Computational Materials Science 152 (2018) 125–133

Contents lists available at ScienceDirect

Computational Materials Science

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

Fast predictive models based on multi-fidelity sampling of properties in molecular dynamics simulations

M. Razi[a,*], A. Narayan[a], R.M. Kirby[a], D. Bedrov[b]

[a] Scientific Computing and Imaging Institute, University of Utah, United States
[b] Department of Materials Science and Engineering, University of Utah, United States

ARTICLE INFO

Keywords:
Model-order reduction
Molecular dynamics
Multi-fidelity models
Parameter estimation
Surrogate modeling

ABSTRACT

In this paper we introduce a novel approach for enhancing the sampling convergence for properties predicted by molecular dynamics. The proposed approach is based upon the construction of a multi-fidelity surrogate model using computational models with different levels of accuracy. While low fidelity models produce result with a lower level of accuracy and computational cost, in this framework they can provide the basis for identification of the optimal sparse sampling pattern for high fidelity models to construct an accurate surrogate model. Such an approach can provide a significant computational saving for the estimation of the quantities of interest for the underlying physical/engineering systems. In the present work, this methodology is demonstrated for molecular dynamics simulations of a Lennard-Jones fluid. Levels of multi-fidelity are defined based upon the integration time step employed in the simulation. The proposed approach is applied to two different canonical problems including (i) single component fluid and (ii) binary glass-forming mixture. The results show about 70% computational saving for the estimation of averaged properties of the systems such as total energy, self diffusion coefficient, radial distribution function and mean squared displacements with a reasonable accuracy.

1. Introduction

Accurate sampling of the evolution of system of interest in molecular dynamics (MD) simulations can be very challenging. Thus, despite their outstanding predictive power and application in different areas of science and engineering, MD simulations can be carried out over a limited timescale. Since there is a demand for running these simulations for much longer timescales, particularly for studying systems with rough energy landscape and long relaxation processes, several approaches have been proposed in the literature to address this problem.

One of the traditional ways of dealing with this issue is to apply transition-state-theory, compute the rate relevant for infrequent events, and eventually obtain an estimation of the long-time MD results [1]. However, prior determination of all important reaction paths is very difficult. Hence, researchers have proposed different modifications of the potential energy models for MD simulations to overcome the energy barriers much faster. In these series of approaches, through raising the potential energy surfaces in the regions of potential minima, in which standard MD simulations spend a large portion of their computational time, faster exploration of potential energy landscape becomes possible. Hence the simulation moves faster over potential barriers and is able to capture infrequent-event transitions and eventually the equilibrium status of the system in much less computational time [2,3]. This group of approaches, which are characterized by interatomic energy manipulation, are known as the accelerated molecular dynamics (AMD) method. Parallel-replica dynamics method using process parallelization, hyperdynamics approach based upon importance sampling, and temperature-accelerated dynamics by adaptive assignment temperature to transitions are among the well known AMD approaches developed [3]. Similar to the hyperdynamics method, other authors have explored the alteration of potential landscape by proposing a bias potential function as a means for accelerating the MD simulation [2,4]. However, implementation of these methods requires an in-depth understanding of underlying molecular dynamics processes and often performing a series of rigorous computational procedures.

Recently, application of predictive algorithms in the area of molecular dynamics has proven to achieve accurate and computationally efficient long timescale analysis. These approaches span from functional uncertainty quantification (UQ) [5] to multi-fidelity machine learning models [6]. In this context, Reeve and Strachan introduced a novel technique to apply functional UQ to Lennard-Jones two-body interaction model for prediction of high order interatomic interactions [5]. In spite of using a perturbative technique for the numerical evaluation of function derivatives, this approach is computationally

* Corresponding author at: Scientific Computing and Imaging Institute, University of Utah, 72 Central Campus Dr, RM: 2654, Salt Lake City, UT 84112, United States.
E-mail addresses: manirazi@sci.utah.edu (M. Razi), akil@sci.utah.edu (A. Narayan), kirby@sci.utah.edu (R.M. Kirby), d.bedrov@utah.edu (D. Bedrov).

https://doi.org/10.1016/j.commatsci.2018.05.029
Received 28 February 2018; Received in revised form 12 May 2018; Accepted 17 May 2018

demanding since a sufficiently large number of samples from low-fidelity models is required. This is mainly to ensure that the low-fidelity model explores enough of phase space of the high-fidelity model. Furthermore, this approach works only when the discrepancies between high-fidelity and low-fidelity potential energy models stay within reasonable bounds.

In another recent research study, multi-fidelity machine learning regression models have been used for the accurate bandgap prediction of solid materials [6]. This approach, which is based upon a Gaussian process (GP) regression framework, involves the application of co-kriging statistical learning on several bandgap predictive models with different levels of fidelity. This requires having sufficient independent data for parameter estimation and the assumption that the Gaussian process properly describes the underlying molecular dynamics. Hence, any deviation of underlying physics and limited availability of data can lead to a poor parameter estimation.

In the present paper, considering the challenges in computational complexity of predictive models and limited availability of even low-fidelity data in practice, a stochastic collocation methodology with multi-fidelity [7,8] is applied to MD simulation in order to built accurate surrogate models. This approach does not require any *a priori* assumption about the probability measure of the underlying physics that the discrepancies between low-fidelity and high-fidelity model can be much larger than what can be recovered by standard predictive surrogate modeling tools such as spectral (polynomial based) or Bayesian surrogates. The examples of such a large discrepancies between the mathematical models of low- and high-fidelity and the effectiveness of similar multi-fidelity surrogate model constructions in other areas including frequency-modulated trigonometric functions [7], heat driven cavity flows [9] and irradiated particle-laden turbulence [10] have been discussed in the literature. This surrogate modeling approach is also designed to work with a limited number of samples and provide an optimal sampling strategy for the high-fidelity MD simulations. Moreover, compared with the commonly used approaches for acceleration of MD simulations, the proposed method works well not only with small data availability but also with significant difference between high- and low-fidelity models. As long as the low-fidelity model reflects the impact of the variation of model parameters in parameter space (not solution space), it can be used in this framework. In order to demonstrate the proposed multi-fidelity predictive method, it is applied to two canonical problems involving output parameter estimation of MD simulation for (i) one- and (ii) two-component systems. In this study, the integration time step of the MD simulation defines its level of fidelity. Here, by using the proposed multi-fidelity surrogate modeling approach, the goal is to accelerate parameter exploration for a given MD simulation setup. In this light, the resultant acceleration of the MD simulation does directly benefit situations when the MD simulation has a longer terminal time and hence provides an effective tool for faster exploration of time-scales in the defined phase space. As such, this approach can be considered as an efficient alternative to AMD methods, particularly when obtaining a series of MD solutions for different sets of parameters (and not only one) is desirable. The results of our experiments with the application of the proposed approach to both test problems indicate both accuracy and computational efficiency of constructed data driven predictive models. The aforementioned canonical problems are considered for the sole purpose of proof of concept and demonstration of the proposed approach, which is used in the area of molecular dynamics simulations for the first time. In this sense, there is no limitation to apply this approach to more complex MD problems and the authors plan to investigate those cases in their future works. It is also worth noting that the processes of uniform random sampling for the low-fidelity model and important sampling for the high-fidelity model in the phase space lend itself nicely to parallel implementation. By taking advantage of this feature, one can expect some additional acceleration in computational speed for the construction of this multi-fidelity predictive model.

The present manuscript is organized as follows. In Section 2, a brief description of selected model system is provided. This follows by a detailed theoretical foundation of the proposed multi-fidelity in the subsequent section. Next, the results and discussion for two canonical problems are presented in Section 4. Finally, concluding remarks are provided as the final section of this paper.

2. Theoretical foundation

2.1. Molecular dynamics simulation model

The classical computational algorithm for performing molecular dynamics simulations has been developed based upon the Lagrangian methodology of tracking particle dynamics based on Newton's equations of motion. Numerical integration of these equations provides an accurate evaluation of the time evolution of a molecular system and hence the system's quantities of interest. While there are many integration schemes available in the literature, the velocity Verlet algorithm [11] is used in the present work due to its popularity, reasonable accuracy and simplicity of implementation. The Lennard-Jones (LJ) potential [11] was used for simulating the interatomic interactions in the MD system:

$$u(r_{ij}) = 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right], \tag{1}$$

where r_{ij} denotes the pairwise distance between particles i and j, ϵ is the potential well depth and σ defines the length scale for this pairwise interatomic interaction model. For two-species interactions this length scale becomes $\sigma = \sqrt{\sigma_1\sigma_2}$. Here, in the second test problem, we consider a two-component MD system with $\frac{\sigma 1}{\sigma_2} = 0.9$ and same potential well depth.

In the process of molecular dynamic simulation using any interatomic interaction model, when the size of the integration time-step is large, the major challenge is going to be the stability of the MD integration scheme. If the time-step is too large, then a one time step integration may predict significant overlap between molecules/atoms which will lead to huge (unphysical) repulsive forces and large displacements on the next time step, which in turn will lead to even larger overlap and more unphysical forces on the next time step. To prevent such divergence of integration scheme, we capped the magnitude of repulsive interactions for closely approaching atoms. For the Lennard-Jones potential, such capping can be implemented straightforwardly by modifying the potential at short distance. As such, for both test cases in this work, a potential energy cap is considered when the ratio of $\frac{\sigma}{r_{ij}}$ exceeds 1.2. For more complex systems which may involve electrostatic interactions, the capping should include modifications of charge-charge, charge-induced dipole, and induced dipole-induced dipole interactions. These modifications would require a slightly more coding and data management, but will be similar to current schemes of excluding or scaling of intramolecular electrostatic interactions between atoms connected by bonds, bends, and dihedrals.

2.2. Bi-fidelity model construction

In the context of MD simulations, the size of time-step for the integration procedure plays a crucial rule in the computational cost of achieving large time-scale predictions. The drawback of choosing a large time-step to accelerate the MD computations is the loss of accuracy and more often numerical stability of the solution. Here, our goal is to build a multi-fidelity surrogate model, for which the levels of fidelity are defined based upon the time-step size of the corresponding MD simulation. The first step in building such a model is to define a region of parameter space where one needs to estimate state variables with a reasonable accuracy. Choosing a large-time step and running the low-fidelity simulation on randomly selected points of the region of interest in the parameter space is the next step. Here, we consider temperature and density of the molecular system as the parameters defining the phase space. For every sample point, quantities of interest are computed using the low-fidelity model (here, the MD simulation with a large time-step). These quantities of interest, which are in the form of either scalars or vectors, are then concatenated as a vector. For instance, when radial distribution function (RDF), mean squared displacement (MSD), averaged total energy and self diffusion coefficient are the quantities of interest, for the *i*th sample point, we can perform the low- and high-fidelity simulation and generate a low- and high-fidelity vectors as

$$g_{i,L} = (R_{i,L}^T\ M_{i,L}^T\ E_{i,L}^T\ D_{i,L}^T)^T,$$
$$g_{i,H} = (R_{i,H}^T\ M_{i,H}^T\ E_{i,H}^T\ D_{i,H}^T)^T, \tag{2}$$

where R, M, E and D denote RDF, MSD, averaged total energy and diffusion coefficient, respectively. In this bi-fidelity setup, all the high-fidelity quantities (subscript H) are associated with an MD simulation using a small time-step, and the low-fidelity quantities (subscript L) are associated with an MD simulation using a large time-step. The quantities R, M, E, and D are defined as follows and have sizes:

$$\begin{aligned}
&R_{i,L} \in \mathbb{R}^{N_{R,L}}, \quad R_{i,H} \in \mathbb{R}^{N_{R,H}}\\
&M_{i,L} \in \mathbb{R}^{N_{M,L}}, \quad M_{i,H} \in \mathbb{R}^{N_{M,H}}\\
&E_{i,L} \in \mathbb{R}^{N_{E,L}}, \quad E_{i,H} \in \mathbb{R}^{N_{E,H}}\\
&D_{i,L} \in \mathbb{R}^{N_{D,L}}, \quad D_{i,H} \in \mathbb{R}^{N_{D,H}}.
\end{aligned} \tag{3}$$

Here, N denotes the number of elements of corresponding vectorized variable. Thus, the vectors $g_{i,L}$ and $g_{i,H}$ have dimensions of

$$g_{i,L} \in \mathbb{R}^{N_L}, \quad N_L = N_{R,L} + N_{M,L} + N_{E,L}, +N_{D,L} \tag{4}$$

$$g_{i,H} \in \mathbb{R}^{N_H}, \quad N_H = N_{R,H} + N_{M,H} + N_{E,H}, +N_{D,H}. \tag{5}$$

However, in order to have equal impact on the estimation made by the resultant multi-fidelity model, the contribution of all elements should be kept equal. For this purpose and with I number of data points, one can define:

$$\begin{aligned}
&\overline{R} = \frac{1}{I}\sum_{i=1}^{I} \|R_{i,L}\|_2^2,\\
&\overline{M} = \frac{1}{I}\sum_{i=1}^{I} \|M_{i,L}\|_2^2,\\
&\overline{E} = \frac{1}{I}\sum_{i=1}^{I} \|E_{i,L}\|_2^2,\\
&\overline{D} = \frac{1}{I}\sum_{i=1}^{I} \|D_{i,L}\|_2^2,
\end{aligned} \tag{6}$$

and hence the corresponding vector for the *i*th sample point becomes:

$$g_{i,L} = \left[\frac{R_i}{\sqrt{N_R\overline{R}}}, \frac{M_i}{\sqrt{N_M\overline{M}}}, \frac{E_i}{\sqrt{N_M\overline{E}}}, \frac{D_i}{\sqrt{N_D\overline{D}}}\right]^T. \tag{7}$$

The next step involves the construction of the Gramian matrix using $g_{i,L}$. This matrix is defined as the symmetric correlation matrix $G_L \in \mathbb{R}^{I\times I}$ with entries

$$(G_L)_{i,j} = \langle g_{i,L}, g_{j,L}\rangle \quad i, j = 1, \ldots, I, \tag{8}$$

where $\langle\cdot,\cdot\rangle$ is the inner product vectors $g_{i,L}$ and $g_{j,L}$. There are a handful of procedures that, given this matrix, choose $n < I$ indices from the full set $\{1, 2, \ldots, I\}$. The common point selection procedures are Orthogonal-triangular (QR) decomposition [7], Cholesky decomposition [8], LU factorization [12], leverage score sampling methods [13], and group matching methods [14]. The first two approaches are equivalent weak greedy procedures for point selection based on linear algebra operations. In their work, Narayan et al. have shown that greedy procedures can be effectively used for selecting the candidate interpolation points [7]. They have also shown that the application of these greedy point selection algorithms on low-fidelity data produces an error which is reasonably close to that of optimal projection using high-fidelity data. These results suggest that greedy procedures are more effective than random (Monte Carlo) sampling of points in parameter space.

In this research, the authors choose to use QR decomposition for this purpose firstly because it is more understandable to general engineering community and also due to its simplicity, speed, effectiveness as well as the availability of its computationally efficient implementations in most programming environments. QR decomposition can produce the ordering information for the V_L matrix, where V_L is a matrix where its *i*th column is vector $g_{i,L}$ (So that $G_L = V_L^T V_L$). This procedure is equivalent to Cholesky factorization of G_L. Hence, if n column indices $(i_1, \ldots, i_n)$ of G_L are selected based on this ordering, i.e.,

$$\{i_1, \ldots, i_n\} \subset \{1, 2, \ldots, I\}, \tag{9}$$

they are algebraically equivalent to as ordered pivots from a Cholesky factorization of G_L.

The multi-fidelity procedure builds the approximation

$$g_{i,H} \approx \widetilde{g}_{i,H} = \sum_{k=1}^{n} g_{i_k,H}\, c_k(i), \quad i = 1, \ldots, I \tag{10}$$

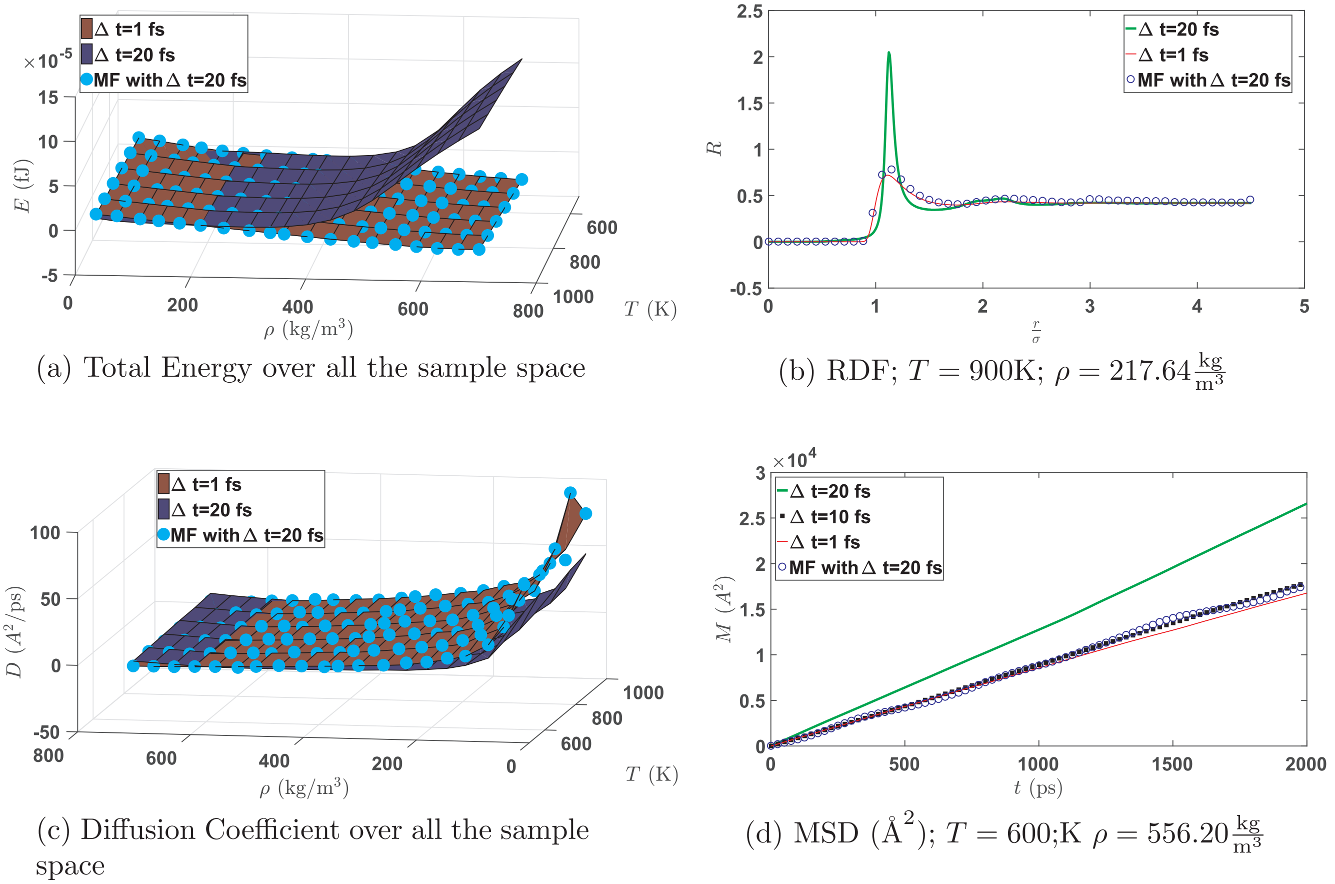
where the coefficients $c_k(i)$ are learned from the low-fidelity data $g_{i,L}$ only and precisely via the vector solution $C(i) = [c_1(i), c_2(i), \ldots, c_n(i)]^T$ to the linear system:

$$\begin{pmatrix} (G_L)_{i_1,i_1} & \cdots & (G_L)_{i_1,i_n}\\ \vdots & \ddots & \vdots\\ (G_L)_{i_n,i_1} & \cdots & (G_L)_{i_n,i_n} \end{pmatrix} C(i) = \begin{pmatrix} \langle g_{i,L}, g_{i_1,L}\rangle\\ \vdots\\ \langle g_{i,L}, g_{i_n,L}\rangle \end{pmatrix}. \tag{11}$$

The solution of Eq. (11) provides the weights for the resultant surrogate predictive model ($\widetilde{g}_{i,H}$; Eq. (10)) and the column index selection process provides the optimal sampling set of input parameters for running the high-fidelity simulation with. This bi-fidelity model can be easily extended to include more levels of fidelity [8], as will be briefly discussed in the next section. Also, this procedure can be carried out even with a few number of samples. Moreover, the computational cost of formation and solution of Eq. (11) is often negligible when compared to the computational cost of producing even one sample of a low-fidelity MD simulation (compare that to the cost of hyperparameter optimization for Gaussian process regression or coefficient recovery for polynomial-based surrogate modeling). More importantly, in order to use this procedure, one only needs to consider a low-fidelity model that can capture the variation in the parameter space (which includes a collection of model snapshots under different sets of parameters), which indeed is a minimum requirement for a model. The resultant predictive model might produce inaccurate results provided that there is a drastic discrepancy between the behaviors of low- and high-fidelity in the phase (parameter) space. It is also important to emphasize that discrepancies that prove adversarial for the bi-fidelity approach would need to manifest in parameter space and should not be confused with the differences in models' responses such as having different number and positions of peaks in the solution space (see Refs. [7,10] for examples of such cases).

2.3. Extension to multi-fidelity model construction

The proposed procedure for the construction of the bi-fidelity approach can be easily extended to the cases, in which three or more models exist. In such cases, the process of selection of important points in the parameter space is performed based on the data obtained from the model with the lowest accuracy and computational cost in the same manner as discussed in the previous section. The models, which are used in this multi-fidelity procedure, should be then organized in a hierarchical manner from the lowest to the highest fidelity level. Next, the process of multi-fidelity model construction proceeds by



(a) Total Energy over all the sample space

(b) RDF; $T = 900\text{K}$; $\rho = 217.64\frac{\text{kg}}{\text{m}^3}$

(c) Diffusion Coefficient over all the sample space

(d) MSD ($\mathring{\text{A}}^2$); $T = 600$;K $\rho = 556.20\frac{\text{kg}}{\text{m}^3}$

Fig. 1. Comparison of properties predicted for single-component system as obtained using high fidelity model (Δt = 1 fs), low fidelity models (Δt = 20 fs and Δt = 10 fs), and multi-fidelity approach based on low fidelity data and few (25) samples from high-fidelity; Test Problem 1.

computation of the coefficients of the linear combination of the next higher fidelity level using the samples from the one level lower fidelity model in the hierarchy. This step by step procedure continues until, the coefficients for the highest fidelity surrogate model are computed. Finally, using Eq. (10) the quantities of interest can be estimated at desired locations in the phase space. In this manuscript, only the cases with bi-fidelity model construction are considered for the purpose of demonstration and the proof of concept. The inclusion of more models with the assumption of the increase in the computational cost with their fidelity levels, is straightforward and may generate even higher accuracy and computational efficiency. A comprehensive and detailed description of this generalization can be found in the manuscript by Zhu et al. [8].

3. Results and discussion

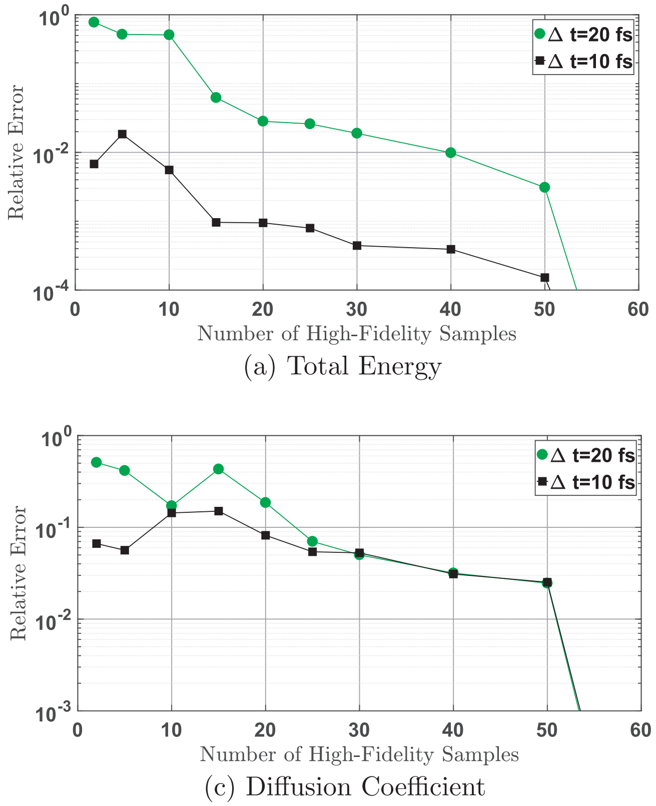
In order to demonstrate the proposed application capability to construct an accurate predictive model for MD simulations, we present two application examples in this section. These examples have been selected such that the computational cost associated with the high-fidelity model is reasonable. Hence, evaluation of the prediction error over a substantially large grid of sample points is tractable in these cases. The error that is shown in the plots in this work is a median over a size-I ensemble of errors. For example, the energy error is computed as follows:

$$\text{error}(E) = \underset{i=1,\dots,I}{\text{median}} \frac{|E_{i,H} - \widetilde{E}_{i,H}|}{E_{i,H}}. \tag{12}$$

While in the following test problems the same number of high- and low-fidelity simulations have been carried out for the sake of the error estimation, it should be noted that the proposed procedure is designed such that one only needs n samples of high-fidelity MD simulations for the accurate prediction of a size-I data set of quantities of interest ($n \ll I$).

3.1. Test Problem 1: Single-component system

The first benchmark example deals with the MD simulation of a one-component LJ system. Here, the interatomic interactions between molecules of one type follows Eq. (1). For this example, we assume σ, ε, m or molecular mass to be 3 Å, 1 $\frac{\text{kcal}}{\text{mol}}$ and 12.01 $\frac{\text{g}}{\text{mol}}$, respectively. Here, the boundary conditions in all sides of cubic simulation box (with the width of 27.05 Å) are considered to be periodic. Also, we consider a uniform grid of temperature and density defined as $T \times \rho$: [500, 1000] K × [36.27, 701.29] $\frac{\text{kg}}{\text{m}^3}$ (ρ^*: [0.05, 0.95]) with 114 sample points, where ρ and ρ^* are density and dimensionless density equal to $\frac{Nm}{V}$ and $\frac{N\sigma^3}{V}$, respectively. Here the mass of each particle (m) and simulation box size ($L = V^{\frac{1}{3}}$) is set to 12.01 g/mol and 27.05 Å, respectively. Hence based on the aforementioned dimensionless density range, the number of molecules is ranging from 36 to 696. The low-fidelity simulations for this example have a temporal increment of 10 and 20 fs. On the other hand, setting the MD simulation time step to 1 fs results in obtaining high-fidelity simulation data. As shown in Fig. 1, the results from 10 fs simulations suffer much less from inaccuracy compared to those of 20 fs simulations. However, as shown in Fig. 2, after the implementation of the proposed bi-fidelity approach, we could predict different properties from MD simulation with a very good accuracy (less than 10% error) in spite of considerable low accuracy of the low-fidelity model (20 fs simulation model). In the construction of these bi-fidelity surrogate models only a few (25) samples from high-fidelity



(a) Total Energy

(b) RDF

(c) Diffusion Coefficient

(d) MSD

Fig. 2. Median error of the multi-fidelity model constructed based upon the results from the low fidelity models (MD simulations with Δt = 20 and Δt = 10 fs) in the prediction of the quantities of interest for the high fidelity model (Δt = 1 fs); Test Problem 1; 114 data points.

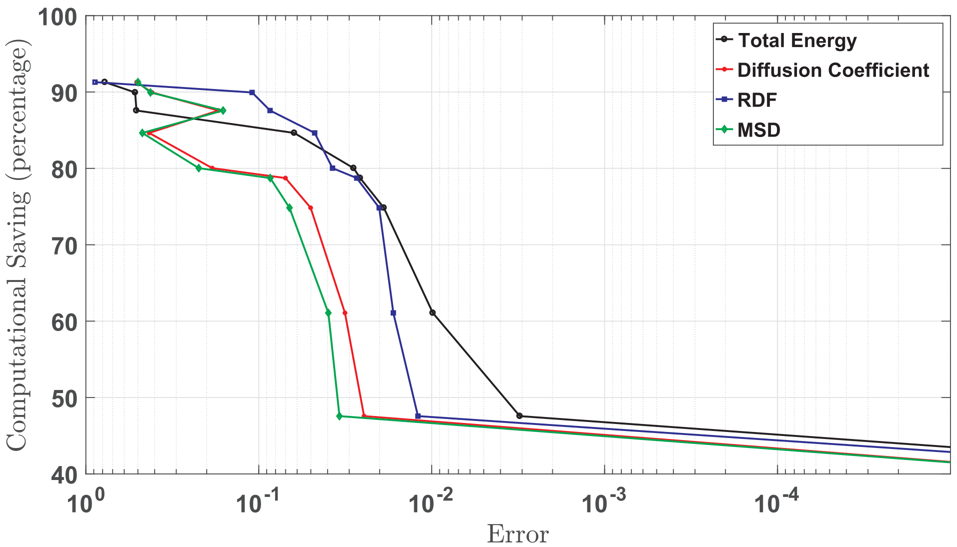
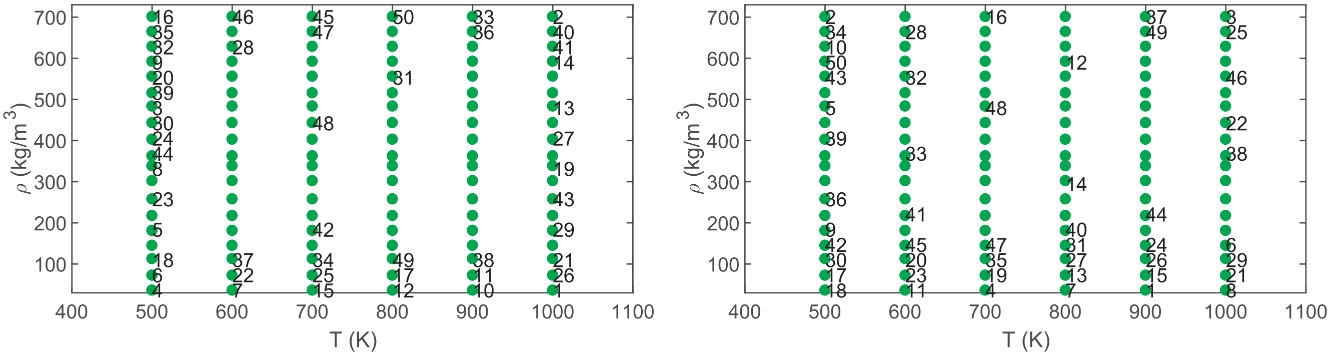


Fig. 3. Resultant computational saving from using the low fidelity model (Δt = 20 fs) in the prediction of the quantities of interest for the high fidelity model (Δt = 1 fs); Test Problem 1; 114 data points.

model are needed to obtain this level of accuracy Here, we tested two equivalent weak greedy procedures for point selection algorithms listed in Ref. [7] and obtained the very same ordering in a fraction of second. Moreover, from our results, it appears that compared to the other MD simulation acceleration approaches, the proposed method works even with significant difference between high- and low-fidelity models responses (as shown in Fig. 1), and works well with small data availability (see Fig. 2). Also, the sharp decline in the error appears to be the results of redundancy in the data with respect to the corresponding variation in the energy, diffusion coefficient, radial distribution function and mean squared displacement of the investigated LJ system.

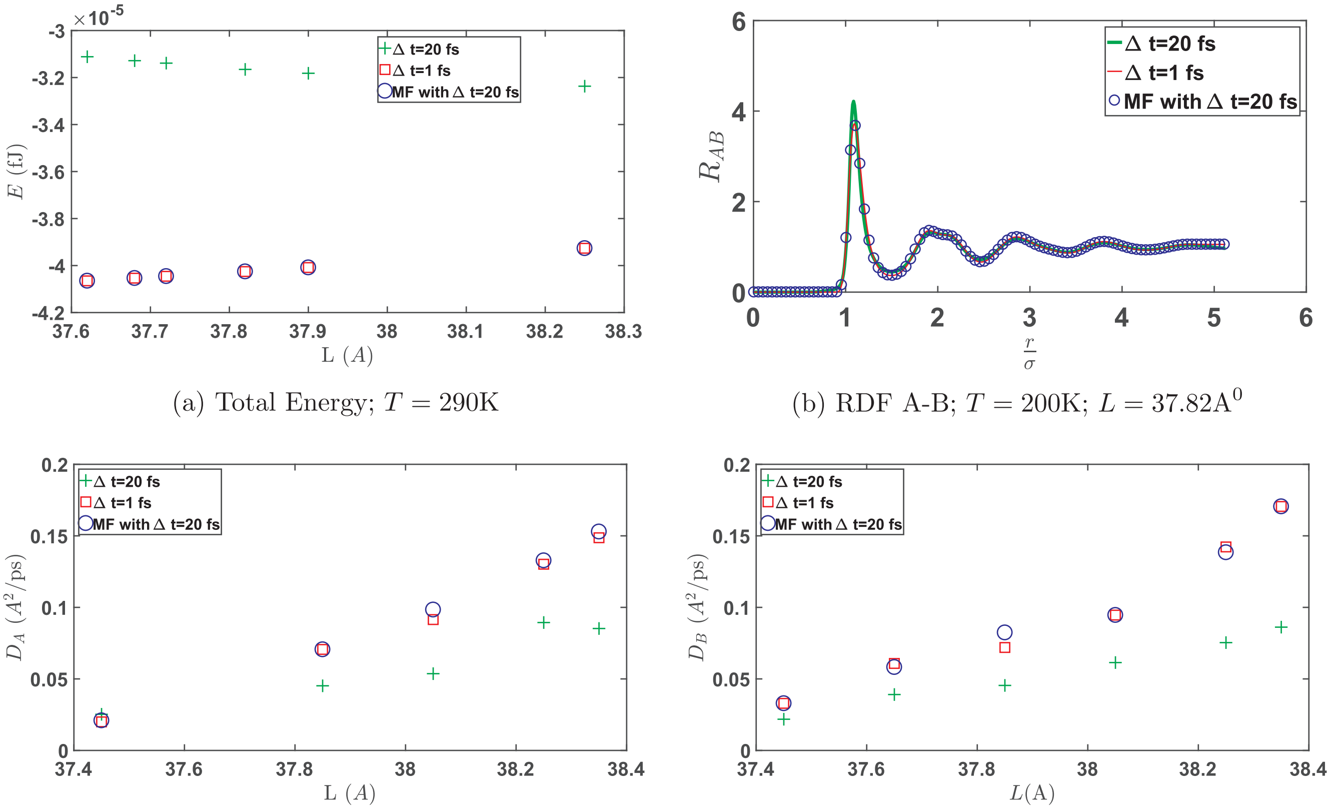
In comparison to the model constructed only by high-fidelity samples, the computational saving of using the proposed approach is very significant. As shown in Fig. 3, this reduction in computational cost is about 70% for achieving less than 10% prediction error. It is also worth noting that the cost of this model construction method is insignificant even compared to the cost of one low-fidelity MD simulation run. As such, the resultant computational saving highlights not only this outstanding feature of the proposed methodology but also its potential capability to extend the application of MD simulation to explore larger



(a) $\Delta t = 20$ fs

(b) $\Delta t = 10$ fs

Fig. 4. Importance sampling based on ordered pivots from the Orthogonal-triangular (QR) decomposition of G_L; Test Problem 1; 114 data points.



(a) Total Energy; $T = 290$K

(b) RDF A-B; $T = 200$K; $L = 37.82A^0$

(c) Diffusion Coefficient A-A $(\frac{A^2}{\text{ps}})$; $T = 250$K

(d) Diffusion Coefficient B-B $(\frac{A^2}{\text{ps}})$; $T = 270$K

Fig. 5. Comparison of properties predicted for two-component system as obtained using high fidelity model (dt = 1 fs), low fidelity model (dt = 20 fs), and multi-fidelity approach based on low fidelity data and few (11) samples from high-fidelity; Test Problem 2.

time-scales.

The other interesting aspect of using this approach is the physical intuition that can be obtained by its process of importance sampling (Gramian matrix column selection). As mentioned in Section 2, here we assess the importance of samples based on ordered pivots from the Orthogonal-triangular (QR) decomposition of V_L. For this canonical problem, we notice that, among first 50 sample points, this process samples more frequently from low density region of the phase space (see Fig. 4). This is more conspicuous if 10 fs MD simulation is considered as the low-fidelity model, due to its better accuracy as compared to 20 fs model. Hence, it appears that in order to construct a surrogate model that can describe the system behavior more accurately, often more data are needed to capture the parametric variation in this region.

3.2. Test Problem 2: Two-component system

The second test problem, which is discussed in this paper, involves the construction of a predictive multi-fidelity model for molecular simulations of a two-component glass-forming system. Here, the length scale for pairwise interatomic interactions between molecules are considered to be $\sigma_A = 3.5$ Å for interactions between molecules of type "A", $\sigma_B = 3.888$ Å for interactions between molecules of type "B" and $\sigma_{AB} = \sqrt{\sigma_A \sigma_B}$ for interaction between both types of molecules. For this experiment, 24 input data points in the density-temperature (parameter) space are considered. The number of molecules for each type is set to be equal to 512. Once again, the boundary conditions are assumed to be periodic for all sides of the simulation box. The temperature and simulation box length for these sample points vary between

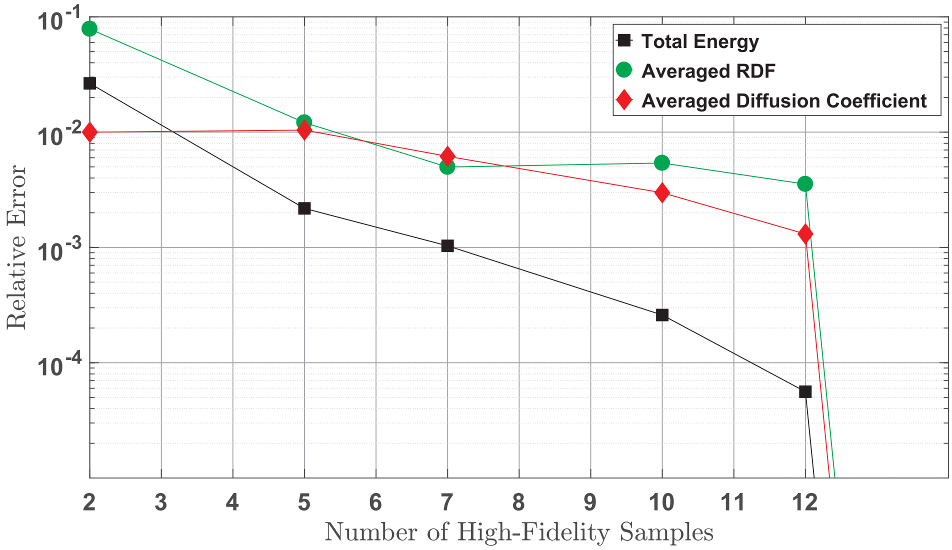


Fig. 6. Median error of the multi-fidelity model constructed based upon the results from the low fidelity model (MD simulations with Δt = 20 fs) in the prediction of the quantities of interest for the high fidelity model (Δt = 1 fs); Test Problem 2; 24 data points; averaged RDF: averaged median error for the prediction of R_A, R_B and R_{AB}; averaged diffusion coefficient: averaged median error for the prediction of D_A and D_B.

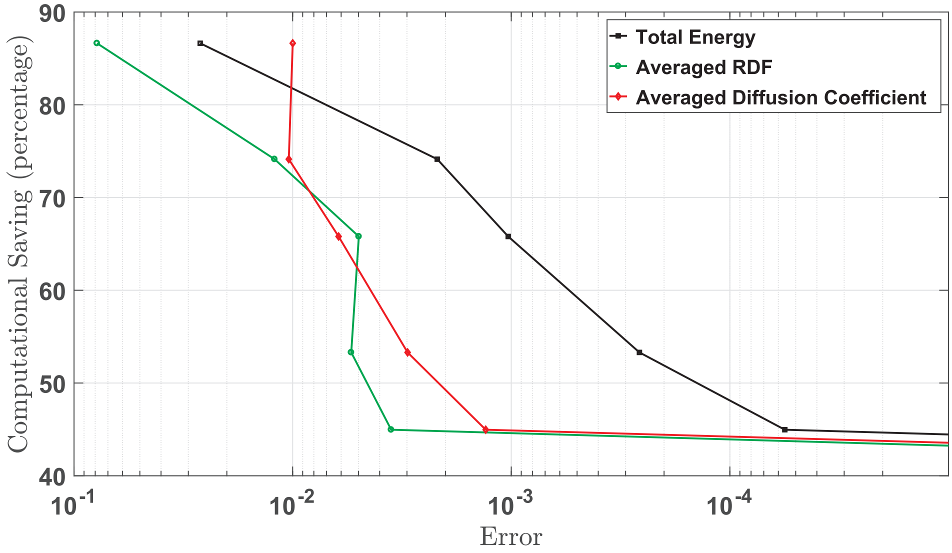
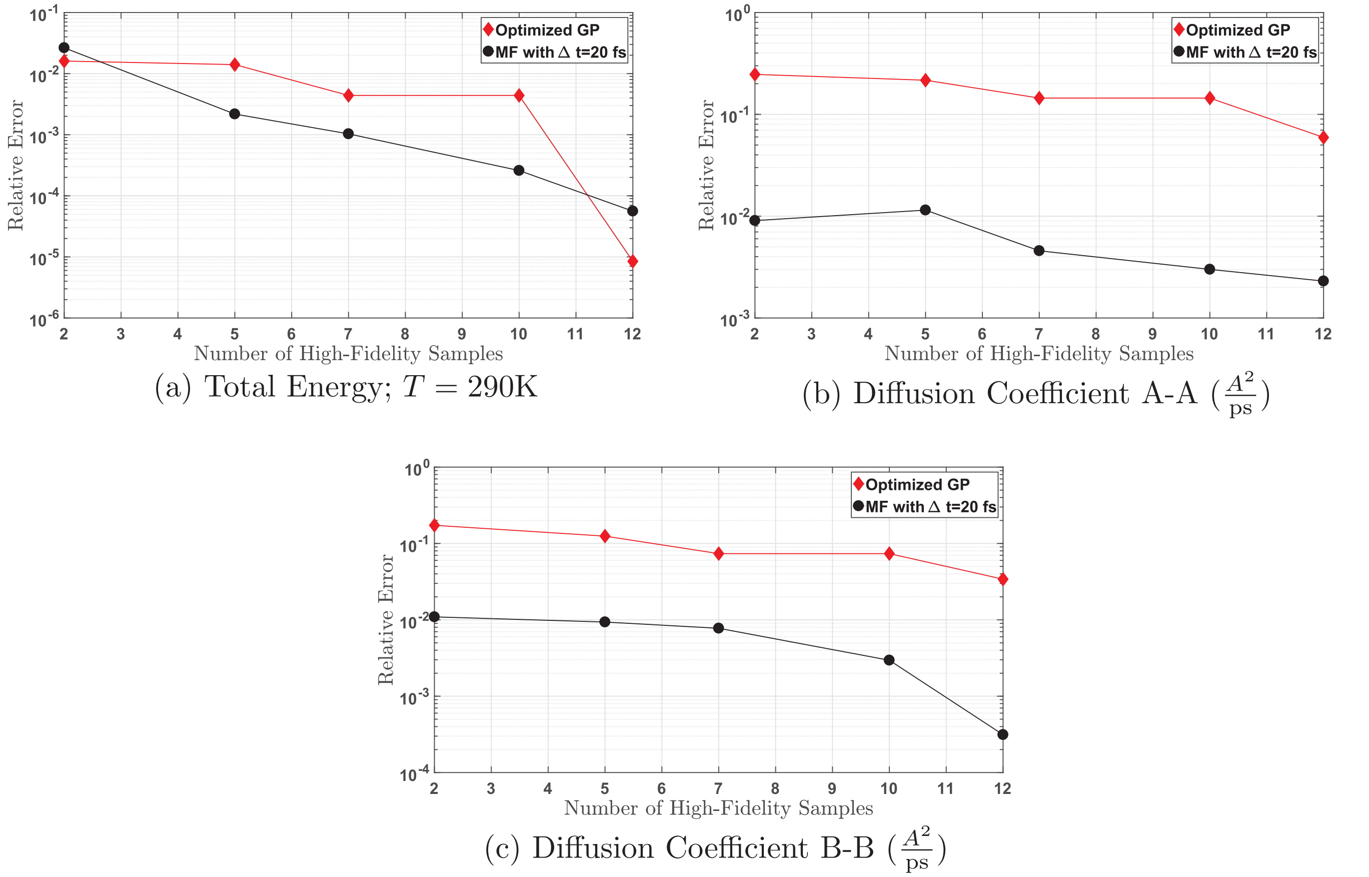


Fig. 7. Resultant computational saving from using the low fidelity model (Δt = 20 fs) in the prediction of the quantities of interest for the high fidelity model (Δt = 1 fs); Test Problem 2; 24 data points; averaged RDF: averaged median error for the prediction of R_A, R_B and R_{AB}; averaged diffusion coefficient: averaged median error for the prediction of D_A and D_B.

180 K and 290 K, and 37.62 Å and 38.35 Å, respectively. Unlike the previous test problem, due to low temperature and slow dynamics at some of these data points, exploring longer time-scales to estimate the self diffusion coefficient is necessary. Here, the vector $g_{i,L}$ is comprised of concatenation of three radial distribution functions involving interactions of molecules of type “A” and “B” (R_{AA}, R_{AB} and R_{BB}) as well as corresponding diffusion coefficients (D_A and D_B) and system’s total energy E. This vector, similar to Eqs. (6) and (7) can be normalized and then the corresponding Gramian matrix (G_L) can be constructed by computing the inner products of the resultant vectors. As discussed previously, the ranks of the sampling points is based on their importance and can be produced by the Cholesky factorization or LU factorization of G_L or QR decomposition of V_L matrix. All of these equivalent linear algebra operations produced the very same ranking in a fraction of second for this case, as we had expected.

In this approach, since samples are drawn from the phase space and hence predictions are made for a desired set of parameters in this space, there is absolutely no need for the low-fidelity snapshots to look similar to the high-fidelity ones at all. This is shown in previous test problem (see Fig. 1b) and once again, as shown in Fig. 5, the inaccuracy of the low-fidelity two-component model (MD simulation with Δt = 20 fs) is reduced using the proposed procedure. Although only the enhancement in the accuracy of R_{AB} is shown in this figure, we observe similar agreement between the bi- and high-fidelity estimation of R_{AA} and R_{BB}. The extents of this error reduction for different MD simulation variables are illustrated in Fig. 6. Similar to the previous case, a sharp decline in the error is observed when the number of high fidelity samples exceeds half size of the low-fidelity simulation training set. This decline leads to a significant computational gain. As can be seen in Fig. 7, applying this approach to reduce the error up to an order of magnitude requires only 25–35% of the computational effort needed for the full high-fidelity based surrogate model and hence one can construct a surrogate model that predicts the quantities of interest with a high accuracy (relative error of less than 0.01) with a significant computational saving.



(a) Total Energy; $T = 290\text{K}$

(b) Diffusion Coefficient A-A $(\frac{A^2}{\text{ps}})$

(c) Diffusion Coefficient B-B $(\frac{A^2}{\text{ps}})$

Fig. 8. Comparison of median error of the multi-fidelity model constructed based upon the results from the low fidelity model (MD simulations with $\Delta t = 20$ fs) and Gaussian process (GP) regression using high-fidelity data in the prediction of the quantities of interest; Test Problem 2; 24 data points.

Moreover, the application of the bi-fidelity approach to this test problem produces better accuracy and computational feasibility for predictive analysis of MD simulation all over the phase space in comparison with standard direct interpolation schemes. Particularly as shown in Fig. 8, when the results of this bi-fidelity analysis are compared with the results of Gaussian process regression for fitting a surrogate model using the high-fidelity MD data, it becomes clear that the proposed approach often produces more accurate predictions by up to two orders of magnitude. Here, after centering and normalization of high-fidelity data, a standard Gaussian process regression algorithm that optimizes hyperparameters along with the common choice of the squared exponential kernel function with isotropic distance measure are used. This algorithm is available in MATLAB® via the GPML® routine [15]. For the purpose of comparison only, a set of randomly selected high-fidelity data is used as the training set. Furthermore, in order to evade local minima for the hyperparameter optimization process, 10 sets of starting points for both hyperparameters and noise are used. It is also worth noting that we have found the process of optimization of hyper parameters for Gaussian process regression to be computationally more costly than the proposed multi-fidelity strategy as the computational speed of Cholesky factorization of the resultant Gramian matrix for the point selection in parameter space clearly is larger than that of even most efficient hyperparameter optimization approaches for this nonconvex optimization problem. Hence, it appears that the proposed approach for the construction of the multi-fidelity surrogate model is superior to Gaussian process regression both with respect to accuracy and computational cost for these particular test problems (the authors observed similar behavior for a comparison on the first test problem; these results are not included in this manuscript for the sake of its brevity). One of the main reasons for this difference is that Gaussian process emulators with optimized hyperparameters are non-adapted approximations. That is, the predicted surface is constructed from a dictionary of kernel functions whose variation may not be predictive of the process being modeled. The bi-fidelity approach prescribed here ameliorates this by using an approximation scheme in parameter space that is adapted since it is built from the low-fidelity model. As such, the bi-fidelity procedure uses the low-fidelity model for two purposes: (1) to guide point selection, and (2) to build an adaptive approximation scheme in parameter space. Hence, even if a Gaussian process emulator uses training data built from intelligent point selection, it cannot boast adaptive approximation properties.

4. Concluding remarks

In order to accelerate molecular simulation and explore larger timescale in a reasonable computational time, one natural approach is to increase the simulation time-step. However, the penalty for this increase is a significant inaccuracy of solutions. In this manuscript, an approach is proposed to address this issue directly. The proposed multi-fidelity predictive modeling method takes advantage of often low computational costs associated with running an MD simulation with a large time-step to identify a set of optimal sampling points. Next, by sampling the solution of MD simulation with a small time-step a quadrature rule for the accurate prediction of quantities of interest is obtained. The capabilities of this approach were demonstrated with two benchmark problems involving the (i) one- and (ii) two-component Lennard-Jones systems. The results for both cases indicate the consistent performance of the proposed multi-fidelity method in accurate estimation of high-fidelity MD simulations. As the number of high-fidelity samples is small, one can gain a significant saving of computational time.

In short, this approach can be characterized by (1) short analysis time even compared with computationally cheap low-fidelity simulations, (2) requirement of relatively small number of high-fidelity model evaluations, (3) effective use of low-fidelity simulation for importance sampling and (4) the possibility of using any model that captures the variations in the parameter space. Hence, its application can pave the way to study more challenging problems in engineering practice, which often demands exploring large-time scales, with MD simulations.

Data availability statement

The authors confirm that the data required to reproduce the findings of this study are available within the article and can be reproduced by molecular dynamic simulations.

Acknowledgments

M. Razi, D. Bedrov, A. Narayan and R. M. Kirby acknowledge that their part of this research was sponsored by ARL under Cooperative Agreement Number W911NF-12-2-0023. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] P. Pechukas, Transition state theory, Annu. Rev. Phys. Chem. 32 (1981) 159–177.

[2] D. Hamelberg, J. Mongan, J.A. McCammon, Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules, J. Chem. Phys. 120 (2004) 11919–11929.

[3] D. Perez, B.P. Uberuaga, Y. Shim, J.G. Amar, A.F. Voter, Accelerated molecular dynamics methods: introduction and recent developments, Annu. Rep. Comput. Chem. 5 (2009) 79–98.

[4] R.A. Miron, K.A. Fichthorn, Accelerated molecular dynamics with the bond-boost method, J. Chem. Phys. 119 (2003) 6210–6216.

[5] S.T. Reeve, A. Strachan, Error correction in multi-fidelity molecular dynamics simulations using functional uncertainty quantification, J. Comput. Phys. 334 (2017) 207–220.

[6] G. Pilania, J.E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids, Comput. Mater. Sci. 129 (2017) 156–163.

[7] A. Narayan, C. Gittelson, D. Xiu, A stochastic collocation algorithm with multifidelity models, SIAM J. Sci. Comput. 36 (2014) A495–A521.

[8] X. Zhu, A. Narayan, D. Xiu, Computational aspects of stochastic collocation with multifidelity models, SIAM/ASA J. Uncertain. Quantif. 2 (2014) 444–463.

[9] J. Hampton, H. Fairbanks, A. Narayan, A. Doostan, Parametric/stochastic model reduction: low-rank representation, non-intrusive bi-fidelity approximation, and convergence analysis, 2017. Available from: arXiv preprint < arXiv:1709.03661 > .

[10] L. Jofre, G. Geraci, H. Fairbanks, A. Doostan, G. Iaccarino, Multi-fidelity uncertainty quantification of irradiated particle-laden turbulence, 2018. Available from: arXiv preprint < arXiv:1801.06062 > .

[11] J.G. Lee, Computational Materials Science: An Introduction, CRC Press, 2016.

[12] D. Anderson, M. Gu, An efficient, sparsity-preserving, online algorithm for low-rank approximation, in: International Conference on Machine Learning, 2017, pp. 156–165.

[13] D.J. Perry, R.T. Whitaker, Augmented leverage score sampling with bounds, Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2016, pp. 543–558.

[14] A. Lozano, G. Swirszcz, N. Abe, Group orthogonal matching pursuit for logistic regression, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 452–460.

[15] C.E. Rasmussen, H. Nickisch, Gaussian processes for machine learning (gpml) toolbox, J. Mach. Learn. Res. 11 (2010) 3011–3015.