Computer Methods in Applied Mechanics and Engineering 431 (2024) 117323

Contents lists available at ScienceDirect

Comput. Methods Appl. Mech. Engrg.

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

Stochastic symplectic reduced-order modeling for model-form uncertainty quantification in molecular dynamics simulations in various statistical ensembles

S. Kounouho [a], R. Dingreville [b], J. Guilleminot [a,*]

[a] Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, United States of America
[b] Center for Integrated Nanotechnologies (CINT), Sandia National Laboratories, Albuquerque, NM, United States of America

ARTICLE INFO

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

ABSTRACT

This work focuses on the representation of model-form uncertainties in molecular dynamics simulations in various statistical ensembles. In prior contributions, the modeling of such uncertainties was formalized and applied to quantify the impact of, and the error generated by, pair-potential selection in the microcanonical ensemble (NVE). In this work, we extend this formulation and present a linear-subspace reduced-order model for the canonical (NVT) and isobaric (NPT) ensembles. The symplectic reduced-order basis is randomized on the tangent space of the Stiefel manifold to provide topological relationships and capture model-form uncertainty. Using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS), we assess the relevance of these stochastic reduced-order atomistic models on canonical problems involving a Lennard-Jones fluid and an argon crystal melt.

1. Introduction

Molecular dynamics (MD) simulations have become an indispensable tool in the materials science and computational mechanics toolbox to study materials at fine scales [1]. They are routinely used to analyze and predict the evolution of a very broad range of physical properties, from elasticity [2], to defect dynamics [3], to phase transformation [4], to fracture behavior [5], and can be coupled with continuum descriptions to enable materials design and discovery across spatial scales [6]. In practice, conducting such simulations requires the selection and application-specific calibration of force fields and interatomic potentials which are used to describe the interactions between atoms. However, their calibration are not uniquely defined, neither in terms of optimal parameters nor with respect to their functional form—the identification of which depend on the specific properties targeted, such as lattice constants, elastic constants, vibrational frequencies, binding energy, etc. (obtained from, e.g., density functional theory simulations or physical experiments) that are fed into the calibration process as ground truth. Non-exhaustive lists of force fields and potentials for elements (or multi-element systems and non-elemental materials) can be found in the Interatomic Potentials Repository, hosted by the National Institute of Standards and Technology (NIST) for instance [7,8]. Such model misspecification can dramatically impact the accuracy of the predictions [9,10], which raises the important challenge of developing appropriate uncertainty quantification (UQ) techniques that enable the integration of relevant forms of uncertainties on property predictions.

Most of the papers related to UQ for MD simulations have focused on the integration of parametric uncertainties in potential parameters, such as depth and distance parameters in a Lennard-Jones potential [11–14]. Other studies have investigated uncertainties induced by finite sampling [15] and reproducibility issues [16], as well as uncertainties in models built using machine-learning

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

https://doi.org/10.1016/j.cma.2024.117323

(ML) techniques [17–21]. The treatment of model-form uncertainties — which are induced by the *functional forms* of the potentials, rather than their parameters — remains comparatively more elusive. This lack of fundamental results mostly stems from the fact that randomizing a functional form is more intricate than building stochastic models for coefficients: the definition of appropriate encoders for model information, the construction of probability measures, and the integration of the mathematical constraints that arise in such formulations all constitute open problems. The consideration of multiple model candidates was addressed through the lens of model selection in prior work [22,23]. In these papers, a methodology was proposed to identify the "best" model candidate under a given validation scenario, using concepts (*e.g.*, likelihood and plausibility) from Bayesian analysis. Other studies have considered the issue of model correction, assuming that predictions are realized with one model that does not exactly match the underlying ground truth model. Such corrections can be achieved using an additive corrector in the solution space (*e.g.*, with Gaussian process regression (GPR) [24]), or functional perturbations [25]. It should be noted that these contributions do not proceed with the construction of a probability measure over the space of model candidates and typically deliver deterministic predictions (potentially augmented with some statistical fluctuations capturing the epistemic uncertainties induced by the chosen class of surrogates for GPR).

The aim of this work is to advance a stochastic formulation enabling the representation of model-form uncertainties in MD simulations, given a set of model candidates identified through domain expertise or learned using machine learning (ML). This paper leverages two sets of results building on prior work [26,27], and extends these results to the case of commonly-employed statistical ensembles. In the former contribution by Soize and Farhat [26], it was shown that model-form uncertainties in nonlinear dynamical systems described by a given unique model — a setting that we refer to as the uni-model setting in this paper — can be captured using a stochastic reduced-order basis formulation. In the work by Zhang and Guilleminot [27], a novel stochastic reduced-order representation was specifically developed for the multi-model setting. While the approach by Zhang and Guilleminot [27] presents noticeable benefits, including a simple and interpretable low-dimensional parameterization, the ability to constraint the mean of the stochastic reduced-order basis, and ease of implementation and propagation, it was formulated for the micro-canonical ensemble; that is, without constraints in the phase space. This appears as a restriction for broad adoption in a wide range of MD simulation setups. The overarching goal of this work is therefore to develop a new stochastic reduced-order representation that ensures stability in arbitrary statistical ensembles (almost surely). As such, we first derive the Hamiltonian formulation for the equations of motion with standard control variables. This critical step allows us to clarify the definition of the snapshot matrix in the model reduction framework (here, a proper symplectic decomposition), for all considered ensembles. We then present the probabilistic formulation, and derive important results pertaining to inference and dynamical behavior. Through numerical examples, we finally demonstrate the capabilities of the approach to ensure control along trajectories and to capture model-form variability in forward simulations.

This paper is organized as follows. In Section 2, we present a unified (deterministic) reduced-order model for MD simulations and specifically define scaling factors for the matrix of snapshots used to compute the projection basis. In Section 3, we summarize the probabilistic model ensuring well-posedness in the almost sure sense. Numerical results are next provided in Section 4, including deterministic results demonstrating stability in control (for pressure and temperature variables) and stochastic results illustrating the integration of model-form uncertainties. Concluding remarks are finally given in Section 5.

The following notation is used throughout this paper.

- d: Dimension of the physical domain.
- n: Reduced-order dimension.
- $\boldsymbol{q}$: Position vector (physical space).
- $\boldsymbol{p}$: Momentum vector (physical space).
- $\boldsymbol{x}$: Phase-space vector (physical space).
- $\boldsymbol{y}$: Position vector (reduced-order space).
- $\boldsymbol{\pi}$: Momentum vector (reduced-order space).
- $\boldsymbol{z}$: Phase-space vector (reduced-order space).
- $\widehat{\mathbf{x}}$: Virtual phase-space vector (physical space).
- $\widehat{\mathbf{z}}$: Virtual phase-space vector (reduced-order space).
- N: Number of degrees of freedom in the system.
- N_0: Number of linear constraints in the system.
- N_a: Number of atoms.
- N_t: Number of timesteps.
- N_s: Number of snapshots.
- $[I_n]$: Identity matrix of size $n \times n$.
- $[0_n]$: Zero matrix of size $n \times n$.
- $\mathbf{0}_n$: Zero vector of length n.

2. Reduced-order modeling for molecular dynamics simulations

2.1. Reduced-order modeling in free space

Let $\boldsymbol{q} \in \mathbb{R}^N$ and $\boldsymbol{p} \in \mathbb{R}^N$ be the position and momentum vectors associated with a set of N_a particles in d dimensions, respectively, with $N = d \times N_a$. Let $\boldsymbol{x} = (\boldsymbol{q}; \boldsymbol{p}) \in \mathbb{R}^{2N}$ denote the phase-space vector, where the semicolon indicates vertical concatenation. Let

$H : \mathbb{R}^{2N} \to \mathbb{R}$ denote the Hamiltonian function characterizing the motion of the particles. Specifically, the evolution of the system is described by the differential equation

$$\dot{\boldsymbol{x}} = [J_{2N}]\nabla_{\boldsymbol{x}} H(\boldsymbol{x}), \tag{2.1}$$

where

$$[J_{2N}] = \begin{bmatrix} [0_N] & [I_N] \\ -[I_N] & [0_N] \end{bmatrix} \tag{2.2}$$

is the (skew-symmetric) Poisson matrix. For later use, the equations of motion (EOM) for the *physical* Hamiltonian is symbolically written as

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}), \tag{2.3}$$

with $\boldsymbol{f} : \mathbb{R}^{2N} \to \mathbb{R}^{2N}$ (under proper initial conditions). We assume that the position vector $\boldsymbol{q}$ satisfies a set of linear constraints, given by

$$[B]^T \boldsymbol{q}(t) = \boldsymbol{0}_{N_0}, \quad \forall t \geq 0, \tag{2.4}$$

where N_0 is the number of constraints in the system and $[B] \in \mathbb{R}^{N \times N_0}$ satisfies $[B]^T[B] = [I_{N_0}]$. In this work, we assume that the above constraints represent homogeneous Dirichlet boundary conditions and that the mass of each particle does not change over time. In this case, the momentum vector $\boldsymbol{p}$ also satisfies the equation

$$[B]^T \boldsymbol{p}(t) = \boldsymbol{0}_{N_0}, \quad \forall t \geq 0. \tag{2.5}$$

In popular MD codes (such as the Large-scale Atomic/Molecular Massively Parallel Simulator [28], LAMMPS in short), the above equations of motion are integrated using a modified velocity-Verlet algorithm, which is symplectic (*i.e* it ensures that certain fundamental properties of a Hamiltonian system, like total energy conservation, are preserved over time during the simulation) and time-reversible.

A reduced-order model (ROM) can then be defined by using a proper symplectic decomposition (PSD); see in a non-exhaustive manner Refs. [29–34] for linear subspace formulations, as well as Sharma et al. [35] for an extension with higher-order terms. Note that the construction of ROMs based on PSDs is a very active research area that is not the focus of the present study.

Definition 1 (*Snapshot Matrix for PSD Without Control*)**.** Let $t_0, \ldots, t_{N_t}$ be a discretization of the time interval [0,T], where $t_j = j\Delta t$ with time step Δt. Let $\mathcal{J} = \{j_1, \ldots, j_{N_s}\} \subset \{1, \ldots, N_t\}$, with $1 \leq N_s \leq N_t$, be the set of indices identifying N_s (possibly non-ordered) snapshots. Let γ be a weighting coefficient that balances accuracy in the reconstructions of the position and momentum vectors after basis truncation. Then the snapshot matrix for a PSD without control variables is defined as

$$[X] = \left[\boldsymbol{q}(t_{j_1}) - \boldsymbol{q}(0), \ldots, \boldsymbol{q}(t_{j_{N_s}}) - \boldsymbol{q}(0), \gamma \boldsymbol{p}(t_{j_1}), \ldots, \gamma \boldsymbol{p}(t_{j_{N_s}})\right]. \tag{2.6}$$

To minimize the projection error for the position, we choose $\gamma = 0$ (see Peng and Mohseni [29] for a discussion). We introduce the singular value decomposition

$$[X] = [U][S][V]^T, \tag{2.7}$$

where the sequence of singular values is nonincreasing. A reduced-order basis (ROB) $[\Phi]$ can be obtained by retaining the n first columns of $[U]$. By construction, $[\Phi]$ satisfies the orthogonality property $[\Phi]^T[\Phi] = [I_n]$, as well as the property

$$[B]^T[\Phi] = [0_{N_0,n}], \tag{2.8}$$

inherited from the aforementioned linear constraints. The symplectic reduced-order basis $[\Psi] \in \mathbb{R}^{2N \times 2n}$ takes the form

$$[\Psi] = \mathrm{diag}([\Phi], [\Phi]) \tag{2.9}$$

and satisfies

$$[\Psi]^T[J_{2N}][\Psi] = [J_{2n}]. \tag{2.10}$$

The Galerkin projection (*i.e.*, the projection from the physical to reduced-order space) is then defined by the linear symplectic lift $\Gamma : \mathbb{R}^{2n} \to \mathbb{R}^{2N}$ given by [35]

$$\Gamma(\boldsymbol{z}) = [\Psi]\boldsymbol{z} + \boldsymbol{x}_0, \tag{2.11}$$

where $\boldsymbol{z} = (\boldsymbol{y}; \boldsymbol{\pi}) \in \mathbb{R}^{2n}$ is the reduced phase-state space vector, with $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{\pi} \in \mathbb{R}^n$ the reduced position and momentum vectors, respectively, and $\boldsymbol{x}_0 = (\boldsymbol{q}(0); \boldsymbol{0}_N) \in \mathbb{R}^{2N}$ is a reference vector. The projection of the Hamiltonian differential equation into the reduced-order space yields

$$\dot{\boldsymbol{z}} = [J_{2n}]\nabla_{\boldsymbol{z}} \mathcal{H}(\boldsymbol{z}), \tag{2.12}$$

where $\mathcal{H} = H \circ \Gamma$ denotes the *reduced* Hamiltonian. Note that due to the symplectic projection, the reduced Hamiltonian $\mathcal{H}$ preserves the symplectic structure in the original Hamiltonian H—a property that is generally lost through standard Galerkin projection [29].

The reduced-order EOM can be derived from the physical EOM by the symplectic lift:

$$\dot{\boldsymbol{z}} = \widetilde{\boldsymbol{f}}(\boldsymbol{z}), \text{ with } \widetilde{\boldsymbol{f}}(\boldsymbol{z}) = [\Psi]^T \boldsymbol{f}(\Gamma(\boldsymbol{z})). \tag{2.13}$$

Since the parameters $[\Psi]$ and $\boldsymbol{x}_0$ in the symplectic lift do not depend on time, the projection can be used with the symplectic integrator while maintaining time reversibility.

2.2. Reduced-order modeling with dynamical control

Control algorithms allow for the modeling of systems at constant temperature (also referred to as NVT or canonical ensemble) and/or pressure (also referred to as NPT or isothermal–isobaric ensemble), with the aim of controlling specific thermodynamic states to better reflect experimental conditions and provide ways to perform virtual experiments. For instance, control in pressure enables the computation of effective elastic parameters, mimicking the action of statically uniform boundary conditions in homogenization for continuum media [36]. Current state-of-the-art dynamical methods for pressure and temperature controls were studied by Andersen [37] and Nosé [38], and later modified by Tuckerman et al. [39]. The general approach introduces a set of virtual coordinates scaled by a set of control variables and their corresponding momenta. The coupling between position, momentum, and control variables is described by an extended Hamiltonian. For example, in the canonical ensemble described in Nosé [38], the extended Hamiltonian takes the form

$$\widehat{H}(\widehat{\boldsymbol{q}}, \widehat{\boldsymbol{p}}, s, p_s) = \frac{1}{s^2}\widehat{\boldsymbol{p}}^T [M]^{-1} \widehat{\boldsymbol{p}} + \phi(\widehat{\boldsymbol{q}}) + \frac{p_s^2}{2Q} + g k_B T_{\text{tgt}} \ln(s), \tag{2.14}$$

where $\widehat{\boldsymbol{q}}$ and $\widehat{\boldsymbol{p}}$ are the virtual position and momentum, s and p_s are the thermostat control variables, Q is the mass of the thermostat, k_B is the Boltzmann's constant, T_{tgt} is the target temperature of the system, and g is the number of degrees of freedom in the physical system ($g = d \times N_a$ in general). A change of variable introduced in [40] allows for the equations of motion in the virtual coordinates to be written in the physical coordinates.

Let $\widehat{\boldsymbol{x}} \in \mathbb{R}^{2N}$ be the time-dependent virtual phase-space vector and let $\boldsymbol{\mu} \in \mathbb{R}^{2k}$ denote the *associated* phase-space vector for k control variables and their corresponding momenta. Let $\widehat{H} : \mathbb{R}^{2N} \times \mathbb{R}^{2k} \to \mathbb{R}$ be the Hamiltonian expressed in the virtual coordinates. The evolution of the system is now described by the coupled system of differential equations

$$\begin{cases} \dot{\widehat{\boldsymbol{x}}} = [J_{2N}] \nabla_{\widehat{\boldsymbol{x}}} \widehat{H}(\widehat{\boldsymbol{x}}, \boldsymbol{\mu}), \\ \dot{\boldsymbol{\mu}} = [J_{2k}] \nabla_{\boldsymbol{\mu}} \widehat{H}(\widehat{\boldsymbol{x}}, \boldsymbol{\mu}), \end{cases} \tag{2.15}$$

where $[J_{2k}]$ is the Poisson matrix for the control variables. The phase-space vector $\widehat{\boldsymbol{x}}$ does not depend directly on the control variables and momenta $\boldsymbol{\mu}$ in virtual coordinates. However, the phase-space vector $\boldsymbol{x}$ (in physical coordinates) can be written as a function of $\widehat{\boldsymbol{x}}$ and $\boldsymbol{\mu}$:

$$\boldsymbol{x}(\widehat{\boldsymbol{x}}, \boldsymbol{\mu}) = [G(\boldsymbol{\mu})] \widehat{\boldsymbol{x}}, \tag{2.16}$$

where the matrix $[G(\boldsymbol{\mu})] = \text{diag}(g_q(\boldsymbol{\mu})[I_N], g_p(\boldsymbol{\mu})[I_N])$. The scaling factors g_q and g_p corresponding to various ensembles are provided in subsequent sections.

The Hamiltonian formalism cannot be readily applied in physical coordinates since the matrix $[G(\boldsymbol{\mu})]$ does not define a symplectic transformation in general. However, the symplectic lift can be applied in virtual coordinates to construct a reduced-order model with dynamical control variables and momenta $\boldsymbol{\mu}$. Consider the Galerkin projection given by the symplectic lift

$$\widehat{\boldsymbol{x}} = \Gamma(\widehat{\boldsymbol{z}}), \tag{2.17}$$

where $\widehat{\boldsymbol{z}} \in \mathbb{R}^{2n}$ denotes the virtual reduced phase-space vector and Γ is defined in Eq. (2.11). Note that the reference vector has the same value in virtual coordinates, *i.e.* $\widehat{\boldsymbol{x}}_0 = \boldsymbol{x}_0$, since there is no scaling of the positions at $t = 0$. The reduced phase-space vector $\boldsymbol{z}$ in physical coordinates is then given by

$$\boldsymbol{z}(\widehat{\boldsymbol{z}}, \boldsymbol{\mu}) = [\mathcal{G}(\boldsymbol{\mu})] \widehat{\boldsymbol{z}}, \tag{2.18}$$

with $[\mathcal{G}(\boldsymbol{\mu})] = [\Psi]^T [G(\boldsymbol{\mu})][\Psi] = \text{diag}(g_q(\boldsymbol{\mu})[I_n], g_p(\boldsymbol{\mu})[I_n])$. Next, let $\widehat{\mathcal{H}} : \mathbb{R}^{2n} \times \mathbb{R}^{2k} \to \mathbb{R}$ be the reduced Hamiltonian expressed in virtual coordinates. The evolution of the reduced-order system is described by

$$\begin{cases} \dot{\widehat{\boldsymbol{z}}} = [J_{2n}] \nabla_{\widehat{\boldsymbol{z}}} \widehat{\mathcal{H}}(\widehat{\boldsymbol{z}}, \boldsymbol{\mu}), \\ \dot{\boldsymbol{\mu}} = [J_{2k}] \nabla_{\boldsymbol{\mu}} \widehat{\mathcal{H}}(\widehat{\boldsymbol{z}}, \boldsymbol{\mu}). \end{cases} \tag{2.19}$$

Using the above definitions, and starting from Eq. (2.17), it follows that

$$\boldsymbol{x} = [\Psi]\boldsymbol{z} + [G(\boldsymbol{\mu})]\boldsymbol{x}_0 = [\Psi]\boldsymbol{z} + (g_q(\boldsymbol{\mu})\boldsymbol{q}(0); \boldsymbol{0}_N). \tag{2.20}$$

This result, obtained by considering virtual coordinates, clarifies the scaling factor in the snapshot matrix and leads to the following

Definition 2 (*Snapshot Matrix for Control-Dependent PSD*)**.** Let $[G(\boldsymbol{\mu}(t))] = \text{diag}(g_q(\boldsymbol{\mu})[I_N], g_p(\boldsymbol{\mu})[I_N])$ denote the block-diagonal matrix of scaling factors such that $g_q(\boldsymbol{\mu}(0)) = g_p(\boldsymbol{\mu}(0)) = 1$. Assume a given time discretization, and let $\boldsymbol{q}_0(t) = g_q(\boldsymbol{\mu}(t))\boldsymbol{q}(0)$ be the scaled initial position vector at time t. Then the snapshot matrix is defined as

$$[X_{\boldsymbol{\mu}}] = \left[\boldsymbol{q}(t_{j_1}) - \boldsymbol{q}_0(t_{j_1}), \ldots, \boldsymbol{q}(t_{j_{N_s}}) - \boldsymbol{q}_0(t_{j_{N_s}}), \gamma \boldsymbol{p}(t_{j_1}), \ldots, \gamma \boldsymbol{p}(t_{j_{N_s}})\right]. \tag{2.21}$$

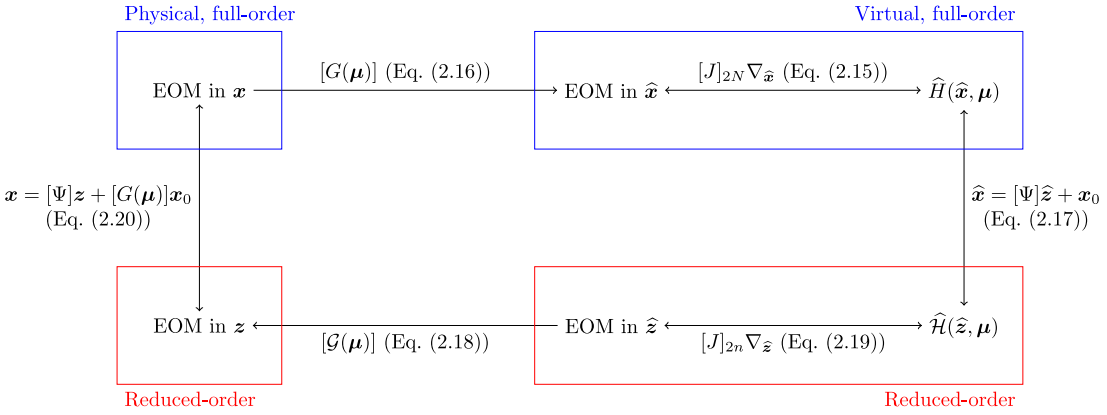


Fig. 1. Derivation of the equations of motion for reduced systems using physical and virtual coordinates.

The different changes of variables, together with the associated EOMs in physical and virtual variables, are summarized in Fig. 1.

In the next sections, we provide the definition of the matrix-valued scaling factor $[G(\boldsymbol{\mu})]$ in the most commonly-employed ensembles in molecular dynamics. For each ensemble, we specifically define the scaling functions defining $[G(\boldsymbol{\mu})]$, *i.e.* $g_q(\boldsymbol{\mu})$ and $g_p(\boldsymbol{\mu})$ in

$$[G(\boldsymbol{\mu})] = \text{diag}(g_q(\boldsymbol{\mu})[I_N], g_p(\boldsymbol{\mu})[I_N])\,, \tag{2.22}$$

and discuss the impact of the scaling factors on the modified snapshot matrix.

2.2.1. Micro-canonical (NVE) ensemble

In the micro-canonical ensemble, no scaling occurs and

$$g_q(\boldsymbol{\mu}) = g_p(\boldsymbol{\mu}) = 1\,. \tag{2.23}$$

The initial positions are not scaled when computing snapshots or mapping between reduced and physical variables. The equations of motion for this ensemble are given in Appendix A.1.

2.2.2. Canonical (NVT) ensemble

Thermostats involve control equations that enable MD simulations at (nearly) constant temperature. The Hamiltonian dynamics in the virtual coordinates is shown to conserve the isothermal conditions [38]. Two common methods for temperature control are the Nosé-Hoover (NH) thermostat and Nosé-Hoover Chain (NHC) thermostat. The Nosé-Hoover thermostat follows directly from the equations of motion in [40], which modifies the extended Hamiltonian in Eq. (2.14) by the change of variables $\xi = \ln s$ and $p_\xi = p_s/s$. Martyna et al. [41] modified the NH thermostat equations to increase stability in small-scale simulations. The modified equations introduce a chain of thermostat variables rather than a single thermostat variable. However, only the first thermostat variable scales the velocity.

We now specify the appropriate choice for $g_q(\boldsymbol{\mu})$ and $g_p(\boldsymbol{\mu})$ for each method. For the NH thermostat, let $\boldsymbol{\mu} = (\xi, p_\xi)$, where ξ and p_ξ are the real-valued thermostat position and momentum, respectively, defined in the equations of motion in [41]. According to the virtual coordinates used in [38], we define

$$g_q(\boldsymbol{\mu}) = 1\,, \quad g_p(\boldsymbol{\mu}) = e^{-\xi}\,. \tag{2.24}$$

For the NHC thermostat, $\boldsymbol{\mu} = (\xi_1, \dots, \xi_m, p_{\xi_1}, \dots, p_{\xi_m})$. The scaling factor matrices only depend on the first thermostat variable:

$$g_q(\boldsymbol{\mu}) = 1\,, \quad g_p(\boldsymbol{\mu}) = e^{-\xi_1(\xi_2,\dots,\xi_m)}\,. \tag{2.25}$$

In both cases, $g_q(\boldsymbol{\mu})$ is 1 and $[X_{\boldsymbol{\mu}}] = [X]$. This result is similar to the one obtained for the micro-canonical ensemble. The equations of motion for the canonical ensemble are provided in Appendix A.2.

2.2.3. Isoenthalpic–isobaric (NPH) ensemble

Barostats allow for MD simulations under (nearly) constant pressure. The formulation introduces a set of virtual coordinates that depend on a volume scaling factor. In the isobaric ensemble (NPH), we consider an isotropic external pressure, denoted by P_{ext}. The dynamics in virtual coordinates are shown to preserve the isobaric conditions [37]. Let $\boldsymbol{\mu} = (V, p_\epsilon)$, where V is the volume and p_ϵ are the real-valued barostat momentum, respectively, defined in [41]. The volume and momentum are related by a barostat mass W:

$$\dot{V} = \frac{dV p_\epsilon}{W}\,. \tag{2.26}$$

Using the virtual coordinates introduced in [37], we define the scaling factor matrices as

$$g_q(\boldsymbol{\mu}) = V^{1/d}\,, \quad g_p(\boldsymbol{\mu}) = V^{-1/d}\,. \tag{2.27}$$

Thus, $[X_{\boldsymbol{\mu}}] \neq [X]$ and the scaled initial position (used to calculate the displacement in the modified snapshot) is given by

$$\boldsymbol{q}_0(t) = V(t)^{1/d}\boldsymbol{q}(0)\,. \tag{2.28}$$

The equations of motion for the isoenthalpic–isobaric ensemble can be found in Appendix A.3.

2.2.4. Isothermal–isobaric (NPT) ensemble

To achieve control in pressure and temperature, consider $\mu = (\xi, V, p_\xi, p_\epsilon)$, where the barostat and thermostat control variables as defined in the two previous sections [38]. The scaling factor matrices are then defined as

$$g_q(\boldsymbol{\mu}) = V^{1/d}\,, \quad g_p(\boldsymbol{\mu}) = e^{-\xi}V^{-1/d}\,. \tag{2.29}$$

Thus, $[X_{\boldsymbol{\mu}}] \neq [X]$ and the scaled initial position is given by

$$\boldsymbol{q}_0(t) = V(t)^{1/d}\boldsymbol{q}(0)\,. \tag{2.30}$$

The equations of motion for the isoenthalpic–isobaric ensemble are listed in Appendix A.4.

The importance of properly rescaling the positions in the matrix of snapshots will be illustrated in Section 4.1.1, which includes a comparison of the approximation error in a Lennard-Jones fluid.

3. Stochastic reduced-order modeling

Consider the common setup where M model candidates for interatomic interactions coexist. These candidates can be defined by selecting different potentials, either constructed using physics-based arguments or learned through machine-learning techniques, or by considering a given potential parameterized by different material parameters. We denote by

$$[\boldsymbol{\Psi}^{(i)}] = \mathrm{diag}([\boldsymbol{\Phi}^{(i)}], [\boldsymbol{\Phi}^{(i)}])$$

the symplectic reduced-order basis defined in Definition 2 for the ith model (with snapshot matrix $[X^{(i)}_{\boldsymbol{\mu}}]$), using the appropriate matrix-valued scaling factor $[G(\boldsymbol{\mu})]$ (as defined in Sections 2.2.1 through 2.2.4 for all classical ensembles). We further introduce the global reduced-order basis $[\boldsymbol{\Psi}^*] = \mathrm{diag}([\boldsymbol{\Phi}^*], [\boldsymbol{\Phi}^*])$, computed from the matrix of snapshots for all models, denoted by $[X^*_{\boldsymbol{\mu}}]$:

$$[X^*_{\boldsymbol{\mu}}] = [[X^{(1)}_{\boldsymbol{\mu}}], \dots, [X^{(m)}_{\boldsymbol{\mu}}]]\,. \tag{3.1}$$

We then consider the randomization of the symplectic reduced-order basis $[\boldsymbol{\Psi}]$, *given* the matrix dataset $\{[\boldsymbol{\Psi}^{(1)}], \dots, [\boldsymbol{\Psi}^{(M)}], [\boldsymbol{\Psi}^*]\}$. Conceptually, this is equivalent to randomizing the underlying model, given a functional set of models chosen based on domain expertise or any ad hoc model selection procedure. This *multi-model setting* was introduced in [27]; see also [26] for seminal derivations in the uni-model setting, as well as [42] for applications to MD simulations.

3.1. Probabilistic modeling

Let $[\boldsymbol{\Psi}] = \mathrm{diag}([\boldsymbol{\Phi}], [\boldsymbol{\Phi}])$ be the stochastic representation of $[\boldsymbol{\Psi}]$, defined on the probability space (Θ, Σ, P). Following derivations in Section 2, the random matrix $[\boldsymbol{\Phi}]$ takes values in the subset of the Stiefel manifold $\mathrm{St}(N, n)$

$$\mathbb{S}(N, n) = \{[A] \in \mathrm{St}(N, n) \mid [B]^T[A] = [0_{N_0 \times n}]\} \subset \mathrm{St}(N, n)\,, \tag{3.2}$$

where $[B]$ is the deterministic matrix introduced in Section 2.1. Modeling on the Stiefel manifold, or a subset thereof, is traditionally achieved on the tangent space to the manifold at a given base point, using appropriate projection and retraction operators. The latter must be chosen accounting for various constraints, such as numerical efficiency and ease of integration for the linear constraints in Eq. (3.2). The use of operators based on a polar decomposition was proposed in [26] in the uni-model setting. The resulting formulation promotes computational efficiency through an ad hoc parameterization on the tangent plane, but relies on nonlinear transformations that complicate inference (in terms of direct estimation for the mean of the reduced-order basis and number of hyperparameters, the latter growing as $\mathcal{O}(n^2)$). An alternative formulation leveraging available information in the multi-model setting was proposed in [27] and makes use of Riemannian projection and retraction operators. This stochastic model is summarized below to make the presentation self-contained.

Following Zhang and Guilleminot [27], the stochastic reduced-order basis is defined as

$$[\boldsymbol{\Phi}] = \exp^{\mathrm{St}}_{[\boldsymbol{\Phi}^*]}\left\{\sum_{i=1}^{m} P_i \log^{\mathrm{St}}_{[\boldsymbol{\Phi}^*]}([\boldsymbol{\Phi}^{(i)}])\right\}, \tag{3.3}$$

where $\exp^{\mathrm{St}}_{[\boldsymbol{\Phi}^*]}$ and $\log^{\mathrm{St}}_{[\boldsymbol{\Phi}^*]}$ denote the Riemannian retraction and projection operators at base point $[\boldsymbol{\Phi}^*]$, respectively, and $\boldsymbol{P}$ is a random vector (with components $P_1, \dots, P_m$) defining stochastic combinations on the tangent space. The use of the above operators is motivated by the consideration of the constraint that $[B]^T[\boldsymbol{\Phi}] = [0_{N_0 \times n}]$ almost surely. Efficient algorithms to evaluate the

Riemannian operators can be found in Zimmermann and Hüper [43]. Note that this evaluation (and in particular, the evaluation of the Riemannian logarithm) may be computationally intensive for large values of n. The derivation of alternative methods for fast and accurate estimations of Riemannian operators are beyond the scope of this paper. In this work, the base point $[\boldsymbol{\Phi}^*]$ is defined as the global reduced-order basis constructed by collecting snapshots for all model candidates, and is assumed to belong to the convex hull defined by these candidates. Alternatively, it may be considered as a model parameter and identified through minimization with respect to some reference (*e.g.*, experimental) data. This choice, however, introduces a large number of hyperparameters and a constrained optimization problem on $\mathbb{S}(N,n)$.

In order to construct the probability measure of $\boldsymbol{P}$, we note that the formulation on the tangent space relaxes structural constraints but must be such that new samples remain "close" to the base point $[\boldsymbol{\Phi}^*]$ — otherwise, standard algorithms to retract onto the manifold may not converge; see [43] for a discussion. Given that a set of model candidates is available, the desire to generate within the range of these models appears as a reasonable choice. This can be achieved, in practice, by assuming that $\boldsymbol{P}$ follows a Dirichlet distribution:

$$\boldsymbol{P} \sim \mathcal{D}(\boldsymbol{\alpha}), \quad f_{\boldsymbol{P}}(\boldsymbol{p}) = \frac{\Gamma(\sum_{i=1}^m \alpha_i)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m p_i^{\alpha_i - 1}, \tag{3.4}$$

where $\boldsymbol{\alpha}$ is the vector of positive concentration parameters, $f_{\boldsymbol{P}}$ is the probability density function of $\boldsymbol{P}$ and Γ is the Gamma function. This implies that $\sum_{i=1}^m P_i = 1$ almost surely, and defines a Riemannian convex combination on the manifold. Note that this choice corresponds to the distribution obtained by entropy maximization in the context of information theory, with repulsion constraints such that $0 \leq P_i \leq 1 \ \forall i \in \{1, \dots, m\}$ almost surely, and has important consequences in terms of inference, as pointed out in the next section.

3.2. Properties of the symplectic reduced-order basis

Based on the construction in Section 3.1, the following properties can be deduced:

(P1) The stochastic symplectic reduced-order basis $[\boldsymbol{\Psi}]$ is of second-order.

(P2) The vector of concentration parameters such that $\mathbb{E}\{[\boldsymbol{\Psi}]\} \approx \mathrm{diag}([\boldsymbol{\Phi}^*], [\boldsymbol{\Phi}^*])$ in the Fréchet sense is the solution to

$$\boldsymbol{\alpha} = \mathrm{argmin}_{\boldsymbol{a} \in \mathbb{R}^m_{>0}} \boldsymbol{a}^T [H] \boldsymbol{a},$$

where $[H]$ is the diagonal matrix with entries $H_{ij} = \langle \log^{\mathrm{St}}_{[\boldsymbol{\Phi}^*]}([\boldsymbol{\Phi}^{(i)}]), \log^{\mathrm{St}}_{[\boldsymbol{\Phi}^*]}([\boldsymbol{\Phi}^{(j)}]) \rangle_F$.

(P3) Under suitable assumptions, $[\boldsymbol{\Psi}]$ takes values in the convex hull defined by the set $\{[\boldsymbol{\Psi}^{(i)}]\}_{i=1}^m$.

A few remarks regarding the above properties are in order. Property (P1) follows from the consideration of convex combinations on the tangent space at base point $[\boldsymbol{\Phi}^*]$ and the (local) continuity of the exponential retraction map. It ensures modeling consistency and stability in the forward propagation problem. The second property (P2) is fundamental in that it allows for the direct calibration of the hyperparameters, given a set of model candidates (and their associated reduced-order bases); see [27] for a proof. Note that the positive definiteness of $[H]$ depends on the dataset and ensures that the solution can be obtained using convex quadratic programming. The last property (P3) holds under the following set of assumptions: (1) $\mathrm{St}(N,n)$ is of constant nonnegative curvature (at least locally); (2) the convex set defined by $\{[\boldsymbol{\Psi}^{(i)}]\}_{i=1}^m$ lies in a ball of radius of at most $r_{\mathrm{cvx}} = (1/2)\min\{\mathrm{inj}\ \mathrm{St}(N,n), \pi/\sqrt{\bar{\rho}}\}$, where $\mathrm{inj}\ \mathrm{St}(N,n)$ is the injectivity radius of $\mathrm{St}(N,n)$ and $\bar{\rho}$ is a lower bound for the sectional curvature of the Stiefel manifold $\mathrm{St}(N,n)$ [44]. The fact that the global curvature of the Stiefel manifold is nonnegative in the canonical metric, for arbitrary N and n, is established; see, *e.g.*, [45]. The other assumptions cannot be proven for arbitrary dimensions and base point but may be conjectured when the reduced-order bases defining the envelope are close to one another (so that curvature can be assumed constant locally, for instance).

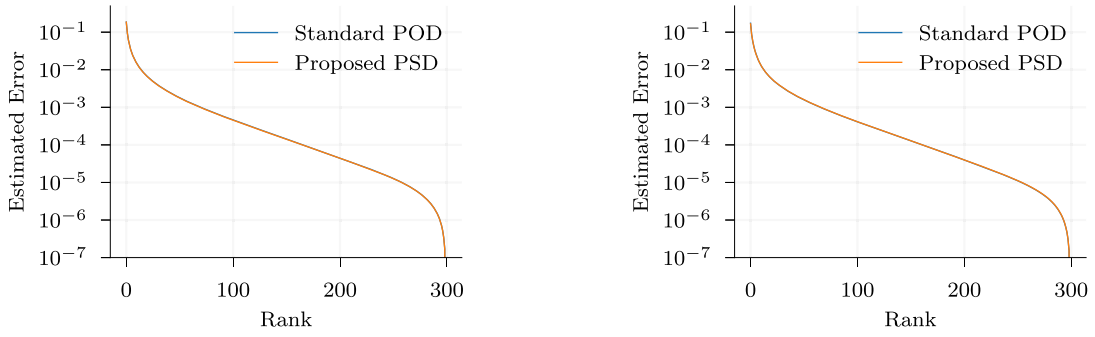
4. Numerical results

4.1. Deterministic reduced-order modeling in the isothermal–isobaric (NPT) ensemble

In this section, we verify the reduced-order model (and specifically, the definition of the scaling matrices) by considering a physical system composed of 864 argon atoms in a cubic simulation box. We follow the simulation setup introduced in Rahman [46], using Lennard-Jones (LJ) units. The simulation box has a side length $L = 10.229\sigma$ and the cutoff distance for the LJ potential is set to $R = 2.25\sigma$, where $\sigma = 3.4$ Å for argon. The full-order simulations are performed under periodic boundary conditions. To initialize the system, the atoms are randomly placed in the simulation box and energy minimization was performed before equilibrating the temperature and pressure for 10,000 timesteps, with $\Delta t = 0.001$ fs.

To perform model verification, we consider two problems using the NPT ensemble, namely a pressure-control test and temperature-control test. These tests are referred to as Problem 1 and Problem 2, respectively, and are defined according to the tests developed in Kim et al. [47]:

- In Problem 1 (with pressure control), the temperature and pressure are initialized to $T_0 = 1.0$ and $P_0 = 0.5$, respectively using LJ units. After 10,000 timesteps, the pressure is raised to $P_1 = 1.0$. After another 10,000 timesteps, the pressure is reduced back to $P_2 = 0.5$.



(a) Problem 1: Pressure Test. (b) Problem 2: Temperature Test.

Fig. 2. Convergence analysis for the POD and PSD reduced-order bases, for (a) the pressure test and (b) the temperature test. Note that while the snapshots and eigenvectors in the POD and PSD are different, the eigenvalues for both methods are very similar. See Appendix A.5 for complementary results.

- In Problem 2 (with temperature control), the temperature and pressure are initialized to $T_0 = 0.8$ and $P_0 = 0.5$, respectively using LJ units. After 10,000 timesteps, the temperature is raised to $T_1 = 1.2$. After another 10,000 timesteps, the temperature is reduced back to $T_2 = 0.8$.

In Section 4.1.1, we compare the error convergence estimates for a standard POD basis and the proposed PSD basis (with a rescaled snapshot matrix). Results obtained with the full-order model, the PSD approach, and the standard POD approach are then provided in Section 4.1.2, along a discussion of the stability of the control algorithms.

4.1.1. Convergence comparison between standard POD and proposed PSD

In this section, the reduced-order basis for both the standard POD and proposed PSD are constructed by taking snapshots every 100 timesteps. Then, the (sorted) eigenvalues from the singular value decomposition are used to estimate the projection error according to

$$\varepsilon(r) = 1 - \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{N} \lambda_i},$$

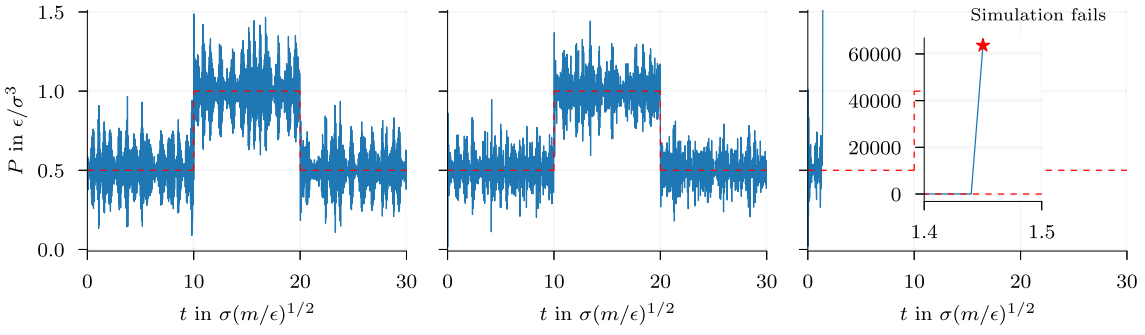
where λ_i is the ith eigenvalue associated with the ith eigenvector (column) of the reduced-order basis and r is the number of modes taken (the reduced dimension n being the smallest integer such that $\varepsilon(n) \leq \varepsilon_0$, with ε_0 a given threshold).

The graph of $r \mapsto \varepsilon(r)$ is shown for both the proposed PSD and standard POD in Fig. 2, for the two problems. It is seen that the two decompositions exhibit similar convergence rates, regardless of the control variable. Indeed, the eigenvalues are nearly identical — though not the same — for the two reduced-order models (POD and PSD) in both problems. Based on these results, $n = 167$ (295 resp.) and $n = 163$ (294, resp.) modes are selected for Problem 1 and Problem 2, respectively, corresponding to a projection error $\varepsilon_0 = 10^{-4}$ (10^{-6}, resp.). Note that these values are expected to ensure stability and reasonable accuracy in the approximation of the full-order model response. We will see, in the next section, that this is not the case for the standard POD method.

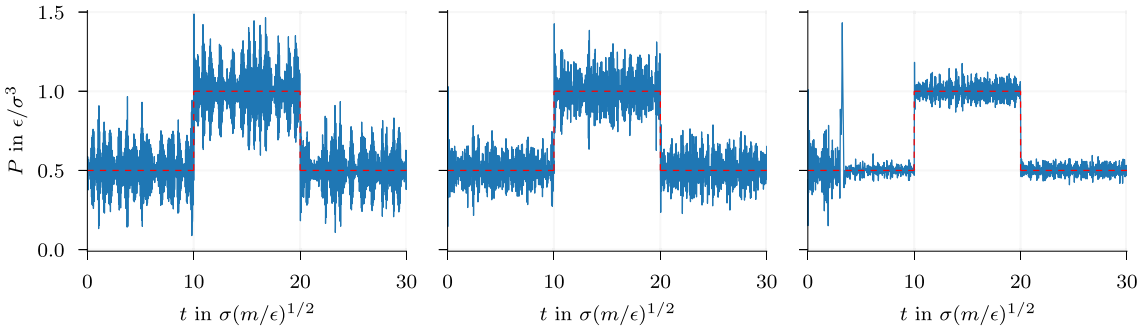
4.1.2. Comparison between reference, standard POD and proposed PSD results

Results for Problem 1 are shown in Fig. 3 (together with the target value in dashed red line), associated with pressure control. Here, the reference model showed an average pressure of 0.5003 during the first 10 time units, 0.9993 during the pressure step, and 0.5002 during the last 10 time units. The range of observed fluctuations using the Nosé-Hoover thermostat-barostat is consistent with the results reported in [47]. The proposed PSD ROM for $n = 167$ (295, resp.) showed an average pressure of 0.5005 (0.5007, resp.) during the first 10 time units, 0.9992 (0.9991, resp.) during the pressure step, and 0.5001 (0.5008, resp.) during the last 10 time units. The target pressure is therefore well controlled for the proposed PSD formulation. In contrast, the simulation based on the standard POD failed on timestep 152 for $n = 167$, with a pressure reaching a maximum value of 60. For $n = 295$, the POD ROM showed an average pressure of 0.5135 during the first 10 time units, 0.9992 during the pressure step, and 0.5009 during the last 10 time units. In comparison to the POD ROM, the proposed PSD ROM provided more accurate pressure averages during the three pressure steps at nearly half the number of modes. Though not shown, the temperature during the simulation was properly maintained at about 1.0 for the reference and proposed PSD simulations.

Similar results are shown in Fig. 4 for the temperature control problem (Problem 2). In this setting, the reference model showed an average temperature of 0.8006 during the first 10 time units, 1.1884 during the temperature step, and 0.8103 during the last 10 time units. As previously observed for Problem 1, the range of fluctuations using the Nosé-Hoover thermostat-barostat is consistent with the results obtained in [47]. The proposed PSD ROM for $n = 163$ (294, resp.) showed an average temperature of 0.7955 (0.7956, resp.) during the first 10 time units, 1.1826 (1.1870, resp.) during the temperature step, and 0.8111 (0.8134, resp.) during the last 10 time units. The reduced simulation also exhibits more variance in temperature during each control segment, as shown in the



(a) $n = 167$, for $\varepsilon = 10^{-4}$



(b) $n = 295$, for $\varepsilon = 10^{-6}$

Fig. 3. Problem 1: reference model (left), proposed PSD ROM (middle), and standard POD ROM (right). The same vertical axis is used for all figures in each row. The target pressure is shown in dashed red. The rightmost figure includes a subpanel with a larger vertical scale to show the point at which the simulation fails. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

middle plot in Fig. 4. This observation may be a result of projection errors introduced into, and thus propagated by errors in, the velocity. Proper control in temperature is thus well achieved. On the contrary, the standard POD ROM simulation for $n = 163$ failed on timestep 751, with the temperature reaching a maximum value of 1.6. For $n = 294$, the POD ROM showed an average temperature of 0.7963 during the first 10 time units, 1.2068 during the temperature step, and 0.8101 during the last 10 time units. In comparison to the POD ROM, the proposed PSD ROM provided more accurate temperature averages during the three temperature steps at nearly half the number of modes. Though not shown, the pressure during the simulation was maintained about 0.5 for the reference and proposed PSD simulations.

In summary, the reduced-order model using the proposed PSD shows stability and better agreement with the reference, full-order model. In comparison, the standard POD becomes unstable within 10 LJ time units at $\varepsilon = 10^{-4}$ for both problems.

4.2. Stochastic reduced-order modeling in the isothermal–isobaric (NPT) ensemble

4.2.1. Description of the setup

In this section, MD simulations of the phase transition from solid to liquid of argon is considered to illustrate the probabilistic framework and its ability to capture model-form uncertainties. It should be pointed out that this setting is challenging in terms of probabilistic representation, due to the transient nature of the phase change. We will show the ability of the proposed approach to capture uncertainties in the volume expansion during the phase transition, while appropriately controlling pressure and temperature.

A Face-Centered Cubic (FCC) lattice of 864 argon atoms is first brought to equilibrium at 50 K and at atmospheric pressure (1 atm). The dimension of the full-order model is thus $864 \times 3 = 2592$. After 10,000 timesteps, the temperature of the system is stepped up to 100 K to produce a phase transition. Snapshots of the simulation system for the solid (a) and liquid (b) phases can be seen in Fig. 5. Note that unlike the setup in Section 4.1, the simulation begins from a crystal FCC lattice structure and is not run in LJ units, allowing for comparison across pair potentials.

To quantify the impact of interatomic potential selection, the following three pair potentials are used.

- Model 1 corresponds to the LJ potential used in the seminal work [46]:

$$\phi_1(r) = 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right],$$

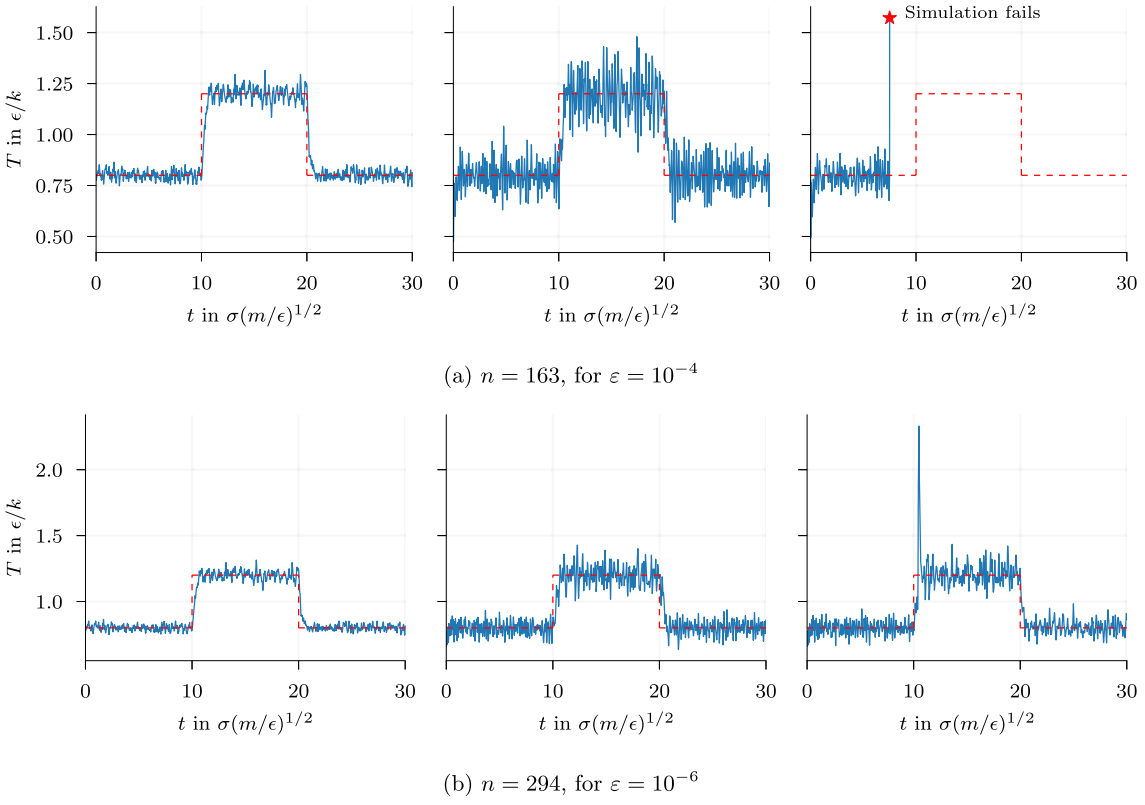
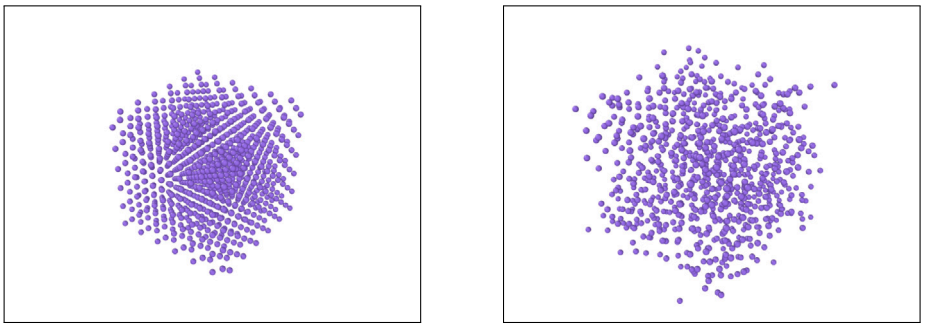


Fig. 4. Problem 2: reference model (left), proposed PSD ROM (middle), and standard POD ROM (right). The same vertical axis is used for all figures in each row. The target temperature is shown in dashed red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Configuration in (FCC) solid phase. (b) Configuration in liquid phase.

Fig. 5. Images of the first and last frames of the simulation system, rendered with Ovito (at the same scale).

where ϵ and σ represent the depth of the potential well and the distance at which the potential energy is zero, respectively.

- Model 2 is the LJ potential defined in Bernardes [48] in the NIST Interatomic Potentials Repository:

$$\phi_2(r) = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] & \text{for } r < r_{cut}\,, \\ 0 \text{ for } r \geq r_{cut}\,, \end{cases}$$

where ϵ and σ are parameters similar to those in Model 1, and r_{cut} is a cutoff distance.

- Model 3 is the Morse potential defined in Jelinek [49] in the aforementioned database:

$$\phi_3(r) = \epsilon \left[e^{-2c(r-r_0)} - 2e^{-c(r-r_0)} \right],$$

Table 1
Model parameters for interatomic potentials.

Model	Parameters
1	$\epsilon/k_B = 120.0$ K, $\sigma = 3.40$ Å
2	$\epsilon/k_B = 120.7$ K, $\sigma = 3.40$ Å, $r_{cut} = 9.23$ Å
3	$\epsilon/k_B = 156.4$ K, $r_0 = 3.786$ Å, $c = 1.545$ Å^{-1}

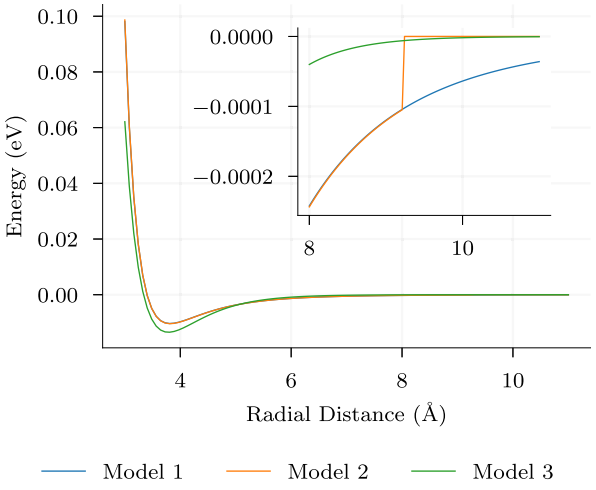


Fig. 6. Comparison of the interatomic potential functions for the three models used above. The figure includes a subpanel with a larger vertical scale to show the difference due to the cutoff radius in Model 2.

where ϵ is the potential well depth, r_0 is the equilibrium distance, and c is a parameter controlling the width of the potential well.

These potentials were accessed using the OpenKIM repository [47,50–52]. Parameters for each potential are provided in Table 1, and energy as a function of radial distance is shown in Fig. 6. The potential well depth ϵ is expressed in terms of the Boltzmann's constant k_B.

Both the full- and reduced-order MD simulations were run in LAMMPS. The reduced dimension is set to $n = 521$, following a convergence analysis on the projection error with $\varepsilon_0 = 10^{-7}$; see Fig. 7(a) for convergence results for all models (recall from Eq. (3.1) that the global reduced-order basis is computed by collecting all snapshots from all three models).

In order to assess the evolution of the error induced by the reduced-order model, the normalized L_2 error

$$\eta(t) = \frac{\|\boldsymbol{q}_{\mathrm{ROM}}(t) - \boldsymbol{q}_{\mathrm{FOM}}(t)\|_{L_2}}{\|\boldsymbol{q}_{\mathrm{FOM}}(t)\|_{L_2}} \tag{4.1}$$

is introduced (where the subscript "FOM" indicates full-order model results). The graph of the error function $t \mapsto \eta(t)$ is shown in Fig. 7(b).

Unsurprisingly, the error increases with simulation time for all model candidates, especially during the melting phase transition. This highlights the challenge of applying a linear subspace method to a crystal melt under dynamical control. This error can be reduced by increasing the number of modes (*i.e.*, the dimension n), or by augmenting the representation with a nonlinear approximation term. Such an extension is classical in the literature of reduced-order modeling and is beyond the scope of this work.

4.2.2. Sampling and forward propagation results

We now turn to forward propagation of model-form uncertainties, realized with the Monte Carlo approach. The stochastic reduced-order models are run using adaptive selection [27]. In this procedure, the interatomic potential used for updating the force term (by lifting in physical space, and by subsequently projecting back in the reduced-order space) is selected by characterizing the distance between the current sample and the reduced-order bases associated with all models (in other words, if the sample is closer to, say, the reduced-order basis $[\boldsymbol{\Phi}^{(1)}]$ computed with Model 1, then the latter is considered as the reference model in the physical space). This selection is visualized in Fig. 8, where a set of 200 additional samples are shown.

We recall $\boldsymbol{P}$ is sampled from a Dirichlet distribution with concentration parameters that can be identified from the quadratic programming problem following the procedure outline in Section 3.2:

$$\boldsymbol{\alpha} = \mathrm{argmin}_{\boldsymbol{a}\in\mathbb{R}^3_{>0}}\, \boldsymbol{a}^T[H]\boldsymbol{a}\,, \tag{4.2}$$

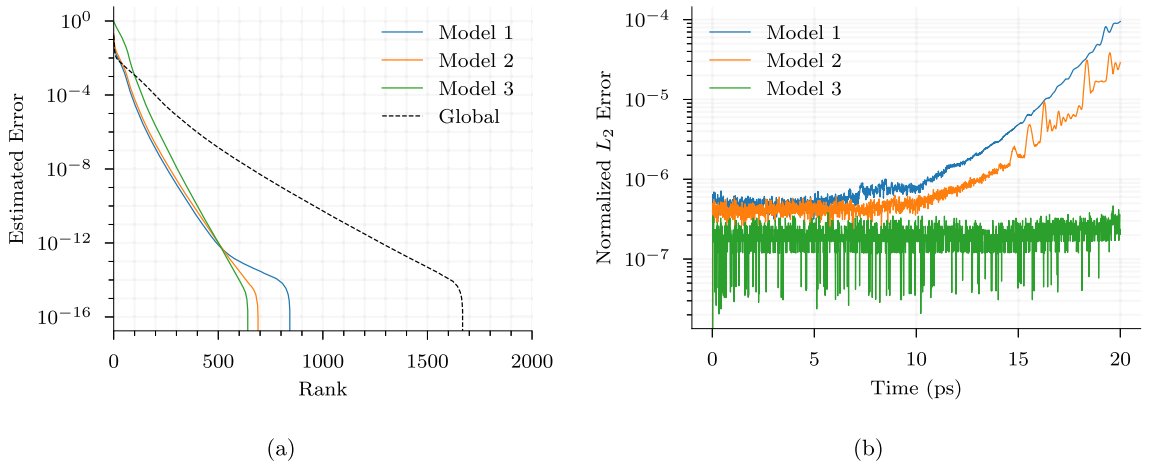


Fig. 7. (a) Error estimated from eigenvalues of the reduced-order bases. (b) The normalized L_2 error for the reduced-order models after taking the first $n = 521$ modes.

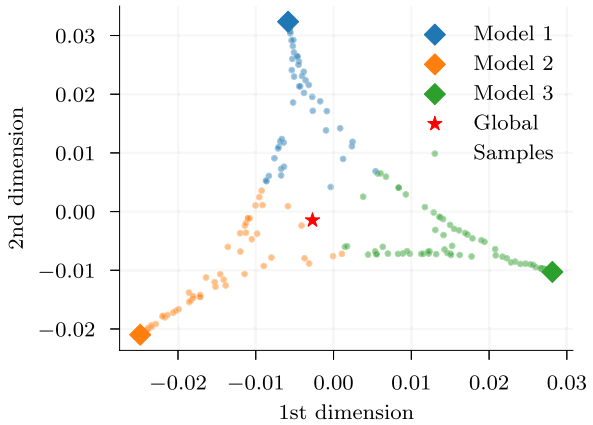


Fig. 8. Spectral embedding of the reduced-order bases for each model (colored diamonds), the global reduced-order basis (red star), and 200 reduced-order basis samples (colored dots). The samples of the stochastic reduced-order basis are colored according to the potential identified through adaptive selection. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where

$$[H] = \begin{bmatrix} 1436.0165 & 2.9569221 & 3.6475679 \\ 2.9569221 & 1479.4282 & 0.18607749 \\ 3.6475679 & 0.18607749 & 1483.5043 \end{bmatrix} \succ 0$$

for the three selected models. The solution is found as $\boldsymbol{\alpha} = (0.33984, 0.33062, 0.32955)$.

Figs. 9 and 10 show the trajectories of the temperature, pressure, and volume in each simulation, with and without model classification, respectively. Probability density functions obtained from these results, extracted at times $t = 10$ [ps] and $t = 20$ [ps] in the isobaric–isothermal regime, are shown in Fig. 11.

As previously indicated, the target value for the pressure is 1 atm. Large fluctuations are observed for this variable in the solid state, when the simulation box is close to the atoms on all sides. Temperature control is also properly achieved at the target values, set to 50 K and 100 K respectively. Smaller fluctuations are obtained for this quantity. A larger spread is observed for the volume, which can be expected given that (1) this quantity is very sensitive to the chosen reference model (as seen in the rightmost panel in Fig. 9), and (2) the dimension of the reduced-order model is small compared to the full-order dimension (with a reduction of about 80%). The probability density functions (see Fig. 11) are centered at the target temperatures in both phases, and it is seen the stochastic simulations span the response of each interatomic potential. Overall, this demonstrates the capability of the probabilistic formulation to capture differences in the volume change during the phase transition, while maintaining control on both the pressure and temperature.

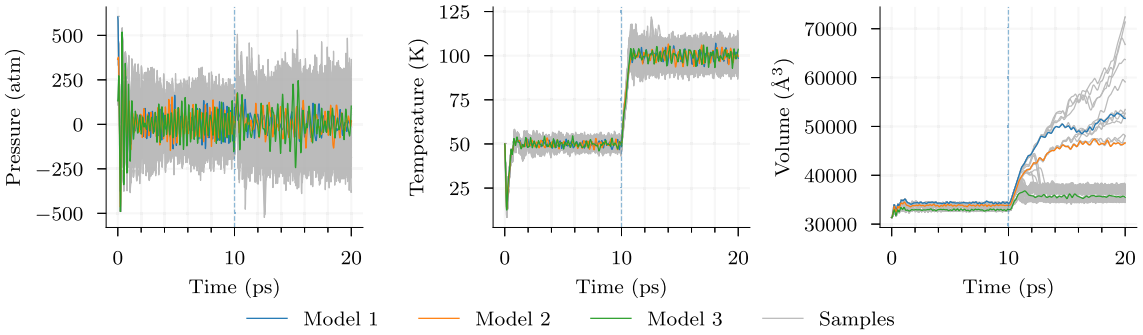


Fig. 9. Trajectories of thermodynamical and simulation quantities for 200 stochastic reduced-order samples, compared to the full-order models. Temperature step from 50 K to 100 K occurs at 10 picoseconds, shown here by the dashed blue line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

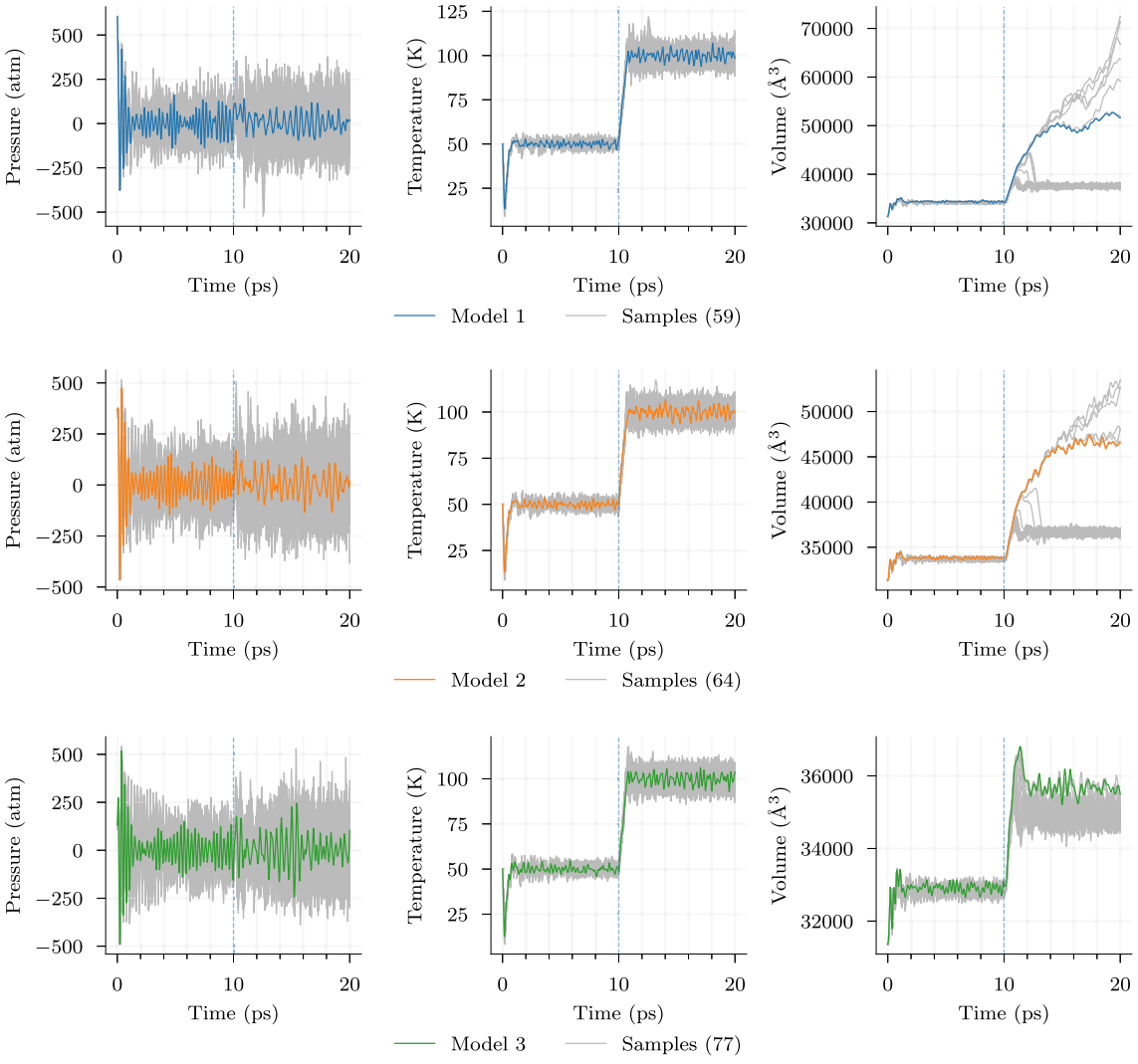
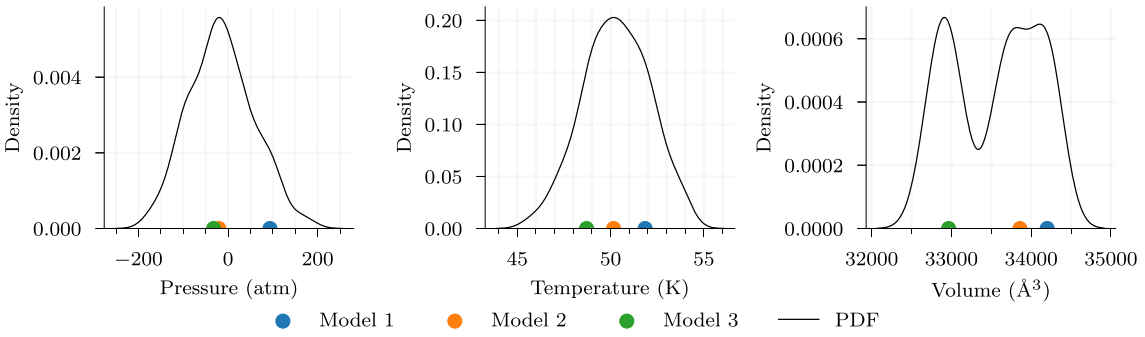
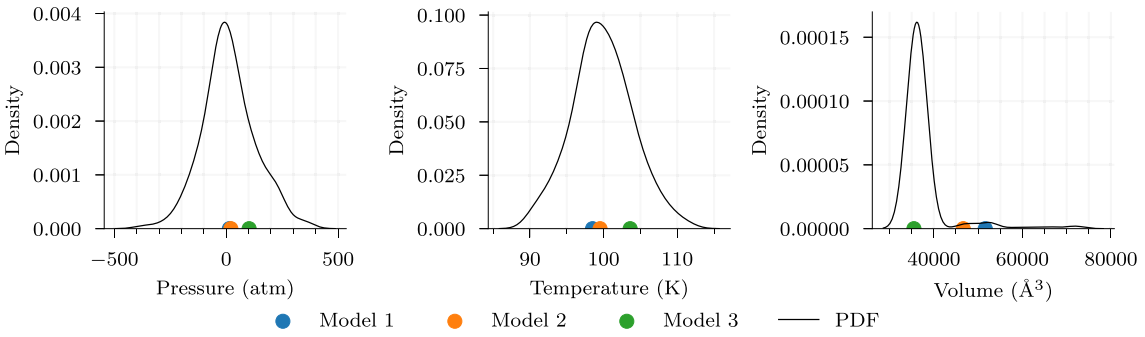


Fig. 10. Plot of trajectory samples, classified according to the reference model identified with the adaptive selection procedure (the number of samples is indicated in parentheses). Trajectories closer to reference trajectories typically correspond to reduced-order basis samples that are the closest to the reference reduced-order basis. Intermediate samples (*e.g.*, equidistant from two reference reduced-order bases) increase the range of the fluctuations after melting has occurred.



(a) Results obtained after equilibration in the solid state ($t = 10$ [ps]).



(b) Results obtained after the transition from solid to gas state ($t = 20$ [ps]).

Fig. 11. Snapshots probability density functions of thermodynamic and simulation quantities for 200 stochastic reduced-order samples compared to the full-order models. Note that the pressure remains centered around approximately 1 atm, and the temperature at 50 K before the sublimation, and 100 K after.

5. Conclusion

This paper develops a new stochastic reduced-order representation to capture model-form uncertainties for all classical statistical ensembles commonly used in MD simulations. The proposed Hamiltonian formulation specifically ensures the stability of the control algorithms in the almost sure sense, owing to appropriate rescaling in the proper symplectic decomposition, and can be easily implemented. The efficiency of the approach was demonstrated on an evaporation problem in the isothermal–isobaric (NPT) ensemble, following seminal work on control algorithms in the field. These results show that the approach allows for proper control in the reduced-order state space—in contrast with the POD operating with non-scaled snapshots, which is seen to produce divergent results regardless of the control variable. The proposed stochastic framework paves the way for broad adoption in a wide range of MD simulation setups involving control variables, such as multiscale methods (where control on pressure can be used to estimate macroscopic effective parameter, for instance). By preserving and building a probabilistic description of an envelope of models, it promotes adaptivity and robustness for, *e.g.*, parameter calibration and model definition in settings where ground truth data are scarce. Avenues for future research include algorithmic developments to efficiently sample high-dimensional reduced-order bases, as well as theoretical extensions to very highly nonlinear, potentially irreducible systems.

CRediT authorship contribution statement

S. Kounouho: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis. **R. Dingreville:** Writing – review & editing, Validation, Supervision, Investigation. **J. Guilleminot:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Johann Guilleminot reports financial support was provided by National Science Foundation. Johann Guilleminot reports

financial support was provided by Army Research Office. If there are other authors, they 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

Data will be made available on request.

Acknowledgments

The work of J. G. was supported by the National Science Foundation, Division of Civil, Mechanical and Manufacturing Innovation, United States, under award CMMI-1942928, and by the Army Research Office, United States under grant W911NF-23-1-0125. This work is supported by the Center for Integrated Nanotechnologies, United States, an Office of Science user facility operated for the U.S. Department of Energy. This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title, and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan.

Appendix. Equations of motion

A.1. Micro-canonical (NVE) ensemble

In the micro-canonical ensemble, no control equation is required and the equations of motions read as

$$\begin{aligned} \dot{\boldsymbol{q}} &= [M]^{-1}\boldsymbol{p}\,, \\ \dot{\boldsymbol{p}} &= \boldsymbol{f}(\boldsymbol{q})\,. \end{aligned} \tag{A.1}$$

A.2. Canonical (NVT) ensemble

In the canonical ensemble, the equations of motion are given by

$$\begin{aligned} \dot{\boldsymbol{q}} &= [M]^{-1}\boldsymbol{p}\,, \\ \dot{\boldsymbol{p}} &= \boldsymbol{f}(\boldsymbol{q}) - \frac{p_\xi}{Q}\boldsymbol{p}\,, \end{aligned} \tag{A.2}$$

and are supplemented with the control equations

$$\begin{aligned} \dot{\xi} &= \frac{p_\xi}{Q}\,, \\ \dot{p_\xi} &= \boldsymbol{p}^T[M]^{-1}\boldsymbol{p} - Nk_BT_{\text{tgt}}\,. \end{aligned} \tag{A.3}$$

A.3. Isoenthalpic–isobaric (NPH) ensemble

In the isoenthalpic–isobaric ensemble, the equations of motion are given by

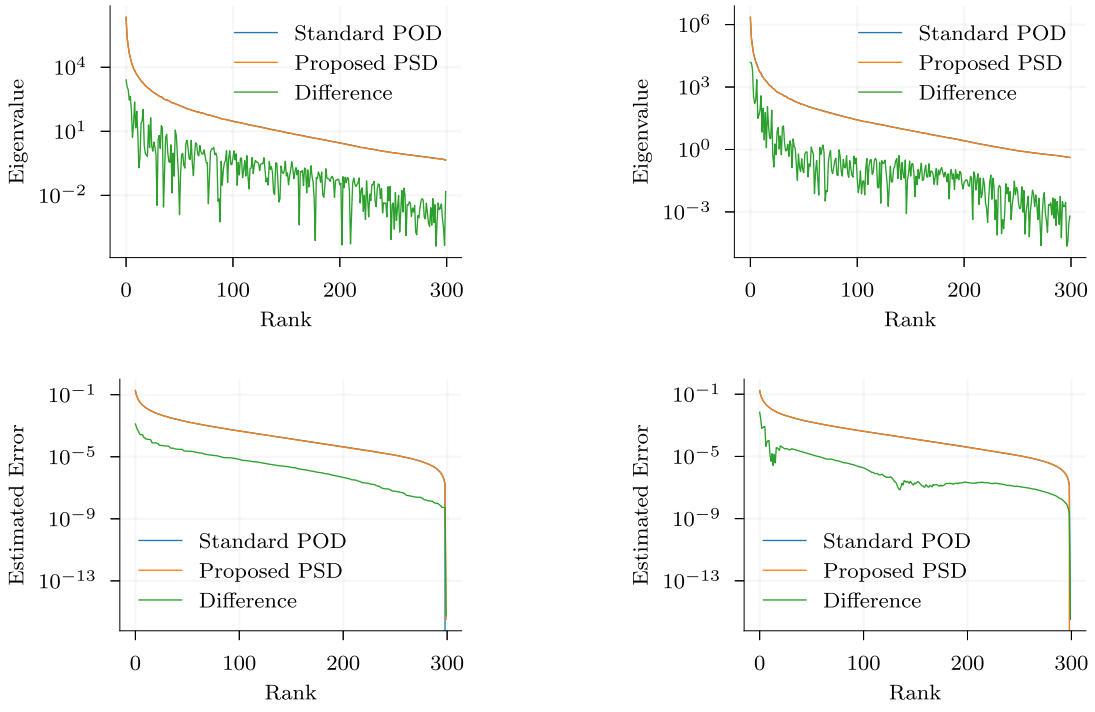
$$\begin{aligned} \dot{\boldsymbol{q}} &= [M]^{-1}\boldsymbol{p} + \frac{p_\epsilon}{W}\boldsymbol{q}\,, \\ \dot{\boldsymbol{p}} &= \boldsymbol{f}(\boldsymbol{q}) - \left(1 + \frac{d}{N}\right)\frac{p_\epsilon}{W}\boldsymbol{p}\,. \end{aligned} \tag{A.4}$$

Control equations are further defined as

$$\begin{aligned} \dot{V} &= \frac{dVp_\epsilon}{W}\,, \\ \dot{p_\epsilon} &= dV(P_{\text{int}} - P_{\text{ext}}) + \frac{d}{N}\boldsymbol{p}^T[M]^{-1}\boldsymbol{p}\,, \end{aligned} \tag{A.5}$$

where P_{ext} is the external pressure and

$$P_{\text{int}} = \frac{1}{dV}\left(\boldsymbol{p}^T[M]^{-1}\boldsymbol{p} + \boldsymbol{q}^T\boldsymbol{f}(\boldsymbol{q}) - (dV)\frac{\partial\phi}{\partial V}\right) \tag{A.6}$$



(a) Problem 1: Pressure Test.

(b) Problem 2: Temperature Test.

Fig. A.12. Difference in eigenvalue decay (first row) and projection error estimate (second row) for POD and PSD reduced-order bases, for (a) the pressure test and (b) the temperature test. Note that the difference between the eigenvalues is roughly two orders of magnitude smaller than the magnitude of the eigenvalues, but not zero.

A.4. Isothermal–isobaric (NPT) ensemble

For the isothermal–isobaric ensemble, the equations of motion are given by

$$\begin{aligned} \dot{\boldsymbol{q}} &= [M]^{-1}\boldsymbol{p} + \frac{p_\epsilon}{W}\boldsymbol{q}\,, \\ \dot{\boldsymbol{p}} &= \boldsymbol{f}(\boldsymbol{q}) - \left(1 + \frac{d}{N}\right)\frac{p_\epsilon}{W}\boldsymbol{p} - \frac{p_\xi}{Q}\boldsymbol{p}\,. \end{aligned} \tag{A.7}$$

In addition, control equations read as

$$\begin{aligned} \dot{V} &= \frac{dVp_\epsilon}{W}\,, \\ \dot{p}_\epsilon &= dV(P_{\text{int}} - P_{\text{ext}}) + \frac{d}{N}\boldsymbol{p}^T[M]^{-1}\boldsymbol{p} - \frac{p_\xi}{Q}p_\epsilon\,, \\ \dot{\xi} &= \frac{p_\xi}{Q}\,, \\ \dot{p}_\xi &= \boldsymbol{p}^T[M]^{-1}\boldsymbol{p} + \frac{p_\epsilon^2}{W} - (N+1)k_B T_{\text{tgt}}\,, \end{aligned} \tag{A.8}$$

where P_{ext} is the external pressure and

$$P_{\text{int}} = \frac{1}{dV}\left(\boldsymbol{p}^T[M]^{-1}\boldsymbol{p} + \boldsymbol{q}^T\boldsymbol{f}(\boldsymbol{q}) - (dV)\frac{\partial\phi}{\partial V}\right) \tag{A.9}$$

A.5. Eigenvalue decay for pressure and temperature test

In Section 4.1.1, the projection error estimates for the POD and PSD bases plotted in Fig. 2 look nearly identical. Indeed, the eigenvalues computed from the standard POD and proposed PSD are very similar but do not coincide. Below, we plot the eigenvalues and projection error estimate along with point-wise differences between the POD and PSD results (see Fig. A.12).

References

[1] D.C. Rapaport, The Art of Molecular Dynamics Simulation, Cambridge University Press, 2004.
[2] W. Shinoda, M. Shiga, M. Mikami, Rapid estimation of elastic constants by molecular dynamics simulation under constant stress, Phys. Rev. B 69 (13) (2004) 134103.
[3] X.W. Zhou, R. Dingreville, R.A. Karnesky, Molecular dynamics studies of irradiation effects on hydrogen isotope diffusion through nickel crystals and grain boundaries, Phys. Chem. Chem. Phys. 20 (1) (2018) 520–534.
[4] T. Schneider, E. Stoll, Molecular-dynamics study of structural-phase transitions. I. One-component displacement models, Phys. Rev. B 13 (3) (1976) 1216.
[5] R. Dingreville, D. Aksoy, D.E. Spearot, A primer on selecting grain boundary sets for comparison of interfacial fracture properties in molecular dynamics simulations, Sci. Rep. 7 (1) (2017) 8332.
[6] S. Xiao, T. Belytschko, A bridging domain method for coupling continua with molecular dynamics, Comput. Methods Appl. Mech. Engrg. 193 (17–20) (2004) 1645–1669.
[7] C.A. Becker, F. Tavazza, Z.T. Trautt, R.A.B. de Macedo, Considerations for choosing and using force fields and interatomic potentials in materials science and engineering, Curr. Opin. Solid State Mater. Sci. 17 (6) (2013) 277–283.
[8] L.M. Hale, Z.T. Trautt, C.A. Becker, Evaluating variability with atomistic simulations: the effect of potential and calculation methodology on the modeling of lattice and elastic constants, Modelling Simul. Mater. Sci. Eng. 26 (5) (2018) 055003.
[9] C.A. Becker, F. Tavazza, Z.T. Trautt, R.A. Buarque de Macedo, Considerations for choosing and using force fields and interatomic potentials in materials science and engineering, Curr. Opin. Solid State Mater. Sci. 17 (6) (2013) 277–283.
[10] L.M. Hale, Z.T. Trautt, C.A. Becker, Evaluating variability with atomistic simulations: the effect of potential and calculation methodology on the modeling of lattice and elastic constants, Modelling Simul. Mater. Sci. Eng. 26 (5) (2018) 055003.
[11] 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.
[12] P.E. Hadjidoukas, P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Π4U: A high performance computing framework for Bayesian uncertainty quantification of complex models, J. Comput. Phys. 284 (2015) 1–21.
[13] F. Rizzi, R.E. Jones, B.J. Debusschere, O.M. 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.
[14] F. Rizzi, R.E. Jones, B.J. Debusschere, O.M. 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.
[15] 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.
[16] S. Wan, R.C. Sinclair, P.V. Coveney, Uncertainty quantification in classical molecular dynamics, Phil. Trans. R. Soc. A 379 (2197) (2021) 20200082.
[17] 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.
[18] M. Kulichenko, K. Barros, N. Lubbers, Y.W. Li, R. Messerly, S. Tretiak, J.S. Smith, B. Nebgen, Uncertainty-driven dynamics for active learning of interatomic potentials, Nature Comput. Sci. 3 (3) (2023) 230–239.
[19] C.I. Yang, Y.-P. Li, Explainable uncertainty quantifications for deep learning-based molecular property prediction, J. Cheminformatics 15 (1) (2023) 13.
[20] S. Thaler, G. Doehner, J. Zavadlav, Scalable Bayesian uncertainty quantification for neural network potentials: Promise and pitfalls, J. Chem. Theory Comput. 19 (14) (2023) 4520–4532.
[21] B.R. Duschatko, J. Vandermause, N. Molinari, B. Kozinsky, Uncertainty driven active learning of coarse grained free energy models, NPJ Comput. Mater. 10 (1) (2024) 9.
[22] 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.
[23] K. Farrell-Maupin, J.T. Oden, Adaptive selection and validation of models of complex systems in the presence of uncertainty, Res. Math. Sci. 4 (1) (2017) 1–15.
[24] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (3) (2001) 425–464.
[25] S.T. Reeve, A. Strachan, Error correction in multi-fidelity molecular dynamics simulations using functional uncertainty quantification, J. Comput. Phys. 334 (2017) 207–220.
[26] 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.
[27] H. Zhang, J. Guilleminot, A Riemannian stochastic representation for quantifying model uncertainties in molecular dynamics simulations, Comput. Methods Appl. Mech. Engrg. 403 (2023) 115702.
[28] 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, et al., LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, Comput. Phys. Comm. 271 (2022) 108171.
[29] L. Peng, K. Mohseni, Symplectic model reduction of Hamiltonian systems, SIAM J. Sci. Comput. 38 (1) (2016) A1–A27.
[30] P. Buchfink, A. Bhatt, B. Haasdonk, Symplectic model order reduction with non-orthonormal bases, Math. Comput. Appl. 24 (2) (2019) 43.
[31] J.S. Hesthaven, C. Pagliantini, N. Ripamonti, Rank-adaptive structure-preserving model order reduction of Hamiltonian systems, ESAIM: M2AN 56 (2) (2022) 617–650.
[32] Y. Gong, Q. Wang, Z. Wang, Structure-preserving Galerkin POD reduced-order modeling of Hamiltonian systems, Comput. Methods Appl. Mech. Engrg. 315 (2017) 780–798.
[33] C. Pagliantini, Dynamical reduced basis methods for Hamiltonian systems, 2021, arXiv:2008.07427.
[34] H. Sharma, Z. Wang, B. Kramer, Hamiltonian operator inference: Physics-preserving learning of reduced-order models for canonical Hamiltonian systems, Physica D 431 (2022) 133122.
[35] H. Sharma, H. Mu, P. Buchfink, R. Geelen, S. Glas, B. Kramer, Symplectic model reduction of Hamiltonian systems using data-driven quadratic manifolds, Comput. Methods Appl. Mech. Engrg. 417 (2023) 116402.
[36] T. Le, J. Guilleminot, 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.
[37] H.C. Andersen, Molecular dynamics simulations at constant pressure and/or temperature, J. Chem. Phys. 72 (4) (1980) 2384–2393.
[38] S. Nosé, A unified formulation of the constant temperature molecular dynamics methods, J. Chem. Phys. 81 (1) (1984) 511–519.
[39] M.E. Tuckerman, J. Alejandre, R. López-Rendón, A.L. Jochim, G.J. Martyna, A Liouville-operator derived measure-preserving integrator for molecular dynamics simulations in the isothermal–isobaric ensemble, J. Phys. A: Math. Gen. 39 (19) (2006) 5629–5651.
[40] W.G. Hoover, Canonical dynamics: Equilibrium phase-space distributions, Phys. Rev. A 31 (1985) 1695–1697, http://dx.doi.org/10.1103/PhysRevA.31.1695, URL https://link.aps.org/doi/10.1103/PhysRevA.31.1695.
[41] G.J. Martyna, M.E. Tuckerman, D.J. Tobias, M.L. Klein, Explicit reversible integrators for extended systems dynamics, Mol. Phys. 87 (5) (1996) 1117–1157.

[42] H. Wang, J. Guilleminot, C. Soize, Modeling uncertainties in molecular dynamics simulations using a stochastic reduced-order basis, Comput. Methods Appl. Mech. Engrg. 354 (2019) 37–55.
[43] R. Zimmermann, K. Hüper, Computing the Riemannian logarithm on the Stiefel manifold: Metrics, methods, and performance, SIAM J. Matrix Anal. Appl. 43 (2) (2022) 953–980.
[44] 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.
[45] R. Zimmermann, J. Stoye, High curvature means low-rank: On the sectional curvature of Grassmann and Stiefel manifolds and the underlying matrix trace inequalities, 2024, arXiv:2403.01879.
[46] A. Rahman, Correlations in the motion of atoms in liquid argon, Phys. Rev. 136 (1964) A405–A411, http://dx.doi.org/10.1103/PhysRev.136.A405, URL https://link.aps.org/doi/10.1103/PhysRev.136.A405.
[47] M. Kim, E. Kim, S. Lee, J.S. Kim, S. Lee, New method for constant-NPT molecular dynamics, J. Phys. Chem. A 123 (8) (2019) 1689–1699.
[48] N. Bernardes, Theory of solid Ne, A, Kr, and Xe at 0°K, Phys. Rev. 112 (1958) 1534–1539.
[49] G.E. Jelinek, Properties of crystalline Argon, Krypton, and Xenon based upon the Born-Huang method of homogeneous deformations. III. The low-temperature limit, Phys. Rev. B 5 (1972) 3210–3217.
[50] E.B. Tadmor, J. Lennard-Jones, Driver for the Lennard-Jones Model Uniformly Shifted to Have Zero Energy at the Cutoff Radius v004, OpenKIM, 2020.
[51] R.S. Elliott, E.B. Tadmor, Knowledgebase of Interatomic Models (KIM) Application Programming Interface (API), OpenKIM, 2011, http://dx.doi.org/10.25950/ff8f563a.
[52] E.B. Tadmor, R.S. Elliott, J.P. Sethna, R.E. Miller, C.A. Becker, The potential of atomistic simulations and the knowledgebase of interatomic models, JOM 63 (7) (2011) 17.