \mathbb{S}_{N,n}$, $[\Delta^{(i)}] = \log_{[W]}^{St}([W^{(i)}])$ for $1 \leq i \leq m$. Then the linear combination $[\tilde{\Delta}] = \sum_{i=1}^m p_i [\Delta^{(i)}]$, with $(p_1, \dots, p_m) \in \mathbb{R}^m$, satisfies the property

$$\exp_{[W]}^{St}([\tilde{\Delta}]) \in \mathbb{S}_{N,n} \subset St(N, n).$$

Proof. The result is immediate using Propositions 1 and 2. $\square$

The above proposition suggests to seek the stochastic representation as

$$[\Phi] := \exp_{[W]}^{St} \left\{ \sum_{i=1}^m P_i \log_{[W]}^{St}([W^{(i)}]) \right\}. \quad (35)$$

This form ensures that $[\Phi]$ takes values in the constrained set $\mathbb{S}_{N,n}$, by construction.

In Eq. (35), the random vector $\mathbf{P} = (P_1, \dots, P_m)^T$ is defined on a probability space $(\Theta, \mathcal{T}, P)$. A natural choice for the probability measure of $\mathbf{P}$ is the Dirichlet distribution with concentration parameter $\boldsymbol{\alpha}$, $\mathbf{P} \sim \mathcal{D}(\boldsymbol{\alpha})$. This choice ensures that $P_i \geq 0$ and $\sum_{i=1}^m P_i = 1$ almost surely, and therefore defines a stochastic Riemannian convex combination on the Stiefel manifold. In practice, this construction leads to samples that belong to the convex hull defined by the reduced-order bases dataset $\{[W^{(i)}]\}_{i=1}^m$ (see [42] for an analysis in a deterministic setting).

Remark 1. With the proposed formulation, uncertainty propagation can be achieved through Monte Carlo simulations, as well as by using state-of-the-art stochastic collocation methods. More specifically, let $\mathbf{Y}$ be the random variable with values in $\mathbb{R}_{>0}^m$ and with independent components, such that $Y_i \sim \mathcal{G}(\alpha_i, 1)$. In this case, $\mathbf{P}$ and $\mathbf{Y}$ are related through

$$P_i = \frac{Y_i}{\sum_{j=1}^m Y_j}, \quad 1 \leq i \leq m. \quad (36)$$

It follows that $[\Phi]$ can equivalently be viewed as a function of Y (that is, $[\Phi] = [\Phi(\mathbf{P})] = [\Phi(Y)]$), which enables the use of, *e.g.*, polynomial chaos expansions in terms of Laguerre polynomials [43,44] to represent, and efficiently identify, stochastic quantities of interest defined through a multiscale operator; see [45,46] for reviews regarding representations and stochastic solvers.

2.2.3. Integrating a constraint on Fréchet mean

The aim of this section is to derive a formulation that allows the empirical mean model associated with the representation (35) to be prescribed. To this end, we assume that the global reduced-order basis $[W]$ belongs to the convex hull of $\{[W^{(i)}]\}_{i=1}^m$, and consider the identification of the concentration parameter α such that

$$\mathbb{E}\{[\Phi]\} \approx [W], \quad (37)$$

where the mean holds in the Fréchet sense (note that in the above equation, the right-hand side may alternatively be defined by exploiting strategies from model averaging). Recall that the Riemannian L^2 center of mass of a dataset $\{[\Phi^{(1)}], \dots, [\Phi^{(q)}]\}$ composed of q samples of $[\Phi]$ (in $St(N, n)$) is defined as the minimizer of

$$h([V]) = \frac{1}{2} \sum_{i=1}^q w_i d([V], [\Phi^{(i)}])^2, \quad (38)$$

where $\{w_i\}_{i=1}^q$ are scalar weights in the $(q-1)$ -dimensional simplex and d is the Riemannian canonical distance. Imposing that the gradient of the objective function vanishes at $[W]$ then yields

$$\sum_{i=1}^q w_i \log_{[W]}^{St} \{[\Phi^{(i)}]\} = [0_{N \times n}]. \quad (39)$$

Using the definition (35) and taking $w_i = 1/q$ for all weights then implies

$$\sum_{j=1}^m \left(\frac{1}{q} \sum_{i=1}^q p_j^{(i)} \right) \log_{[W]}^{St}([W^{(j)}]) = [0_{N \times n}], \quad (40)$$

where $p_j^{(i)}$ denotes the i th realization of the component P_j of $\mathbf{P} \sim \mathcal{D}(\alpha)$ ($p_j^{(i)} = p_j(\theta_i)$, $\theta_i \in \Theta$). Since

$$\frac{1}{q} \sum_{i=1}^q p_j^{(i)} \approx \frac{\alpha_j}{\sum_{i=1}^m \alpha_i} \quad (41)$$

for q sufficiently large, it can be deduced that the concentration parameters must satisfy

$$T_{[W]} St(N, n) \ni \sum_{j=1}^m \alpha_j \log_{[W]}^{St}([W^{(j)}]) = [0_{N \times n}]. \quad (42)$$

The above property can hence be enforced by imposing the constraint

$$\left\| \sum_{j=1}^m \alpha_j \log_{[W]}^{St}([W^{(j)}]) \right\|_F = 0. \quad (43)$$

In practice, α can be evaluated as

$$\alpha = \operatorname{argmin}_{\alpha \in \mathbb{R}_{>0}^m} \left\| \sum_{i=1}^m \alpha_i \log_{[W]}^{St}([W^{(i)}]) \right\|_F^2, \quad (44)$$

which is recast, for implementation purposes, as

$$\alpha = \operatorname{argmin}_{\alpha \in \mathbb{R}_{>0}^m} \mathbf{a}^T [H] \mathbf{a}, \quad (45)$$

where $[H]$ is the symmetric positive-definite matrix in $\mathbb{R}^{m \times m}$, the entries of which are given by

$$H_{ij} = \operatorname{tr}(\log_{[W]}^{St}([W^{(i)}])^T \log_{[W]}^{St}([W^{(j)}])). \quad (46)$$

This problem can be solved by any conventional quadratic programming algorithm. In this work, the built-in MATLAB function `quadprog` is used for the sake of illustration.

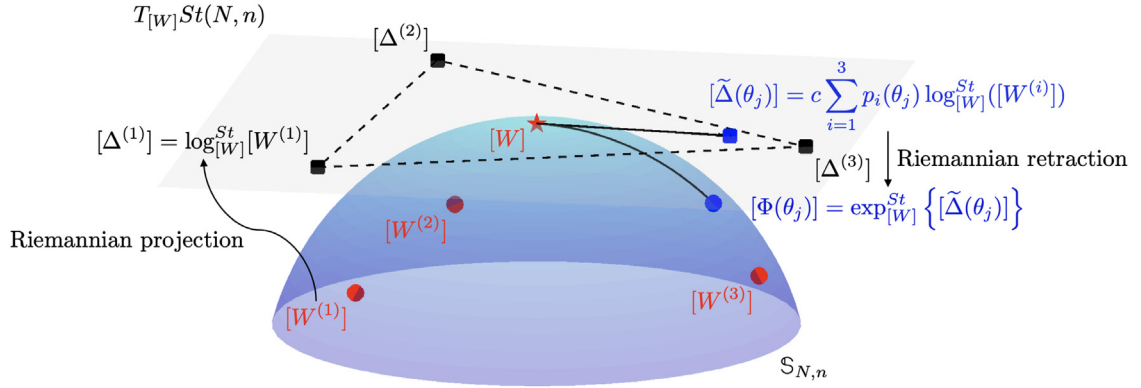


Fig. 1. Schematic view of the proposed stochastic modeling strategy on $\mathbb{S}_{N,n}$, with $m = 3$. Riemannian operators are defined with respect to the canonical metric. Note that geometries of sets are illustrative and may not reflect actual structures.

2.2.4. Scaling fluctuations on the tangent space

Defining the stochastic reduced-order basis as

$$[\Phi] = \exp_{[W]}^{St} \left\{ \sum_{i=1}^m P_i \log_{[W]}^{St}([W^{(i)}]) \right\}, \quad \mathbf{P} \sim \mathcal{D}(\alpha), \quad (47)$$

restricts statistical fluctuations in the convex hull of $\{[W^{(i)}]\}_{i=1}^m$. In order to increase fluctuations, a scaling parameter $c \geq 1$ is introduced to scale variations on the tangent space:

$$[\Phi] = \exp_{[W]}^{St} \left\{ c \sum_{i=1}^m P_i \log_{[W]}^{St}([W^{(i)}]) \right\}. \quad (48)$$

It should be noticed that the calibration strategy for the concentration parameter α is insensitive to multiplicative scaling (see Section 2.2.3). Consequently, considering $c > 1$ may lead to a shift in the Fréchet mean that is all the more pronounced that the distance between the Fréchet mean taken over the dataset $\{[W^{(i)}]\}_{i=1}^m$ and the global reduced-order basis $[W]$ is important. The value of c may be calibrated in practice solving a statistical inverse problems on microscopic or macroscopic quantities of interest (see Section 3.3.3 for an example).

2.3. Summary of the proposed approach

The main steps of the proposed modeling framework are listed below and are schematically illustrated in Fig. 1. Recall that $\{[W^{(1)}], \dots, [W^{(m)}]\}$ are the reduced-order bases in $\mathbb{S}_{N,n}$ computed through a proper orthogonal decomposition with snapshots associated with the full dynamical system for m different input candidate models, and that $[W]$ denotes the global reduced-order basis obtained by gathering all snapshots for all models. The global ROB is taken as base point and target Fréchet mean in the sampling procedure.

Step 1: Compute the tangent vectors $\{[\Delta^{(1)}], \dots, [\Delta^{(m)}]\}$ using the Riemannian projection operator, $[\Delta^{(i)}] = \log_{[W]}^{St}([W^{(i)}])$ with $[\Delta^{(i)}] \in T_{[W]}St(N, n)$.

Step 2: Compute the concentration parameters $\alpha = (\alpha_1, \dots, \alpha_m)$ by solving the quadratic programming problem defined in Eq. (45).

Step 3: Draw ν samples $\{p(\theta_j)\}_{j=1}^\nu$ of $\mathbf{P} \sim \mathcal{D}(\alpha)$, $\theta_j \in \Theta$ for $1 \leq j \leq \nu$.

Step 4: Compute the associated samples $\{[\Phi(\theta_j)]\}_{j=1}^\nu$ of $[\Phi]$ as

$$[\Phi(\theta_j)] = \exp_{[W]}^{St} \left\{ c \sum_{i=1}^m p_i(\theta_j) \log_{[W]}^{St}([W^{(i)}]) \right\} \in \mathbb{S}_{N,n} \subset St(N, n), \quad (49)$$

where $c = 1$ for stochastic Riemannian convex combinations or $c \geq 1$ to enforce fluctuations beyond the convex hull of the dataset.

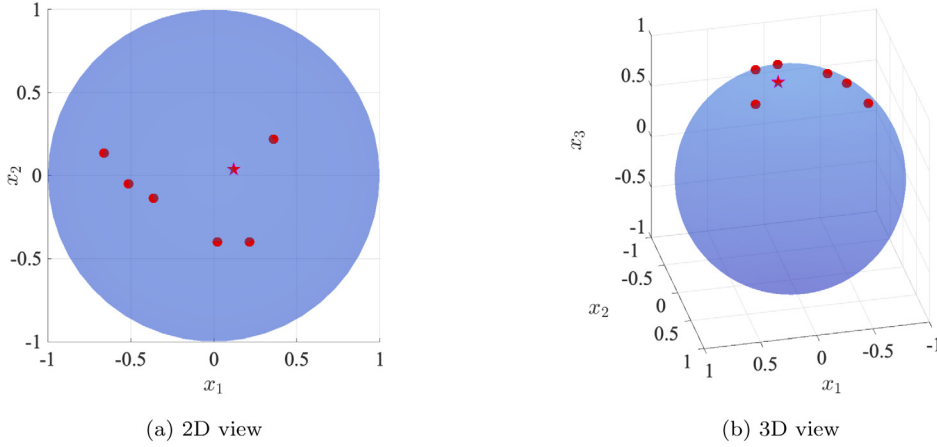


Fig. 2. Dataset on the unit sphere in $\mathbb{R}^3$: base point (red star) and vertices (red points).

In the next section, we deploy the proposed approach on a variety of applications. The case of the unit sphere is first presented in Section 3.1 to illustrate the approach with standard visualization in $\mathbb{R}^3$. Applications to molecular dynamics simulations on graphene-based systems are then discussed in Sections 3.2 and 3.3, with focus on microscopic and macroscopic responses respectively. The open-source package LAMMPS [47] is used for both full-order and reduced-order MD simulations.

3. Applications

3.1. Illustrative example: Sampling on (a subset of) the unit sphere $St(3, 1)$

3.1.1. Sampling without linear constraints

In this first example, we consider sampling on the half unit sphere (that is, without the linear constraints defined by the matrix $[B]$, see Eq. (2)). The dataset consists of seven points randomly distributed on the sphere ($m = 6$), with one base point included in the convex hull defined by the remaining points; see Fig. 2.

Two specific choices are made at this point. In a first setting, all concentration parameters are set to the same value, $\alpha_i = 0.2$ for $i \in \{1, \dots, 6\}$. In the second configuration, concentration parameters are calibrated such that the Fréchet mean is as close as possible to the aforementioned base point, following the strategy proposed in Section 2.2.3. Here, c is set to 1 so that only stochastic Riemannian convex combinations are used. Fig. 3 shows a set of 2,000 samples for both cases. It is seen that uniform sampling in the convex hull can be achieved by setting all concentration parameters equal to a small value (see the left panel in Fig. 3). The Fréchet mean of the samples, computed using the algorithm detailed in the Appendix, then lies far away from the chosen base point. In contrast, determining the concentration parameters by solving the quadratic programming problem defined in Eq. (45) allows for the Fréchet mean to be constrained to the neighborhood of the base point.

We next consider $c = 3$, using the same two configurations for the concentration parameters. Samples can be seen in Fig. 4. As expected, the generated samples are distributed beyond the convex hull defined by the given vertices (red dots). In addition, the distance between the Fréchet mean and the base point substantially increases for unconstrained sampling (left panel in Fig. 4), while remaining small for constrained sampling (recall that the concentration parameters are not adjusted a posteriori since the calibration formulation is insensitive to multiplicative scaling). In fact, the later observation strongly depends on the positions of the vertices relative to the target mean: when the distance between the center of mass of the vertices and the target mean is sufficiently small, scaling on the tangent space generally leads to a small drift in the Fréchet mean. On the contrary, a large distance implies a “lack of symmetry” in the definition of the sampling domain, in which case the mean is affected more significantly.

3.1.2. Sampling with linear constraints

We now turn to the proper integration of linear constraints on the half unit sphere. The only relevant case corresponds to $N_{CD} = 1$, other values leading to overconstrained problems that are not appropriate in terms of

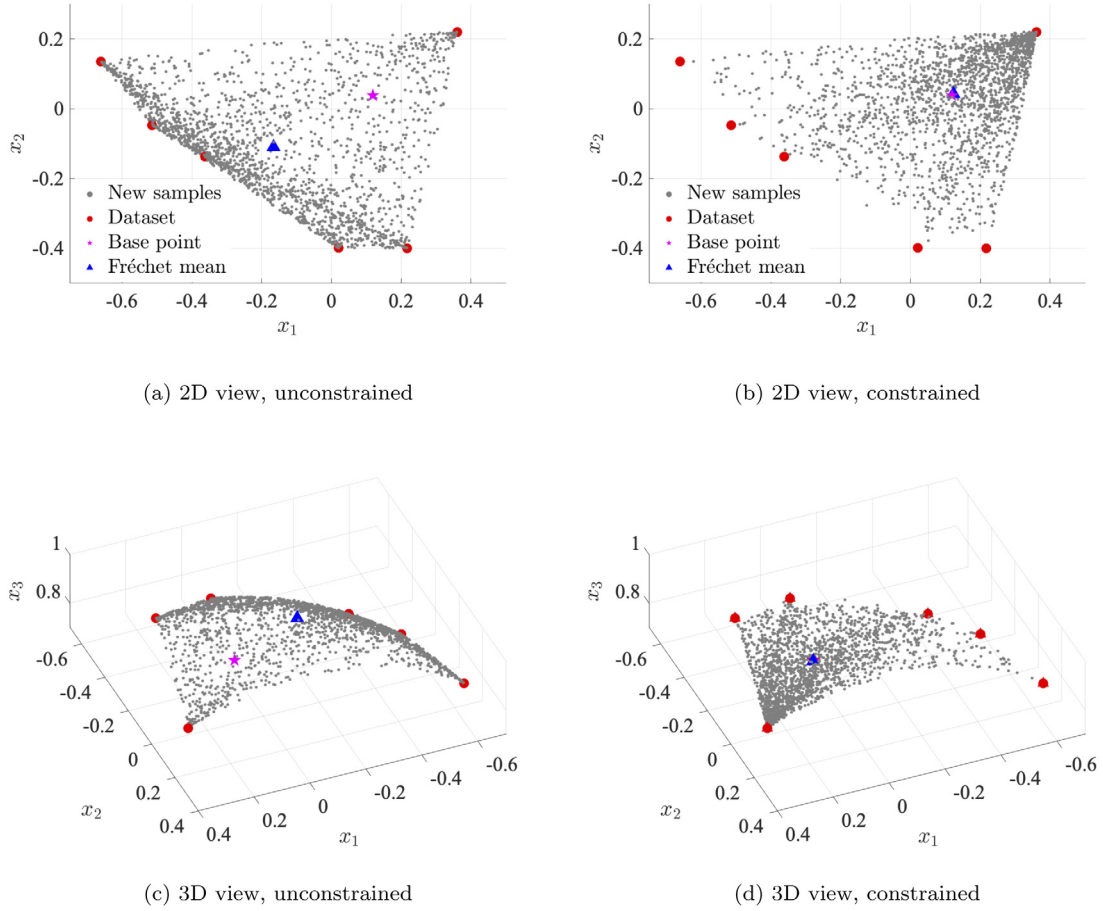


Fig. 3. Left panel (unconstrained sampling): samples obtained for $c = 1$ and $\alpha_i = 0.2$ for $i \in \{1, \dots, 6\}$. Right panel (constrained sampling): samples obtained for $c = 1$ and $\alpha = (0.2357, 0.2000, 0.1820, 0.1805, 1.4151, 0.2866)^T$.

sampling. Let $[B] = [b_1, b_2, b_3]^T \in \mathbb{R}^{3 \times 1}$, with $\|[B]\| = 1$ (see Eq. (3)), and consider $[\Phi] = [\Phi_1, \Phi_2, \Phi_3]^T \in \mathbb{S}_{3 \times 1} \subset St(3, 1)$. In this case, $\mathbb{S}_{3,1}$ defines a semi-ellipse (as the intersection of the unit sphere and an arbitrary plane) embedded in $\mathbb{R}^3$. Without loss of generality, six points are randomly chosen through uniform sampling on a semi-ellipse ($m = 6$), and one base point is selected near the middle of the curve defined by these points. Concentration parameters are chosen as $\alpha_i = 0.2$ for $i = 1, \dots, 6$, and samples are shown in Fig. 5 for $c = 1$. It is seen that all samples are distributed on the ellipse in a uniform manner (given the choice of the concentration parameters), which qualitatively shows that the linear constraint is properly satisfied.

3.2. Single graphene sheet subjected to harmonic excitation

3.2.1. System description

In this section, we apply the approach to molecular dynamics simulations on a single graphene sheet (in $\mathbb{R}^3$), composed of 272 carbon atoms, see Fig. 6. A zero Dirichlet boundary condition is applied on the left side of the structure (hence defining the linear constraints and matrix $[B]$), while a harmonic excitation force is applied on the right side according to

$$f_{ext}(t) = A \sin(2\pi\omega t) e^2, \quad t \geq 0, \quad (50)$$

with $A = 6$ [kcal mol $^{-1}$ Å $^{-1}$] and $\omega = 20 \times 10^9$ [rad/s]. Model uncertainties arise from the selection of the interatomic potentials governing the evolution of the system, and six different potentials commonly employed

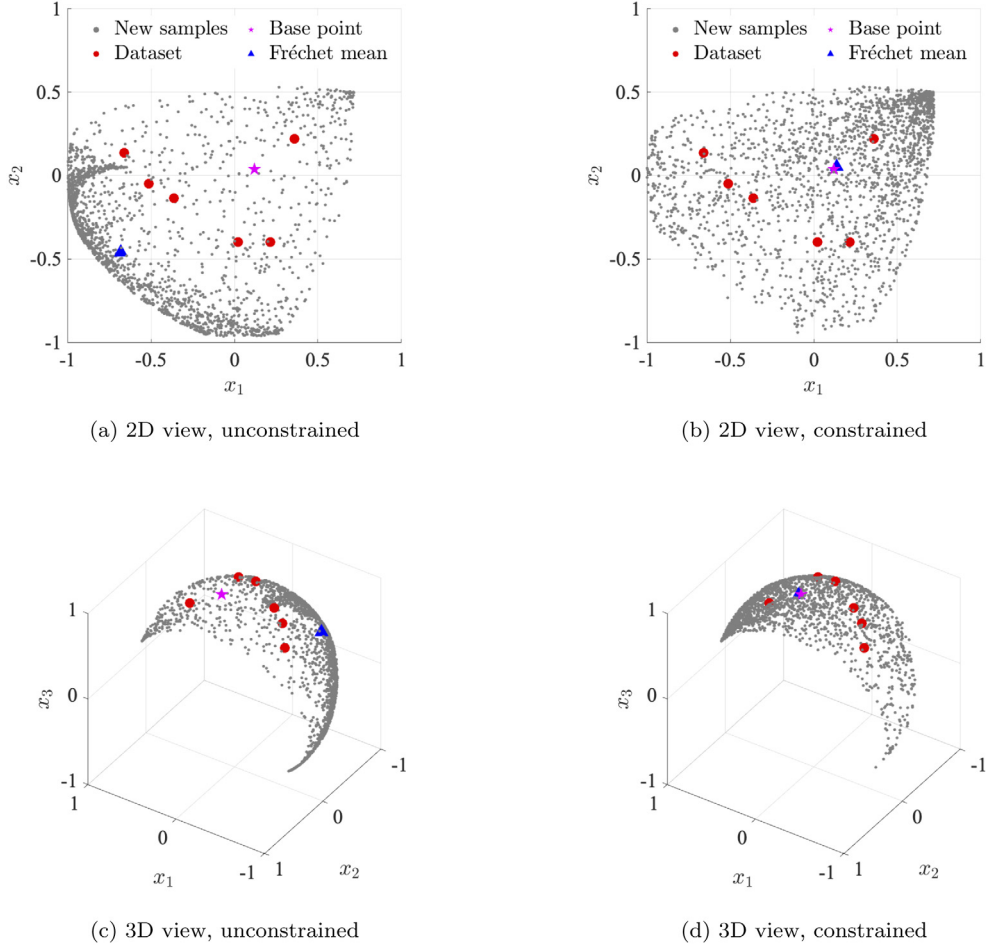


Fig. 4. Left panel (unconstrained sampling): samples obtained for $c = 3$ and $\alpha_i = 0.2$ for $i \in \{1, \dots, 6\}$. Right panel (constrained sampling): samples obtained for $c = 3$ and $\alpha = (0.24, 0.20, 0.18, 0.18, 1.42, 0.27)^T$.

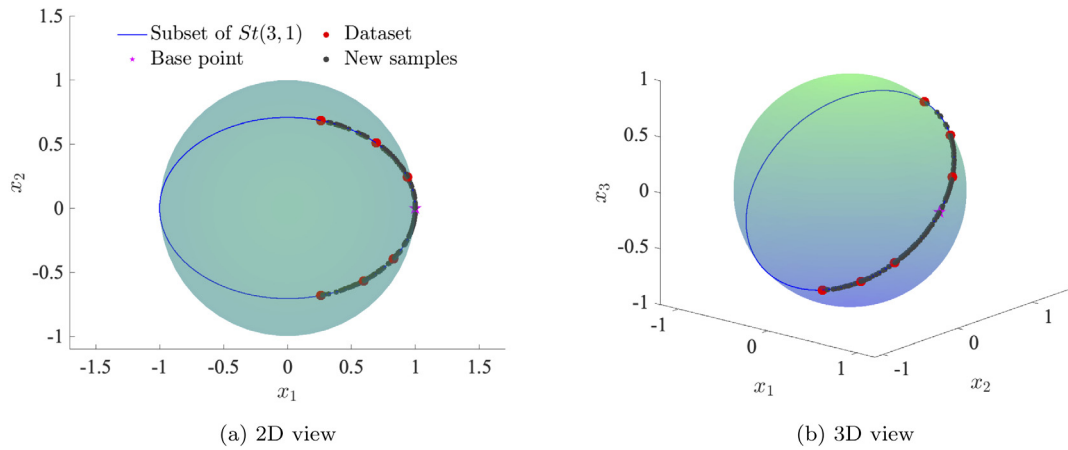


Fig. 5. Samples obtained for $c = 1$ and $\alpha_i = 0.2$ for $i \in \{1, \dots, 6\}$ (no constraint on Fréchet mean). Observe that all samples properly belong to $\mathbb{S}_{3,1}$, owing to the use of the Riemannian projection and retraction operators.

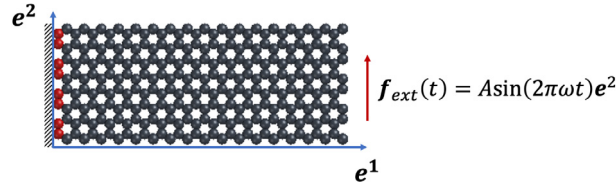


Fig. 6. Single-layer graphene sheet model composed of 272 atoms. A zero homogeneous Dirichlet boundary condition is applied on the left edge, and a harmonic excitation force is applied on the right edge.

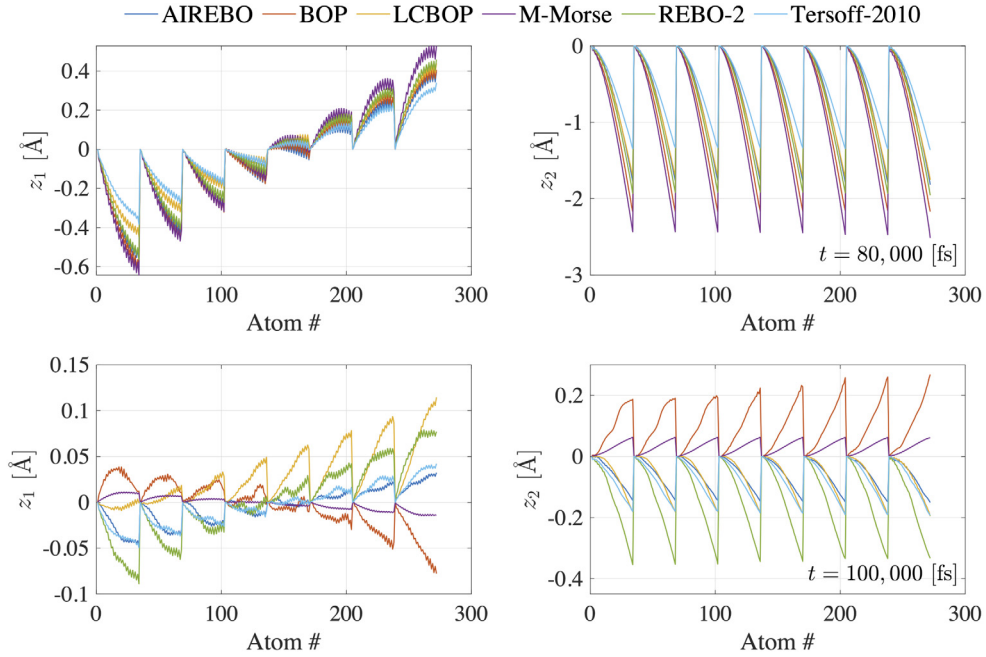


Fig. 7. Snapshots of horizontal and vertical displacements at $t = 80$ [ps] (top row) and 100 [ps] (bottom row), computed with six different interatomic potentials.

to model graphene-based systems are considered, namely AIREBO [11], BOP [12], LCBOP [14], Modified-Morse [48], REBO-2 [13], and Tersoff-2010 [49]. Atom displacement is chosen as the quantity of interest to study the influence of model-form uncertainties in the graphene system. Relaxation is performed through energy minimization before the external force $\mathbf{f}_{ext}$ is applied. Sampling is conducted in the microcanonical ensemble (NVE), with a time step set to 1 [fs] (1×10^{-15} [s]).

3.2.2. Forward simulations and model reduction

The impact of model selection, viewed from the perspective of model uncertainties, is illustrated in Fig. 7 where horizontal and vertical displacements for all atoms are displayed at $t = 80,000$ and $t = 100,000$ [fs], respectively, for the six considered potentials. It is seen that the choice of the potential has a significant impact on the fine-scale dynamics of the graphene system, motivating the use of the proposed approach to quantify and propagate model-form uncertainties at relevant scales.

The POD approach is next employed to construct the reduced-order bases $\{[W^{(i)}]\}_{i=1}^6$ (the bases $\{[W^{(1)}], \dots, [W^{(6)}]\}$ are associated with AIREBO, BOP, LCBOP, Modified-Morse, REBO-2, and Tersoff-2010 potentials, respectively). For each MD configuration (choice of interatomic potential), 1,000 displacement snapshots are collected with a time interval between consecutive snapshots set to 200 [fs] to promote independence (see Eq. (4)). Recall that the global ROB $[W]$ is obtained by concatenating the displacement snapshots for all MD configurations. A singular value decomposition is used to identify the reduced dimension (taken as the minimum over all

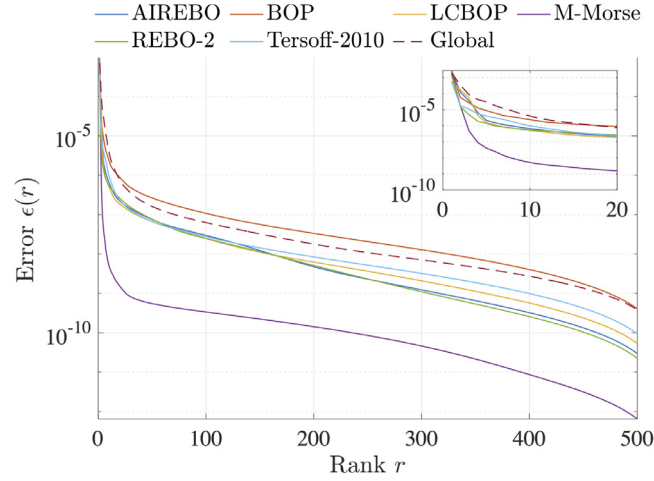


Fig. 8. Graph of the L^2 error function for all scenarios.

configurations) and the associated projection bases. Using a threshold of 1×10^{-4} , we identify the reduced dimension, $n = 5$ (see Fig. 8), and therefore consider stochastic modeling in $\mathbb{S}_{816,5} \subset St(816, 5)$. Note that the dimension of $St(816, 5)$ is: $816 \times 5 - \frac{1}{2} \times 5 \times (5 + 1) = 4,065$.

3.2.3. Sampling results

The proposed stochastic model and sampling procedure are then deployed to generate reduced-order basis samples on $\mathbb{S}_{816,5}$. The scaling parameter c is taken as 1, meaning that only Riemannian convex combinations are considered, and the concentration parameters are computed by solving the quadratic programming problem defined in Eq. (45) (to reduce the distance between the global reduced-order basis $[W]$ and the center of mass of the generated samples):

$$\alpha = (0.494, 0.601, 0.006, 0.236, 0.421, 0.242)^T. \quad (51)$$

In this example, the smallest eigenvalue of $[H]$ (in Eq. (45)) is 0.482, which shows the well-posedness of the quadratic programming problem.

To visualize the dataset and the 2,500 generated samples in a low-dimensional space (here, a two-dimensional space), several commonly used non-linear dimension reduction techniques were tested, including spectral embedding [50], t-SNE [51], UMAP [52], and PACMAP [53]. It was found through extensive numerical experiments that the spectral embedding approach typically delivers representations that can be interpreted more easily, in terms of structure; see Fig. 9. This figure illustrates the fact that all samples are generated inside the convex hull defined by the dataset, with curved edges owing to the use of the non-linear reduction technique (data compression). It is also observed that the Fréchet mean computed with the samples appears close to the global reduced-order basis $[W]$, demonstrating the efficiency of the proposed methodology to identify the concentration parameters based on a Fréchet mean constraint in a molecular dynamics setting.

Such visualization techniques and results should, however, be handled and interpreted with caution, due to the reduction process. A comparative study about such representations is beyond the scope of this work. Their use in the context of reduced-order modeling for dynamical systems, in particular, is an interesting topic that is left for future work.

3.2.4. Forward propagation of model uncertainties

In this section, model-form uncertainties are propagated through Monte Carlo simulations with the stochastic reduced-order model corresponding to the graphene system subjected to harmonic excitation. This step necessitates the selection of the interatomic potential used after pullback in the physical space (to evaluate forces). Two strategies can be pursued at this point. In a first scenario, the same potential is used for all simulations, regardless of the reduced-order basis sample. This potential may be chosen, in practice, as the one minimizing the distance

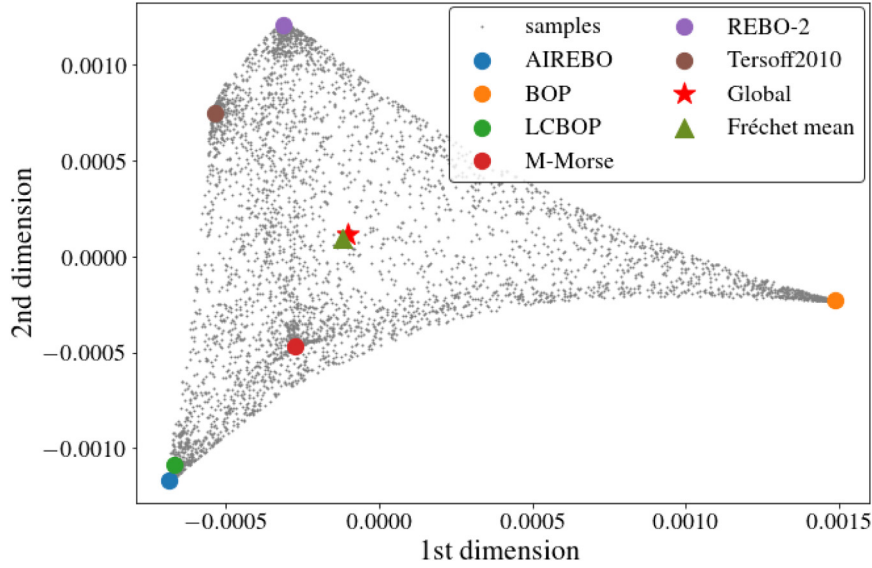


Fig. 9. First two dimensions after dimension reduction using the spectral embedding approach [50]. A total number of 5,000 samples are generated and shown, together with the original samples and the Fréchet mean.

to the mean behavior. A second strategy consists in performing selection for each sample of the reduced-order basis, retaining the potential (in the physical space) that is the closest to the sample under consideration in the reduced-order space. In this case, the potential can be identified by computing relative distances between the sample $[\Phi(\theta_j)]$ and all elements in the dataset (i.e., $[W^{(1)}], \dots, [W^{(m)}]$), using the canonical metric, or by leveraging the definition through a convex combination. Specifically, let $\mathcal{I}_j$, with $1 \leq \mathcal{I}_j \leq m$, be the integer such that $p_{\mathcal{I}_j}(\theta_j) = \max\{p_1(\theta_j), \dots, p_m(\theta_j)\}$. The sample $[\Phi(\theta_j)]$ is then located closer to $[W^{(\mathcal{I}_j)}]$, so that the $(\mathcal{I}_j)$ th potential may be used in the physical space.

Results obtained with the above two strategies are shown in Fig. 10. In this example, the BOP potential [12] is used in the first strategy, and 200 samples are generated using the values given in Section 3.2.3. Snapshots of the vertical displacement (along e^2) are displayed at $t = 22$ and $t = 25$ [ps]. Trajectories computed for the 200 samples of the stochastic reduced-order basis are shown, together with the trajectories corresponding to full-order MD simulations with all six reference potentials. It is seen that both strategies yield fairly similar results in terms of spread. The domain defined by the set of full-order simulations is properly captured by the sampled trajectories, which indicates that model uncertainty has been successfully encoded into stochastic modeling process. It is worth mentioning that the zero Dirichlet boundary condition is also preserved across all samples and full-order models.

Fine-scale uncertainties generated by model error can also be observed using confidence intervals and probability distributions. The mean trajectories and 90% confidence intervals are shown in Fig. 11 for the two selection strategies. The estimated probability density functions for the vertical displacement of atom #100 at $t = 22$ and $t = 25$ [ps] are also shown in Fig. 12. These results show that the choice of the selection strategies does not significantly impact predictions. It should however be noticed that sample-based selection allows to better differentiate between contributions in the dataset; see, e.g., the peak observed for the BOP potential in the bottom-right figure (as compared to the top-right figure) in Fig. 12. Moreover, this strategy does not generate additional computational cost, and does not rely on *a priori* selection. For these reasons, the sample-based selection approach will be used in subsequent calculations and in particular, in the multiscale results presented in Section 3.3.

3.3. Single graphene sheet subjected to tension

3.3.1. System description

We finally model and quantify the impact of model-form uncertainties in both fine- and coarse-scale predictions on a graphene sheet under tension. The graphene sheet is composed of 1,008 carbon atoms with an overall in-plane

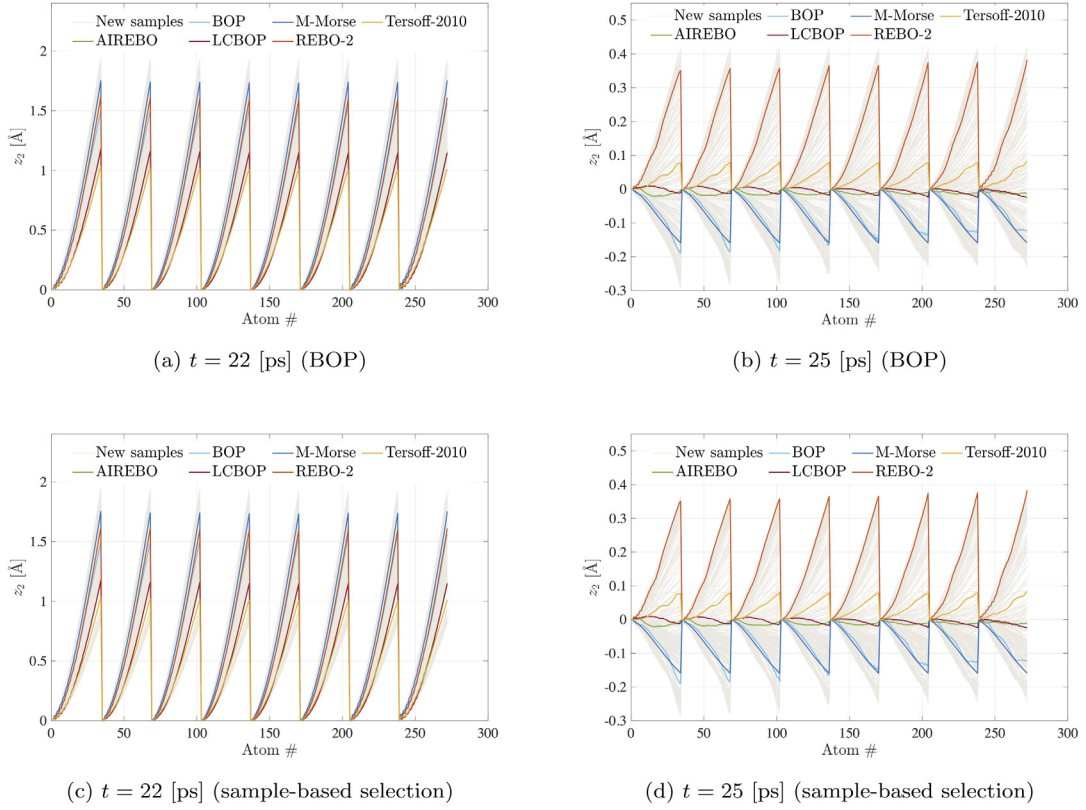


Fig. 10. Trajectories predicted by the reduced-order MD simulations using 200 ROB samples (gray solid lines), together with full-order MD simulation results (colored solid lines) at simulation times $t = 22$ [ps] (left panels) and $t = 25$ [ps] (right panels), respectively. In the top row, the BOP potential is used for all the reduced-order MD simulations, while sample-based selection is carried out in the bottom row. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

size of 50.17×49.74 [Å], a size that is large enough to produce size-independent coarse-scale tensile test results (see [54]). The carbon bond length is selected as 1.418 [Å], in accordance with [55]. The tensile test is conducted in both zigzag and armchair directions, as shown in Fig. 13. In each virtual tensile test, a zero Dirichlet boundary condition is applied to the atoms located on one edge of the sheet (e.g., to the atoms satisfying $x_1 = 0$ for the zigzag direction) and a stretching force is applied to the atoms on the opposite edge. The time step is set to 1 [fs], with a total simulation time of 20 [ps] for both directions. Load stepping is used with increments prescribed every 100 time steps in order to ensure proper relaxation. The loads are specifically defined such that the largest engineering strain rate is equal to 0.22 for the tensile test in the zigzag direction, and to 0.19 in the armchair direction.

3.3.2. Deterministic forward simulations

The uncertainty resulting from the selection of the interatomic potential is evaluated using full-order MD simulations and the AIREBO, BOP, LCBOP, REBO-2, and Tersoff-2010 potentials (see Section 3.2.1).

Two quantities of interest are considered. First, a fine-scale characterization is obtained by analyzing the displacements along the e^1 and e^2 directions. Second, the impact on a coarse-scale property, namely the apparent strain energy, is illustrated. For the sake of comparison, results obtained with the continuum-mechanics-based model presented in [56] are also reported as complementary reference. The relationships between the strain energy and the engineering strain, denoted by u and ϵ respectively, are given by

$$\begin{aligned} u_{zz} &= \frac{1}{2} \frac{E}{1-\nu^2} \epsilon^2 + \frac{1}{6} C_{111} \epsilon^3, \\ u_{ac} &= \frac{1}{2} \frac{E}{1-\nu^2} \epsilon^2 + \frac{1}{6} C_{222} \epsilon^3, \end{aligned} \quad (52)$$

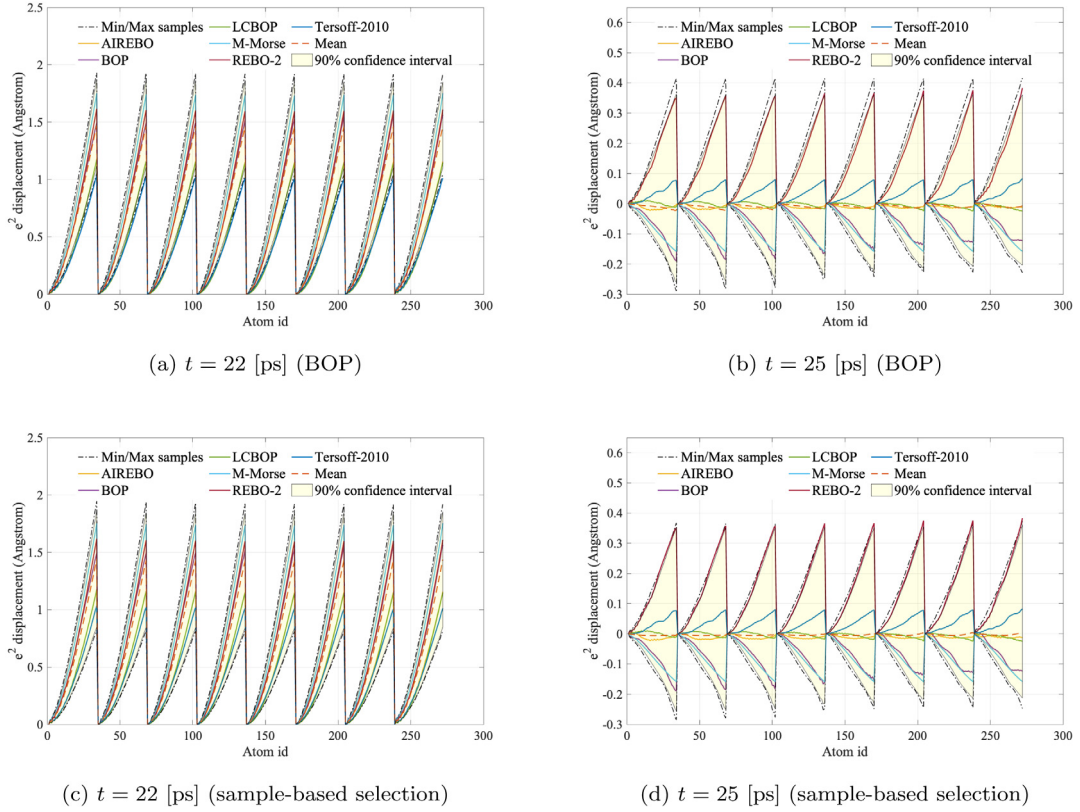


Fig. 11. Mean/extreme trajectory and confidence interval for the vertical displacement of all atoms at $t = 22$ [ps] (left panels) and $t = 25$ [ps] (right panels). In the top row, the BOP potential is used for all the reduced-order MD simulations, while sample-based selection is carried out in the bottom row.

where the subscripts “zz” and “ac” refer to the zigzag and armchair directions, $E = 312$ [N/m] denotes the Young’s modulus, $\nu = 0.31$ is the Poisson ratio, $C_{111} = -1689.2$ [N/m] and $C_{222} = -1487.7$ [N/m] are the elastic constants [56].

The evolution of the strain energy in both directions and for all potentials is shown in Fig. 14. Very large variations induced by model-form uncertainties can be observed. Specifically, the discrepancy in strain energy for a 15%-strain is 26.3% in the zigzag direction, and 40.6% for the armchair direction. Such discrepancies may generate substantial fluctuations when propagated through multiscale approaches [57], which underpins the need to properly capture such variability and perform uncertainty quantification within MD simulations.

3.3.3. Stochastic modeling and forward propagation of model uncertainties

In order to apply the proposed modeling framework, a total number of 500 snapshots are collected for each tensile direction and all interatomic potentials. The five reduced-order bases $\{[W^{(1)}], \dots, [W^{(5)}]\}$ (associated with AIREBO, BOP, LCBO, REBO-2, and Tersoff-2010 potentials, respectively), together with the global reduced-order basis $[W]$, are then calculated using the POD approach. Selecting $n = 10$ modes leads to a truncation error that is less than 10^{-4} for all candidates, so we consider sampling on $\mathbb{S}_{3024,10} \subset St(3024, 10)$.

Model-form uncertainties can then be propagated using the modeling strategy summarized in Section 2.3, combined with a Monte Carlo approach. The concentration parameters are determined by solving the quadratic programming problem given by Eq. (45) to ensure that Fréchet mean of the generated samples are close to the global ROB, which is the base point to define the tangent space. These coefficients are found to be

$$\alpha_{zz} = (0.89, 0.80, 0.45, 0.78, 1.08)^T \quad (53)$$

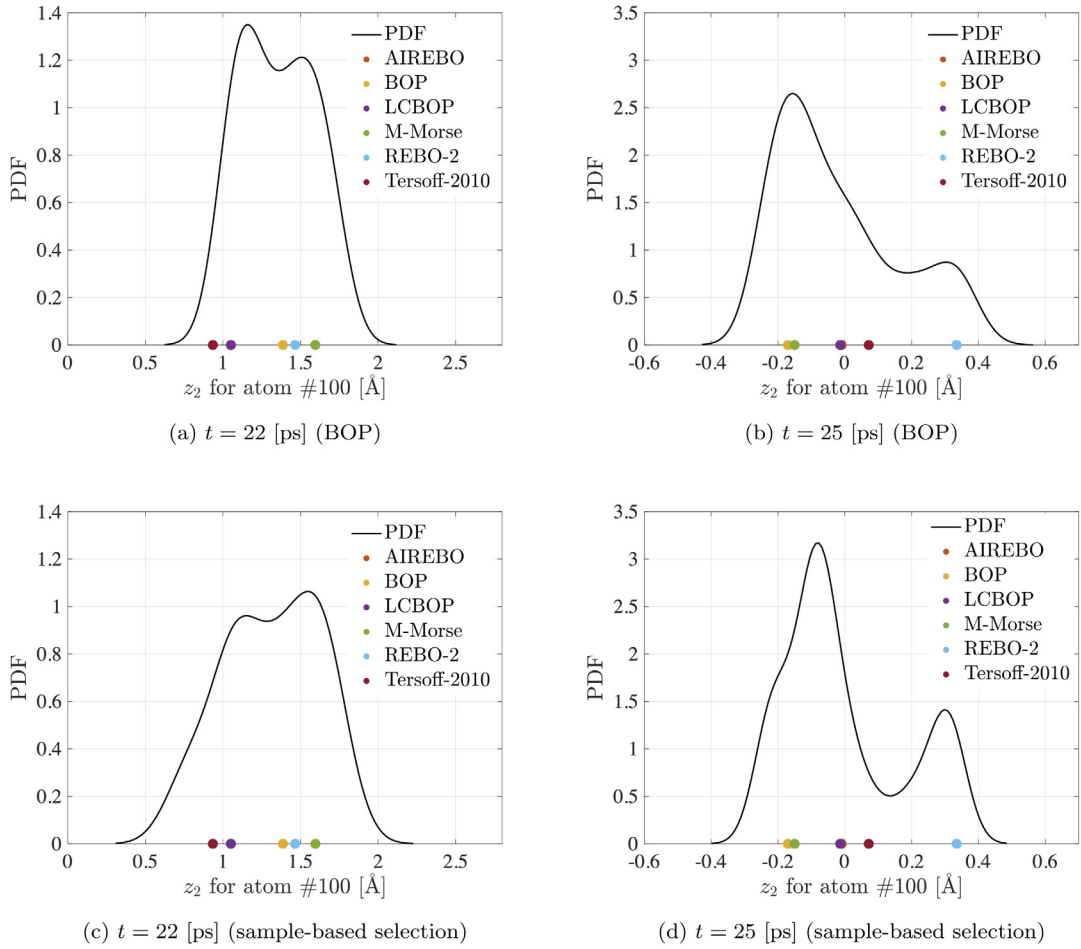


Fig. 12. Estimated probability density function (pdf) for the vertical displacement of atom #100 at $t = 22$ [ps] (left panels) and $t = 25$ [ps] (right panels). Full-order MD simulation results are also reported for the sake of comparison. Note that the markers associated with the AIREBO and LCBOP potentials are quite close to one another. In the top row, the BOP potential is used for all the reduced-order MD simulations, while sample-based selection is carried out in the bottom row.

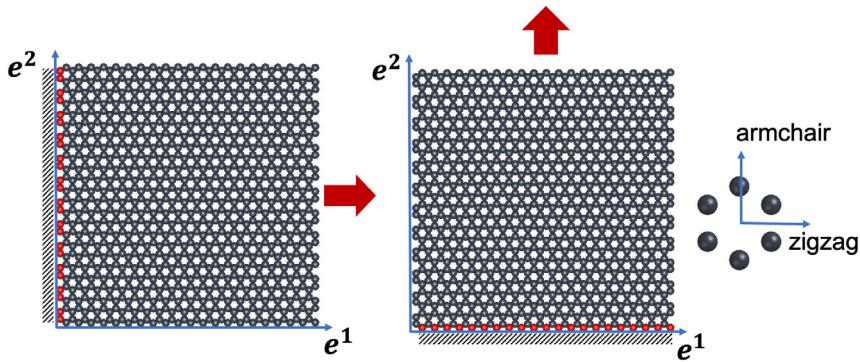


Fig. 13. Single-layer graphene sheet with 1,008 carbon atoms. Tensile test is conducted in zigzag (left subfigure) and armchair (middle subfigure) directions. A zero Dirichlet boundary condition is applied on one edge of the sheet and a driving force is applied to the atoms located on the opposite edge.

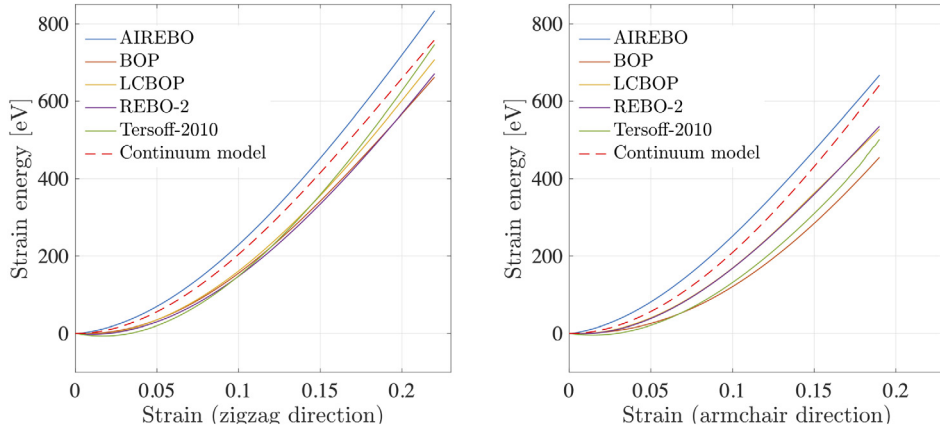


Fig. 14. Evolution of the strain energy (in [eV]) as a function of applied strain, for all potentials and the continuum-mechanics-based model, in the zigzag (left) and armchair (right) direction.

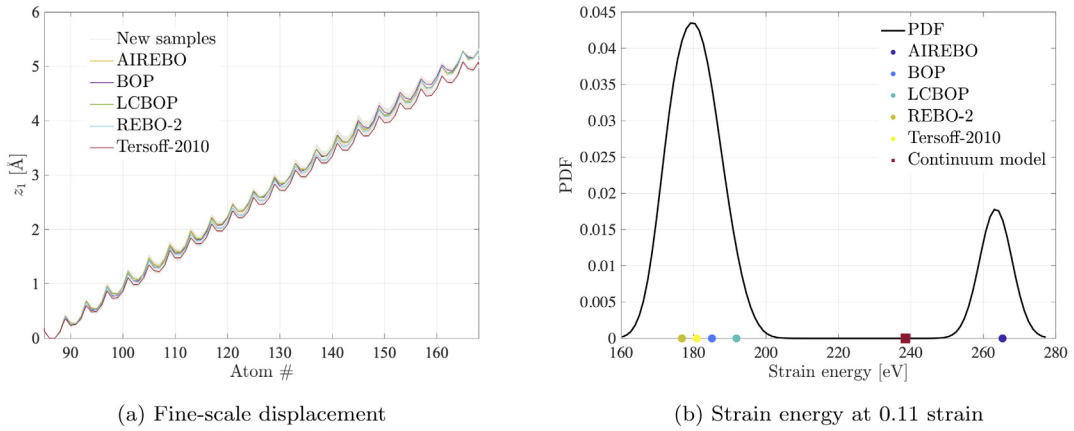


Fig. 15. Fine-scale and coarse-scale predictions for tensile test in the zigzag direction, obtained for $c = 1$ (stochastic Riemannian convex combination) and the concentration parameters given in Eq. (53). The interatomic potential in physical space is chosen using sample-based selection.

for the zigzag direction, and

$$\alpha_{ac} = (0.82, 0.80, 1.14, 0.62, 0.62)^T \quad (54)$$

for the armchair direction.

Fine-scale and coarse-scale stochastic predictions for the tensile test in the zigzag direction are shown in Fig. 15, using Riemannian convex combinations ($c = 1$) and 200 samples. It is seen that while the sampled trajectories are evenly distributed within the region defined by the potential candidates (Fig. 15, left subfigure), the distribution of the coarse-scale properties is limited to a small region around the original full-order-model results (Fig. 15, right subfigure). In order to increase the range of coarse-scale fluctuations (if required based on the application), scaling of the fluctuations can be performed in the tangent space; see Section 2.2.4. A simple and natural way to identify the additional parameter c is to impose that the range of observed values for a given coarse-scale quantity of interest is included in the confidence region predicted by the stochastic model. Other strategies to solve statistical inverse problems can also be deployed, depending on the availability and nature of coarse-scale data.

A total number of 200 reduced-order basis samples are generated on $\mathbb{S}_{3024,10}$ for both the zigzag and armchair directions. The scaling factor is set to $c = 8$ for the zigzag direction, and to $c = 4$ for the armchair direction.

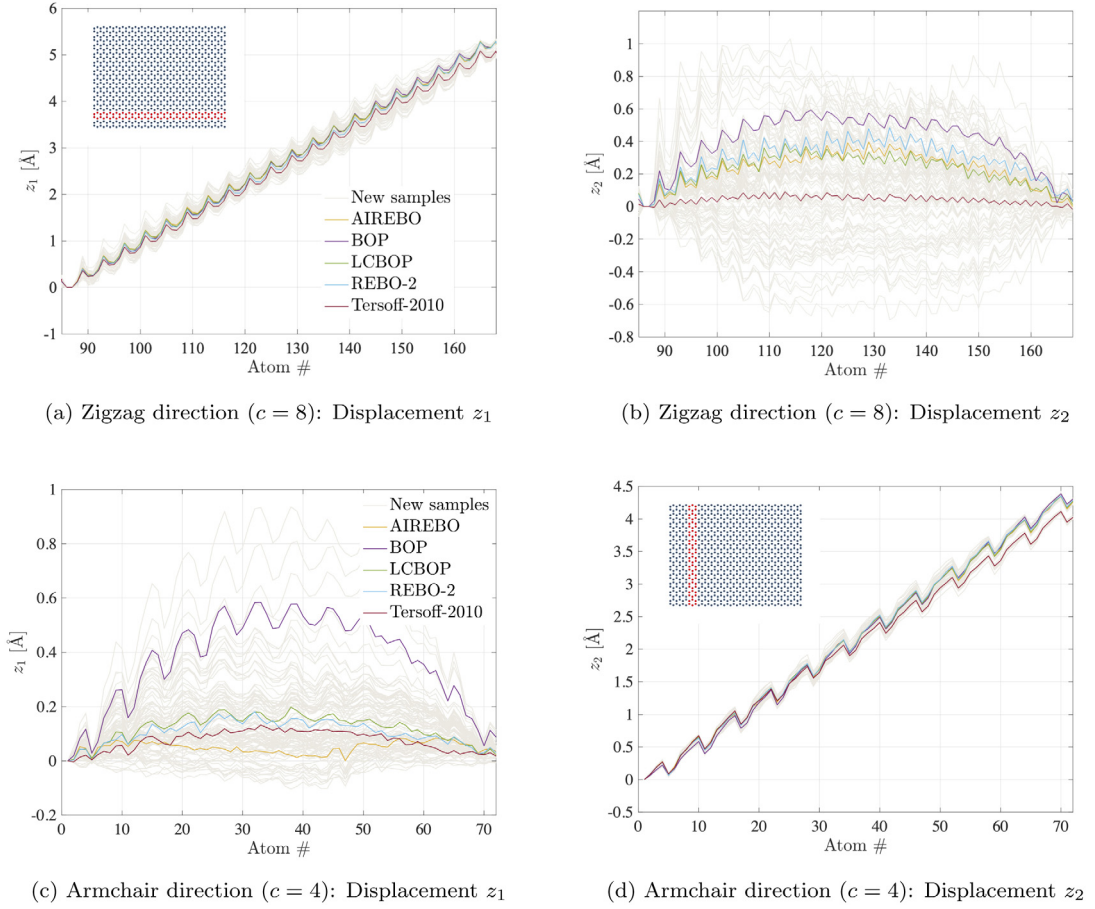


Fig. 16. Atom displacements for tension in the zigzag direction (first row, with scale factor $c = 8$) and armchair direction (second row, with scale factor $c = 4$). Selected atoms are displayed in red in the top-left and bottom-right subfigures. Gray curves correspond to samples obtained with the stochastic reduced-order MD simulations, while the colored curves are associated with full-order MD simulations.

Fine-scale results, in the form of displacements for a few selected atoms, are first shown in Fig. 16 (using sample-based selection for the potential in physical space). Displacement along $\mathbf{e}^1$ and $\mathbf{e}^2$ are collected at simulation time $t = 10$ [ps]. As expected, it is observed that the range of displacements becomes much larger as c increases, hence highlighting the sensitivity to this parameter.

The impact of model uncertainties can also be quantified on the distribution of the coarse-scale strain energy. Fig. 17 shows the 90% confidence interval of the strain energy with regard to the engineering strain in the two stretching directions. The estimated probability density function for the strain energy is shown in Fig. 18 for the two different stretching directions. In these figures, the strain is chosen as 0.11 for the zigzag direction, and as 0.10 for the armchair direction. These results demonstrate the capability of the proposed formulation to capture model-form uncertainties at fine scale and to propagate them on a coarse-scale quantity of interest. Such uncertainties can be properly encoded into the ROB samples such that the distribution of the quantity of interest can be analyzed in a multi-scale pipeline.

4. Conclusion

A Riemannian stochastic representation of model-form uncertainties in molecular dynamics was proposed. The approach relies on a stochastic reduced-order model, defined through a randomized projection basis on a subset of the Stiefel manifold. It was shown that the use of Riemannian projection and retraction operators allows linear constraints, relevant to Dirichlet boundary conditions for instance, to be preserved. This fundamental property

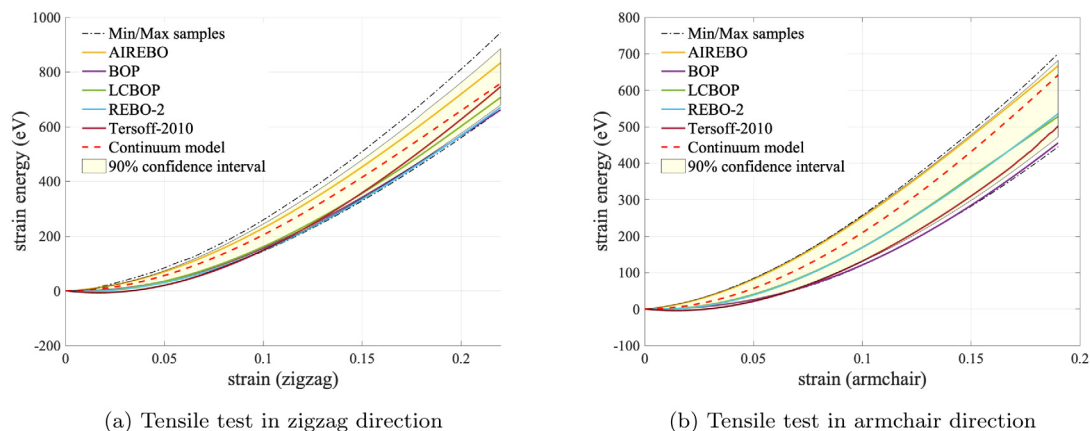


Fig. 17. 90% confidence interval for the coarse-scale strain energy, together with the strain energy obtained with the full-order MD simulations, the continuum model, and the extreme samples.

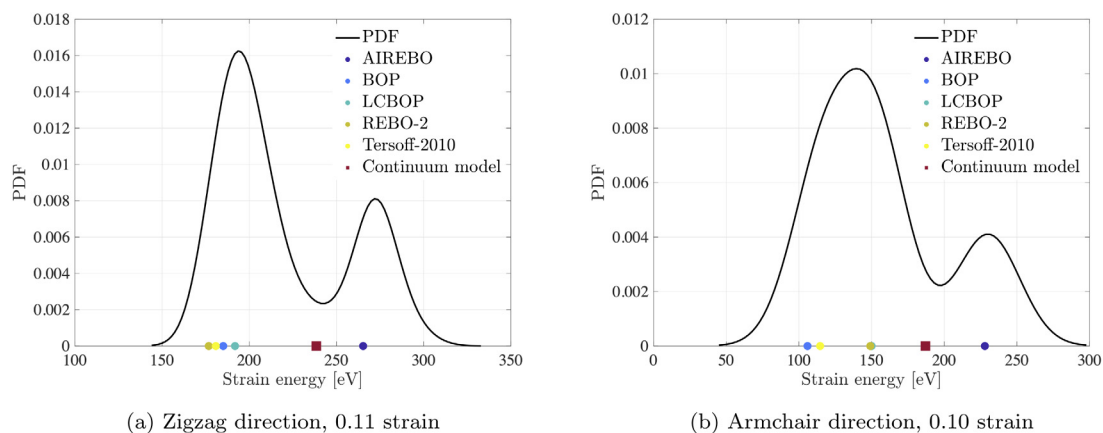


Fig. 18. Estimated probability density function for the strain energy in zigzag direction (left) and armchair direction (right). Values obtained with full-order MD simulations (with all potential candidates) and the continuum-mechanics-based model are also reported.

enables the consideration of convex Riemannian combinations on the tangent space. The proposed formulation offers several advantages, including a simple and interpretable low-dimensional parameterization, the ability to constraint the Fréchet mean solving a quadratic programming problem, and ease of implementation and propagation through stochastic collocation methods. The relevance of the proposed modeling framework was finally demonstrated on various applications, including sampling on the unit sphere and multiscale simulations on graphene-based systems. Perspectives for future work include the application of the proposed stochastic framework to more complex molecular systems, as well as model refinements exploring possible connections with model averaging and model selection strategies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The work of the J.G. was supported by the National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation, under award CMMI-1942928.

Appendix. Computation of the fréchet mean

The algorithm to compute the Fréchet mean on a Stiefel manifold, denoted by $\mathbb{M}$, based on a set of samples is given in Alg. 1 (see [58]).

Algorithm 1 Calculation of Fréchet Mean Based on a Gradient Descent Method

Input: set of samples $\{[Y^{(i)}] \in \mathbb{M}\}_{i=1}^q$, stepsize t , convergence threshold ϵ , algorithms to compute the projection and retraction operators

- 1: Choose a initial guess, denoted by $[Y_{(0)}] \in \mathbb{M}$, for the Fréchet mean, and set $\text{Err} = \text{Inf}$ and $k = 0$
- 2: **while** $\text{Err} > \epsilon$ **do**
- 3: Calculate: $\nabla h([Y_{(k)}]) = -\sum_{i=1}^q P_{[Y_{(k)}]}([Y^{(i)}])$
- 4: Calculate the error: $\text{Err} = \|\nabla h([Y_{(k)}])\|_2$
- 5: Update the Fréchet mean: $[Y_{(k+1)}] = R_{[Y_{(k)}]}(-t^k \nabla h([Y_{(k)}]))$
- 6: $k + 1 \leftarrow k$
- 7: **end while**

Output: Fréchet mean: $[Y] = [Y_{(k)}]$

References

- [1] W.L. Jorgensen, Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water, *J. Am. Chem. Soc.* 103 (2) (1981) 335–340, <http://dx.doi.org/10.1021/ja00392a016>.
- [2] H. Berendsen, J. Postma, W. Van Gunsteren, J. Hermans, Interaction models for water in relation to protein hydration, *Intermolecular Forces* (1981).
- [3] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, *J. Chem. Phys.* 79 (2) (1983) 926–935, <http://dx.doi.org/10.1063/1.445869>.
- [4] H. Berendsen, J. Grigera, T. Straatsma, The missing term in effective pair potentials, *J. Phys. Chem.* 91 (24) (1987) 6269–6271, <http://dx.doi.org/10.1021/j100308a038>.
- [5] J.D. Bernal, R.H. Fowler, A theory of water and ionic solution, with particular reference to hydrogen and hydroxyl ions, *J. Chem. Phys.* 1 (8) (1933) 515–548, <http://dx.doi.org/10.1063/1.1749327>.
- [6] S. Izadi, R. Anandakrishnan, A.V. Onufriev, Building water models: a different approach, *J. Phys. Chem. Lett.* 5 (21) (2014) 3863–3871, <http://dx.doi.org/10.1021/jz501780a>.
- [7] W.L. Jorgensen, J. Tirado-Rives, The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin, *J. Am. Chem. Soc.* 110 (6) (1988) 1657–1666, <http://dx.doi.org/10.1021/ja00214a001>.
- [8] P.A. Olsson, E. Andreasson, E. Bergvall, E.P. Jutemar, V. Petersson, G.C. Rutledge, M. Kroon, et al., All-atomic and coarse-grained molecular dynamics investigation of deformation in semi-crystalline lamellar polyethylene, *Polymer* 153 (2018) 305–316, <http://dx.doi.org/10.1016/j.polymer.2018.07.075>.
- [9] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, *J. Am. Chem. Soc.* 118 (45) (1996) 11225–11236, <http://dx.doi.org/10.1021/ja9621760>.
- [10] S. Sæther, M. Falck, Z. Zhang, A. Lervik, J. He, Thermal transport in polyethylene: the effect of force fields and crystallinity, *Macromolecules* 54 (13) (2021) 6563–6574, <http://dx.doi.org/10.1021/acs.macromol.1c00633>.
- [11] S.J. Stuart, A.B. Tutein, J.A. Harrison, A reactive potential for hydrocarbons with intermolecular interactions, *J. Chem. Phys.* 112 (14) (2000) 6472–6486, <http://dx.doi.org/10.1063/1.481208>.
- [12] X. Zhou, D.K. Ward, M.E. Foster, An analytical bond-order potential for carbon, *J. Comput. Chem.* 36 (23) (2015) 1719–1735, <http://dx.doi.org/10.1002/jcc.23949>.
- [13] D.W. Brenner, O.A. Shenderova, J.A. Harrison, S.J. Stuart, B. Ni, S.B. Sinnott, A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons, *J. Phys.: Condens. Matter* 14 (4) (2002) 783–802, <http://dx.doi.org/10.1088/0953-8984/14/4/312>.
- [14] J.H. Los, A. Fasolino, Intrinsic long-range bond-order potential for carbon: Performance in Monte Carlo simulations of graphitization, *Phys. Rev. B* 68 (2003) 024107, <http://dx.doi.org/10.1103/PhysRevB.68.024107>.
- [15] A.C. Van Duin, S. Dasgupta, F. Lorant, W.A. Goddard, Reaxff: a reactive force field for hydrocarbons, *J. Phys. Chem. A* 105 (41) (2001) 9396–9409, <http://dx.doi.org/10.1021/jp004368u>.

- [16] J. Tersoff, Modeling solid-state chemistry: Interatomic potentials for multicomponent systems, *Phys. Rev. B* 39 (8) (1989) 5566, <http://dx.doi.org/10.1103/PhysRevB.39.5566>.
- [17] T.E. Gartner III, A. Jayaraman, Modeling and simulations of polymers: a roadmap, *Macromolecules* 52 (3) (2019) 755–786, <http://dx.doi.org/10.1021/acs.macromol.8b01836>.
- [18] M. Salloum, J.A. Templeton, Inference and uncertainty propagation of atomistically informed continuum constitutive laws, part 2: Generalized continuum models based on gaussian processes, *Int. J. Uncertain. Quantif.* 4 (2) (2014) <http://dx.doi.org/10.1615/Int.J.UncertaintyQuantification.2014008154>.
- [19] T.-T. Le, J. Guillemot, C. Soize, Stochastic continuum modeling of random interphases from atomistic simulations. Application to a polymer nanocomposite, *Comput. Methods Appl. Mech. Engrg.* 303 (2016) 430–449, <http://dx.doi.org/10.1016/j.cma.2015.10.006>.
- [20] H. Wang, J. Guillemot, C. Soize, Modeling uncertainties in molecular dynamics simulations using a stochastic reduced-order basis, *Comput. Methods Appl. Mech. Engrg.* 354 (2019) 37–55, <http://dx.doi.org/10.1016/j.cma.2019.05.020>.
- [21] M. Koslowski, A. Strachan, Uncertainty propagation in a multiscale model of nanocrystalline plasticity, *Reliab. Eng. Syst. Saf.* 96 (9) (2011) 1161–1170, <http://dx.doi.org/10.1016/j.res.2010.11.011>.
- [22] S. Liu, A. Gerisch, M. Rahimi, J. Lang, M.C. Böhm, F. Müller-Plathe, Robustness of a new molecular dynamics–finite element coupling approach for soft matter systems analyzed by uncertainty quantification, *J. Chem. Phys.* 142 (10) (2015) 104105, <http://dx.doi.org/10.1063/1.4914020>.
- [23] B. Kouchmeshky, N. Zabarar, Microstructure model reduction and uncertainty quantification in multiscale deformation processes, *Comput. Mater. Sci.* 48 (2) (2010) 213–227, <http://dx.doi.org/10.1016/j.commatsci.2010.01.001>.
- [24] R. Bostanabad, B. Liang, J. Gao, W.K. Liu, J. Cao, D. Zeng, X. Su, H. Xu, Y. Li, W. Chen, Uncertainty quantification in multiscale simulation of woven fiber composites, *Comput. Methods Appl. Mech. Engrg.* 338 (2018) 506–532, <http://dx.doi.org/10.1016/j.cma.2018.04.024>.
- [25] L. Mehrez, J. Fish, V. Aitharaju, W.R. Rodgers, R. Ghanem, A PCE-based multiscale framework for the characterization of uncertainties in complex systems, *Comput. Mech.* 61 (1) (2018) 219–236, <http://dx.doi.org/10.1007/s00466-017-1502-4>.
- [26] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework, *J. Chem. Phys.* 137 (14) (2012) 144103, <http://dx.doi.org/10.1063/1.4757266>.
- [27] P.E. Hadjidoukas, P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, *II4U*: A high performance computing framework for Bayesian uncertainty quantification of complex models, *J. Comput. Phys.* 284 (2015) 1–21, <http://dx.doi.org/10.1016/j.jcp.2014.12.006>.
- [28] F. Rizzi, R. Jones, B. Debusschere, O. Knio, Uncertainty quantification in MD simulations of concentration driven ionic flow through a silica nanopore. I. Sensitivity to physical parameters of the pore, *J. Chem. Phys.* 138 (19) (2013) 194104, <http://dx.doi.org/10.1063/1.4804666>.
- [29] F. Rizzi, R. Jones, B. Debusschere, O. Knio, Uncertainty quantification in MD simulations of concentration driven ionic flow through a silica nanopore. II. Uncertain potential parameters, *J. Chem. Phys.* 138 (19) (2013) 194105, <http://dx.doi.org/10.1063/1.4804669>.
- [30] C. Kim, O. Borodin, G.E. Karniadakis, Quantification of sampling uncertainty for molecular dynamics simulation: Time-dependent diffusion coefficient in simple fluids, *J. Comput. Phys.* 302 (2015) 485–508, <http://dx.doi.org/10.1016/j.jcp.2015.09.021>.
- [31] G. Imbalzano, Y. Zhuang, V. Kapil, K. Rossi, E.A. Engel, F. Grasselli, M. Ceriotti, Uncertainty estimation for molecular dynamics and sampling, *J. Chem. Phys.* 154 (7) (2021) 074102, <http://dx.doi.org/10.1063/5.0036522>.
- [32] K. Farrell, J.T. Oden, D. Faghihi, A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems, *J. Comput. Phys.* 295 (2015) 189–208, <http://dx.doi.org/10.1016/j.jcp.2015.03.071>.
- [33] K. Farrell-Maupin, J. Oden, Adaptive selection and validation of models of complex systems in the presence of uncertainty, *Res. Math. Sci.* 4 (1) (2017) 1–15, <http://dx.doi.org/10.1186/s40687-017-0104-2>.
- [34] S.T. Reeve, A. Strachan, Error correction in multi-fidelity molecular dynamics simulations using functional uncertainty quantification, *J. Comput. Phys.* 334 (2017) 207–220, <http://dx.doi.org/10.1016/j.jcp.2016.12.039>.
- [35] C. Soize, C. Farhat, A nonparametric probabilistic approach for quantifying uncertainties in low-dimensional and high-dimensional nonlinear models, *Internat. J. Numer. Methods Engrg.* 109 (6) (2017) 837–888, <http://dx.doi.org/10.1002/nme.5312>.
- [36] C. Farhat, R. Tezaur, T. Chapman, P. Avery, C. Soize, Feasible probabilistic learning method for model-form uncertainty quantification in vibration analysis, *AIAA J.* 57 (11) (2019) 4978–4991, <http://dx.doi.org/10.2514/1.J057797>.
- [37] P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, USA, 2007.
- [38] Q. Rentmeesters, *Algorithms for data fitting on some common homogeneous spaces* (Ph.D. thesis), Université Catholique de Louvain, 2013.
- [39] R. Zimmermann, Manifold interpolation and model reduction, 2019, arXiv preprint [arXiv:1902.06502](https://arxiv.org/abs/1902.06502).
- [40] R. Zimmermann, A matrix-algebraic algorithm for the Riemannian logarithm on the Stiefel manifold under the canonical metric, *SIAM J. Matrix Anal. Appl.* 38 (2) (2017) 322–342, <http://dx.doi.org/10.1137/16M1074485>.
- [41] A. Edelman, T.A. Arias, S.T. Smith, The geometry of algorithms with orthogonality constraints, *SIAM J. Matrix Anal. Appl.* 20 (2) (1998) 303–353, <http://dx.doi.org/10.1137/S0895479895290954>.
- [42] B. Afsari, R. Tron, R. Vidal, On the convergence of gradient descent for finding the Riemannian center of mass, *SIAM J. Control Optim.* 51 (3) (2013) 2230–2260, <http://dx.doi.org/10.1137/12086282X>.
- [43] D. Xiu, G.E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.* 24 (2) (2002) 619–644, <http://dx.doi.org/10.1137/S1064827501387826>.
- [44] C. Soize, R. Ghanem, Physical systems with random uncertainties: Chaos representations with arbitrary probability measure, *SIAM J. Sci. Comput.* 26 (2) (2004) 395–410, <http://dx.doi.org/10.1137/S1064827503424505>.
- [45] R. Ghanem, D. Higdon, H. Owahdi, *Handbook of Uncertainty Quantification*, Vol. 6, Springer, 2017.
- [46] O. Le Maître, O.M. Knio, *Spectral Methods for Uncertainty Quantification: With Applications To Computational Fluid Dynamics*, Springer Science & Business Media, 2010.

- [47] A.P. Thompson, H.M. Aktulga, R. Berger, D.S. Bolintineanu, W.M. Brown, P.S. Crozier, P.J. in't Veld, A. Kohlmeyer, S.G. Moore, T.D. Nguyen, R. Shan, M.J. Stevens, J. Tranchida, C. Trott, S.J. Plimpton, LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, *Comput. Phys. Comm.* 271 (2022) 108171, <http://dx.doi.org/10.1016/j.cpc.2021.108171>.
- [48] T. Belytschko, S.P. Xiao, G.C. Schatz, R.S. Ruoff, Atomistic simulations of nanotube fracture, *Phys. Rev. B* 65 (2002) 235430, <http://dx.doi.org/10.1103/PhysRevB.65.235430>.
- [49] L. Lindsay, D.A. Broido, Optimized Tersoff and brenner empirical potential parameters for lattice dynamics and phonon thermal transport in carbon nanotubes and graphene, *Phys. Rev. B* 81 (2010) 205441, <http://dx.doi.org/10.1103/PhysRevB.81.205441>.
- [50] U. Von Luxburg, A tutorial on spectral clustering, *Stat. Comput.* 17 (4) (2007) 395–416, <http://dx.doi.org/10.1007/s11222-007-9033-z>.
- [51] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* 9 (11) (2008) URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- [52] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, 2018, <http://dx.doi.org/10.48550/ARXIV.1802.03426>.
- [53] Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik, Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization, *J. Mach. Learn. Res.* 22 (201) (2021) 1–73, URL <http://jmlr.org/papers/v22/20-1061.html>.
- [54] Z. Ni, H. Bu, M. Zou, H. Yi, K. Bi, Y. Chen, Anisotropic mechanical properties of graphene sheets from molecular dynamics, *Physica B* 405 (5) (2010) 1301–1306, <http://dx.doi.org/10.1016/j.physb.2009.11.071>.
- [55] J. Slonczewski, P. Weiss, Band structure of graphite, *Phys. Rev.* 109 (2) (1958) 272, <http://dx.doi.org/10.1103/PhysRev.109.272>.
- [56] E. Cadelano, P.L. Palla, S. Giordano, L. Colombo, Nonlinear elasticity of monolayer graphene, *Phys. Rev. Lett.* 102 (23) (2009) 235502, <http://dx.doi.org/10.1103/PhysRevLett.102.235502>.
- [57] J. Fish, G.J. Wagner, S. Keten, Mesoscopic and multiscale modelling in materials, *Nature Mater.* 20 (6) (2021) 774–786, <http://dx.doi.org/10.1038/s41563-020-00913-0>.
- [58] D. Bryner, Endpoint geodesics on the stiefel manifold embedded in euclidean space, *SIAM J. Matrix Anal. Appl.* 38 (4) (2017) 1139–1159, <http://dx.doi.org/10.1137/16M1103099>.