Fusing heterogeneous data for the calibration of molecular dynamics force fields using hierarchical Bayesian models

Stephen Wu, Panagiotis Angelikopoulos, Gerardo Tauriello, Costas Papadimitriou, and Petros Koumoutsakos

Citation: J. Chem. Phys. **145**, 244112 (2016); doi: 10.1063/1.4967956
View online: http://dx.doi.org/10.1063/1.4967956
View Table of Contents: http://aip.scitation.org/toc/jcp/145/24
Published by the American Institute of Physics

Fusing heterogeneous data for the calibration of molecular dynamics force fields using hierarchical Bayesian models

Stephen Wu,[1,a)] Panagiotis Angelikopoulos,[1] Gerardo Tauriello,[1] Costas Papadimitriou,[2] and Petros Koumoutsakos[1,b)]
[1]*Computational Science and Engineering Laboratory, ETH-Zurich, Clausiusstrasse 33, CH-8092 Zurich, Switzerland*
[2]*Department of Mechanical Engineering, University of Thessaly, 38334 Volos, Greece*

(Received 30 May 2016; accepted 5 November 2016; published online 30 December 2016)

We propose a hierarchical Bayesian framework to systematically integrate heterogeneous data for the calibration of force fields in Molecular Dynamics (MD) simulations. Our approach enables the fusion of diverse experimental data sets of the physico-chemical properties of a system at different thermodynamic conditions. We demonstrate the value of this framework for the robust calibration of MD force-fields for water using experimental data of its diffusivity, radial distribution function, and density. In order to address the high computational cost associated with the hierarchical Bayesian models, we develop a novel surrogate model based on the empirical interpolation method. Further computational savings are achieved by implementing a highly parallel transitional Markov chain Monte Carlo technique. The present method bypasses possible subjective weightings of the experimental data in identifying MD force-field parameters. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4967956]

I. INTRODUCTION

Recent advances in high performance computing have enabled Molecular Dynamics (MD) simulations to become an effective "computational microscope" for understanding living organisms and for the design of new materials.[1] The calibration of force-field parameters is a critical aspect of the predictive capabilities of MD simulations. A fundamental problem in such calibrations is that the data are obtained from different types of experiments (e.g., neutron magnetic resonance and cryo-scanning tunneling microscope) or quantum mechanical calculations.[2,3] Furthermore, the data may represent different thermodynamical states as well as molecular and bulk level properties of the system under consideration. The identification of force-field parameters can be cast as a single or multi-objective optimisation problem to reproduce different kinetic and thermodynamic measurements. However, the use of heterogeneous data often leads to inconsistent results that may be attributed to the uncorrelated physical properties expressed by the data as well as to the various sources of uncertainties, such as measurement noise and modeling inadequacy.

The modeling inadequacy reflects that all functional forms of MD potentials and their respective simulation protocols are incapable to capture simultaneously all thermodynamic properties. Hence, one may find large discrepancies in the identified parameter values across calibrations based on the different data. One example is the water-carbon interaction parameters. There is extensive literature on choosing the quantities to calibrate when parameterizing water force-fields.[4–6] However, even when using the exact same water model, one can find drastically different values for the oxygen-hydrogen interaction parameters depending on the measurement selected for the calibration. It is evident that the force-fields parameters hinge upon the type of measurements while the combination of different experimental data is largely treated empirically.

One approach for calibrating systematically the parameters based on heterogeneous data is to use multi-objective optimization.[7–9] A drawback in these approaches is the rather arbitrary weighing of each data set. The weights are empirically chosen to represent the relative contribution of each data set. Furthermore, multi-objective optimization approaches only report sets of Pareto-optimal force-field parameter values, without quantifying their uncertainties.

Here we propose a fusion of heterogeneous data using a hierarchical Bayesian framework for the calibration and uncertainty quantification of MD force-field parameters. Similar hierarchical models have been used for accounting cross-experiment uncertainty.[10,11] In this study, the contribution of each data set to the calibration is handled through the evidence (marginal likelihood) in the Bayesian framework. The evidence incorporates a trade-off between data-fitting and information gain from each experimental data set and guards against over-fitting. This approach weighs the contribution from each data set based on its training error sensitivity to the parameter values.

A hierarchical Bayesian analysis requires a large number of function evaluations, which implies a significant computational cost when applied to MD simulations. To alleviate the computational cost, we develop an efficient approximation scheme that combines the use of a two-level surrogate model, based on the Empirical Interpolation Method (EIM),[12] with an efficiently parallelized[13] Transitional Markov Chain

[a)]Current address: Institute of Statistical Mathematics, Tokyo, Japan.
[b)]Electronic mail: petros@ethz.ch; http://www.cse-lab.ethz.ch/

Monte Carlo (TMCMC) method.[14] We validate our proposed methodology for the calibration of MD force-field parameters of water molecules. We use experimental measurements of three properties: (1) diffusivity (a kinematic property), (2) density (a macroscopic physical property), and (3) radial distribution function (RDF) (a microscopic structural property). These data are used to calibrate the Lennard-Jones potential parameters and the Coulombic partial charges of the water model.

This paper is structured as follows: In Section II, we present the mathematical formulation for the hierarchical Bayesian framework. In Section III, we describe the construction of the two-level EIM surrogate. In Section IV, we illustrate the MD model and the setup of the TMCMC to be used with the EIM surrogate. We present our inference results in Section V and conclude in Section VI.

II. HIERARCHICAL BAYESIAN FRAMEWORK FOR HETEROGENEOUS DATA

In the Bayesian framework,[15] calibration and uncertainty quantification of parameters $\vec{\theta}$ of a model $\mathcal{M}$ imply the computation of the posterior distribution ($p(\vec{\theta}|\mathcal{D}, \mathcal{M})$) of $\vec{\theta}$ given a set of data $\mathcal{D}$. The posterior distribution is proportional to the product of a likelihood function $p(\mathcal{D}|\vec{\theta}, \mathcal{M})$, which describes how likely the data are observed given $\vec{\theta}$, and a prior distribution $p(\vec{\theta}|\mathcal{M})$,

$$p(\vec{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\vec{\theta}, \mathcal{M})p(\vec{\theta}|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}, \tag{1}$$

where the evidence $p(\mathcal{D}|\mathcal{M})$ is a normalizing constant that quantifies the trade-off between how well $\mathcal{M}$ fits $\mathcal{D}$ and how much information is extracted from $\mathcal{D}$ in order for $\mathcal{M}$ to fit $\mathcal{D}$.[16] The posterior distribution is used to predict a quantity of interest y based on the total probability theorem and the assumption that prediction of y does not depend on the data when $\vec{\theta}$ is known,

$$p(y|\mathcal{D}, \mathcal{M}) = \int p(y|\vec{\theta}, \mathcal{M})p(\vec{\theta}|\mathcal{D}, \mathcal{M})\, \mathrm{d}\vec{\theta}. \tag{2}$$

When there are multiple sets of heterogeneous data ($\mathcal{D} = \{D_i | i = 1, \ldots, N_D\}$), one approach is to apply the same Bayes' theorem and choose a likelihood function that accounts for all data sets simultaneously. For example, we can use a weighted sum of the likelihood for each data set, in a fashion similar to classical multi-objective optimization methods. Another possibility is to assume that the model parameters $\vec{\theta}$ predict the quantity of interest y_i for each data set (D_i) independently (see Figure 1(a) for a Bayesian network representation). This results in a product of the likelihood functions of all data sets. We note that this approach integrates data sets based on data-fitting represented by the likelihood functions.

We propose an alternative approach that fuses the heterogeneous data using an extra layer in the Bayesian network. Our approach is based on a hierarchical probabilistic model that includes a chosen structure for the prior distribution, parameterized by $\vec{\psi}$. We assume that the predictions y_i of each data set D_i may depend on different model parameter values $\vec{\theta}_i$, and there is an underlying distribution that describes all $\vec{\theta}_i$ (see Figure 1(b)). We compare the hierarchical model, denoted as

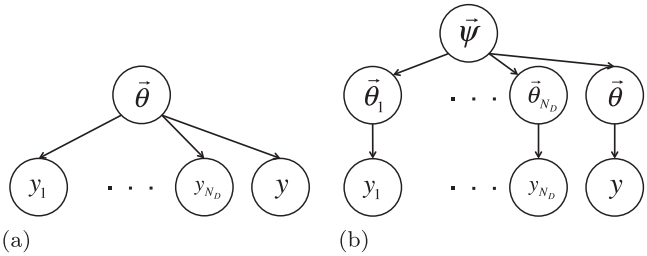

FIG. 1. Bayesian networks for two different models on heterogeneous data sets.

$\mathcal{M}_{HB}$, with a non-hierarchical model $\mathcal{M}_{NB}$ (Figure 1(a)) for water.

A. Bayesian inference for a non-hierarchical model

We integrate the N_D data sets with MD simulations with input x (e.g., temperature, pressure, and initial conditions of the atoms) and force-field parameters $\vec{\theta}$ so that their output y_i (for multiple predictions) can be expressed in terms of a function $f_i(x, \vec{\theta})$ and ϵ_{y_i} is the additive error term,

$$y_i = f_i(x, \vec{\theta}) + \epsilon_{y_i} \quad i = 1, \ldots, N_D. \tag{3}$$

The error term ϵ_{y_i} is typically chosen to be a Gaussian distribution $N_{\epsilon_{y_i}}(0, \sigma_{y_i})$ with zero mean and σ_{y_i} standard deviation. We denote the corresponding model for D_i as M_i. Here, σ_{y_i} becomes an extra parameter to be inferred from the data under the Bayesian framework. We group all σ_{y_i} into a single vector $\vec{\sigma}_y = \{\sigma_{y_i} | i = 1, \ldots, N_D\}$. Based on Bayes' theorem,

$$p(\vec{\theta}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{NB}) = \frac{p(\mathcal{D}|\vec{\theta}, \vec{\sigma}_y, \mathcal{M}_{NB})p(\vec{\theta}, \vec{\sigma}_y|\mathcal{M}_{NB})}{p(\mathcal{D}|\mathcal{M}_{NB})}. \tag{4}$$

We can use a Markov Chain Monte Carlo (MCMC) method to sample from the posterior distribution $p(\vec{\theta}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{NB})$.

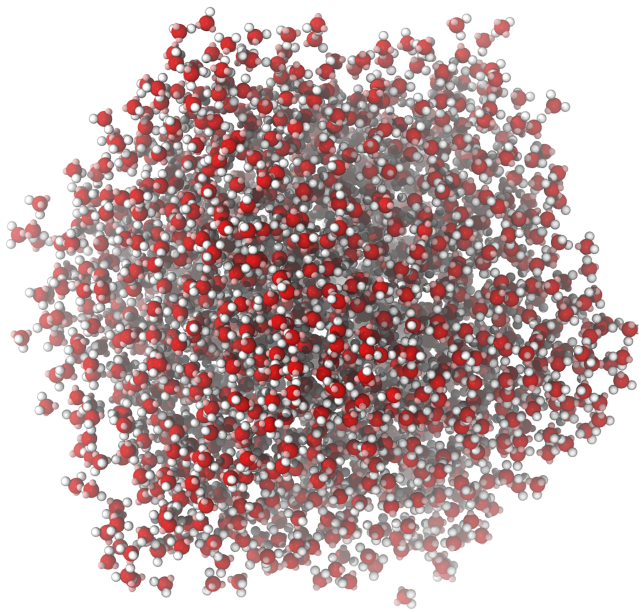

FIG. 2. TIP5P-E water system representation used in the simulations. Oxygen is depicted red, hydrogen white, and the two fictitious charge points pink.

TABLE I. Upper and lower bounds of all parameters.

Parameter	Min.	Max.	Units
ϵ_{LJ}	0.4	1.2	kJ/mol
q	0.20	0.32	e
$\sigma_y^{(1)}$	0.007 5	0.75	10^{-9} m^2/s
$\sigma_y^{(2)}$	2.5	250.0	kg/m^3
$\sigma_y^{(3)}$	0.004	0.40	
μ_ϵ	0.6	1.0	kJ/mol
σ_ϵ	0.000 8	0.24	kJ/mol
μ_q	0.22	0.30	e
σ_q	0.000 25	0.0625	e
ρ	−0.99	0.99	

To obtain an estimate for the marginalized posterior distributions, e.g., $p(\vec{\theta}|\mathcal{D}, \mathcal{M}_{NB})$ or $p(\vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{NB})$, we take the corresponding components of the samples and neglect the other. As mentioned before, this model implies that the likelihood function

$$p(\mathcal{D}|\vec{\theta}, \vec{\sigma}_y, \mathcal{M}_{NB}) = \prod_{i=1}^{N_D} p(D_i|\vec{\theta}, \sigma_{y_i}, M_i) \tag{5}$$

and the evidence term will be used to quantify the plausibility of the model,

$$p(\mathcal{D}|\mathcal{M}_{NB}) = \int p(\mathcal{D}|\vec{\theta}, \vec{\sigma}_y, \mathcal{M}_{NB}) p(\vec{\theta}, \vec{\sigma}_y|\mathcal{M}_{NB})\, \mathrm{d}\vec{\theta}\, \mathrm{d}\vec{\sigma}_y. \tag{6}$$

B. Bayesian inference for a hierarchical model

The hierarchical approach requires that we propose a prior distribution parameterized by $\vec{\psi}$. We choose a Gaussian prior of $\vec{\theta}$, thus $\vec{\psi}$ includes the parameters that define the mean vector and covariance matrix. Then, the posterior distribution is calculated by

$$p(\vec{\theta}|\mathcal{D}, \mathcal{M}_{HB}) = \int p(\vec{\theta}|\vec{\psi}, \mathcal{M}_{HB}) p(\vec{\psi}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{HB})\, \mathrm{d}\vec{\psi}\, \mathrm{d}\vec{\sigma}_y, \tag{7}$$

where $p(\vec{\theta}|\vec{\psi}, \mathcal{M}_{HB})$ is the chosen Gaussian prior of $\vec{\theta}$ and $p(\vec{\psi}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{HB})$ is the posterior distribution of $\vec{\psi}$ and $\vec{\sigma}_y$, which can be calculated using Bayes' theorem,

$$p(\vec{\psi}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{HB}) = \frac{p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB}) p(\vec{\psi}, \vec{\sigma}_y|\mathcal{M}_{HB})}{p(\mathcal{D}|\mathcal{M}_{HB})}. \tag{8}$$

Given $\vec{\psi}$ and the independence of predictions allows us to express the likelihood function $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$ by

$$p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB}) = \prod_{i=1}^{N_D} p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB}). \tag{9}$$

To understand the meaning of $p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB})$, we first look at the posterior distribution of $\vec{\theta}_i$ in this model,

$$p(\vec{\theta}_i|D_i, \vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB}) = \frac{p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB}) p(\vec{\theta}_i|\vec{\psi}, \mathcal{M}_{HB})}{p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB})}, \tag{10}$$

where $p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB})$ is the same likelihood function as $p(D_i|\vec{\theta}, \sigma_{y_i}, M_i)$. We observe that $p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB})$ is the evidence term in this Bayesian inference problem for given $\vec{\psi}$ and σ_{y_i}. Beck[16] proves that the evidence term in Bayesian inference quantifies the trade-off between data-fitting and information gain from the data. This suggests that higher values of $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$ correspond to a prior distribution of $\vec{\theta}$ that overlaps more with the likelihood functions of all the data sets. If a likelihood function is peaked, which implies a low uncertainty of the data set, the prior distribution is highly constrained by this likelihood. This is because a prior that does not cover the peaked likelihood will lose a significant contribution to the final evidence $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$. Hence, our approach automatically assigns a higher weight to the more "confident" data. Meanwhile, this confidence toward a data set depends on the estimated data noise as well as the sensitivity of the data to the model parameters.

A major challenge of using this model is the high computational demand required to calculate $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$, which

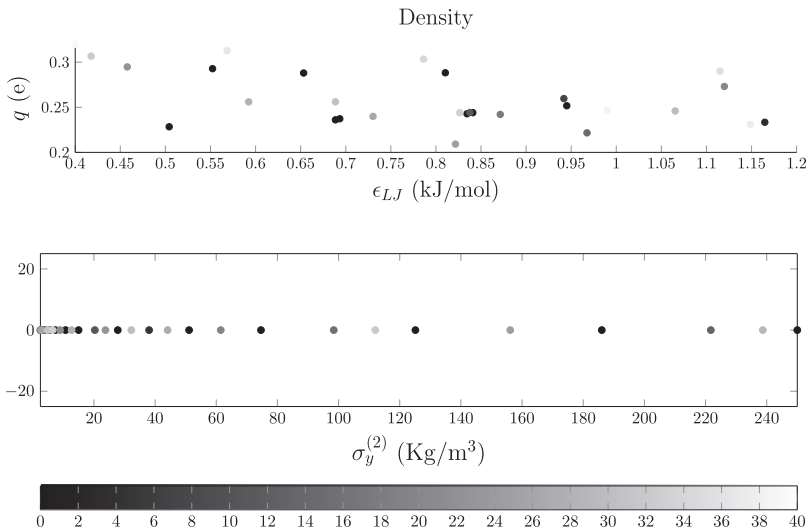

FIG. 3. Selected 30 values of ϵ_{LJ}, q, and $\sigma_{y,1}$ for EIM surrogate model of density likelihood (see Figures 4 and 5 for diffusion and RDF). The color scale indicates the sequence of selection from earliest to latest (black to white). For example, the first basis function chosen for density takes the values of $\epsilon = 0.84$, $q = 0.24, \sigma_y = 7.36$, and the last (30th) basis function chosen for density takes the values of $\epsilon = 0.40, q = 0.32, \sigma_y = 6.47$.

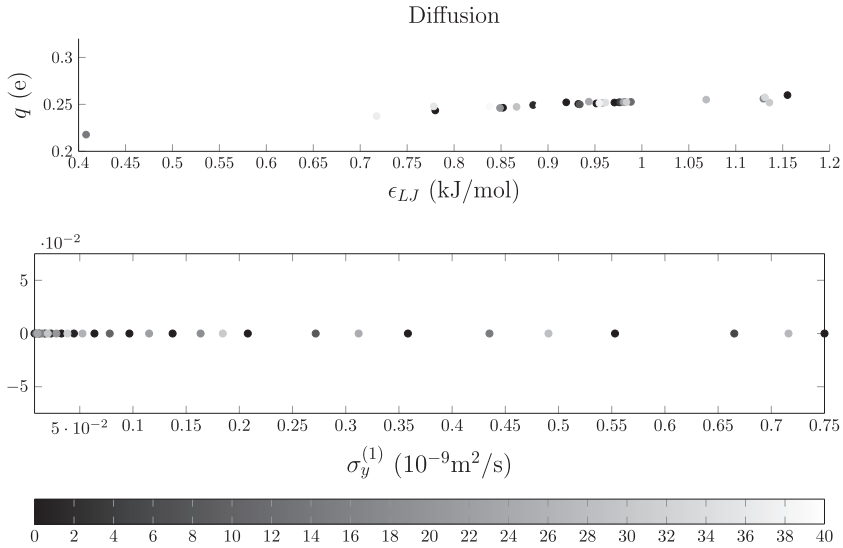

FIG. 4. Selected 30 values of ϵ_{LJ}, q, and $\sigma_{y,1}$ for EIM surrogate model of diffusion likelihood. The color scale indicates the sequence of selection from earliest to latest (black to white).

involvesmultiple integral evaluations that typically have no analytical solutions. In Sec. III, we propose a novel method to approximate $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$ based on EIM that is combined with the TMCMC method.

III. SURROGATES FOR A HIERARCHICAL BAYESIAN FRAMEWORK

The evidence for all models M_i is needed in order to construct the likelihood function $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$. As a result, the inference process involves a nested Monte Carlo evaluation: (i) sampling $\vec{\psi}$ and $\vec{\sigma}_y$ and (ii) sampling $\vec{\theta}_i$ for each $(\vec{\psi}, \vec{\sigma}_y)$ sample. This can become computationally intractable due to the quadratic growth of the number of samples needed. To remedy the computational cost, we develop an efficient approximation method by combining TMCMC with a novel, two-level surrogate model.

We use the parallelized TMCMC algorithm to draw samples from the posterior distribution of $\vec{\psi}$ and $\vec{\sigma}_y$, $p(\vec{\psi}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{HB})$, in order to perform uncertainty quantification of the model parameters and predictions. We note that when using full MD simulations, one evaluation of the evidence term $p(D_i|\vec{\psi}, \vec{\sigma}_y, \mathcal{M}_{HB})$ is very computationally intensive even with this efficient TMCMC method.[17] Building upon the EIM idea, we construct the two-level surrogate model to achieve orders of magnitude reduction in the total computational demand.

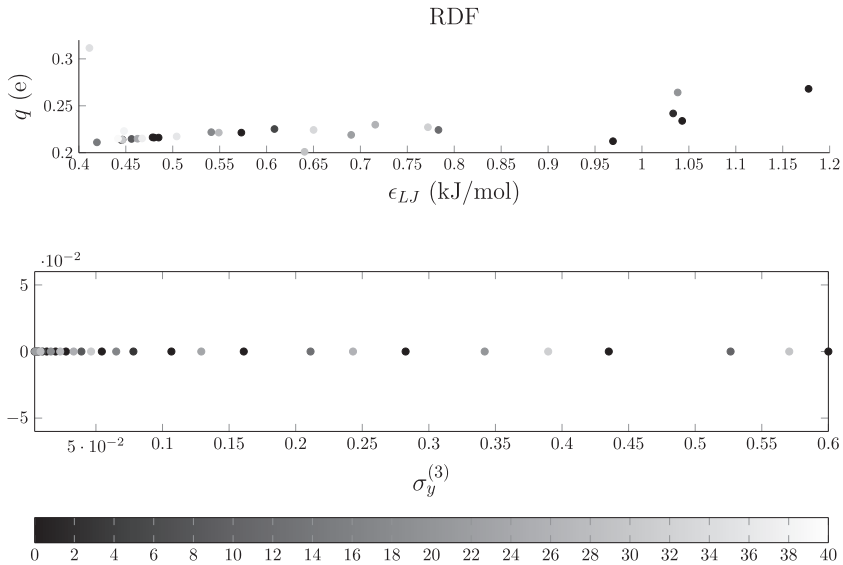

FIG. 5. Selected 30 values of ϵ_{LJ}, q, and $\sigma_{y,1}$ for EIM surrogate model of RDF likelihood. The color scale indicates the sequence of selection from earliest to latest (black to white).

A. TMCMC on a two-level surrogate

We estimate Equation (8) using the parallelized TMCMC with the likelihood function (Equation (9)). The evidence terms $p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB})$ for $i = 1, \ldots, N_D$ are calculated by

$$p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB}) = \int p(D_i|\vec{\theta}_i, \sigma_{y_i}, M_{HB}) p(\vec{\theta}_i|\vec{\psi}, \mathcal{M}_{HB})\, \mathrm{d}\vec{\theta}_i. \tag{11}$$

We note that $p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB})$ is the same likelihood function as $p(D_i|\vec{\theta}, \sigma_i, M_i)$, and $p(\vec{\theta}|\vec{\psi}, \mathcal{M}_{HB})$ is the chosen Gaussian prior for $\vec{\theta}$ parameterized by $\vec{\psi}$.

If the deterministic functions f_i is a linear function of $\vec{\theta}$, the integral of the Gaussian likelihood model with a Gaussian prior has an analytical solution. Such linear dependencies are not common, so an alternative is to have a surrogate model for the likelihood function based on Gaussian radial basis functions. However, this is also impractical because the likelihood is a function of σ_{y_i}, which is part of the sampling space. As a result, we will need a large amount of surrogate models to cover all possible σ_{y_i} values. Wepropose to use EIM as the first level surrogate to decouple the parameters $\vec{\theta}$ and σ_{y_i} using L basis functions $q_{l,I}$,

$$p(D_i|\vec{\theta}_i, \sigma_{y_i}, M_{HB}) \approx \sum_{l=1}^{L} \alpha_{l,i}(\sigma_{y_i}) q_{l,i}(\vec{\theta}_i), \tag{12}$$

where $\alpha_{l,i}$ are the weights computed depending on σ_{y_i} as explained in Appendix A. Then, as a second level surrogate, we approximate each basis function $q_{l,i}$ with a linear combination of Gaussian radial basis functions,

$$q_{l,i}(\vec{\theta}_i) \approx \sum_{j=1}^{N_{l,i}} w_{l,i}^{(j)} N_{\vec{\theta}_i}(\mu_{l,i}^{(j)}, \Sigma_{l,i}^{(j)}), \tag{13}$$

where $w_{l,i}^{(j)}$ is the weight for the Gaussian basis $N_{\vec{\theta}_i}(\mu_{l,i}^{(j)}, \Sigma_{l,i}^{(j)})$ with mean $\mu_{l,i}^{(j)}$ and covariance matrix $\Sigma_{l,i}^{(j)}$.

As a result, we obtain an analytical approximation to the desired integral in Equation (11) using the property of Gaussian integrals,

$$\begin{aligned} p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB}) &\approx \sum_{l=1}^{L} \alpha_{l,i}(\sigma_{y_i}) \sum_{j=1}^{N_{l,i}} w_{l,i}^{(j)} \int N_{\vec{\theta}_i}(\mu_{l,i}^{(j)}, \Sigma_{l,i}^{(j)}) N_{\vec{\theta}_i}(\mu_\theta(\vec{\psi}), \Sigma_\theta(\vec{\psi}))\, \mathrm{d}\vec{\theta} \\ &= \sum_{l=1}^{L} \alpha_{l,i}(\sigma_{y_i}) \sum_{j=1}^{N_{l,i}} w_{l,i}^{(j)} N_{\mu_{l,i}^{(j)}}\left(\mu_\theta(\vec{\psi}), \Sigma_{l,i}^{(j)} + \Sigma_\theta(\vec{\psi})\right), \end{aligned} \tag{14}$$

where μ_θ and Σ_θ are the mean and covariance matrix of the Gaussian prior of $\vec{\theta}$ determined by $\vec{\psi}$. Based on the EIM, $w_{l,i}^{(j)}$, $\mu_{l,i}^{(j)}$, and $\Sigma_{l,i}^{(j)}$ are predetermined and stored, whereas $\alpha_{l,i}$ are calculated "online," i.e., during the TMCMC sampling. The Gaussian radial basis approximation is performed by empirically fixing a grid of $\mu_{l,i}^{(j)}$ and determining a diagonal covariance matrix $\Sigma_{l,i}^{(j)}$. Then, we find the weights $w_{l,i}^{(j)}$ using the least-square method. We note that, as an alternative for building surrogate models from Gaussian radial basis functions, one can use any standard method, such as the Relevance Vector Machine that can induce sparsity robustly.[18] For building the bases $q_{l,i}$ in the EIM, we use a greedy algorithm that is explained in Appendix A.

IV. HIERARCHICAL MD SIMULATIONS OF WATER

A. Data and models

We use a 5-site water model — TIP5P[19] — and we resolve long range electrostatics using the particle-mesh-ewald method. The calibration data consist of three distinct subsets:

- D_1: self diffusion coefficient of water at 4 thermodynamic states ($P = 1$ bar, $T = 288, 300, 330, 350$ K),[20]
- D_2: density of water at the same 4 thermodynamic states as in D_1[21]
- D_3: the values of the first 3 local maxima and 1 minimum of the RDF of oxygen-oxygen in bulk water, at a single thermodynamic state $(P, T) = (1$ bar, 300 K$)$.[22]

Each evaluation of a sample requires a full MD-simulation run that was performed using the MD-code GROMACS.[23] We performed our simulations on the compute nodes of the Piz Daint Cray XC30 cluster at the Swiss National Supercomputing Center CSCS in Lugano. We use version 5.0.2 of

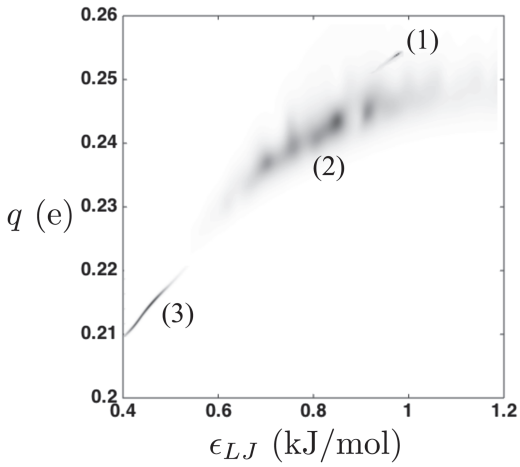

FIG. 6. Posterior distribution of model parameters given individually the three data sets: (1) diffusion coefficient, (2) density, and (3) RDF.

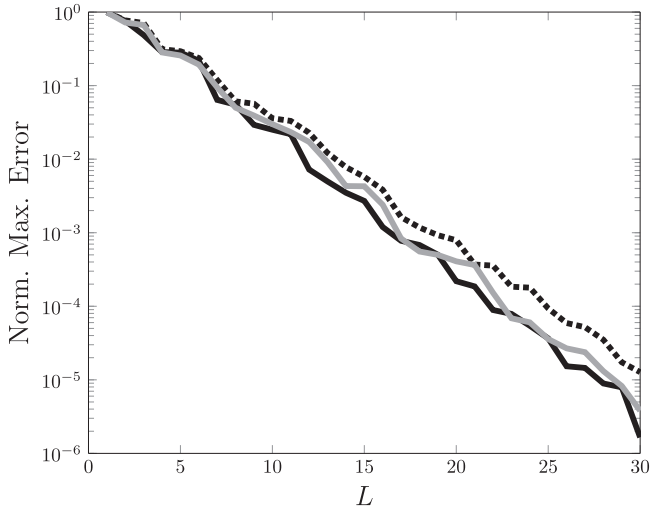

FIG. 7. Relative maximum error value of EIM surrogate model with L basis functions normalized by the maximum value in the training data set. Black solid line denotes diffusion (D_1), gray solid line denotes density (D_2), and black dashed line denotes RDF ($D3$).

GROMACS, compiled with GPU support. The system contains 1603 TIP5P water atoms (see Figure 2 for a snapshot). Normal periodic boundary conditions are used throughout. The computational workflow includes equilibration using the steepest descent of the system for at least 20 000 steps, subsequent equilibration for a further 2 ns using a time step of $\Delta t = 0.5$ fs based on the Berendsen barostat and thermostat in the NPT ensemble, and finally a 5 ns production run in the NPT ensemble for each run using a time step of $\Delta t = 1.0$ fs and Parrinello-Rahman pressure coupling. A single evaluation on the parameter space took approximately 1 h on one node, for a total computational cost of the calibration campaign of 3500 Piz Daint node hours.

Then, we extract the predicted oxygen-oxygen RDF ($g(r)$) averaged over the last 2.5 ns of the production part of the trajectory sampled every 100 fs. We detect the first 3 local maxima and the first local minimum of the $g(r)$ function. Similarly, we extract the ensemble average of the self diffusion coefficient using the Einstein formulation, as well as the density.

We choose to vary the Lennard–Jones parameter pertaining to the oxygen-oxygen interaction ϵ_{LJ}, as well as the partial charge of the TIP5P water model q. We keep σ_{LJ} fixed at its nominal value. The nominal values provided by Rick[19] for TIP5P-E are $(\hat{\epsilon}_{LJ}, \hat{\sigma}_{LJ}, \hat{q}) = (0.669\ \text{kJ/mol}, 0.334\ \text{nm}, 0.241\ q)$.

We assume that the predictions for the data points within a data set D_i are independent of each other. The likelihood models for each data point in data set D_i are defined as the Gaussian distribution with mean f_i, which is the MD simulation result corresponding to the data point, and standard deviation σ_{y_i}, for $i = 1, \ldots, 3$. We note that for the likelihood function of D_3, the standard deviation $\sigma_y^{(3)}$ is normalized by the nominal value of each local optimum. The nominal values are determined based on the experimental data. All priors are chosen to be uniformly distributed, except the specific Gaussian prior of $\vec{\theta}_i$ in $\mathcal{M}_{HB}$. This Gaussian prior is parameterized by $\vec{\psi} = \{\mu_\epsilon, \sigma_\epsilon, \mu_q, \sigma_q, \rho\}$, where μ_ϵ and σ_ϵ are the mean and standard deviation of ϵ_{LJ}, μ_q and σ_q are the mean and standard deviation of q, and ρ is the coefficient of correlation between ϵ_{LJ} and q. Moreover, the error reported for the experimental data is very small when compared to the model uncertainties and we have omitted it in our study. One may add an experimental error term during the likelihood construction to capture this quantity, but the results are expected to be qualitatively the same.

B. Algorithm setup and efficiency check

Based on our empirical study, we selected the region of interest for all parameters: $\vec{\theta} = \{\epsilon_{LJ}, q\}, \vec{\sigma}_y = \{\sigma_y^{(1)}, \sigma_y^{(2)}, \sigma_y^{(3)}\}$,

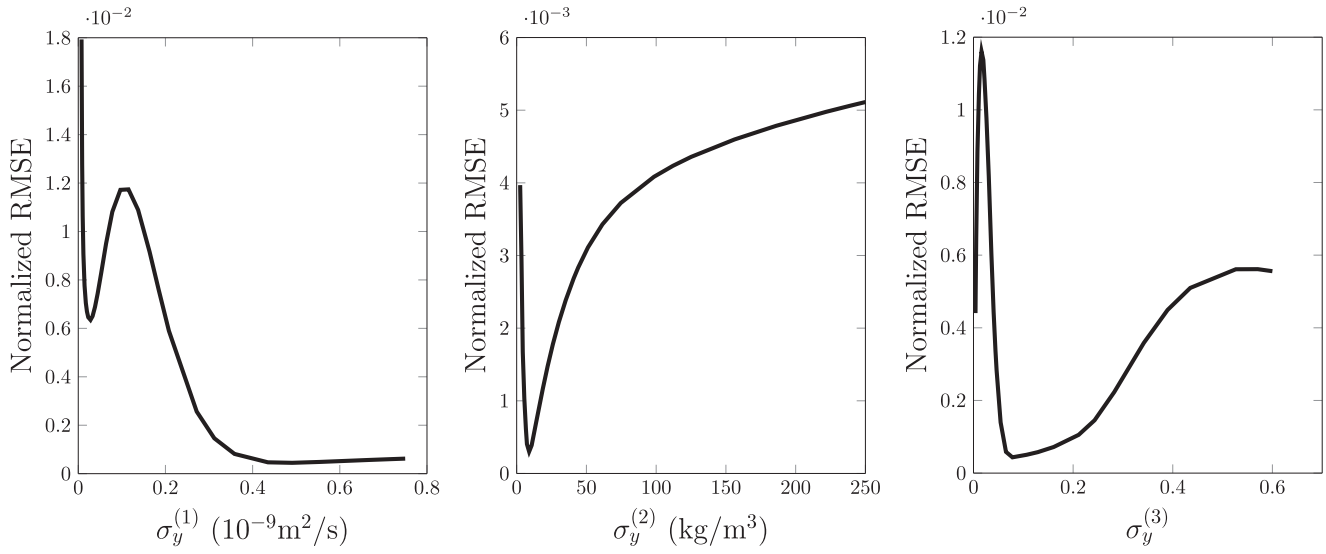

FIG. 8. Relative root mean square error of Gaussian radial basis function surrogate model (normalized by the maximum value in the training data set) as a function of model uncertainty parameters σ_{y_i} (left to right): diffusion, density, and RDF. The same training data set used for the EIM surrogate is used here.

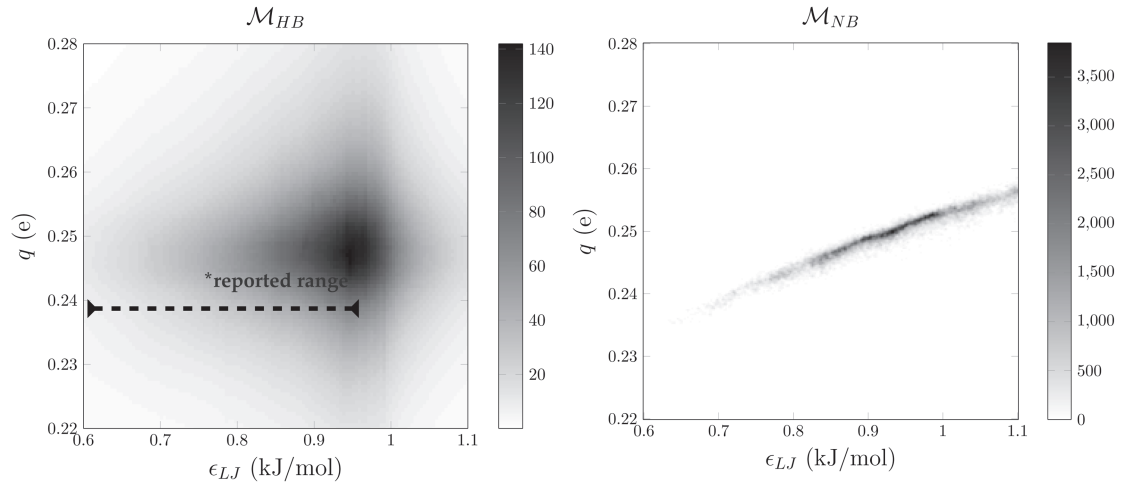

FIG. 9. Posterior distribution of model parameters based on hierarchical Bayesian framework $\mathcal{M}_{HB}$ and normal Bayesian framework $\mathcal{M}_{NB}$. The dashed line shows the range of ε_{LJ} from different calibration studies reported in the work of Chaplin.[25]

and $\vec{\psi} = \{\mu_\epsilon, \sigma_\epsilon, \mu_q, \sigma_q, \rho\}$. The boundaries are summarized in Table I.

In our study, we perform Bayesian inference five times with all priors chosen to be uniformly distributed with the boundaries listed in Table I:

- **1st to 3rd times:** posterior distribution of ϵ_{LJ}, q, and σ_{y_i} is calculated based on individual data set D_i, for $i = 1, \ldots, 3$. The Bayesian inference follows the section on non-hierarchical Bayesian model, but using only one data set for the likelihood $p(\mathcal{D}|\vec{\theta}, \vec{\sigma}_y, \mathcal{M}_{NB})$ at a time.
- **4th time:** posterior distribution of ϵ_{LJ}, q, and $\vec{\sigma}_y$ is calculated based on $\mathcal{M}_{NB}$. The Bayesian inference follows the section on non-hierarchical Bayesian model.
- **5th time:** posterior distribution of ϵ_{LJ}, q, $\vec{\sigma}_y$, and $\vec{\psi}$ is calculated based on $\mathcal{M}_{HB}$. The Bayesian inference follows the section on hierarchical Bayesian model.

Results of the Bayesian inference for individual data sets are used as a reference when comparing between $\mathcal{M}_{NB}$ and $\mathcal{M}_{HB}$.

The parallelized TMCMC algorithm[17] is used for both $\mathcal{M}_{NB}$ and $\mathcal{M}_{HB}$. TMCMC generates posterior samples from prior distribution based on multiple stages of MCMC sampling that follow an annealing schedule for the likelihood function.[14] The number of samples per stage, which is also the final number of posterior samples, used in our study is 20 000 for $\mathcal{M}_{NB}$ and 50 000 for $\mathcal{M}_{HB}$. The total number of stages depends on a parameter τ_{CV} that controls the speed of converging to the actual likelihood function. We use the suggested value of $\tau_{CV} = 100\%$ from the work of Ching and Chen.[14] Another important parameter, β_{COV} controls the step size of the random walk in each MCMC chain. We use $\beta_{COV} = 0.4$, which is slightly higher than the suggested value in the work of Ching and Chen,[14] as we need the random walk samples to explore a larger area in the space. As a result, a total of 8 and 7 stages are used for $\mathcal{M}_{NB}$ and $\mathcal{M}_{HB}$, respectively.

For the EIM greedy algorithm, we initialize $\sigma_{y_i}^{(1)} = \text{argmax}_{\sigma_{y_i}^{(1)}} \left\{\max_{\vec{\theta}_i} p(D_i|\vec{\theta}_i, \sigma_{y_i}^{(1)})\right\}$ for $i = 1, \ldots, 3$ to improve the performance of the algorithm. Because each MD simulation requires a long computation time, it is impractical to train the EIM surrogate model by performing the optimization on $\vec{\theta}_i$ directly. Instead, we use a 500 × 500

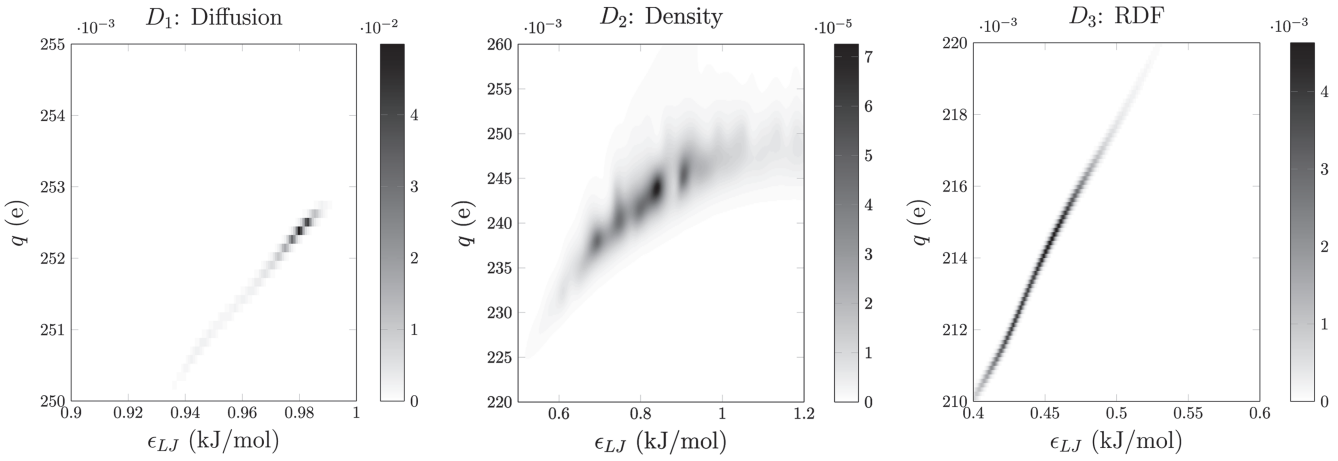

FIG. 10. Zoom-in of the posterior distribution of model parameters given individually the three data sets: diffusion coefficient, density, and RDF (left to right).

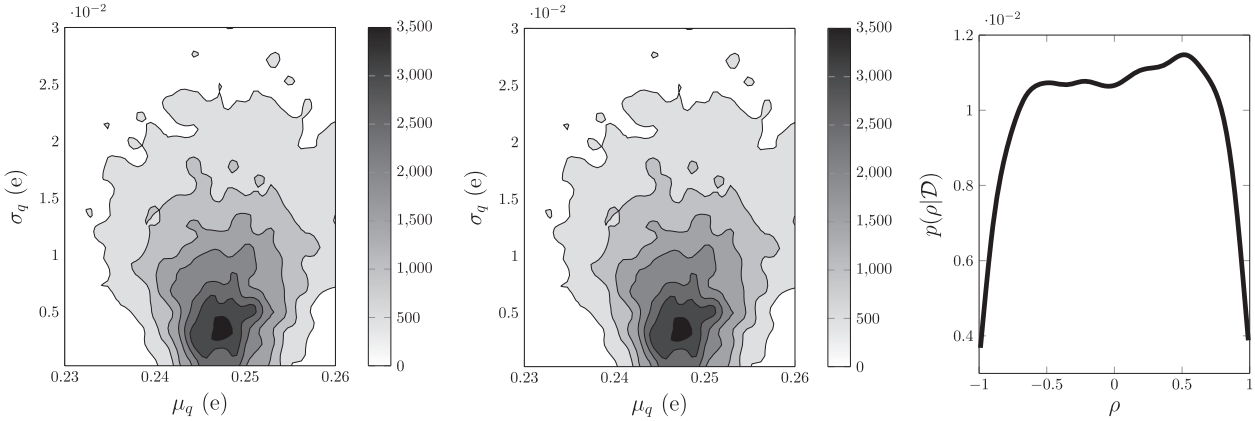

FIG. 11. Posterior distribution of $\vec{\psi}$ (left to right): $p(\mu_\epsilon, \sigma_\epsilon|\mathcal{D}, \mathcal{M}_{HB})$, $p(\mu_q, \sigma_q|\mathcal{D}, \mathcal{M}_{HB})$, and $p(\rho|\mathcal{D}, \mathcal{M}_{HB})$.

grid of interpolation points evaluated based on 780 MD simulations for four different temperatures as a training set to perform the optimization. A total of 30 basis functions (L = 30) are used to reach at most 10^{-5} error (see Sec. V A).

Finally, for the Gaussian radial basis function surrogate model, basis functions are centered at each point in the training set with standard deviation values equal to the square root (because of 2-D grid) of the grid's average step size. The weights of each basis function are calculated by solving the linear system of equations for all grid points. Weights that are smaller than $0.000\,01 \times \left(\max_j(w_{l,i}^{(j)})\right)$ are removed from the surrogate model.

During the training stage of the two level surrogate model, a total of 3120 MD simulations were performed. After the EIM training process, which took around 1 h on one CPU, the efficient approximation took less than 1 h on a 12-core CPU, which is around one MD simulation in our example. A hierarchical Bayesian analysis based on a nested TMCMC approach and without any surrogate models will require a total number of MD simulations in the order of 10^{10}. Our approach makes the originally intractable problem feasible.

V. RESULTS AND DISCUSSIONS

A. Performance of surrogate models

The two-level surrogate model approach requires careful monitoring of the surrogate modeling error. In Figures 3–5, it is shown that the selected 30 values of ϵ_{LJ}, q, and σ_{y_i} serve as the basis functions for the EIM surrogate models. The locations of the chosen values correspond well to the high likelihood value region of each data set (Figure 6). The normalized maximum error is controlled to be at most 0.001% for all three data sets (Figure 7).

On the other hand, the normalized root mean square error is controlled to be around 1% for the Gaussian radial basis function surrogate models (Figure 8). The region of high error mainly comes from the extremely low values of $\sigma_{y,1}$ and $\sigma_{y,3}$. Overall, the error is small enough to give us the confidence in our analyses (see Appendix B for further verification of the results).

B. Posterior distribution of parameters and models

In Figure 9, it is shown that $\mathcal{M}_{HB}$ and $\mathcal{M}_{NB}$ both have a similar optimal model parameter values. However, the

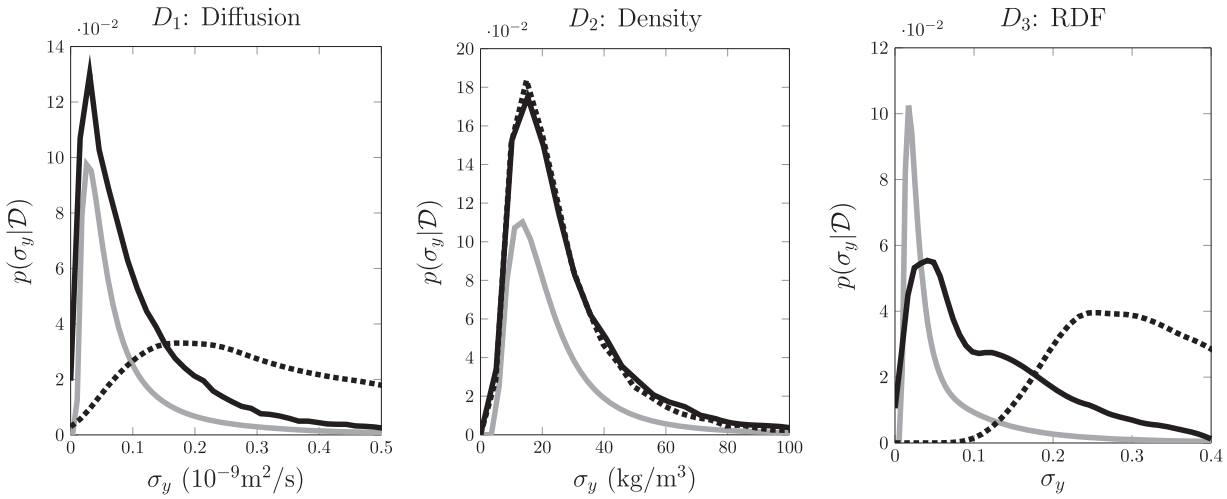

FIG. 12. Posterior distribution of model uncertainties $\vec{\sigma}_y$ of the three data sets (left to right): diffusion, density, and RDF. Gray solid lines are individual data set results, black solid lines are hierarchical Bayesian $\mathcal{M}_{HB}$ results, and black dashed lines are normal Bayesian $\mathcal{M}_{NB}$ results.

uncertainties associated with the parameters are significantly different. To understand the differences, we investigate the posterior distribution of $\vec{\theta}$ for each individual data set (Figure 6 and a zoom-in version in Figure 10). We note that the posterior probability values are directly proportional to the likelihood function values of the data sets because the priors are uniformly distributed. We conclude from the likelihood values that the diffusion data will dominate the likelihood function in $\mathcal{M}_{NB}$. Although the RDF data also have relatively high likelihood values, the region of high values is too far away from the peak value region of the diffusion data so that its influence becomes minimal. Indeed, we observe the dominance of the diffusion data in distribution shape and peak values of the model parameters in $\mathcal{M}_{NB}$. This is not a surprising result because the MD model and simulation protocol model used can capture the diffusion data at the investigated temperature regimes more accurately than the density or the RDF. This occurs as the statistical thermodynamic assumptions involved in the MD model (e.g., pairwise interactions between the atoms) affect less the diffusion calculation than the pair correlation functions.[24]

On the other hand, $\mathcal{M}_{HB}$ integrates the heterogeneous data based on the evidence values of each data set. Although the peak values of the model parameters are still controlled by the diffusion data, there is a larger uncertainty associated with the parameters.

In order to understand the difference in distribution shape between $\mathcal{M}_{HB}$ and $\mathcal{M}_{NB}$, we investigate the posterior distribution of $\vec{\psi}$ in $\mathcal{M}_{HB}$. From Figure 11, one can observe that the smaller mean values of ϵ_{LJ} tend to have a higher uncertainty. This contributes to the asymmetric posterior distribution of $\vec{\theta}$ along the ϵ_{LJ} axis. Furthermore, the posterior distribution of ρ is approximately uniform. This implies that the given heterogeneous data are actually not sufficient to confirm any correlation between ϵ_{LJ} and q. This is not intuitive based on the posterior distribution shown in Figure 6. Further investigation may be needed to confirm this conclusion.

One important strength of the hierarchical Bayesian framework is the ability to accurately quantify the model uncertainty σ_{y_i} for each data set D_i. Figure 12 shows the posterior distribution of model uncertainties of the three data sets obtained from different Bayesian analysis approaches. One can observe that the results from $\mathcal{M}_{HB}$ coincide well with the actual model uncertainty values (results of the Bayesian analysis on each of the individual data sets). However, results from $\mathcal{M}_{NB}$ tend to overestimate the model uncertainties. This is because $\mathcal{M}_{HB}$ separates the uncertainty coming from the heterogeneous data from the actual model uncertainties. On the other hand, $\mathcal{M}_{NB}$ integrates all uncertainties to the model uncertainties $\vec{\sigma}_y$.

A major benefit of our approach is that the final evidence of $\mathcal{M}_{HB}$ is estimated as a by-product of the parallelized TMCMC algorithm. We can use this evidence value to perform model selection with other approaches, e.g., $\mathcal{M}_{NB}$ in our example. We evaluate the posterior probability of a model $\mathcal{M}$ using Bayes' theorem,

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})}. \tag{15}$$

TABLE II. Bayesian model selection between hierarchical and non-hierarchical models in the Bayesian framework.

Model	ln(Evidence)	$P(\text{Model}\|\mathcal{D})$
$\mathcal{M}_{HB}$	−20.44	0.98
$\mathcal{M}_{NB}$	−24.51	0.02

Typically, a uniform prior for $p(\mathcal{M})$ is chosen to avoid bias on any model before observing any data. As a result, $p(\mathcal{M}|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M})$, which is the final evidence of $\mathcal{M}$. Table II shows that $\mathcal{M}_{HB}$ is a more probable approach than $\mathcal{M}_{NB}$ for the given heterogeneous data $\mathcal{D}$ in this example.

This method is capable of the calibration of MD simulations for other materials using heterogeneous experimental data. In general, application of this method to other materials does not require any changes. However, if the number of model parameters increases, the dimension of $\vec{\psi}$ will be higher as well. Extra care is needed when developing the Gaussian radial basis function surrogate model in a high dimension space.

C. Robust posterior predictions

We perform a full Bayesian robust prediction to demonstrate the uncertainties propagated from the model parameters and the model uncertainty σ_y. The posterior prediction distribution $p(y|\mathcal{D}, \mathcal{M}_{HB})$ of a property of interest y can be estimated using the N_s posterior samples $\vec{\psi}^{(k)}$ and $\vec{\sigma}_y^{(k)}$ as follows:

$$\begin{aligned} p(y|\mathcal{D}, \mathcal{M}_{HB}) &= \int p(y|\vec{\psi}, \vec{\sigma}, \mathcal{M}_{HB}) p(\vec{\psi}, \vec{\sigma}|\mathcal{D}, \mathcal{M}_{HB})\, \mathrm{d}\vec{\psi}\mathrm{d}\vec{\sigma} \\ &\approx \frac{1}{N_s}\sum_{k=1}^{N_s} p(y|\vec{\psi}^{(k)}, \vec{\sigma}^{(k)}, \mathcal{M}_{HB}), \end{aligned} \tag{16}$$

where $p(y|\vec{\psi}^{(k)}, \vec{\sigma}^{(k)}, \mathcal{M}_{HB})$ is a special case of $p(D_i|\vec{\psi}^{(k)}, \vec{\sigma}^{(k)}, \mathcal{M}_{HB})$ with only one value in D_i. More specifically, for an output quantity of interest y (e.g., diffusion), the $p(y|\vec{\psi}^{(k)}, \vec{\sigma}^{(k)}, \mathcal{M}_{HB})$ quantifies the statistical error computed from an MD simulation. Herein this is selected to be a Gaussian distribution with mean and standard deviation computed from the MD simulations.

The results help us understand how the hierarchical model merges contribution from the heterogeneous data set from a new aspect. In Figure 13, we show that diffusion has the largest prediction uncertainty, while density is the second and RDF is the last. This observation illustrates why the hierarchical model shows a stronger influence of the diffusion data on the parameter calibration: a larger prediction uncertainty for a given parameter distribution implies a higher sensitivity of the property prediction from the parameter variation and thus attracts more attention during the calibration step as compared to other less sensitive properties. In this aspect, the Bayesian hierarchical model presented in this study results in calibration that naturally balances between the prediction sensitivities of the heterogeneous data set. This is another way to express the concept of data fusion using uncertainty information under the Bayesian framework. On the other hand, for the normal Bayesian model, the same phenomenon results

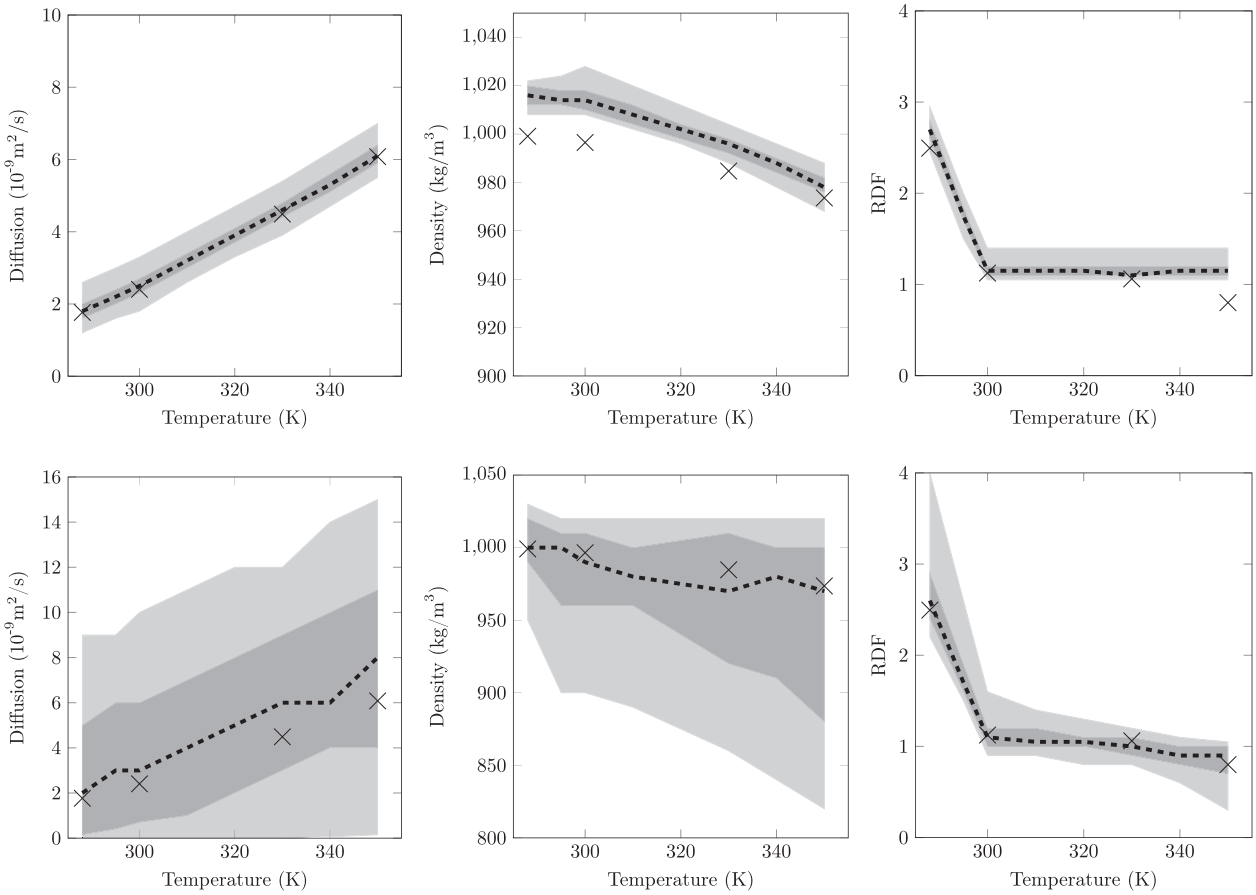

FIG. 13. Bayesian robust prediction through uncertainty propagation for diffusion, density, and RDF; the hierarchical model (left to right), for the normal Bayesian model (top) and the hierarchical Bayesian model (bottom). The dashed black line is the mean prediction with actual data indicated as cross marker. The two levels of gray regions indicate the 25%–75% and 5%–95% quantiles. The non-smooth curves are resulting from a coarse grid on the temperature during prediction simulations with full MD runs, which is limited by our computational expense available.

in a better prediction on the diffusion data with a misleading uncertainty quantification for the density and RDF predictions.

VI. CONCLUSIONS

We propose a new approach for calibrating force-field parameters in MD simulations fusing heterogeneous data through a hierarchical Bayesian framework. This approach automatically integrates the contribution of each data set without any additional parameter tuning. The fusion of the data sets is based on their evidence values, obtained through a classical Bayesian analysis, and provides a new perspective of how to combine heterogeneous data. We show analytically that our approach automatically includes the uncertainty information of each data set in the data fusion process, thus increasing its robustness.

We tackle the computational cost associated with the hierarchical Bayesian framework, by introducing an efficient approximation based on a two-level surrogate model and a parallelized TMCMC algorithm. The two-level surrogate comprises an EIM-based model and a Gaussian radial basis surrogate model. The present surrogate model allows for an order of magnitude reduction in the total computational demand compared to a hierarchical Bayesian model without surrogates.

To showcase our method, we calibrated two oxygen interaction parameters for MD simulations of water using

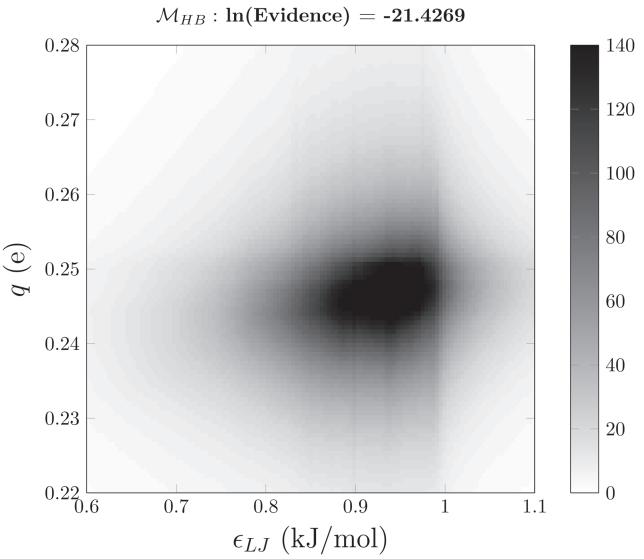

FIG. 14. Posterior distribution of model parameters for $\mathcal{M}_{HB}$ using a nested TMCMC approach. Note that the evidence value is within 5 of the EIM surrogate one, providing a very good agreement within the variability of TMCMC.

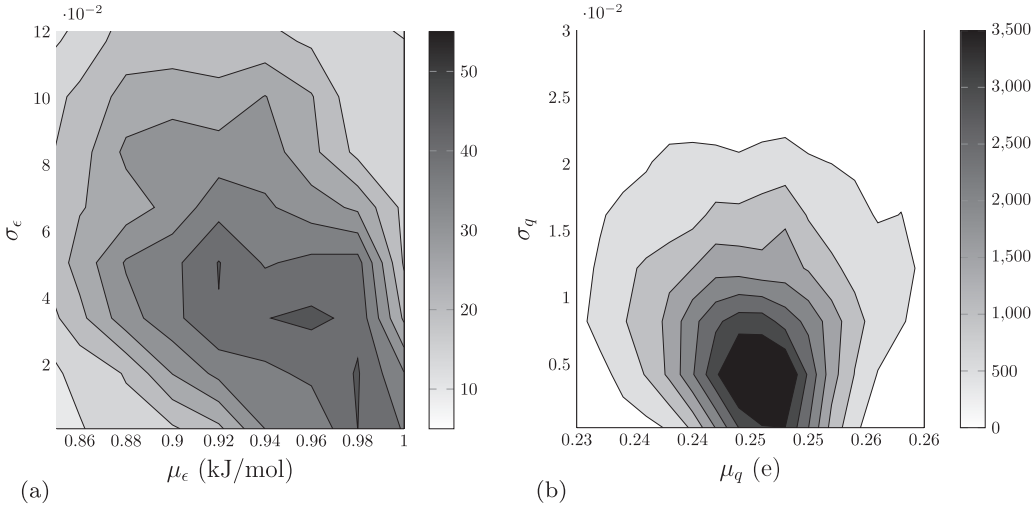

FIG. 15. Posterior distribution of $\vec{\psi}$ using a nested TMCMC approach.

experimental data of three types of significantly different material properties: diffusivity, density, and the RDF. The results demonstrate that our new approach captures the model uncertainties accurately. A single set of optimal model parameter values is found with high uncertainty due to the inconsistent calibration results from each data set, and our results motivate further investigation of the correlation between ϵ_{LJ} and q. Instead of promoting the use of a single set of optimal parameters and trying to develop methodology for finding the optimal values, our framework provides a quantitative evidence for whether a given model is reliable for predictions across a specific set of heterogeneous data.

In this work we show that by using a single water TIP5P force field, it is not possible to predict simultaneously three different liquid properties. This observation may serve as a warning for the transferability of commonly used force-field models. At the same time, the present hierarchical Bayesian framework is shown to provide robust predictions for all three liquid properties at the expense of higher uncertainty in the model parameters. In turn, this may provide a framework for improving the development of force fields.

We can use the uncertainties reflected from the posterior distribution of the parameters and the predictions to guide our understanding of the relationship between the model and the data set. Ongoing investigations aim to extend the present hierarchical Bayesian framework to MD simulations of other materials as well as to other models of physical processes for which heterogeneous data are available. We envision new capabilities for MD simulations through the data driven quantification of model uncertainties.

ACKNOWLEDGMENTS

We acknowledge support from ETH Zurich and computational time at the Swiss National Supercomputing Center CSCS under Project No. s448. P.K. and P.A. acknowledge support from the European Research Council (Advanced Investigator Award No. 341117).

APPENDIX A: EIM-BASED ALGORITHM

We want to find L values of $\vec{\theta}_i$ and σ_{y_i} to be the bases of the approximation in Equation (12). Then, we can solve the following system of linear equations for $\alpha_{l,i}$, where $l = 1, \ldots, L$, to obtain the approximation during "online" operations:

$$\begin{aligned} &p(D_i|\vec{\theta}_i^{(n)}, \sigma_{y_i}, \mathcal{M}_{HB}) = P_n^{(i)}(\sigma_{y_i}) \\ &= \sum_{l=1}^{L} \alpha_{l,i}(\sigma_{y_i}) q_{l,i}(\vec{\theta}_i^{(n)}), \ 1 \le n \le L \\ &\text{or in a matrix form: } [q_{nl}^{(i)}]\{\alpha_{l,i}\} = \{P_n^{(i)}\}, \end{aligned} \tag{A1}$$

where $\{P_n^{(i)}\}$ is the vector of the actual $p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB})$ evaluated at σ_{y_i} and the nth basis values of $\vec{\theta}_i$, denoted as $\vec{\theta}_i^{(n)}$; $[q_{nl}^{(i)}]$ is the matrix with components $q_{nl}^{(i)}$ that denotes the lth basis function $q_{l,i}$ evaluated at $\vec{\theta}_i^{(n)}$; $\{\alpha_{l,i}\}$ is the unknown vector of $\alpha_{l,i}$ to be calculated from solving the linear system of equations.

We use the greedy (Algorithm 1) to find L basis functions $q_{l,i}(\vec{\theta}_i)$. We choose $q_{l,i}$ to be the original likelihood function $p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB})$ with some given values of σ_{y_i} and $\vec{\theta}_i$ (denoted as $\sigma_{y_i}^{(l)}$ and $\vec{\theta}_i^{(l)}$ in Algorithm 1).

We note that L can be determined based on a predetermined threshold on the error function $e_{l,i}(\vec{\theta}_i; \sigma_{y_i}^{(l)})$, i.e., the FOR-loop in Algorithm 1 can be substituted with a WHILE-loop conditional on the error function exceeding the threshold. Also, the choice of $\sigma_{y_i}^{(1)}$ is random, but we found that the given choice improved the performance of the algorithm. We furthermore point out that the original EIM paper by Barrault *et al.*[12] suggests using normalized error functions as bases recursively built by

$$\begin{aligned} &\tilde{q}_{l,i}(\vec{\theta}_i) = \frac{e_{l-1,i}(\vec{\theta}_i; \sigma_{y_i}^{(l)})}{e_{l-1,i}(\vec{\theta}_i^{(l)}; \sigma_{y_i}^{(l)})}, \\ &\text{and } \tilde{q}_{1,i}(\vec{\theta}_i) = \frac{p(D_i|\vec{\theta}_i, \sigma_{y_i}^{(1)})}{p(D_i|\vec{\theta}_i^{(1)}, \sigma_{y_i}^{(1)})}. \end{aligned} \tag{A2}$$

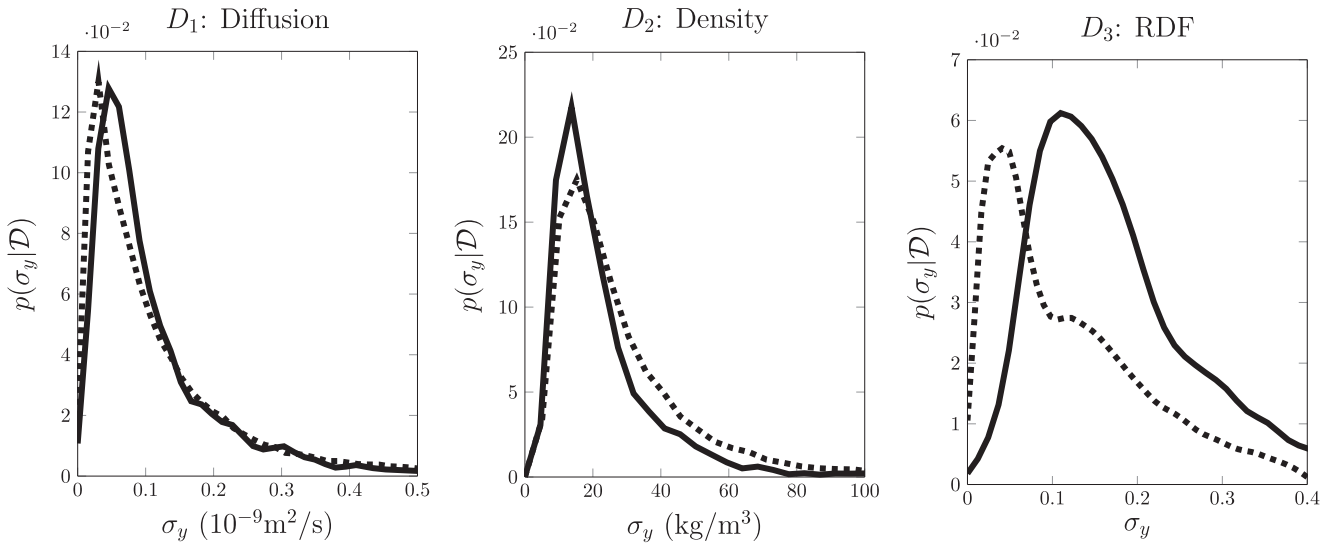

FIG. 16. Posterior distribution of model uncertainties $\vec{\sigma}_y$ of the three data sets (left to right): diffusion, density, and RDF. Solid lines are results from the two-level surrogate model and dashed lines are results from the nested TMCMC method.

This set of bases allows the computation of the solution of Equation (A1) to be more efficient during the online stage. However, the error functions are usually not smooth. Since we need a second level surrogate to approximate the bases, the original likelihood function is preferred.

APPENDIX B: VERIFICATION USING NESTED TMCMC

To further verify the results obtained from the two-level surrogate model, we perform a surrogate-free estimation on $\mathcal{M}_{HB}$ using a nested TMCMC approach. When a sample is drawn from the posterior distribution $p(\vec{\psi}, \vec{\sigma}_y|\mathcal{D}, \mathcal{M}_{HB})$ using TMCMC (denote as the outer TMCMC), we need to evaluate the likelihood $p(\mathcal{D}|\vec{\psi}, \vec{\sigma}, \mathcal{M}_{HB})$ by Equations (9) and (11). We use TMCMC again to draw samples from $p(\vec{\theta}_i|D_i, \vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB})$ (Equation (10)) and $p(D_i|\vec{\psi}, \sigma_{y_i}, \mathcal{M}_{HB})$ is estimated as a by-product of the algorithm (denoted as the inner TMCMC). We use the same TMCMC parameter values and prior settings as in Sec. IV B for both of the TMCMC usages here. Observing from Figure 11 that ρ is unidentifiable with the given data, we remove ρ from $\vec{\psi}$, i.e., $\vec{\psi} = \{\mu_\epsilon, \sigma_\epsilon, \mu_q, \sigma_q\}$. We use 500 samples per stage for the inner TMCMC because it is only in a 2D parameter space ($\vec{\theta}$), and we use 5000 samples per stage for the outer TMCMC, which is in a 7D parameter space ($\vec{\psi}$ and $\vec{\sigma}$). Figures 14–16 show the results using this nested TMCMC approach. Comparing with Figures 9, 11, and 12, we observe that the results from the two-level surrogate model approach agree with the results from the nested TMCMC approach very well. Hence, we are confident on the accuracy of the results from our surrogate model approach.

ALGORITHM 1. Greedy algorithm

Initialize: $\sigma_{y_i}^{(1)} = \mathrm{argmax}_{\sigma_{y_i}} \left\{ \max_{\vec{\theta}_i} p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB}) \right\}$
and $\vec{\theta}_i^{(1)} = \mathrm{argmax}_{\vec{\theta}_i} \left\{ p(D_i|\vec{\theta}_i, \sigma_{y_i}^{(1)}) \right\}$
for $l = 2$ **to** L do
$e_{l-1,i}(\vec{\theta}_i; \sigma_{y_i})$
$= p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB}) - \sum_{k=1}^{l-1} \alpha_{k,i}(\sigma_{y_i}) q_{k,i}(\vec{\theta}_i)$
$\sigma_{y_i}^{(l)} = \mathrm{argmax}_{\sigma_{y_i}} \left\{ \max_{\vec{\theta}_i} \left\{ e_{l-1,i}(\vec{\theta}_i; \sigma_{y_i}) \right\} \right\}$
$\vec{\theta}_i^{(l)} = \mathrm{argmax}_{\vec{\theta}_i} \left\{ e_{l-1,i}(\vec{\theta}_i; \sigma_{y_i}^{(l)}) \right\}$
$q_{l,i}(\vec{\theta}_i) = p(D_i|\vec{\theta}_i, \sigma_{y_i}^{(l)}, \mathcal{M}_{HB})$
end for

APPENDIX C: ERROR NORMALIZATION

In Figures 6 and 7, we normalized the maximum error by the maximum value of the estimated quantity found in the training set, in order to estimate the relative error for consistent comparison,

$$\text{Figure 4 error at stage } l = \frac{\max_{\vec{\theta}_i, \sigma_{y_i}} \left\{ e_{l,i}(\vec{\theta}_i; \sigma_{y_i}) \right\}}{\max_{\vec{\theta}_i, \sigma_{y_i}} p(D_i|\vec{\theta}_i, \sigma_{y_i}, \mathcal{M}_{HB})}, \tag{C1}$$

$$\text{Figure 5 error} = \frac{\text{RMSE of Gaussian radial basis function}}{\max_{\vec{\theta}_i, \sigma_{y_i}} p(D_i|\vec{\theta}_i, \sigma_{y_i})}, \tag{C2}$$

where $\vec{\theta}_i$ and σ_{y_i} belong to the pre-selected training set.

[1]E. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, "Discovery through the computational microscope," Structure **17**, 1295–1306 (2009).
[2]K. Palmo, B. Mannfors, N. G. Mirkin, and S. Krimm, "Potential energy functions: From consistent force fields to spectroscopically determined polarizable force fields," Biopolymers **68**, 383–394 (2003).
[3]J. H. Walther, R. Jaffe, T. Halicioglu, and P. Koumoutsakos, "Carbon nanotubes in water: Structural characteristics and energetics," J. Phys. Chem. B **105**, 9980–9987 (2001).
[4]T. D. Iordanov, G. Schenter, and B. Garrett, "Sensitivity analysis of thermodynamic properties of liquid water: A general approach to improve empirical potentials," J. Phys. Chem. A **110**, 762–771 (2005).
[5]W. L. Jorgensen, C. Jenson, and M. W. Mahoney, "Liquid water models: Beyond TIP4P and the density (t) problem," Abstr. Pap. Am. Chem. Soc. **218**, U314 (1999).
[6]F. Rizzi, H. Najm, B. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, and O. Knio, "Uncertainty quantification in MD simulations. Part II: Bayesian inference of force-field parameters," Multiscale Model. Simul. **10**, 1460–1492 (2012).

[7]C. M. Handley and R. J. Deeth, "A multi-objective approach to force field optimization: Structures and spin state energetics of d6 Fe(II) complexes,"J. Chem. Theory Comput. **8**, 194–202 (2012).
[8]H. Larsson, A. T. van Duin, and B. Hartke, "Global optimization of parameters in the reactive force field ReaxFF for SiOH," J. Comput. Chem. **34**, 2178–2189 (2013).
[9]A. Jaramillo-Botero, N. Saber, and W. A. Goddard, "General multiobjective force field optimization framework, with application to reactive force fields for silicon carbide," J. Chem. Theory Comput. **10**, 1426–1439 (2014).
[10]F. Cailliez and P. Pernot, "Statistical approaches to forcefield calibration and prediction uncertainty in molecular simulation," J. Chem. Phys. **134**, 054124 (2011).
[11]S. Wu, P. Angelikopoulos, C. Papadimitriou, R. Moser, and P. Koumoutsakos, "A hierarchical Bayesian framework for force field selection in molecular dynamics simulations," Philos. Trans. R. Soc., A **374**(2060), 20150032 (2016).
[12]M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera, "An 'empirical interpolation' method: Application to efficient reduced-basis discretization of partial differential equations," C. R. Math. **339**, 667–672 (2004).
[13]P. E. Hadjidoukas, P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos, "Π4U: A high performance computing framework for Bayesian uncertainty quantification of complex models," J. Comput. Phys. **284**, 1–21 (2015).
[14]J. Y. Ching and Y. C. Chen, "Transitional markov chain monte carlo method for Bayesian model updating, model class selection, and model averaging," J. Eng. Mech. Div., Am. Soc. Civ. Eng. **133**, 816–832 (2007).
[15]J. L. Beck and L. S. Katafygiotis, "Updating models and their uncertainties. I: Bayesian statistical framework," J. Eng. Mech. Div., Am. Soc. Civ. Eng. **124**, 455–461 (1998).
[16]J. L. Beck, "Bayesian system identification based on probability logic," Struct. Control Health Monit. **17**, 825–847 (2010).
[17]P. Angelikopoulos, C. Papadimitriou, and P. Koumoutsakos, "Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework," J. Chem. Phys. **137**, 144103 (2012).
[18]M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," J. Mach. Learn. Res. **1**, 211–244 (2001).
[19]S. Rick, "A reoptimization of the five-site water potential (TIP5P) for use with Ewald sums," J. Chem. Phys. **120**, 6085–6093 (2004).
[20]M. Holz, S. R. Heil, and A. Sacco, "Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements," Phys. Chem. Chem. Phys. **2**, 4740–4742 (2000).
[21]F. E. Jones and G. L. Harris, "Its-90 density of water formulation for volumentric standards calibration," J. Res. Natl. Inst. Stand. Technol. **97**, 335–340 (1992).
[22]A. K. Soper, "The radial distribution functions of water as derived from radiation total scattering experiments: Is there anything we can say for sure?," ISRN Phys. Chem. **2013**, 279463.
[23]S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "Gromacs 4.5: A high-throughput and highly parallel open source molecular simulation toolkit," Bioinformatics **29**, 845–854 (2013).
[24]J. P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 3rd ed. (Academic Press, 2006).
[25]M. Chaplin, Water structure and science, http://www1.lsbu.ac.uk/water/water_structure_science.html, 2000.