Bayesian calibration of force fields for molecular simulations 6

Fabien Cailliez[1], *Pascal Pernot*[1], *Francesco Rizzi*[2], *Reese Jones*[2], *Omar Knio*[3], *Georgios Arampatzis*[4], *Petros Koumoutsakos*[4]

[1]Laboratoire de Chimie Physique, CNRS, University Paris-Sud, Université Paris-Saclay, Orsay, France; [2]Sandia National Laboratories, Livermore, CA, United States; [3]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia; [4]Computational Science and Engineering Laboratory, ETH Zürich, Zürich, Switzerland

6.1 Introduction

Over the last three decades, molecular simulation has become ubiquitous in scientific fields ranging from molecular biology to chemistry and physics. It serves as a tool to rationalize experimental results, providing access to the dynamics of a system at the atomistic and molecular level [1], and predictions of macroscopic properties of materials [2]. As computational hardware and software capabilities increase, molecular simulations are becoming increasingly more important as a tool to complement experiments and have become an invaluable asset for insight, prediction, and decision making by scientists and engineers. This increased importance is associated with an ever-increasing need to interpret quality of the predictions of the complex molecular systems. In the context of *Virtual Measurements*, as proposed by Irikura et al. [3], we remark that *for the output of a molecular simulation to be considered equivalent to an experimental measurement, it must include both a value of the quantity of interest (QoI) and a quantification of its uncertainty*. In turn, uncertainty can be defined [4] as a "*parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand.*" Uncertainty quantification (UQ) is essential for building confidence in model predictions and helping model-based decisions [5]. Monitoring uncertainties in computational physics/chemistry has become a key issue [6], notably for multiscale modeling [7]. Reliable predictions at coarser scales imply the rational propagation of uncertainties from finer scales [8,9]; however, accounting for such uncertainties requires access to a significant computational budget.

The sources of uncertainty in molecular simulations can be broadly categorized as follows:

- *Modeling* (e.g., the choice of the particular force fields)
- *Parametric* (e.g., the 6−12 exponents and the σ, ε parameters in the Lennard-Jones (LJ) potentials)

Uncertainty Quantification in Multiscale Materials Modeling. https://doi.org/10.1016/B978-0-08-102941-1.00006-7

- *Computational* (e.g., the effects of the particular thermostats and non-Hamiltonian time integrators)
- *Measurement* (involving the stochastic output of the simulations for various QoIs).

In this chapter, we focus on Modeling and Parametric uncertainties introduced by the force fields employed in molecular simulations. A force field is a model to represent the energy of the system as a function of the atomic coordinates and empirically determined parameters. In other words, the force field defines the potential energy surface (PES) on which the system evolves in a molecular simulation. In recent years, mathematical descriptions of force fields have also been proposed in nonexplicit form by employing machine learning algorithms such as neural networks [10−12].

Macroscopic properties of systems studied through molecular simulations are obtained using the laws of statistical mechanics through the use of algorithms, mainly Monte Carlo or Molecular Dynamics, that sample the PES of the system. Molecular simulations may involve large numbers of molecules (up to a few trillions of atoms currently) and may reach time scales (up to few microseconds currently) that are presently inaccessible with quantum mechanics. At the same time the output of molecular simulations hinges on the effective representation of the (electronic) degrees of freedom that are removed from the quantum mechanics simulations.

One of the major challenges of molecular simulation is the specification of the interparticle interaction potentials, both in terms of the functional shape and the respective parameter values. Even for force fields where the parameters have a physical meaning (for example, the power of six in the LJ potential), their values are often not directly accessible by experiments, nor by computational chemistry. They need to be calibrated on a set of reference (experimental or calculated) properties. Force field calibration is an exacting and intuitive task and is sometimes considered as an "art." Most of the time, the previous experience of the researcher is required to provide a reasonable initial guess of the parameters, which are then refined locally, either by trial and error, or using optimization methods [13−15]. Calibration is also difficult because the simplicity of the mathematical expressions used in force fields for efficient computations often leads to the impossibility to correctly fit different properties with a unique set of parameters [16,17]. The "best" parameter set is then often the result of a subjective compromise between representing the various data chosen for the calibration. Validation of the force field thus obtained is made by computing properties not used for calibration and comparing them with experimental data [14,18].

The effect of parameters on property predictions is sometimes estimated through sensitivity analysis [14,19−21]. Until very recently, there was no attempt to compute the uncertainties on those properties that are due to the force field parameters. This lack of interest is due in part to the belief that *measurement* uncertainties of molecular simulation (inherited from the stochastic nature of Monte Carlo and molecular dynamics simulations) were greater than *parametric* uncertainties. However, the increase in computational power has greatly reduced the former, without affecting the latter, thus making the parametric errors a significant contribution to the uncertainty budget [17]. A second reason for the scarcity of literature on parametric uncertainties in molecular simulation is the difficulty in estimating the uncertainties on the force field parameters. This requires an extensive exploration of the parameter space, which is often inaccessible due to limited computational resources. Force field calibration was, and still is, mostly based on deterministic least-squares optimization [22].

Uncertain quantities can be represented by probability density functions (PDFs) [4,23], which grounds uncertainty management in the theory of probabilities and provides a sound mathematical framework for their consistent treatment. Parameter calibration is an inference process, for which probability theory provides a computation framework through Bayes' rule. When compared to least-squares−based calibration procedures, the Bayesian approach presents several decisive advantages: (a) it exposes clearly the statistical assumptions underlying the calibration process; (b) it enables the implementation and calibration of complex statistical models (e.g., hierarchical models, model selection …); and (c) it provides directly information on parameter identification through the shape of the PDF.

In the context of molecular simulations, Bayesian inference has been introduced in 2008 by Cooke and Schmidler [24], who calibrated the protein dielectric constant to be used in electrostatic calculations in order to reproduce helicities of small peptides. Since then, several research groups have brought significant contributions to the use of Bayesian calibration aiming at the estimation of force field parameter uncertainties and their impact on property predictions.

This chapter begins with an overview of Bayesian calibration in Section 6.2, focusing on the "standard" approach (Section 6.2.1), its limitations (Section 6.2.2), as well as advanced schemes (Section 6.2.3). Then, in Section 6.3, we discuss computational aspects focusing on metamodels (Section 6.3.2) and approximation of intractable posteriors (Section 6.3.3) necessary to make Bayesian strategies compatible with the high cost of molecular simulations. The main topics treated in the previous sections are summarized in Fig. 6.1. In Section 6.4, we describe representative applications to show what has been learned during the last decade. Finally, in Section 6.5, we present conclusions and perspectives.

6.2 Bayesian calibration

This section introduces Bayesian inference and the general concepts used in Bayesian data analysis [25−27]. It presents also some basic and commonly used statistical models and hypotheses, only to better show their limitation in the context of force field calibration and the necessity to design more advanced calibration schemes.

6.2.1 The standard Bayesian scheme

6.2.1.1 Bayes' theorem

In general, a force field calibration problem involves searching for the value(s) of the parameters $\boldsymbol{\vartheta} = \{\vartheta_i\}_{i=1}^{N_\vartheta}$ of a given computational model $F(x;\vartheta)$ that minimizes the difference between model predictions and a set of reference data $\boldsymbol{D} = \{d_i\}_{i=1}^{N_D}$ in specified (macroscopic, observable) physical conditions (temperature, pressure, etc.) defined by $\boldsymbol{X} = \{x_i\}_{i=1}^{N_D}$.

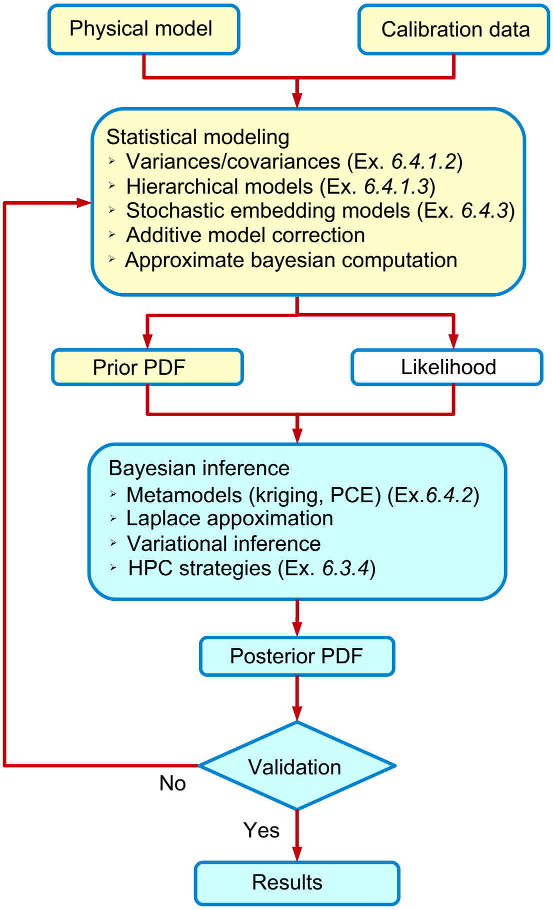
Physical model
Calibration data
Statistical modeling
Variances/covariances (Ex. *6.4.1.2*)
Hierarchical models (Ex. *6.4.1.3*)
Stochastic embedding models (Ex. *6.4.3*)
Additive model correction
Approximate bayesian computation
Prior PDF
Likelihood
Bayesian inference
Metamodels (kriging, PCE) (Ex.*6.4.2*)
Laplace appoximation
Variational inference
HPC strategies (Ex. *6.3.4*)
Posterior PDF
Validation
No
Yes
Results

Figure 6.1 Flowchart of the Bayesian calibration process, underlining the main topics developed in this chapter. The section numbers of application examples are given in italics.

In contrast to deterministic least-squares fitting resulting in a single set of parameter values, in a Bayesian perspective the parameters are considered as random variables $\boldsymbol{\vartheta}$ with associated PDFs that incorporate both prior knowledge and constraints by reference data through the model. Bayesian calibration gives the estimation of the conditional distribution of the parameters that fit the available data given a choice of calibration model M, $p(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},\boldsymbol{M})$. By M, we denote the full calibration model under consideration, i.e., the computational model F, and a statistical model presented below.

Bayes' rule relates the data and *prior* assumptions on the parameters into the *posterior* density of the target parameters. The *posterior* PDF of the parameters, knowing the data and model, is

$$p(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},M) = \frac{p(\boldsymbol{D}|\vartheta,\boldsymbol{X},M)p(\vartheta|M)}{p(\boldsymbol{D}|\boldsymbol{X},M)}, \tag{6.1}$$

where $p(\boldsymbol{D}|\boldsymbol{\vartheta},\boldsymbol{X},\boldsymbol{M})$ is the *likelihood* of observing the data given the parameters and model, $p(\boldsymbol{\vartheta}|\boldsymbol{M})$ is the *prior density* of the parameters reflecting our knowledge *before* incorporating the observations, and $p(\boldsymbol{D}|\boldsymbol{X},M)$ is a normalization factor, called the *evidence*. The denominator is typically ignored when sampling from the posterior since it is a constant, independent of $\boldsymbol{\vartheta}$; however, this factor has to be estimated if one wishes to compare different models (see Section 6.2.1.3).

The choice of a prior PDF is an important step in Bayesian data analysis, notably when the data provide weak constraints on (some of) the parameters. Data influence the resulting posterior probability only through the likelihood $p(\boldsymbol{D}|\boldsymbol{\vartheta},\boldsymbol{X},\boldsymbol{M})$, which involves the difference between the reference data $\boldsymbol{D}$ and the model predictions $F(\boldsymbol{X};\boldsymbol{\vartheta})$.

In general, given the complexity of the model, the posterior density is not known in closed form and one has to resort to numerical methods to evaluate it (see Section 6.3.1).

From the model to the likelihood. A widely adopted approach in data analysis is to express the difference between a reference value of the data d_i and the respective model prediction $F_i(\boldsymbol{\vartheta}) \doteq \boldsymbol{F}(\boldsymbol{x_i};\boldsymbol{\vartheta})$ using an additive noise model

$$d_i = F_i(\boldsymbol{\vartheta}) + \varepsilon_i, \tag{6.2}$$

where ε_i is a zero-centered random variable. This expression is a *measurement model*, expressing that the observed datum is a random realization of a generative process centered on the model prediction. This formulation is assuming that the model $F(x;\boldsymbol{\vartheta})$ *accurately* represents the true, physical process occurring with fixed, but unknown, parameters. This is a strong assumption, which is usually wrong, because every model of a physical process involves some approximation. This is one of the main deficiencies of this approach, which will be treated at length in the following sections. Nevertheless, this is a commonly used method due to its simplicity.

The next common modeling assumption, especially if the data come from various experiments, is to assume the errors to be independent normal random variables with zero mean, i.e., $\varepsilon_i \sim \mathcal{N}\left(0,\sigma_i^2\right)$, where σ_i^2 is the variance of the errors at x_i. Based on Eq. (6.2), an equivalent formulation is $d_i \sim \mathcal{N}\left(F_i(\boldsymbol{\vartheta}),\boldsymbol{\sigma_i^2}\right)$, which yields the following expression for the likelihood

$$\begin{aligned}
p\left(\boldsymbol{D}|\boldsymbol{\vartheta},\boldsymbol{X},M) = \prod_{i=1}^{N_D} p(d_i|\boldsymbol{\vartheta},x_i,M\right), \\
= \prod_{i=1}^{N_D}\left(2\pi\sigma_i^2\right)^{-1/2}\exp\left(-\frac{(d_i - F_i(\boldsymbol{\vartheta}))^2}{2\sigma_i^2}\right), \\
= \left[\prod_{i=1}^{N_D} 2\pi\sigma_i^2\right]^{-1/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{N_D}\frac{(d_i - F_i(\boldsymbol{\vartheta}))^2}{\sigma_i^2}\right).
\end{aligned} \tag{6.3}$$

The value of σ_i depends on the error budget and the available information. As ε_i is the difference of two quantities, its variance is the sum of the variances of d_i and $F_i(\boldsymbol{\vartheta})$ [4]. Typically, one would write

$$\sigma_i^2 = u_{d_i}^2 + u_{F_i(\boldsymbol{\vartheta})}^2, \tag{6.4}$$

where $u_{d_i} \doteq u(d_i)$ is the uncertainty attached to d_i, and $u_{F_i(\boldsymbol{\vartheta})}$ is the *measurement* uncertainty for model prediction $F_i(\boldsymbol{\vartheta})$. In this formulation, the only parameters are those of the model F.

When no value of u_{d_i} is available, it is convenient to make the assumption that the reference data uncertainty is unknown and identical for all data of a same observable. Depending on the heterogeneity level of the reference dataset, one then has one or several additional parameters $\boldsymbol{\tau} = \{\tau_i\}_{i=1}^{N_\tau}$ to be identified, and the likelihood becomes

$$p(\boldsymbol{D}|\boldsymbol{\vartheta}, \boldsymbol{\tau}, \boldsymbol{X}, M) = \left[\prod_{i=1}^{N_D} 2\pi\sigma_i^2(\boldsymbol{\tau})\right]^{-1/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{N_D} \frac{(d_i - F_i(\boldsymbol{\vartheta}))^2}{\sigma_i^2(\boldsymbol{\tau})}\right), \tag{6.5}$$

with

$$\sigma_i^2(\boldsymbol{\tau}) = \tau_{j=I(i)}^2 + u_{F_i(\boldsymbol{\vartheta})}^2, \tag{6.6}$$

where $I(i)$ is a pointer from the datum index i to the adequate index in the set of unknown uncertainty parameters $\boldsymbol{\tau}$.

Although this is a very convenient and commonly used setup, akin to the ordinary least-squares procedure for regression with unknown data variance [28], one should be aware that it can become problematic, especially if the model is inadequate, i.e., it is not able to fit properly the data (more on this in Section 6.2.2.1).

By noting $\boldsymbol{R}(\boldsymbol{\vartheta})$ the vector of N_D differences between the model and data, and $\boldsymbol{\Sigma}_R$ the corresponding covariance matrix, one gets a compact notation for the likelihood in the case of a normal additive noise:

$$p(\boldsymbol{D}|\boldsymbol{\vartheta}, \boldsymbol{X}, M) \propto |\boldsymbol{\Sigma}_R|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{R}(\boldsymbol{\vartheta})^T \boldsymbol{\Sigma}_R^{-1} \boldsymbol{R}(\boldsymbol{\vartheta})\right), \tag{6.7}$$

where the proportionality symbol means that a multiplicative constant has been omitted. $\boldsymbol{\Sigma}_R$ is the sum of the covariance matrix for the reference data $\boldsymbol{\Sigma}_D$ and the covariance matrix of the measurement errors of the model $\boldsymbol{\Sigma}_F$, and $|\boldsymbol{\Sigma}_R|$ is its determinant. So far, especially in Eqs. (6.3) and (6.5), $\boldsymbol{\Sigma}_R$ has been considered as diagonal, with elements $\Sigma_{R,ij} = \sigma_i^2 \delta_{ij}$. Whenever available, covariances of the reference data should be included in the nondiagonal part of $\boldsymbol{\Sigma}_D$.

The prior PDF. In order to complete the definition of the posterior PDF, one needs to define the prior PDF, encoding all the available information on the parameters not conditioned to the data to be analyzed. Mathematical and physical constraints (e.g., positivity) are introduced here through the choice of adapted PDFs [27].

The most common choice in absence of any information on a parameter might be a uniform distribution, or a log-uniform one in case of a positivity constraint (so-called *noninformative* priors). A normal distribution would typically be used to encode a known mean value and uncertainty [23]. An essential consideration at this stage is to ensure that the prior PDF captures intrinsic correlations between parameters (e.g., a sum-to-zero constraint).

Estimation and prediction. The posterior PDF is used to generate statistical summaries of the target parameters. Point estimations can be obtained by the mode of the posterior PDF, or *Maximum* a posteriori (MAP),

$$\widehat{\boldsymbol{\vartheta}} = \underset{\boldsymbol{\vartheta}}{\operatorname{argmax}}\ p(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},M), \tag{6.8}$$

and/or the *mean value* of the parameters,

$$\overline{\boldsymbol{\vartheta}_i} = \int d\boldsymbol{\vartheta}\vartheta_i p(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},M), \tag{6.9}$$

which are different for nonsymmetric PDFs.

The parameter variance, $u_{\boldsymbol{\vartheta}^2}$, and covariances, $\mathrm{Cov}(\vartheta_i,\vartheta_j)$, are also derived from the posterior PDF:

$$\mathrm{Cov}(\vartheta_i,\vartheta_j) = \int d\boldsymbol{\vartheta}\left(\vartheta_i - \overline{\vartheta_i}\right)\left(\vartheta_j - \overline{\vartheta_j}\right)p(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},M), \tag{6.10}$$

$$u_{\vartheta_i} = \mathrm{Cov}(\vartheta_i,\vartheta_i)^{1/2}. \tag{6.11}$$

If the posterior PDF has a shape different from the ideal normal multivariate distribution, other statistical summaries might be useful, but one should also consider contour or density plots of the PDF, which are important diagnostics to assess problems of parameter identification (multimodality, nonlinear correlations, etc.).

Predictions of a QoI $A(\boldsymbol{\vartheta})$ is made through the estimation of the PDF of the QoI averaged over the posterior PDF of the parameters

$$p(A=a|\boldsymbol{D},\boldsymbol{X},M) = \int d\boldsymbol{\vartheta}\ p(A=a|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},M), \tag{6.12}$$

where $p(A=a|\boldsymbol{\vartheta})$ is a PDF describing the dependence of the QoI on $\boldsymbol{\vartheta}$. If A is a deterministic function of the parameters, then $p(A=a|\boldsymbol{\vartheta}) = \delta(\boldsymbol{a} - A(\boldsymbol{\vartheta}))$, where δ is the Dirac delta function. The posterior-weighted integrals used for estimation and prediction are generally evaluated by the arithmetic mean on a representative sample of the posterior PDF (Monte Carlo integration).

6.2.1.2 Validation

As the Bayesian calibration process will always produce a posterior PDF independently of the quality of the fit, it is essential to perform validation checks, notably of the statistical hypotheses used to build $p(\boldsymbol{D}|\boldsymbol{\vartheta}, \boldsymbol{X}, \boldsymbol{M})$. A posterior PDF failing these tests should not be considered for further inference. In particular, the uncertainties on the parameters and their covariances would be unreliable.

Residuals analysis. In a valid statistical setup, *the residuals at the MAP,* $\boldsymbol{R}\left(\widehat{\boldsymbol{\vartheta}}\right)$*, should not display serial correlation* along the control variable(s), which is usually assessed by visual inspection of plots of the residuals [17] and their autocorrelation function. Serial correlation in the residuals is a symptom of model inadequacy and should not be ignored.

Moreover, *the variance of the residuals should be in agreement with the variance of the data and model.* Ideally, the Birge ratio at the MAP,

$$r_B\left(\widehat{\boldsymbol{\vartheta}}\right) = \left(\frac{1}{N_D - N_\vartheta}\boldsymbol{R}\left(\widehat{\boldsymbol{\vartheta}}\right)^T \boldsymbol{\Sigma}_R^{-1}\boldsymbol{R}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)^{1/2}, \tag{6.13}$$

should be close to 1 [29]. The Birge ratio might be too large when the model does not fit the data or when the variances involved in $\boldsymbol{\Sigma}_R$ are underestimated, but it can also be too small when these variances are overestimated.

Note that if the calibration model contains adjustable uncertainty parameters ($\boldsymbol{\tau}$ in Eq. 6.5), their optimization might ensure that $r_B\left(\widehat{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\tau}}\right) \simeq 1$, but would not guarantee that the residuals have no serial correlation [30].

Posterior predictive statistics and plots. The posterior predictive density for the value $\widetilde{d}$ at a new point $\widetilde{x}$ is [27]

$$p\left(\widetilde{d}\middle|\widetilde{x}, \boldsymbol{D}, \boldsymbol{X}, M\right) = \int d\vartheta\, p\left(\widetilde{d}\middle|\widetilde{x}, \vartheta, M\right) p(\vartheta|\boldsymbol{D}, \boldsymbol{X}, M). \tag{6.14}$$

$\widetilde{x}$ is used to generate high-probability (typically 0.9 or 0.95) prediction bands from which one can check the percentage of data effectively recovered by the model predictions [31]. Plots of high-probability bands for the model's residuals as function of the control variable(s) are particularly informative for model validation [30].

6.2.1.3 Model selection

Let us consider a set of alternative models $\boldsymbol{M} = \left\{M^{(i)}\right\}_{i=1}^{N_M}$, parameterized by $\boldsymbol{\vartheta}^{(i)}$, for which one wants to compare the merits in reproducing the reference data $\boldsymbol{D}$. The posterior probability of model $M^{(i)}$ is estimated by applying Bayes' rule

$$p\left(M^{(i)}\middle|\boldsymbol{D}, \boldsymbol{X}\right) = \frac{p\left(\boldsymbol{D}\middle|\boldsymbol{X}, M^{(i)}\right)p\left(M^{(i)}\right)}{\sum_{i=1}^{N_M} p\left(\boldsymbol{D}\middle|\boldsymbol{X}, M^{(i)}\right)p\left(M^{(i)}\right)}, \tag{6.15}$$

where $p\left(M^{(i)}\right)$ is the prior probability of model $M^{(i)}$ and the evidence $p\left(\boldsymbol{D}\middle|\boldsymbol{X},M^{(i)}\right)$ is obtained by

$$p\left(\boldsymbol{D}\middle|\boldsymbol{X},M^{(i)}\right)=\int d\boldsymbol{\vartheta}^{(i)}p\left(\boldsymbol{D}\middle|\boldsymbol{\vartheta}^{(i)},\boldsymbol{X},M^{(i)}\right)p\left(\boldsymbol{\vartheta}^{(i)}\middle|M^{(i)}\right). \tag{6.16}$$

This approach has been used in Refs. [32–36] to compare the performances of different models. Computation of the evidence terms is costly, and high-performance computing (HPC) is generally necessary. On the other hand, the TMCMC algorithm [37], which will be described in Section 6.3.1, offers an estimator for the evidence term.

6.2.2 Limitations of the standard scheme

As in all calibration process, the underlying statistical hypotheses have to be checked, and the Bayesian approach offers no guarantee against model misspecification. In particular, the common hypothesis of i.i.d. errors should be carefully scrutinized.

In fact, the simple likelihood scheme presented above (Eq. 6.7) is often unable to deal properly with the specificities of the calibration of force field parameters, i.e., the corresponding posterior PDF does not pass some of the validation tests. These tests might help to point out the deficiency sources(s), which concern the calibration dataset and its covariance matrix (improper Birge ratio values) or the force field model (serial correlation of the residuals) or both.

6.2.2.1 Modeling of the error sources

A convenient feature of the standard model is the possibility to infer uncertainty parameter(s) ($\boldsymbol{\tau}$ in Eq. 6.5) in order to ensure that $r_B\left(\widehat{\vartheta},\widehat{\boldsymbol{\tau}}\right)\simeq 1$, i.e., to obtain a unit variance of the weighted residuals.

The applicability of this approach relies essentially on the independence of the errors, to be validated by the absence of serial correlation in the residuals. Otherwise, $\boldsymbol{\tau}$ is absorbing model errors in addition to data uncertainty. In the absence of dominant measurement uncertainty, model errors present strong serial correlations, which is in conflict with the standard scheme's i.i.d. hypothesis. In these conditions, using the uncertainty parameters $\boldsymbol{\tau}$ as "error collectors" should not be expected to produce reliable results. It is essential to devise a detailed scheme of error sources in order to get unambiguous identification and modeling of all contributions.

6.2.2.2 Data inconsistency

Experimental data. In force field calibration, one is often confronted with multiple versions of reference data, produced by different teams in similar or overlapping experimental conditions. It is frequent that some measurement series are inconsistent, in the sense that values measured with different methods, instruments, or by different teams (*reproducibility conditions* [4]) are not compatible within their error bars. This might be due to an underestimation of measurement uncertainty, for instance, taking into account only the *repeatability component*, and ignoring nonrandom instrumental error sources.

Depending on the context, this problem can be dealt with in several (nonexclusive) ways [17]:

- pruning the dataset, which should be reserved to experts in the specific data measurements fields;
- scaling the data covariance matrix $\mathbf{\Sigma}_D$ by factor(s) which might be parameter(s) of the calibration process. This assumes that the initial uncertainty assessments are incorrect (this approach is a common practice in the metrology of interlaboratory comparisons [29,38,39]); or
- using data shifts, to be calibrated along with ϑ, in order to reconciliate discrepant data series by compensation of measurement biases [29,30].

Theoretical data. Data might also come from deterministic reference theoretical models (e.g., equations-of-state, as used by the NIST database for the properties of fluids [40]), in which case they are not affected by random errors, and the uncertainty statement issued by the data provider quantifies the representative amplitude of errors of this model with respect to its own calibration data [40,41]. A reductio ad absurdum in this case would be to fit the model's results with their declared error bars by the generative model itself, which would produce numerically null residuals, invalidating the Birge ratio test.

This type of data violates the errors independence hypothesis. One way to take it into account would be to design a data covariance matrix $\mathbf{\Sigma}_D$, but there is generally no available information to establish it reliably.

To our knowledge, this point has generally been overlooked in the force field calibration literature, probably because the uncertainty budget is often dominated by other error sources (numerical simulation uncertainty, and/or model inadequacy). It might, however, readily occur when uncertainty scaling is used to compensate for data inconsistency [29] or model inadequacy [30]. Besides, as the quality of force fields and computational power increase, the problem will eventually emerge in the standard calibration framework.

6.2.2.3 Model inadequacy/model errors

Considering the approximate nature of force fields, model inadequacy has to be expected as a typical feature of the calibration problem [42]. For instance, force field approximations make molecular simulation unable to fit a property over a large range of physical conditions [16,17]. This can be somewhat overcome by explicitly modeling the dependence of the parameters on the control variable(s) (for instance, by using temperature-dependent LJ parameters [43–45]). This kind of approach, i.e., force field *improvement*, is in fact a change of model F. Similarly, for LJ-type potentials, a unique set of parameters is typically unable to fit several observables (e.g., the liquid and vapor densities of Argon [35]), which would call for observable-dependent force field parameters, and the loss of parameter transferability for the prediction of new properties.

Using the standard calibration scheme (Eq. 6.3) in presence of model inadequacy leads to statistical inconsistencies. Within this setup, parameter uncertainty is decreasing when the number of calibration data is increased, which means that prediction uncertainty of the calibrated model is also decreasing [46,47]. On the contrary, model errors are rather expected to increase—at best to stay constant—when new data are added to the calibration set. Therefore, parameter uncertainty as provided by the standard calibration scheme is intrinsically inadequate to account for model errors. It is thus necessary to devise alternative calibration schemes. There has been recently a marked interest in statistical solutions enabling to integrate model errors into parameter uncertainty [42,47–52]. These solutions, based on Bayesian inference, are treated in the next section.

6.2.3 Advanced Bayesian schemes

One has shown above that there are several causes, notably model inadequacy, to reject the standard force field calibration model. Alternative schemes which have been proposed in the literature to deal with these shortcomings are presented in this section.

6.2.3.1 Additive model correction

We consider here a solution which improves model predictions without involving a change of force field model. Model inadequacy can be solved with an additive term to the original model:

$$d_i = F_i(\boldsymbol{\vartheta}) + \delta F_i(\boldsymbol{\vartheta}_{\delta F}) + \varepsilon_i, \tag{6.17}$$

where the discrepancy function δF has its own set of parameters, $\boldsymbol{\vartheta}_{\delta F}$.

The representation of δF by a Gaussian process (GP) has been popularized by Kennedy and O'Hagan [53]. It has many advantages over, for instance, polynomial-based functions, but, by construction, δF can correct any error due to a misspecification of F, which weakens considerably the constraints of $\boldsymbol{D}$ on ϑ. The GP approach might therefore be subject to severe identification problems, if the parameters of F and δF are optimized simultaneously without strong prior information [46,50,54,55].

A two-staged solution, proposed by Pernot and Cailliez [30], is to constrain $\boldsymbol{\vartheta}$ with the posterior PDF resulting from an independent calibration of F. In this case, δF is designed to fit the residuals of $F(x;\boldsymbol{\vartheta})$ by a GP of mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\delta F}$, with elements $\boldsymbol{\Sigma}_{\delta F,ij} = k(x_i, x_j)$, based on a Gaussian kernel $k(u,v) = \alpha^2 \exp\left(-\beta^2(u-v)^2\right)$. The kernel's parameters $\boldsymbol{\vartheta}_{\delta F} = \{\alpha, \beta\}$ have to be estimated in addition to $\boldsymbol{\vartheta}$.

The predictive posterior PDF has a closed form expression [56]

$$p\left(\widetilde{d}|\widetilde{x}, \boldsymbol{D}, \boldsymbol{X}, M\right) = \mathcal{N}\left(\widetilde{d}\,\middle|\,\Omega^T \boldsymbol{\Sigma}_R^{-1} \boldsymbol{D}, \alpha^2 - \Omega^T \boldsymbol{\Sigma}_R^{-1} \Omega\right), \tag{6.18}$$

where $\Omega = \Omega(\boldsymbol{X})$ and $\Omega_i = k(x_i, \widetilde{x})$.

The GP correction method is very efficient [30], but suffers from a major drawback, inherent to all additive correction methods: the discrepancy function δF is not transferable to other observables than the one it was calibrated with, nor can the GP correction be used for extrapolation out of the range of the calibration control variables.

6.2.3.2 Hierarchical models

There is a wealth of heterogeneity when considering the data used for calibrating potentials of molecular simulations. As discussed in Section 6.2.2.2, it is often the case that different experimental groups provide different values for quantities of interest, for example, diffusion constants, even when the experiments are performed in the same thermodynamic conditions. Even more, calibration of molecular systems may often require matching different experimental data ranging from structural properties like radial distribution functions (RDFs) to transport properties like diffusivity. Uncertainties due to different measurement techniques, facilities, and experimental conditions are often reflected in the values of such data.

When model inadequacy arises from the use of a unique parameter set for different observables or experimental conditions, hierarchical models may enable to derive more robust parameter sets [57,58].

In a hierarchical model, the data are being gathered into N_H groups, each containing N_i data, $\boldsymbol{D} = \{\boldsymbol{d}_i\}_{i=1}^{N_H}$ and $\boldsymbol{d}_i = \{d_{i,j}\}_{j=1}^{N_i}$. For each group, a different parameter set $\boldsymbol{\vartheta_i}$ is considered, and all the parameters are controlled by hyperparameters $\boldsymbol{\kappa}$. The structure of this relation is given in Fig. 6.2 in plate notation. The likelihood is now written as

$$p(\boldsymbol{D}|\boldsymbol{\kappa}, \boldsymbol{X}, M) = \prod_{i=1}^{N_H} \int d\boldsymbol{\vartheta}_i \; p(\boldsymbol{D}_i|\boldsymbol{\vartheta}_i, M) p(\boldsymbol{\vartheta}_i|\boldsymbol{\kappa}, M). \tag{6.19}$$

For example, a specific choice for the prior PDF on the $\boldsymbol{\vartheta_i}$ is the normal distribution

$$p(\boldsymbol{\vartheta}_i|\boldsymbol{\kappa}, M) = \mathcal{N}(\vartheta_i; \boldsymbol{\mu}_\vartheta, \boldsymbol{\Sigma}_\vartheta), \tag{6.20}$$

where $\boldsymbol{\kappa} = \{\boldsymbol{\mu}_\vartheta, \boldsymbol{\Sigma}_\vartheta\}$ are the parameters of an *overall* normal distribution, which have to be inferred along with the local values of $\boldsymbol{\vartheta_i}$.

Hierarchical models enable the robust inference of multiple parameter sets with global constraints (prior PDF on the hyperparameters). However, the uncertainty on the hyperparameters $\boldsymbol{\kappa}$ of the overall distribution is conditioned by the number of subsets in the calibration data N_H. When this number is small, strong prior information on the hyperparameters should be provided to help their identification [30,57].

This scheme has been recently applied to the calibration of LJ parameters from various sets of experimental data obtained in different thermodynamic conditions. In Ref. [52], the dataset is split into subsets corresponding to different control temperatures and pressures. A hierarchical model is used for the inference of the

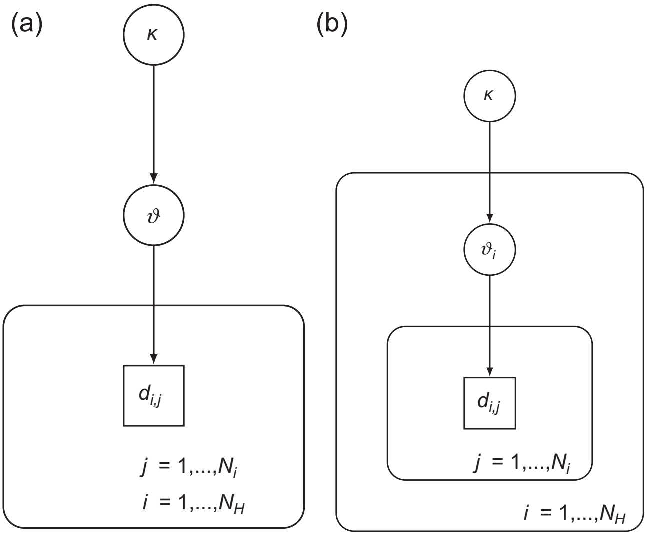

Figure 6.2 Two approaches of grouping the data in a Bayesian inference problem. (a) For the left graph, one parameter $\boldsymbol{\vartheta}$ will be inferred using all the available data. (b) In the right graph, the data are gathered into N_H groups each containing N_i data. Each group has a different parameter $\boldsymbol{\vartheta}_i$. All the parameters are being linked through the hyperparameter $\boldsymbol{\kappa}$.

hyperparameters describing the LJ potentials (model $\mathcal{M}_{H2}$ in Ref. [52]). Using a different data partition, the same authors used a parameter hierarchy to solve data inconsistency (model $\mathcal{M}_{H1}$ in Ref. [52]). Hierarchical models have also been used to accommodate heterogeneous datasets [59]. Pernot and Cailliez [30] introduced a hierarchical model on data shifts to solve data inconsistency.

Depending on the data partition scheme, predictions are made either from the locally adapted parameters ϑ_i or from their overall distribution [57,58]. For instance, when the data partition has been made along the control variables $\boldsymbol{X}$ or according to the nature of data, local parameters can be addressed unambiguously for prediction in any of the identified subset conditions. However, if the partition has been made to separate inconsistent data series, or if one wishes to predict a new property, the local sets cannot be addressed, and the overall distribution has to be used. In general, hierarchical Bayesian inference tries to accommodate information across distinct datasets and, as such, results in much larger prediction uncertainty than what can be inferred by using distinct local parameters sets. In some cases, the prediction uncertainty resulting from the overall distribution is too large for the prediction to be useful [30,59].

6.2.3.3 Stochastic Embedding models

A recent addition to the methods dealing with model inadequacy is based on the idea of replacing the parameters of the model by stochastic variables. This provides an additional variability source to the model's predictions which may be tuned to compensate for model inadequacy. It is important to note, as underlined by Pernot and Cailliez

[30], that this approach, mostly based on the tweaking of the parameters covariance matrix, cannot reduce the gap between model predictions and reference data. Its impact is on *marginal*, or individual, model prediction uncertainties, which can be enlarged sufficiently to cover the difference with the corresponding calibration data.

In the stochastic embedded (SEm) models approach [60,61], the model's parameters, ϑ, are defined as stochastic variables, with a PDF conditioned by a set of hyperparameters $\boldsymbol{\kappa}$, typically their mean value vector $\boldsymbol{\mu}_\vartheta$ and a covariance matrix $\boldsymbol{\Sigma}_\vartheta$, defining a multivariate (normal) distribution $p(\vartheta|\boldsymbol{\mu}_\vartheta, \boldsymbol{\Sigma}_\vartheta, M)$.

Such stochastic parameters can be handled in the Bayesian framework either at the model or at the likelihood level, defining two classes of methods. Both suffer from degeneracy problems, because of the strong covariance of model predictions over the control variable range [42,51].

Model averaging. Statistical summaries (μ_F, $\boldsymbol{\Sigma}_F$) of predictions of the model with stochastic parameters are first estimated and inserted into the likelihood:

$$p(\boldsymbol{D}|\boldsymbol{\mu}_\vartheta, \boldsymbol{\Sigma}_\vartheta, M) \propto |\boldsymbol{\Sigma}_T|^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{R}^T \boldsymbol{\Sigma}_T^{-1} \boldsymbol{R}\right), \tag{6.21}$$

where $R_i = d_i - \mu_F(x_i)$ and $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_D + \boldsymbol{\Sigma}_F$, and the mean values $\mu_F(x_i)$ and covariance matrix $\boldsymbol{\Sigma}_F$ have to be estimated by forward uncertainty propagation (UP), such as linear UP (or combination of variances) [4], or polynomial chaos UP (see below).

When the number of parameters is smaller than the number of data points, $\boldsymbol{\Sigma}_F$ is singular (non-positive definite), causing the likelihood to be degenerate, and the calibration to be intractable [51]. In presence of model inadequacy, the data covariance matrix (and the model measurement uncertainties for stochastic models) are too small to alleviate the degeneracy problem. As a remedy, it has been proposed to replace the multivariate problem by a set of univariate problems (marginal likelihoods [51]), i.e., to ignore the covariance structure of model predictions

$$\Sigma_{T,ij} = \Sigma_{D,ij} + \Sigma_{F,ij}\delta_{ij}. \tag{6.22}$$

It is then possible to modulate the shape of the prediction uncertainty bands by designing the $\boldsymbol{\Sigma}_\vartheta$ matrix [42], notably through a judicious choice of prior PDFs for the hyperparameters.

Likelihood averaging. Integration of the initial likelihood over the model's stochastic parameters provides a new likelihood, conditioned on the hyperparameters $\boldsymbol{\kappa}$ [47,51].

$$p(\boldsymbol{D}|\boldsymbol{\kappa}, \boldsymbol{X}, M) = \int d\boldsymbol{\vartheta}\, p(\boldsymbol{D}|\boldsymbol{\vartheta}, \boldsymbol{X}, M) p(\boldsymbol{\vartheta}|\boldsymbol{\kappa}, \boldsymbol{X}, M). \tag{6.23}$$

The integrated likelihood is in general degenerate, and it has been proposed to replace it by a tractable expression involving summary statistics of the model predictions, to be compared to similar statistics of the data [51]. This approach, called approximate Bayesian computation (ABC), is presented in the next section.

6.2.3.4 Approximate Bayesian Computation

The definition of the likelihood is a centerpiece of Bayesian inference. In certain cases, the analytical statistical models that have been examined above are not suitable for describing such likelihoods. One such example is computational models that generate outputs over several iterations or employ complex simulations. In such cases, ABC [62] has been introduced that bypass the calculation of the likelihood by using a metric to compare the output of the model and the observations. ABC methodologies have received significant attention in fields such as genetics [63], epidemiology [64], and psychology [65]. The ABC approach is often referred to as a *likelihood-free* approach.

The ABC algorithm [66] aims at sampling the joint posterior distribution $p(\vartheta, \boldsymbol{Y}|\boldsymbol{D}, \boldsymbol{X}, M)$, where $\boldsymbol{Y}$ follows $p(\,\bullet|\vartheta, M)$, i.e., $\boldsymbol{Y}$ is a sample from the forward model. Applying Bayes' theorem to the joint distribution we get,

$$p(\boldsymbol{\vartheta}, \boldsymbol{Y}|\boldsymbol{D}, \boldsymbol{X}, M) \propto\; p(\boldsymbol{D}|\boldsymbol{\vartheta}, \boldsymbol{Y}, \boldsymbol{X}, M)\; p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{X}, M)\; p(\boldsymbol{\vartheta}|M). \tag{6.24}$$

The function $p(\boldsymbol{D}|\boldsymbol{Y}, \boldsymbol{\vartheta}, \boldsymbol{X}, \boldsymbol{M})$ gives higher values when $\boldsymbol{Y}$ is close to $\boldsymbol{D}$ and small values in the opposite case. The idea of the ABC algorithm is to sample $\boldsymbol{\vartheta}$ from the prior $p(\boldsymbol{\vartheta}|\boldsymbol{M})$, then sample $\boldsymbol{Y}$ from the forward model $p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{X}, \boldsymbol{M})$, and finally accept the pair $(\boldsymbol{\vartheta}, \boldsymbol{Y})$ when $\boldsymbol{Y} \;= \boldsymbol{D}$.

Since the space that $\boldsymbol{Y}$ lies in is usually uncountable, the event $\boldsymbol{Y} = \boldsymbol{D}$ has zero probability. The basic form of the ABC algorithm introduces a metric ρ and a tolerance parameter δ and accepts $(\boldsymbol{\vartheta}, \boldsymbol{Y})$ when $\rho(\boldsymbol{Y}, \boldsymbol{D}) < \delta$ and rejects it otherwise. In another variant of the algorithm, a summary statistic S is used to compare $\boldsymbol{Y}$ and $\boldsymbol{D}$ through a different metric ρ. In this case, the pair is accepted when $\rho(S(\boldsymbol{Y}), S(\boldsymbol{D}))<\delta$. Note that the ABC algorithm approximates the true posterior $p(\boldsymbol{\vartheta}|\boldsymbol{D}, \boldsymbol{X}, \boldsymbol{M})$ with the density

$$p_\delta(\boldsymbol{\vartheta}|\boldsymbol{D}, \boldsymbol{X}, M) \propto I_{A_\delta(\boldsymbol{D})}(\boldsymbol{Y}) p(\boldsymbol{Y}|\boldsymbol{\vartheta}, \boldsymbol{X}, M) p(\vartheta|M), \tag{6.25}$$

where $A_\delta(\boldsymbol{D}) := \{y : \rho(S(y), S(\boldsymbol{D})) < \delta\}$ and the indicator function $I_{A_\delta(\boldsymbol{D})}(\boldsymbol{Y}) = 1$ when $\boldsymbol{Y} \in A_\delta(\boldsymbol{D})$ and zero otherwise.

Then, the approximation to the target marginal posterior is given by

$$p_\delta(\boldsymbol{\vartheta}|\boldsymbol{D}, \boldsymbol{X}, M) = \int\limits_{A_\delta(\boldsymbol{D})} d\boldsymbol{Y} p(\boldsymbol{Y}|\vartheta, \boldsymbol{X}, M) p(\vartheta|M). \tag{6.26}$$

Notice that the integration over $A_\delta(\boldsymbol{D})$ needs not to be done explicitly since only the $\boldsymbol{\vartheta}$ component is kept from the pair $(\vartheta, \boldsymbol{Y})$.

The parameter δ controls the amount of computational effort that needs to be spent for the generation of $\boldsymbol{\vartheta}$ samples. For large values of δ, a large error is introduced in the approximation, but less computational effort is needed to accept a sample ϑ. On the other hand, small δ leads to a better approximation but the event $\boldsymbol{Y} \in A_\delta(\boldsymbol{D})$ becomes a rare event and most of the $\boldsymbol{Y}$ samples will be rejected.

Various algorithms have been proposed for the acceleration of the original plain ABC algorithm. In Ref. [66], the ABC-MCMC algorithm has been proposed where the rejection step has been replaced by a Markov chain Monte Carlo (MCMC) step leading to an increase of the acceptance ratio. In Ref. [67], the ABC-SubSim algorithm has been proposed, where the subset simulation, an algorithm for sampling rare events, has been combined with ABC. In Ref. [68], more accurate estimates of the posterior distribution are obtained by incorporating the error in the ABC approximation.

Recently, the ABC-SubSim [69] has been used in force field identification of molecular dynamics simulations. In Ref. [70], ABC has been used to approximate the likelihood in the force field calibration of a water model. In Refs. [30,47,51], the ABC algorithm has been used in order to approximate the intractable likelihood in SEm models discussed in Section 6.2.3.3.

6.3 Computational aspects

The implementation of Bayesian calibration for molecular simulations presents challenging aspects due to the necessity to generate a large representative sample of the posterior PDF for the estimation of parameters statistics that is incompatible with the high computational cost of molecular simulations.

6.3.1 Sampling from the posterior PDF

Except for simple toy models, there is no direct method to sample from the posterior PDF. The sampling is usually done by a random walk in parameters space based on the Metropolis algorithm [71] as generalized by Hastings [72]. This is the basis for a family of algorithms named MCMC. There are many flavors of MCMC, and the reader is referred to reference textbooks for more information [73–75].

For the calibration of force fields and model selection, Angelikopoulos et al. developed a version of transitional MCMC (TMCMC), which is well adapted to HPC [32,34,76]. If the evaluation of the model is not computationally heavy or model selection is not required, non-HPC versions of MCMC can be used. For instance, Pernot and Cailliez [17,30,77] used out-of-the-box functions available in the `R` [78] and `stan` [79] programming languages.

Whichever the type of MCMC algorithm, the comprehensive exploration of parameters space by a Markov chain requires many thousands to millions of evaluations of the model. One cannot envision to perform as many molecular simulations with limited resources, and it is mandatory to use alternative strategies involving computationally cheap emulators of the simulation model, also known as *surrogate models* or *metamodels*. Before treating metamodels, we give a brief presentation of the TMCMC algorithm.

Transitional Markov chain Monte Carlo (TMCMC). TMCMC is a sequential MCMC type algorithm for sampling the posterior distribution. It is based on a sequence of intermediate distributions controlled by an annealing scheme:

$$p_j(\boldsymbol{\vartheta}|\boldsymbol{D},\boldsymbol{X},M) \propto p(\boldsymbol{D}|\boldsymbol{\vartheta},\boldsymbol{X},M)^{\gamma_j} p(\boldsymbol{\vartheta}|M), \tag{6.27}$$

for $j = 1, \ldots, m$ and $0 = \gamma_1 < \ldots < \gamma_m = 1$, that leads to the posterior distribution when $\gamma_m = 1$.

The algorithm first draws N_1 samples from the prior distribution and at the $j+1$ stage uses N_j samples from the distribution p_j to obtain N_{j+1} samples from the distribution p_{j+1}. Let $\Theta_j = \{\boldsymbol{\vartheta}_{j,k}\}_{k=1}^{N_j}$ be the samples obtained at the j-th step from p_j. The following procedure gives samples from p_{j+1}:

1. Draw N_{j+1} samples from the set Θ_j with probability of $\boldsymbol{\vartheta}_{j,k}$ equal to

$$\widehat{w}_{j,k} = \frac{w_{j,k}}{\sum_{k=1}^{N_i} w_{j,k}}, \tag{6.28}$$

where $w_{j,k} = p\left(\boldsymbol{D}\middle|\boldsymbol{\vartheta}_{j,k}\right)^{\gamma_{j+1}-\gamma_j}$. Put the new samples in the set $\widetilde{\Theta}_{j+1}$ and set

$$S_j = \frac{1}{N_j} \sum_{k=1}^{N_j} w_{j,k}. \tag{6.29}$$

2. For each sample in $\widetilde{\Theta}_{j+1}$ perform TMCMC with Gaussian proposal distribution and covariance matrix $\beta^2 \Sigma_s^{(j)}$. Here, β is a scaling parameter and $\Sigma_s^{(j)}$ is the sample covariance at the j-th stage given by,

$$\Sigma_s^{(j)} = \sum_{k=1}^{N_j} \widehat{w}_{j,k} \left(\boldsymbol{\vartheta}_{j,k} - \overline{\boldsymbol{\vartheta}}_j\right)\left(\boldsymbol{\vartheta}_{j,k} - \overline{\boldsymbol{\vartheta}}_j\right)^{\top}, \tag{6.30}$$

where $\overline{\boldsymbol{\vartheta}}_j = \sum_{k=1}^{N_j} \widehat{w}_{j,k} \boldsymbol{\vartheta}_{j,k}$. Set the chain length equal to a predefined parameter ℓ_{max}.

In Algorithm 6.1, the pseudocode of a reduced bias variant of TMCMC, namely the BASIS algorithm [52], is presented.

A key advantage of TMCMC is that it can be efficiently parallelized since the likelihood evaluation is independent for each sample. An additional computational benefit introduced by the BASIS algorithm is that all MCMC chains have equal length and thus the work load can be balanced among the processors. Moreover, an important by-product of the algorithm is that the evidence of the data is estimated by Ref. [37]

$$p(\boldsymbol{D}|M) \approx \prod_{j=1}^{m-1} S_j. \tag{6.31}$$

Algorithm 6.1. BASIS (TMCMC)

```
Input: Likelihood function $p(\boldsymbol{D}|\boldsymbol{\vartheta}, \boldsymbol{X}, \boldsymbol{M})$, prior distribution $p(\boldsymbol{\vartheta}|\boldsymbol{M})$
    $N_j, N_{max}$—number of samples per stage, maximum number of stages
    $\gamma, \beta$—threshold parameter, scale parameter
Output: $\Theta_{final}$—a set of samples from $p(\boldsymbol{\vartheta}|\boldsymbol{D}, \boldsymbol{X}, \boldsymbol{M})$
    S—estimation for the evidence $p(\boldsymbol{D}|M)$
Draw initial sample set $\Theta_1 = \{\vartheta_{1,k}\}_{k=1}^{N_1}$ from prior
Initialize $j \leftarrow 1$, $\gamma_1 \leftarrow 0$, $S \leftarrow 1$
repeat
    Choose $\gamma_{j+1}$ such that the coefficient of variation of $w_{j,k} < \gamma$ and $\gamma_{j+1} \leq 1$
    Calculate $w_{j,k}$ with the chosen $\gamma_{j+1}$
    $S \leftarrow S \cdot \frac{1}{N_j}\sum_{j=1}^{N_j} w_{j,k}$
    Obtain $\widetilde{\Theta}_{j+1}$ by drawing $N_{j+1}$ samples from the set $\Theta_j$ with probabilities $\propto w_{j,k}$
    Set $\Sigma$ the weighted covariance given by Eq. (6.30)
    for each sample in $\widetilde{\Theta}_{j+1}$ do
        Perform MCMC with length equal to $\ell_{max}$ and proposal distribution $q(\cdot|\boldsymbol{\vartheta}) = \mathcal{N}(\cdot|\boldsymbol{\vartheta}, \beta^2\Sigma)$
        Add resulting samples in $\Theta_{j+1}$
    end for
    $j \leftarrow j + 1$
until $\gamma_j = 1$ or $j > N_{max}$
$\Theta_{final} \leftarrow \Theta_j$
```

6.3.2 Metamodels

Metamodels are essential to reduce the computational cost of Bayesian inference. When employed as surrogates of the actual computationally expensive models, the accuracy of the Bayesian inference hinges on the accuracy of the metamodels. Metamodels are a familiar entity to human decision making and handling of uncertainty. Scarcely, we take a step or swim with complete knowledge of the mechanics associated with these processes. Humans are well capable of creating effective models of their environment and at the same time operate under uncertainty implied by these models. The use of metamodels is inherent to modeling procedures and a key element in Bayesian inference.

Over the last years, computational frameworks using metamodels have been devised to overcome the cost of simulations required by Bayesian calibration of force fields [32,34,76,77].

Metamodels can be built either from the physical laws of the system under study or as pure mathematical expressions capturing the dependence/behavior of the original model's outputs as a function of the input variables (force field parameters and control variables). Recently, Messerly et al. [80] proposed a third option, configuration sampling–based surrogate models.

Van Westen et al. [81] used PC-SAFT equations as physics-based models to fit simulation results for the calibration of LJ parameters for n-alkanes. A similar approach has been used recently to parameterize Mie force fields [82,83]. Such physics-based models are unfortunately confined to a restricted set of applications, and behavior-based models have been devised for a more general scope.

In the force field calibration literature, GPs, also called kriging [32,34,35,69,77], and polynomial chaos expansions (PCEs) [84] have been used to build metamodels replacing molecular simulations.

6.3.2.1 Kriging

We describe here shortly the principle of kriging metamodels. More details will be found in Refs. [85–87]. Let $\boldsymbol{Y} = \left\{ y^{(i)} \right\}_{i=1}^{N}$ be a set of N values of a QoI at force field parameters $\Theta = \left(\boldsymbol{\vartheta}^{(\boldsymbol{1})}, \ldots, \boldsymbol{\vartheta}^{(N)} \right)$, where $y^{(i)} = F\left(x; \boldsymbol{\vartheta}^{(\boldsymbol{i})} \right) \pm u_F^{(i)}(x)$. In the *universal kriging* framework, $\boldsymbol{Y}$ is assumed to be of the form

$$y(\boldsymbol{\vartheta}) = \sum_{i=1}^{p} b_i f_i(\boldsymbol{\vartheta}) + Z(\boldsymbol{\vartheta}), \tag{6.32}$$

where the f_i are known basis functions and Z is a GP of mean zero with a covariance kernel $k(\boldsymbol{\vartheta}, \boldsymbol{\vartheta}') = \alpha^{\boldsymbol{2}} \boldsymbol{r}(\boldsymbol{\vartheta}, \boldsymbol{\vartheta\prime; \beta})$, based on a correlation function r with parameters $\boldsymbol{\beta}$. This setup is more general than the one considered in Section 6.2.3.1.

The kriging best predictor and covariance at any point in parameter space are given by

$$\widehat{y}(\boldsymbol{\vartheta}) = \boldsymbol{f}(\boldsymbol{\vartheta})^t \, \widehat{\boldsymbol{b}} + \boldsymbol{k}(\boldsymbol{\vartheta})^t \, \boldsymbol{k}^{-1} \left(\boldsymbol{Y} - \boldsymbol{F}\widehat{b} \right), \tag{6.33}$$

$$\begin{aligned} c_y(\boldsymbol{\vartheta}, \vartheta') = {} & k(\boldsymbol{\vartheta}, \vartheta') - \boldsymbol{k}(\boldsymbol{\vartheta})^t \boldsymbol{K}^{-1} \boldsymbol{k}(\vartheta') \\ & + \left(\boldsymbol{f}(\boldsymbol{\vartheta})^t - \boldsymbol{k}(\boldsymbol{\vartheta})^t \boldsymbol{K}^{-1} \boldsymbol{F} \right)^t \left(\boldsymbol{F}^t \boldsymbol{K}^{-1} \boldsymbol{F} \right)^{-1} \left(\boldsymbol{f}(\vartheta')^t - \boldsymbol{k}(\vartheta')^t \boldsymbol{K}^{-1} \boldsymbol{F} \right), \end{aligned} \tag{6.34}$$

where $\boldsymbol{k}(\boldsymbol{\vartheta}) = \left[\boldsymbol{k}\left(\boldsymbol{\vartheta}^{(\boldsymbol{1})}, \boldsymbol{\vartheta} \right), \ldots, \boldsymbol{k}\left(\boldsymbol{\vartheta}^{(N)}, \boldsymbol{\vartheta} \right) \right]$, $\boldsymbol{K}$ is the GP's variance–covariance matrix with elements $K_{ij} = k\left(\boldsymbol{\vartheta}^{(i)}, \boldsymbol{\vartheta}^{(j)} \right)$, $\boldsymbol{f}(\boldsymbol{\vartheta}) = \left(\boldsymbol{f_1}(\boldsymbol{\vartheta}), \ldots, \boldsymbol{f_p}(\boldsymbol{\vartheta}) \right)$ is a vector of basis functions, $\boldsymbol{F} = \left[\boldsymbol{f}\left(\boldsymbol{\vartheta}^{(\boldsymbol{1})} \right)^t, \ldots, \boldsymbol{f}\left(\boldsymbol{\vartheta}^{(N)} \right)^t \right]^t$, and $\widehat{\boldsymbol{b}} = \left(\boldsymbol{F}^t \boldsymbol{K}^{-1} \boldsymbol{F} \right)^{-1} \boldsymbol{F}^t \boldsymbol{K}^{-1} \boldsymbol{Y}$ is the best linear unbiased estimate of $\boldsymbol{b}$. The kriging prediction uncertainty is $u_y(\boldsymbol{\vartheta}) = \boldsymbol{c_y}(\boldsymbol{\vartheta}, \boldsymbol{\vartheta})^{\boldsymbol{1/2}}$.

$\boldsymbol{K}$, $\boldsymbol{k}$, and $\widehat{\boldsymbol{b}}$ depend implicitly on the parameters $\{\alpha, \boldsymbol{\beta}\}$ of the covariance kernel, which have to be calibrated on $\boldsymbol{Y}$, by maximum likelihood or Bayesian inference.

Different covariance structures are possible [85,87]. A common choice is the exponential family

$$k(\boldsymbol{\vartheta}, \vartheta') = \alpha^2 \prod_{j=1}^{N_\vartheta} \exp\left(- \beta_j \left|\vartheta_j - \vartheta'_j|^\gamma\right.\right), \tag{6.35}$$

with parameters $\alpha > 0$, $\beta_j \geq 0\ \forall j$, and $0 < \gamma \leq 2$. The γ parameter can be optimized or fixed, for instance, at $\gamma = 2$ (Gaussian kernel), providing an interpolator with interesting smoothness properties.

In the Bayesian calibration of force fields, the computationally limiting step is the estimation of the likelihood, and kriging metamodels have been used to provide efficient interpolation functions, either at the likelihood level or at the property level. To account for the uncertainty of molecular simulation results, one has to add the simulation variances on the diagonal of the variance–covariance matrix, which becomes $K_{ij} = k\left(\boldsymbol{\vartheta}^{(i)}, \boldsymbol{\vartheta}^{(j)}\right) + \boldsymbol{u}_{\boldsymbol{F}}^{(\boldsymbol{i})\boldsymbol{2}} \delta_{\boldsymbol{ij}}$ [87].

6.3.2.2 Adaptive learning of kriging metamodels

The construction of a kriging metamodel requires an ensemble of simulations for a set of parameter values Θ, which is to be kept as small as possible to reduce computational charge. The most economical scheme is to start with a small number of simulations, and run new ones at strategically chosen values of the parameters, usually in a sequential manner. This is called adaptive learning or infilling. In force field calibration, two approaches have been used to build metamodels of the likelihood function, either directly during the Bayesian inference algorithm (on-the-fly/synchronous learning) [32,34] or as a preliminary stage (asynchronous learning) [77].

Synchronous learning. In synchronous learning, the surrogate model is built in parallel with the sampling algorithm. The first samples of the sampling algorithm are produced by running the original model. After enough samples have been generated, a surrogate model is constructed. The sampling proceeds by using the surrogate or the exact model according to some criterion.

In Ref. [34], the synchronous learning approach was combined with the TMCMC sampling algorithm [37]. The algorithm is named K-TMCMC and the surrogate model used is kriging. In order to control the size of the error introduced by the approximation of the likelihood, the following rules have been used:

1. The training set consists only of points that have been accepted using the exact and not the surrogate model.
2. Given a point $\boldsymbol{\vartheta_c}$, the training set for the construction of the surrogate consists on the closest (in a chosen metric) n_{neigh} points. Moreover, $\boldsymbol{\vartheta_c}$ must be contained in the convex hull of those points. The parameter n_{neigh} is user-defined and its minimum value depends on the dimension of the sample space.

3. The estimate is checked to verify whether its value is within the lower 95% quantile of all posterior values of the points accounted so far with full model simulations.

4. The relative error $\sqrt{c_y(\boldsymbol{\vartheta_c},\boldsymbol{\vartheta_c})/\widehat{\boldsymbol{y}}(\boldsymbol{\vartheta_c})}$ should be less than a user-defined tolerance ε, see Eqs. (6.33) and (6.34).

The algorithm has been applied on a structural dynamics problem.

Asynchronous learning. It is also possible to build a metamodel of the likelihood before performing Bayesian inference. In this case, one starts with a small design, typically based on a Latin hypercube sample [88], and add new points while searching for the maximum of the likelihood function, or equivalently, for the minimum of $-\log p(\boldsymbol{D}|\boldsymbol{\vartheta},\boldsymbol{X},\boldsymbol{M})$ [77].

It has been shown that optimizing directly a metamodel $\widetilde{y}$ is not very efficient [89], the risk being of getting trapped in minima of $\widehat{y}$ resulting from its approximate nature. More reliable strategies have been defined for metamodel-based optimization: the *Efficient Global Optimization* (EGO) algorithm [89,90] and its variants [91–94]. The advantage of EGO is to provide an optimal infilling scheme, starting from sparse initial designs. In this context, metamodels such as low-order polynomials are too rigid to enable the discovery of new minima, and higher-order polynomial would require too large designs for their calibration. In contrast, kriging metamodels handle easily this issue and are generally associated with EGO [85,95,96].

EGO is initialized by computing the function y to be minimized for a sample of inputs $\Theta = \left(\boldsymbol{\vartheta}^{(\mathit{1})},\ldots,\boldsymbol{\vartheta}^{(N)}\right)$. A first GP $\widehat{y}$ is built that reproduces the value of $\boldsymbol{Y}$ for the design points. Outside of the design points, $\widehat{y}(\boldsymbol{\vartheta})$ is a prediction of $y(\boldsymbol{\vartheta})$, with an uncertainty $u_y(\boldsymbol{\vartheta})$. As $\widehat{y}$ is only an approximation of y, its optima do not necessarily coincide with those of y. The metamodel is thus improved by performing a new evaluation of y for a new parameter set $\boldsymbol{\vartheta}^{(N+\mathit{1})}$ that maximizes a utility function $\Gamma(\boldsymbol{\vartheta})$. This utility function measures the improvement of the metamodel expected upon the inclusion of $\boldsymbol{\vartheta}^{(N+\mathit{1})}$ into the sampling design. The process is iterated until $\max[\Gamma(\boldsymbol{\vartheta})]$ is below a user-defined threshold. At the end of the EGO, $\widehat{y}$ can be used as a good estimator of y, especially in the neighborhood of its minima.

In the original version of EGO [90], dedicated to the optimization of deterministic functions, $\Gamma(\boldsymbol{\vartheta})$ is the expected improvement (EI) defined as

$$EI(\boldsymbol{\vartheta}) = \mathbb{E}\left[\max\left(y(\boldsymbol{\vartheta}^*) - \widehat{y}(\boldsymbol{\vartheta}), 0\right)\right], \tag{6.36}$$

where $\boldsymbol{\vartheta}^*$ is the point of the sampling design for which y is minimum. When $\widehat{y}$ is a kriging metamodel, EI can be computed analytically which makes the search for $\boldsymbol{\vartheta}^{(N+\mathit{1})}$ very efficient. When dealing with the minimization of a noisy function (which is the case when y is computed from molecular simulation data), the relevance of EI as defined above is questionable [92]. Many adaptations of EGO have been proposed, that differ by the definition of the utility function $\Gamma(.)$, which consists of a trade-off between minimizing $\widehat{y}(\boldsymbol{\vartheta})$ and reducing its prediction uncertainty $u_y(\boldsymbol{\vartheta})$. For a recent review of the EGO variants adapted to noisy functions and their relative merits, the reader is referred to Ref. [93].

6.3.2.3 Polynomial Chaos expansions

A PCE is a spectral representation of a random variable [97–99]. Any real-valued random variable Y with finite variance can be expanded in terms of a PCE representation of the form

$$Y = \sum_{|I|=0}^{\infty} Y_I \Psi_I(\xi_1, \xi_2, \ldots), \tag{6.37}$$

where $\{\xi_i\}_{i=1}^{\infty}$ are i.i.d. standard random variables, Y_I are the coefficients, $I = (I_1, I_2, \ldots)\ \forall I_j \in \mathbb{N}_0$ is an infinite-dimensional multi-index, $|I| = I_1 + I_2 + \ldots$ is the ℓ_1 norm, and Ψ_I are multivariate normalized orthogonal polynomials. The PCE in Eq. (6.37) converges to the true random variable Y in the mean-square sense [99,100]. The basis functions can be written as products of univariate orthonormal polynomials as

$$\Psi_I(\xi_1, \xi_2, \ldots) = \prod_{j=1}^{\infty} \psi_{I_j}(\xi_j). \tag{6.38}$$

The univariate functions ψ_{I_j} are I_j-th order polynomials in the independent variable ξ_j orthonormal with respect to the probability density $p(\xi_j)$, i.e., they satisfy

$$\int \psi_{I_j}(\xi)\psi_{I_k}(\xi)\mathrm{d}p(\xi) = \delta_{jk}. \tag{6.39}$$

For instance, if the germ ξ_j is a standard Gaussian random variable, then the PCE is built using Hermite polynomials. Different choices of ξ_j and ψ_m are available via the generalized Askey family [100]. For computational purposes, the infinite dimensional expansion (Eq. 6.37) must be truncated

$$Y = \sum_{I \in \mathscr{I}} Y_I \Psi_I(\xi_1, \xi_2, \ldots, \xi_{n_s}), \tag{6.40}$$

where $\mathscr{I}$ is some index set, and n_s is some finite stochastic dimension that typically corresponds to the number of stochastic degrees of freedom in the system. For example, one possible choice for $\mathscr{I}$ is the total-order expansion of degree p, where $\mathscr{I} = \{I : |I| \leq p\}$.

To understand the applicability of a PCE to predictive modeling and simulation, assume that we have a target model of interest, $F(\vartheta)$, where ϑ is a single parameter.

The model yields a prediction for the quantity of interest $Q = F(\vartheta)$. If we expand the input ϑ like in (40), the PCE for a generic observable Q can then be written in a similar form

$$Q(\xi) = F(\vartheta(\xi)) = \sum_{I\in\mathscr{I}} c_I \Psi_i(\xi_1, \xi_2, \dots, \xi_{n_s}). \tag{6.41}$$

To compute the PC coefficients c_I with $I\in\mathscr{I}$, we can identify two classes of methods, namely intrusive and nonintrusive [99]. The former involves substituting the expansions into the governing equations and applying orthogonal projection to the resulting equations, resulting in a larger system for the PCE coefficients. This approach is applicable when one has access to the full forward model and can thus modify the governing equations. The nonintrusive approach is more generally applicable, because it involves finding an approximation in the subspace spanned by the basis functions by evaluating the original model many times. This nonintrusive approach basically treats the simulator for the forward model as a black-box, and it does not require any modification of the governing equations or the simulator itself.

One example of nonintrusive methods relies on orthogonal projection of the solution

$$c_I = \mathbb{E}[F(\vartheta)\Psi_I] = \int_\Xi F(\vartheta(\xi))\Psi_I(\xi)p(\xi)d\xi. \tag{6.42}$$

and is known as nonintrusive spectral projection (NISP). In general, this integral must be estimated numerically, using Monte Carlo or quadrature techniques [98,99]. Monte Carlo methods are insensitive to dimensionality, but it is well known that their convergence is slow with respect to the number of samples. For a sufficiently smooth integrand, quadrature methods converge faster, but they are affected by the curse of dimensionality. The number of dimensions that define the threshold for when the problem becomes unaffordable cannot be set a priori, and it is obviously problem dependent. However, one can guess that most physical problems of interest, due to their high computational cost, become intractable even for a small number of dimensions. Sparse grids can mitigate the curse of dimensionality, but they can lead to issues due to negative weights.

An alternative nonintrusive method is regression, which involves solving the linear system:

$$\underbrace{\begin{bmatrix} \Psi_{I^1}\left(\xi^{(1)}\right) & \cdots & \Psi_{I^K}\left(\xi^{(1)}\right) \\ \vdots & & \vdots \\ \Psi_{I^1}\left(\xi^{(K)}\right) & \cdots & \Psi_{I^K}\left(\xi^{(K)}\right) \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} c_{I^1} \\ \vdots \\ c_{I^K} \end{bmatrix}}_{c} = \underbrace{\begin{bmatrix} F\left(\vartheta\left(\xi^{(1)}\right)\right) \\ \vdots \\ F\left(\vartheta\left(\xi^{(K)}\right)\right) \end{bmatrix}}_{F}, \tag{6.43}$$

where Ψ_{J^n} is the n-th basis function, c_{J^n} is the coefficient corresponding to that basis, and $\xi^{(m)}$ is the m-th regression point. In the regression matrix A, each column corresponds to a basis element and each row corresponds to a regression point from the training set.

The NISP or a fully tensorized regression approach is suitable when the data being modeled are not noisy. In the presence of noisy data, a straightforward regression or NISP would yield a deterministic answer, thus losing any information about the noise. A suitable approach to tackle these problems is Bayesian inference to infer the coefficients, which allows one to capture the uncertainty in the data in a consistent fashion [101]. This approach is convenient also because it allows a certain flexibility in the sampling method of the stochastic space, since no specific rule is required a priori. An obvious constraint, however, is that the number of sampling points should be adequate to the target order of the expansion to infer. The result of the Bayesian approach is an *uncertain* PC representation of a target observable, i.e., the vector of PC coefficients is not a deterministic vector, but it is a random vector defined by a joint probability density containing the full noise information. The Bayesian method to infer the PC coefficients consists of three main steps: collecting a set of the observations of the QoIs, formulating the Bayesian probabilistic model, and, finally, sampling the posterior distribution.

When collecting the observations, one should choose wisely the sampling points over the space. A possible option would be to use the same sampling points one would use for the NISP approach, but this would constrain how to choose the points. One possibility that would help with the curse of dimensionality is to use nested grids, which would benefit approaches like adaptive refinement. One example of this class of points is Fejér nodes [102]. Leveraging the nested nature of these grids, one can explore an adaptive sampling technique to build the target set of observations of the QoIs. Further details of this approach will be discussed below in Section 6.4.2.1.

Once a set of observations $\boldsymbol{F}$ is obtained, one needs to formulate the likelihood for the coefficients c_I with $I \in \mathscr{J}$. Using a standard Gaussian additive model to capture the discrepancy between each data point, f_i, and the corresponding model prediction yields the well-known Gaussian likelihood and the following the joint posterior distribution, of the PC coefficients and noise variance as

$$p(c_I|\boldsymbol{F}, M) \propto p(\boldsymbol{F}|c_I, M)p(c_I),$$

with $I \in \mathscr{J}$, and $p(c_I)$ denoting the prior on the PC coefficients. Once a proper prior distribution is chosen, sampling from this high-dimensional posterior can be performed using, e.g., MCMC methods.

6.3.3 Approximation of intractable posterior PDFs

The Laplace method. This is a technique for the approximation of integrals of the form

$$\int_\Theta \mathrm{e}^{Nf(\vartheta)} d\boldsymbol{\vartheta}.$$

The technique is based on approximating the integrand with a Taylor expansion around the unique maximum of the function *f*. The approximation is valid for large values of *N*.

The same idea can be applied for the approximation of intractable posterior distributions. We expand the logarithm of the posterior distribution, denoted by $\mathscr{L}(\boldsymbol{\vartheta})$, around the MAP estimate of Eq. (6.8), $\widehat{\boldsymbol{\vartheta}}$,

$$\mathscr{L}(\boldsymbol{\vartheta}) \approx \mathscr{L}\left(\widehat{\boldsymbol{\vartheta}}\right) + \frac{1}{2}\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right)^{\top} \nabla\nabla^{\top} \mathscr{L}\left(\widehat{\boldsymbol{\vartheta}}\right)\left(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}\right). \tag{6.44}$$

The posterior distribution is then approximated by

$$p\left(\boldsymbol{\vartheta}|\boldsymbol{D}, M\right) \approx c\left(\widehat{\boldsymbol{\vartheta}}\right)\mathscr{N}\left(\boldsymbol{\vartheta}|\widehat{\boldsymbol{\vartheta}}, \Sigma\left(\widehat{\boldsymbol{\vartheta}}\right)\right), \tag{6.45}$$

where

$$c\left(\widehat{\boldsymbol{\vartheta}}\right) = \left((2\pi)^N\left|\Sigma\left(\widehat{\boldsymbol{\vartheta}}\right)\right|\right)^{-\frac{1}{2}} p\left(\widehat{\boldsymbol{\vartheta}}\middle|M\right) \tag{6.46}$$

is the normalization constant and

$$\Sigma\left(\widehat{\boldsymbol{\vartheta}}\right) = \left(\nabla\nabla^{\top}\mathscr{L}\left(\widehat{\boldsymbol{\vartheta}}\right)\right)^{-1} \tag{6.47}$$

is the inverse of the Hessian of $\mathscr{L}$ at $\widehat{\boldsymbol{\vartheta}}$. The methodology has been applied in Ref. [103] for the approximation of the posterior distribution in a hierarchical Bayesian model (see Section 6.2.3.2).

Variational inference. This is another method for the approximation of intractable posterior distributions. We define a family $\mathscr{D}$ of densities over the space $\boldsymbol{\Theta}$ of parameters, e.g., exponential functions, GPs, neural networks. From this family, we choose the member that is closest to the posterior distribution based on the Kullback—Leibler (or relative entropy) divergence,

$$\widehat{q} = \underset{q \in \mathscr{D}}{\text{argmax}} D_{KL}(q(\,\cdot\,) \parallel p(\,\cdot\,|\boldsymbol{D}, M)), \tag{6.48}$$

D_{KL} where the *KL* divergence is defined by

$$D_{KL}(q(\,\cdot\,) \parallel p(\,\cdot\,|\boldsymbol{D}, M)) = \int_{\Theta} \log\frac{q(\boldsymbol{\vartheta})}{p(\boldsymbol{\vartheta}|\boldsymbol{D}, M)} q(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \tag{6.49}$$

which can be also written as

$$D_{KL}(q(\cdot) \parallel p(\cdot|\boldsymbol{D},M)) = -(\mathbb{E}_q[\log(p(\boldsymbol{D}|\boldsymbol{\vartheta},M)p(\boldsymbol{\vartheta}|M)) - \mathbb{E}_q[\log(q(\boldsymbol{\vartheta}))]) + \log p(\boldsymbol{D}|M). \tag{6.50}$$

The last term, $\log p(\boldsymbol{D}|M)$, is intractable but does not depend on q and thus can be ignored in the optimization process. An usual assumption on the family $\mathscr{D}$ is that it contains product functions,

$$q(\boldsymbol{\vartheta}) = \prod_{i=1}^{N_\vartheta} q_i(\boldsymbol{\vartheta}_i). \tag{6.51}$$

This special case is called *mean-field* inference. One way to solve the optimization problem in this case is by coordinate ascent. At the k-th step of the iteration, we solve the problems

$$\widehat{q}_j^{k+1} = \underset{q_j^{k+1}\in \mathscr{D}_i}{\operatorname{argmax}} \; D_{KL}\left(q_1^{k+1},\ldots,q_{j-1}^{k},q_j^{k+1},q_{j+1}^{k},\ldots,q_{N_\vartheta}^{k+1} \parallel p(\,\cdot\,|\boldsymbol{D},M)\right), \tag{6.52}$$

for $j = 1,\ldots,N_\vartheta$.

In Ref. [104], the variational approach has been applied for the inference of the forcing parameters and the system noise in diffusion processes. In Ref. [105], the authors applied this approach in the path-space measure of molecular dynamics simulations, in order to obtain optimized coarse-grained molecular models for both equilibrium and nonequilibrium simulations.

6.3.4 High-performance computing for Bayesian inference

Bayesian inference can be computationally costly, in particular when MCMC methods are used to sample the posterior distribution. This cost is further exacerbated when the model evaluations are expensive, which is often the case with large-scale molecular simulations. Performing such high throughput simulations in massively parallel HPC architectures may offer a remedy for the computational cost. However, this implementation introduces new challenges as MD simulations performed with different sets of parameters, each distributed on different nodes, may exhibit very different execution times. In such cases, the sampling of the posterior PDF will result in load imbalance as nodes who have completed their tasks would be idling, thus reducing the parallel efficiency of the process. One remedy in this situation is the introduction of task-based parallelism.

A number of frameworks have been proposed to address this situation. PSUADE [106], developed in Lawrence Livermore National Laboratory, is an HPC software written in C++ for UQ and sensitivity analysis. VECMA [107] is a multipartner project under development that aims at running UQ problems on exascale environments. Another relevant software is SPUX [108], developed in EAWAG, Switzerland, aimed for UQ in water research written in Python. The CSE Laboratory at ETHZ has developed Korali, a framework for nonintrusive Bayesian UQ of complex and computationally demanding physical models on HPC architectures.[1] Korali builds on the framework Π4U [76] where Bayesian tools were expressed as engines upon the layer of the TORC [109] tasking library. TORC works by defining a set of workers, each spanning multiple processor cores, distributed throughout a computing cluster and assigning them a fixed set of model evaluations as guided by the stochastic method. However, TORC has two drawbacks: (i) its design is very strongly coupled with each UQ method that requires a problem-specific interface and (ii) its fixed work distribution strategy can cause load imbalance between the worker teams, causing cores to idle. In turn, the parallel implementation of Korali follows a producer—consumer paradigm to distribute the execution of many model evaluations across a computer cluster.

Upon initialization, the Korali runtime system instantiates multiple workers, each comprising one or more processor cores. During execution, Korali keeps workers busy by sending them a new batch of model evaluations to compute. In turn, as soon a worker finishes an evaluation, it returns its partial results to Korali's runtime system. Korali prevents the detrimental effects of load imbalance, where a worker may finish before others and thus remaining idle, by distributing small work batches at a time. Communication between Korali and its worker tasks is entirely asynchronous. That is, each worker will communicate partial results to the runtime system without a reciprocal receive request from the latter. Korali's runtime system employs remote procedure calls (RPCs) to handle worker requests opportunistically, eliminating the need for barriers or other synchronization mechanisms. Korali uses the UPC++ communication library [110] as its back-end for RPC execution.

Finally, Korali allows users to use their own code to simulate the computational model (e.g., a computational fluid dynamics simulation) to perform model evaluations by providing a simple C/C++/Python function call interface. The user code can embrace inter and intranode parallelism expressed with MPI communication, OpenMP directives, or GPUs (e.g., via CUDA).

6.4 Applications

Applications of Bayesian methods to force fields calibration are still few, and mostly focused on problems with a small number of parameters (typically less than 10): simple potentials (LJ, Mie) and coarse-grained potentials [111—113].

In this section, we present several applications at various levels of complexity in order to display the extent of successes and difficulties arising in this area.

[1] https://cselab.github.io/korali/.

6.4.1 Introductory example: two-parameter Lennard-Jones fluids

6.4.1.1 The Lennard-Jones potential

LJ fluids are one of the simplest systems studied in molecular simulation. They are considered a good model (qualitatively and for some applications quantitatively) for rare gases (Ar, Kr) or small "spherical" molecules (CH_4, N_2, etc.). In LJ fluids, two particles, separated by a distance r, interact according to the LJ potential:

$$V_{LJ}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]. \tag{6.53}$$

The LJ potential is the sum of two terms: a r^{-12} term repulsive at short distance and a r^{-6} attractive term. The two parameters, σ and ε, that control this interaction have a physical understanding: σ is related to the size of the particle (its "radius"), whereas ε controls the energetic strength of the interaction. The use of the LJ potential is not limited to the study of monoparticular systems and is indeed one of the most commonly encountered term in common force fields to represent non-Coulombic interactions between two atoms in molecular systems. Note that the LJ potential is sometimes used in the framework of dimensionless units for the sake of generality. However, when one needs to study real fluids, one has to manipulate dimensioned properties in order to compare to experimental data.

The LJ potential has been the major target for force field calibration within the Bayesian framework [17,30,32,42,52,69,114]. This is due to its simplicity, with only two parameters to be calibrated, and to the fact that some experimental properties for LJ fluids (second virial coefficient, diffusion coefficient, etc.) can be obtained using analytical formulae. This enables (a) to get the likelihood "for free" and thus to perform a thorough exploration of the parameter space without the need to use advanced methodologies and (b) to get rid of the effects of simulation parameters, such as the cut-off radius [32], and of simulation numerical/measurement uncertainty in the calibration process ($u_{F_i} = 0$ in Eq. 6.4).

6.4.1.2 Bayesian calibration

As an illustration of the theoretical points presented in the previous sections, we calibrated the LJ potential for Ar over three different phase equilibrium properties (saturated liquid and vapor densities ρ_{liq} and ρ_{vap}, and saturated vapor pressure P_{sat}) at 12 temperatures ranging from 90 to 145 K. Calibration data were recovered from the NIST website [40], and property computation is made through the use of Eqs. (6.9)–(6.11) of Werth et al. [83], with $n = 12$. The NIST website reports upper limits on uncertainty for densities of 0.02% and 0.03% for pressures. The exact meaning of these uncertainties on equations-of-state results has been discussed earlier in this chapter. However, they are much smaller than the modeling errors expected from an LJ model, so that we do not need to be concerned by possible correlations. In all the following calibrations, uniform priors have been used for the LJ parameters σ and ε.

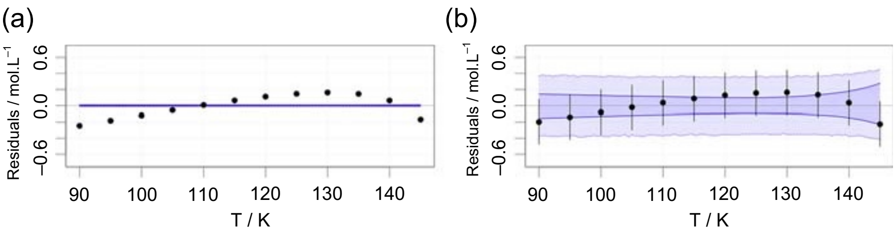

Figure 6.3 Calibration of LJ parameters for Ar over ρ_{liq} by the standard Bayesian scheme: (a) using fixed reference uncertainties or (b) calibrating uncertainty τ along with σ and ε to deal with model inadequacy. The points show the residuals at the MAP with the error bars representing 2τ (in (a), the error bars are smaller than the point size). The shaded areas represent posterior confidence (dark) and prediction (light) 95% probability intervals.

Fig. 6.3(a) shows the residuals of the fit at the MAP for calibration over ρ_{liq}, with a likelihood based on Eq. (6.7), where $\boldsymbol{\Sigma}_R$ is a diagonal matrix with elements $\Sigma_{R,ij} = u_{d_i}^2 \delta_{ij}$, with reference data uncertainties u_{d_i} set to the NIST recommended values. The calibration is clearly unsuccessful. Although the residuals are "small" (roughly 1% of the QoI values), they fall outside of the 95% prediction probability intervals and are clearly correlated. These two observations are symptomatic of model inadequacy: the physics contained in the LJ potential is not sufficient to enable accurate estimation of Ar-saturated liquid density over this range of temperature.

We will discuss here a simple way to overcome this limitation which is to consider $\Sigma_{R,ij} = \tau^2 \delta_{ij}$, where τ is a parameter to be calibrated along with σ and ε, in order to ensure a Birge ratio at the MAP close to 1 (Eq. 6.13). Note that τ is given a half-normal prior PDF with a large standard deviation of 0.5, in order to ensure its positivity and to disfavor unsuitably large values. The residuals of the calibration are displayed in Fig. 6.3(b). The values of parametric uncertainties (given in Table 6.1) are now consistent with the magnitude of the residuals.

Note that the residuals are still correlated due to model inadequacy, so that the calibration should not in principle be considered successful. However, for pedagogical

Table 6.1 Results of the calibration of an LJ force field for Argon.

Calibration data	σ (Å)	ε (K)	$\tau_{\rho_{liq}}$(mol.L^{-1})	$\tau_{\rho_{vap}}$(mol.L^{-1})	$\tau_{P_{sat}}$(MPa)
ρ_{liq}	3.40(1)	115.3(3)	0.17(4)		
ρ_{vap}	3.73(4)	111.8(4)		0.09(2)	
P_{sat}	3.45(3)	115.7(6)			0.009(2)
Consensus	3.41(1)	116.5(1)	0.26(6)	0.24(6)	0.010(3)
Hierarchical	3.53(15)	114.2(2.1)	0.12(2)	0.07(2)	0.009(2)

The mean values of the parameters are given, as well as their marginal uncertainties in parenthetic notation.

purposes, we will ignore this issue for now and postpone discussion about advanced schemes to deal with model inadequacy to Section 6.4.3.

Results of the calibration over ρ_{vap} or over P_{sat} are reported in Table 6.1 and Fig. 6.4. Values obtained for the parameters are quite different, both in terms of mean values and uncertainties (see Table 6.1). Fig. 6.4 shows posterior PDF samples for σ and ε obtained from the three calibrations. The three PDFs do not overlap, the consequence being that it is not possible to reproduce quantitatively a property on which the parameters have not been calibrated, taking into account parameters uncertainties. This is true also when model errors are taken into account, as is done here through the calibration of $\boldsymbol{\tau}$ values. This has already been shown in previous studies [17,35,59]. The very different values obtained in the calibration using ρ_{vap} or ρ_{liq} certainly originate from the fact that they deal with different states of matter: the two-body nature of the LJ potential is not appropriate to deal with condensed phase and its success relies on incorporation of many-body interactions into the LJ parameters, leading to values much different than those in gas phase. The choice of the reference data is thus of paramount importance when calibrating force field parameters, because their values will integrate, in an effective way, part of the physics not described in the mathematical expression of the force field.

Fig. 6.4 also displays a sample of the posterior PDF of LJ parameters calibrated over the three observables together, using three different uncertainty parameters $\boldsymbol{\tau} = \left\{\tau_{\rho_{liq}}, \tau_{\rho_{vap}}, \tau_{P_{sat}}\right\}$. This consensus situation leads to values of the parameters that are dominated by ρ_{liq} and P_{sat}, which is certainly due to the very high sensitivity of ρ_{liq} to σ. This can be acknowledged by the very small uncertainty on σ obtained from the calibration over ρ_{liq}.

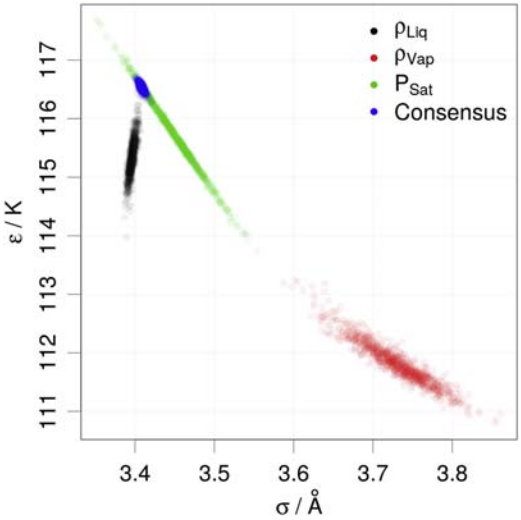

Figure 6.4 Samples of the posterior PDF of LJ parameters for Argon, calibrated on saturated liquid density (black), saturated vapor density (red), saturated vapor pressure (green), or using the three observables simultaneously as calibration data (blue).

It is interesting to note that the parameter uncertainties are greatly reduced in the "consensus" calibration, as can be seen in Fig. 6.4 and Table 6.1. This is balanced by an increase of $\tau_{\rho_{liq}}$ and $\tau_{\rho_{vap}}$, ensuring that the calibration remains statistically sound (when considering only the Birge ratio). As mentioned earlier, adding data to the reference set decreases the quality of the fit (larger residuals and values of τ_i), but simultaneously decreases the uncertainty on the parameters, and by consequence the uncertainty of model predictions. This unsuitable behavior justifies the development of more advanced statistical formulations to deal with model inadequacy.

Fig. 6.4 shows a strong correlation between σ and ε, whatever calibration data are used (although the correlation is not identical in all cases). Such information is crucial when one wants to perform UP. It is thus very important when reporting force field parameter uncertainties not only to limit to the sole marginal uncertainty on each parameter but also to include their correlation, or provide the full posterior PDF.

6.4.1.3 Hierarchical model

Calibration. A hierarchical model (see Section 6.2.3.2) was built to calibrate a unique set of LJ parameters on the three observables used above.

To define the overall PDF with a bivariate normal distribution, one has to calibrate five parameters (two for the center of the distribution, two for the uncertainties, and one for correlation) on six values (the coordinates of the three local LJ parameters). To formulate it differently, the question is "From which bivariate Gaussian distribution are these three points sampled?". One has thus to expect a large statistical uncertainty on the overall distribution. To compensate for the sparse data, one introduces an exponential prior on the uncertainties to constrain them to be as small as possible. A few tests have shown that a too strong constraint affects the local LJ parameters, resulting in a bad fit of ρ_{vap}. The solution shown in Fig. 6.5(a) is the best compromise (i.e., the most compact overall distribution preserving reasonable local LJ parameters) we were able to achieve. The parameters are reported in Table 6.1 for comparison with the standard calibration. The correlation parameter is poorly identified −0.56(37).

Prediction. The overall distribution was used to predict a new property, the second virial coefficient, using the analytical formula by Vargas et al. [115]. The 95% confidence interval is represented in Fig. 6.5(b) in comparison to experimental data [116]. The mean prediction is in very good agreement with the reference data up to 350 K. However, the prediction uncertainty is much larger than the experimental ones. For instance, at 200 K, the prediction uncertainty is more than 10 times larger than the experimental one (0.0040 vs. 0.0003 $\mathrm{L.mol^{-1}}$). Although statistically consistent, hierarchical calibration leads, in this case, to predictions that may not be precise enough to replace experiments.

6.4.1.4 Uncertainty propagation through molecular simulations

UP is typically done by drawing samples from the parameters posterior PDF and computing quantities of interest with the model for each parameter set value. Following this procedure for the LJ parameters for Argon (calibrated over second virial

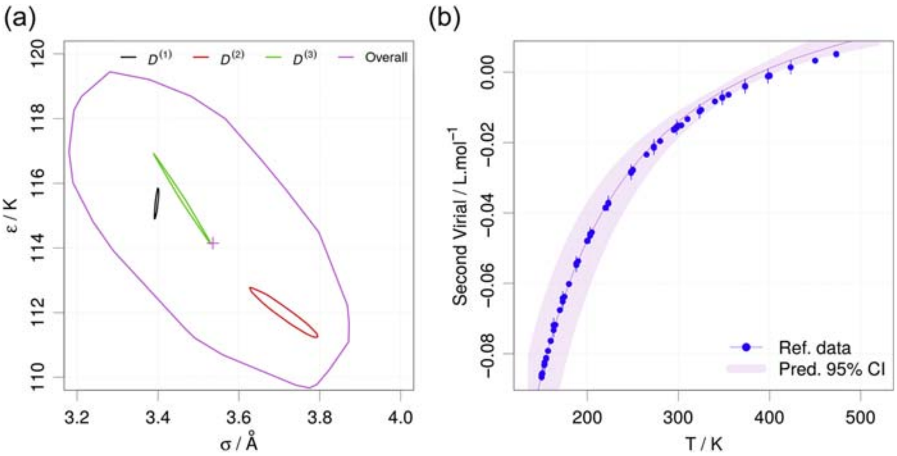

Figure 6.5 Calibration prediction with a hierarchical model: (a) 95% probability contour lines of the posterior PDF of LJ parameters for Ar: local parameters for saturated liquid density ($D^{(1)}$; black), saturated vapor density ($D^{(2)}$; red), saturated vapor pressure ($D^{(3)}$; green), and overall distribution (purple); (b) Comparison of the predicted second virial coefficients (purple) to experimental data (blue).

coefficient data), Cailliez and Pernot computed uncertainties obtained from MD and MC simulations [17]. They showed that molecular simulations amplify parametric uncertainties. Although relative uncertainties on σ and ε were around 0.1%, the output uncertainties were roughly 0.5%–2%, depending on the computed property. More importantly, when decomposing the uncertainties into their numerical (sum of computational and measurement uncertainties as defined in the Introduction) and parametric components, they observed that the latter was the most significant part of the total uncertainty budget. Similar observations have since been made in the literature in the case of other force fields [59,77]. This means that parametric uncertainties should not be overlooked anymore when reporting molecular simulation results.

6.4.1.5 Model improvement and model selection

As described earlier, the LJ potential is unable to reproduce various QoIs with a unique parameter set or even, in some cases, one quantity over a large range of physical conditions. One way to overcome this limitation is to improve the physical description of the interactions by modifying the form of the force field (model improvement). A "simple" improvement of the LJ potential is to add a new parameter to be calibrated. A straightforward choice is the repulsive exponent p (the value of 12 in the LJ potential has no physical grounds). The modified LJ potential (hereafter referred to as LJ_p) becomes

$$V_{LJ_p}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^p - \left(\frac{\sigma}{r}\right)^6\right]. \tag{6.54}$$

The LJ$_p$ potential has been calibrated by the standard scheme for each of the three observables from the previous sections using the metamodel of Werth et al. [83]. Calibration summaries are presented in Table 6.2 and samples of the posterior PDFs are presented in Fig. 6.6. One sees that the situation is better for the densities, which are close to having an intersection but they do not overlap in the three parameter dimensions and there is no reconciliation with P_{Sat}. The additional parameter is not sufficient to offer a complete data reconciliation.

Fit quality. The improvement of fit quality of the new potential can be assessed by comparing the optimized uncertainty parameters τ_i for each QoI in Tables 6.1 and 6.2. For ρ_{liq} the value has diminished from 0.17 to 0.10 mol L^{-1}, indicating a reduction of almost a factor two of the residuals amplitude. For ρ_{vap} the effect is even larger, from 0.09 to 0.02 mol L^{-1}, whereas the fit quality of P_{sat} has not been significantly affected.

Transferability. It also should be noted that the spread of the PDFs is sometimes much larger for σ and ε than for the LJ potential and ε has shifted to much larger values, which indicates that the LJ$_p$ potential might be overparameterized for ρ_{Liq}. The additional parameter therefore improves the representation of the densities (through the metamodel used here) and helps to reconcile them. There remains, however, the impossibility to fit the three QoIs with a single set of these three parameters.

Fitting of LJ$_p$ on the radial distribution functions. Kulakova et al. [35] calibrated LJ and LJ$_p$ potentials in order to reproduce RDFs of argon in six different conditions: five liquid states at different temperatures and pressures (labeled L_1 to L_5) and one vapor state (labeled V). Some of the results from Ref. [35] are reproduced in Table 6.3.

As seen previously, adding one degree of freedom in the potential enables better representation of experimental data: the calibration of the LJ$_p$ potential succeeded for two experimental conditions (L1 and L2) in which the LJ potential failed. Here again, the best values of the LJ parameters are modified when p is optimized. This is due to the fact that the best value for p is in all cases very different from 12 (between 6.15 and 9.84). This is especially true for ε, due to a strong anticorrelation between parameters ε and p observed in this study as well as in the previous example and in Ref. [83].

The added value of the force field improvement can be measured using the model selection criterion described in Section 6.2.1.3. Using equal a priori probabilities for models LJ and LJ$_p$, the preferred model is the one exhibiting the largest evidence. These are reported in Table 6.3. In all liquid conditions, the LJ$_p$ model is shown to be a valuable improvement over the LJ potential. However, in vapor conditions, the result was opposite. In such conditions the LJ$_p$ potential is overparameterized.

Although both examples considered above reach similar conclusions on the study of the LJ$_p$ potential, one can see in the details that the model used (metamodel vs. simulation) and the choice of calibration QoIs have a strong impact on the numerical values of the parameters. For the metamodel of Werth calibrated on ρ_{liq}, ρ_{vap}, and P_{sat}, the optimal values of p lie between 19 and 27, while for simulations calibrated on RDF these lie between 6.15 and 9.84. Another study on Ar, with different model and data, reached the conclusion that the best value for p was close to 12 [117].

Table 6.2 Results of the calibration of an LJ$_p$ force field for Argon.

Calibration data	**σ (Å)**	**ε (K)**	***p***	$\tau_{\rho_{liq}}$**(mol.L^{-1})**	$\tau_{\rho_{vap}}$**(mol.L^{-1})**	$\tau_{P_{sat}}$**(MPa)**
ρ_{liq}	3.51(3)	162.1(9.1)	26.4(4.5)	0.10(3)		
ρ_{vap}	3.48(2)	145.0(1.3)	19.5(4)		0.02(1)	
P_{sat}	4.08(19)	136.6(5.3)	21.7(3.4)			0.007(2)

The mean values of the parameters are given, as well as their marginal uncertainties in parenthetic notation.

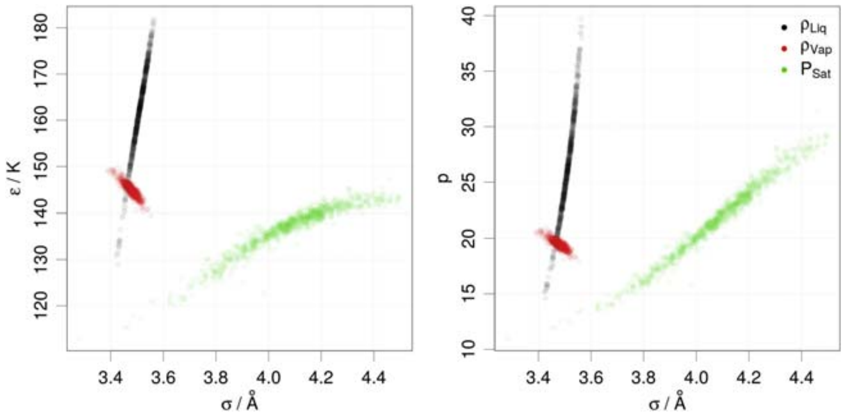

Figure 6.6 Projections of samples of the posterior PDF of the three-parameter LJ$_p$ model, calibrated on saturated liquid density (in black), saturated vapor density (in red), and saturated vapor pressure (in green).

Table 6.3 Results of the calibration of LJ and LJ$_p$ force field for Argon on radial distribution function.

Experimental data	Force field model	$q(\sigma)$(Å)	$q(\varepsilon)$(K)	$q(p)$	log(evidence)
L_2	LJ	–	–	[6.29, 8.31]	–
	LJ$_p$	[3.32, 3.43]	[358.3, 2222.2]		−14.8
L_3	LJ	[3.25, 3.33]	[142.9, 133.9]	[6.33, 7.07]	−9.72
	LJ$_p$	[3.30, 3.40]	[459.9, 1358.7]		2.81
L_4	LJ	[3.30, 3.36]	[127.8, 133.9]	[6.90, 9.84]	5.10
	LJ$_p$	[3.33, 3.40]	[162.0, 533.4]		5.18
V	LJ	[3.03, 3.21]	[41.8, 92.6]	[6.15, 6.92]	−3.83
	LJ$_p$	[3.04, 3.12]	[314.5, 1817.1]		−4.94

For each force field parameter, the 5%−95% quantiles are given in brackets.
Values are taken from L. Kulakova, G. Arampatzis, P. Angelikopoulos, P. Hadjidoukas, C. Papadimitriou, P. Koumoutsakos, Data driven inference for the repulsive exponent of the Lennard-Jones potential in molecular dynamics simulations, Sci. Rep. 7 (2017) 16576. https://doi.org/10.1038/s41598-017-16314-4.

6.4.1.6 Summary

Exploration of the calibration of an LJ potential for Argon, one of the simplest problems in the field, has revealed several difficulties. Mainly the excessive simplicity of this popular potential function leads to model inadequacy, notably for condensed phase conditions. The Bayesian approach provides an unambiguous diagnostic for such problems, that would be difficult to assess with calibration methods ignoring parametric uncertainty. Bayesian inference also offers advanced methods (e.g., hierarchical modeling (HM)) to treat model inadequacy; however, we have seen that the solutions might be of a limited practical interest due to very large prediction uncertainties. The reduction of model inadequacy by improvement of the interaction potential is clearly a necessary step if one wishes to establish predictive models for multiple QoIs. Apparently, the LJ_p model, although a notable improvement over LJ, is not able to solve all the inadequacy problems.

6.4.2 Use of surrogate models for force field calibration

For most force fields and properties, no analytical formulae are available, and one has to resort to molecular simulations in order to compute properties to calibrate the force field parameters. For extensive exploration of parameter space, metamodels are needed to overcome the prohibitive cost of simulations. This section presents examples of PCE- and GP-based metamodels.

6.4.2.1 Polynomial Chaos expansions

An example of using PCE as surrogate models for calibrating force field parameters can be found in Refs. [84,118]. The work focuses on isothermal, isobaric MD simulations of water at ambient conditions, i.e., $T = 298$K and $P = 1$ atm. In the first part, Rizzi et al. describe the forward problem, i.e., how to build a PC surrogate for target macroscale quantities of interest comparing a NISP against a Bayesian approach. The latter, as discussed, is more suited for noisy data because it yields an expansion capturing both the parametric uncertainty stemming from the force field parameters as well as the sampling noise inherent to MD computations. In the second part, Rizzi et al. show how to use these PC expansions as a surrogate model to infer small-scale, atomistic parameters, based on data from macroscale observables.

MD system and uncertain parameters. In this example, the computational domain is a cubic box of side length equal to 37.2 Å with periodic boundary conditions along each axis, containing 1728 water molecules. The water molecule is modeled using the TIP4P representation [119]. The potential is a combination of an LJ component modeling dispersion forces, and Coulomb's law to model electrostatic interactions. The complete force field is defined by seven parameters: three defining the geometry of the water molecule, two for the charges of the hydrogen (H) and oxygen (O) atoms, and, finally, the two parameters for the LJ parameters defining the dispersion forces between molecules. Data for three target observables, namely density (ρ), self-diffusivity (ϕ), and enthalpy (H), are collected during the steady-state part of each run via time averaging. The goal is to characterize how intrinsic sampling noise and uncertainty affecting a subset of the input parameters influence the MD predictions for the selected observables.

The study is based on a stochastic parametrization of the LJ characteristic energy, ε, and distance, σ, as well as the distance, d, from the oxygen to the massless point where the negative charge is placed in the TIP4P model as

$$\begin{aligned}\varepsilon(\xi_1) &= 0.147 + 0.043\xi_1, \quad \text{kcal/mol},\\ \sigma(\xi_2) &= 3.15061 + 0.021\xi_2, \quad \text{Å},\\ d(\xi_3) &= 0.14 + 0.035\xi_3, \quad \text{Å},\end{aligned} \tag{6.55}$$

where $\{\xi_i\}_{i=1}^{3}$ are independent and identically distributed (i.i.d.) standard random variables (RVs) uniformly distributed in the interval $(-1,1)$. This reformulation reflects an "uncertain" state of knowledge about these parameters and is based on means and standard deviations extracted from the following sources: [119−123]. All the remaining parameters are set to their values commonly used for computational applications of TIP4P water, see, e.g., Refs. [119,120].

Collection of observations. Given N points $\{\xi^i\}_{i=1,\dots,N}$ in the parameter space $(-1,1)^3$, and considering four realizations of the MD system at each sampling point, the three sets of $N \times 4$ observations (one for each observable ρ, ϕ, H) can be written as

$$D_k = \left\{d_k^{i,j}\right\}_{i=1,\dots,N}^{j=1,\dots,4}, \quad k = 1,2,3, \tag{6.56}$$

where $\left\{d_k^{i,j}\right\}^{j=1,\dots,4}$ represents the four values obtained for the k-th observable at the i-th sampling point $\xi^i = \left(\xi_1^i, \xi_2^i, \xi_3^i\right)$. The authors demonstrate how to leverage nested Fejér grids [102] to sample the stochastic space. In one dimension, each grid level, l', is characterized by $n_{l'} = 2^{l'} - 1$ points in the interval $(-1,1)$, corresponding to the abscissae of the maxima of Chebyshev polynomials of different orders. Extensions to higher-dimensional spaces can be readily obtained by tensorization. Leveraging the nested nature of these grids, an adaptive technique is used to build a set of observations of the quantities of interest for the Bayesian inference. More specifically, at a given level l'', the density of sampling points is increased only in the regions of the domain where the "convergence" of the PC expansions inferred at levels $l' < l''$ is slower. Compared to fully tensored grids, this yields a considerable reduction in the computational cost without penalizing the accuracy.

The likelihood for the PC coefficients of the observables of interest is formulated by expressing the discrepancy between each data point, $d_k^{i,j}$, and the corresponding model prediction, $F_k(\xi^i)$, as

$$d_k^{i,j} = F_k(\xi_i) + \gamma_k^{i,j}, k = 1,2,3, i = 1,\dots,N, j = 1,\dots,4, \tag{6.57}$$

where $d_k^{i,j}$ represents the j-th data point obtained for the k-th observable at the i-th sampling point, ξ^i, $F_k(\xi^i)$ denotes the value of the PC representation of the k-th observable evaluated at ξ^i, and $\gamma_k^{i,j}$ is a random variable capturing their discrepancy. Based on central limit arguments one can argue that, in the present setting, as the number of atoms in the system and the number of time-averaged samples become large, the distribution of $\mathbf{D}_k$, $k = 1,2,3$, around the true mean tends to a Gaussian. A suitable and convenient choice is to assume each $\gamma_k^{i,j}$ to be *i.i.d.* normal RVs with mean zero and variance $\widetilde{\sigma}_k^2$, i.e., $\gamma_k^{i,j} \sim \mathcal{N}(0, \widetilde{\sigma}_k^2)$, $k = 1,2,3$, $i = 1,\ldots,N$, $j = 1,\ldots,4$. The variances $\{\widetilde{\sigma}_k^2\}_{k=1}^3$ are treated as hyperparameters, i.e., they are not fixed but become part of the unknowns.

These considerations yield the likelihood function:

$$p\left(D_k \middle| \left\{c_l^{(k)}\right\}_{l=0}^P, \widetilde{\sigma}_k^2, F_k\right) = \prod_{i=1}^N \prod_{j=1}^4 \frac{1}{\sqrt{2\pi\widetilde{\sigma}_k^2}} \exp\left(-\frac{\left[d_k^{i,j} - F_k(\xi^i)\right]^2}{2\widetilde{\sigma}_k^2}\right),$$

$$k = 1,2,3, \tag{6.58}$$

where $d_k^{i,j}$ is the j-th observation obtained at the i-th sampling point ξ^i for the k-th observable, while $F_k(\xi^i)$ denotes the value of the PC representation of the k-th observable evaluated at the i-th sampling point ξ^i, and $\left\{c_l^{(k)}\right\}_{l=0}^P$ is the set of PC coefficients for the k-th observable. The set has $P+1$ terms, based on total order, where $P+1 = (3+p)!/3!/p!$, where p is the target order of the expansion. For instance, $p = 1$ for a linear expansion, $p = 2$ for a quadratic, etc.

Using Bayes' theorem, the joint posterior distribution of the PC coefficients and noise variance for the k-th observable can be expressed as

$$p\left(\left\{c_l^{(k)}\right\}_{l=0}^P, \widetilde{\sigma}_k^2 \,|D_k, F_k\right) \propto p\left(D_k \middle| \left\{c_l^{(k)}\right\}_{l=0}^P, \widetilde{\sigma}_k^2, F_k\right) p(\widetilde{\sigma}_k^2) \prod_{l=0}^P \widehat{p}_k\left(c_l^{(k)}\right),$$

$$k = 1,2,3, \tag{6.59}$$

where $p(\widetilde{\sigma}_k^2)$ and $\widehat{p}_k\left(c_l^{(k)}\right)$ denote the presumed independent priors of the noise variance and the l-th PC coefficient, respectively. Uniform priors are assumed on the coefficients, and a posterior sampling algorithm based on MCMC is employed.

Results. We illustrate a strategy based on adaptive selection of sampling points by analyzing, for a given observable, the convergence of the associated PC expansions at successive approximation levels. The idea behind this method can be summarized in two steps. Firstly, we infer, for each observable, the corresponding PC expansion at the resolution levels $l' = 1$ and $l' = 2$ using the *full* grids of Fejér points. Secondly, rather than using the full grid also at level $l' = 3$, we select only a subset of nodes by identifying the regions of the domain where the differences between the expansions obtained at levels $l' = 1$ and 2 exceed a target threshold. This approach can be extended to higher-order levels $l' > 3$.

Specifically, we rely on the difference

$$Z_k^{(l'=1,2)}(\xi) = \left| F_k^{(l'=1,p=0)}(\xi) - F_k^{(l'=2,p=2)}(\xi) \right|, \quad k = 1, 2, 3, \tag{6.60}$$

where $F_k^{(l'=1,p=0)}$ represents the zeroth-order ($p = 0$) expansion of the k-th observable inferred at level 1 (level 1 includes, in fact, a single sampling point, so one can only build a zero-order PC representation), while $F_k^{(l'=2,p=2)}$ represents the second-order ($p = 2$) expansion of the k-th observable inferred at level 2 (quadratic is the maximum order for a well-posed problem at level 2). Also, rather than developing the analysis using the full joint posterior of the coefficients, we simplify the approach and rely on the MAP estimates of their marginalized posteriors which, as discussed in Ref. [118], can be justified here.

Fig. 6.7(a) shows the contours of $Z_k^{(l'=1,2)}(\xi)$ obtained for density. The minima of $Z_k^{(l'=1,2)}$ identify the central region of the space as the region where there is close agreement between the representations inferred at levels 1 and 2, while the maxima of $Z_k^{(l'=1,2)}$ localize near the corners $(-1, -1, -1)$ and $(1, 1, 1)$. The results for self-diffusivity and enthalpy are similar and are omitted here for brevity [118].

By analyzing the distribution of $Z_k^{(l'=1,2)}$, $k = 1, 2, 3$, we can identify which points of the full grid at $l' = 3$ are characterized by the highest values of $Z_k^{(l'=1,2)}$. Fig. 6.7(b)

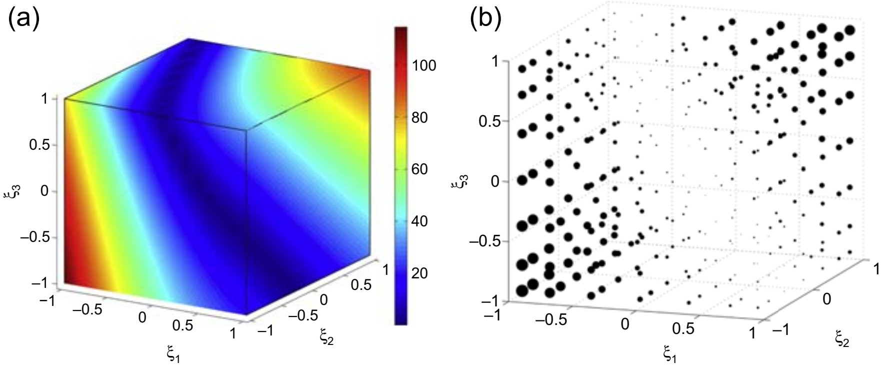

Figure 6.7 (a) Contours of $Z_k^{(l'=1,2)}$ obtained for density. (b) Distributions of Fejér grid points $\{\xi_i\}_{i=1}^{316}$ at level 3 (omitting the subset shared with levels 1 and 2), obtained for density, with each point ξ_i represented by a marker whose size is scaled according the corresponding value of $Z_k^{(l'=1,2)}(\xi_i)$.
This figure is reproduced from [118] with permission.

shows the grid of 316 points exclusively belonging to the third approximation level $l' = 3$ (i.e., we omit those shared with levels $l' = 1$ and $l' = 2$), depicted such that the size of the marker associated with the i-th node ξ^i is scaled according to the local value of $Z_k^{(l'=1,2)}\left(\xi^i\right)$. These plots give a first visual intuition of which subset will be selected and which will be neglected.

To define a quantitative metric to select subset of points, we first nondimensionalize the "error" $Z_k^{(l'=1,2)}(\xi)$ for each observable as

$$\widehat{Z}_k(\xi) = \frac{Z_k^{(l'=1,2)}(\xi)}{\max_\Omega\left(Z_k^{(l'=1,2)}\right)}, \quad k = 1, 2, 3, \tag{6.61}$$

where the normalization factor corresponds to the maximum value of $Z_k^{(l'=1,2)}$ in the parameter domain $\Omega = (-1, 1)^3$. Using a tolerance λ ($0 \leq \lambda \leq 1$), we define $A_k^{(\lambda)}$ to be the set of *new* sampling nodes for the k-th observable at level 3 according to

$$A_k^{(\lambda)} = \left\{\xi^1, \ldots, \xi^{N_k^{(\lambda)}}\right\} \doteq \left\{\xi^i : \widehat{Z}_k\left(\xi^i\right) \geq \lambda, \quad i = 1, \ldots, 316\right\}, \quad k = 1, 2, 3, \tag{6.62}$$

where $N_k^{(\lambda)}$ represents the number of points in the resulting reduced grid, which depends on the type of observable and on the value of λ, while the index i enumerates the 316 points that belong exclusively to the full grid at $l' = 3$. Evidently, for any given λ, we must have $N_k^{(\lambda)} \leq 316$, and $N_k^{(\lambda)} = 316$ when $\lambda = 0$.

We explore the following values of the tolerance: $\lambda = 0$, 0.25, and 0.4. The sets of observations at level $l' = 3$ for different values of λ, i.e., $\lambda = 0$, 0.25, and 0.40, can be in turn exploited within the Bayesian inference framework as discussed before. We focus our attention on inferring for each observable a third-order PC expansion and investigate the dependence of the MAP estimates of the coefficients on λ. Fig. 6.8 shows the normalized difference $\left|c_l^{(k)} - c_{l,\lambda}^{(k)}\right| \Big/ \left|c_0^{(k)}\right|$, $l = 0, \ldots, 19$, where $c_l^{(k)}$ represents the MAP estimate of the l-th PC coefficient for the k-th observable inferred at $l' = 3$ using the full grid, i.e., $\lambda = 0$, whereas $c_{l,\lambda}^{(k)}$ is the MAP estimate of the corresponding coefficient inferred at $l' = 3$ using $\lambda = 0.25$ or $\lambda = 0.40$. The figure shows that for density the differences are minimal, of the order $\sim O\left(10^{-4}\right)$. Another interesting observation is that for density the discrepancy does not substantially vary with the order of the coefficients.

We stop at the third level for this test, but the grid adaptation process described above can be further extended to higher levels if necessary. A possible measure to monitor the impact of refinement can be based on the normalized relative error

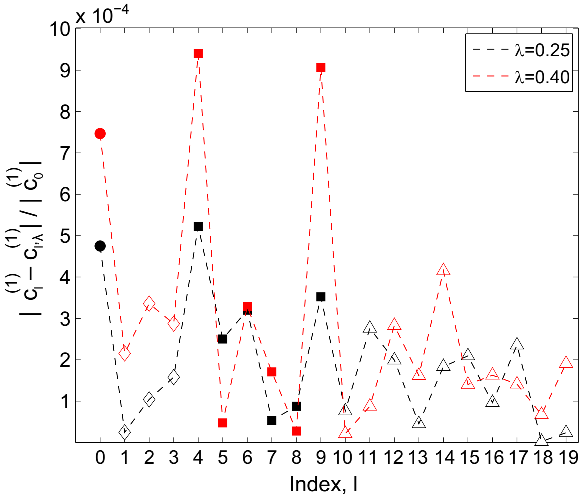

Figure 6.8 Normalized “discrepancy” $\left\{\left|c_l^{(k)} - c_{l,\lambda}^{(k)}\right| \Big/ \left|c_0^{(k)}\right|\right\}_{l=0}^{19}$, for ρ, where $c_{l,\lambda}^{(k)}$ is the MAP estimate of the l-th PC coefficient inferred at $l' = 3$ using the set of observations derived for $\lambda = 0.25$ or $\lambda = 0.40$, while $\left|c_l^{(k)}\right|$ is the absolute value of the MAP estimate of the corresponding coefficient obtained using $\lambda = 0$. Subsequent orders are identified by the following markers: zeroth- (•), first- (◆), second- (■), and third-order coefficients (Δ). This figure is reproduced from F. Rizzi, H.N. Najm, B.J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, O.M. Knio, Uncertainty quantification in MD simulations. Part I: forward propagation, Multiscale Model. Sim. 10 (2012) 1428−1459. https.//doi.org/10.1137/110853169. with permission.

$$\frac{\left\|Z_{k,\lambda}^{(l',l'+1)}\right\|_2}{\left|c_0^{(k)}\right|}, \quad k = 1, 2, 3, \tag{6.63}$$

where $\left|c_0^{(k)}\right|$ is the leading term of the expansion at level l', and the numerator is a global measure of the error defined as

$$\left\|Z_k^{(l'=1,2)}\right\|_2 = \left(\int_\Omega \left|F_k^{(l'=1,p=0)}(\xi) - F_k^{(l'=2,p=2)}(\xi)\right|^2 \frac{1}{8} d\xi\right)^{1/2}, \quad k = 1, 2, 3. \tag{6.64}$$

Exploiting the fact that the integrand is a fourth-order polynomial, the above integrals can be computed exactly using the cubature.

6.4.2.1.1 Calibration using an uncertain PC surrogate model

The inverse problem discussed in Ref. [84] involves the inference of force field parameters for TIP4P water model given a set of observations of one or more macroscale observables of water. We focus on a synthetic problem where fixed values of the TIP4P force field parameters are used to run isothermal, isobaric MD simulations at ambient conditions, $T = 298\text{K}$ and $P = 1$ atm and collect a set of noisy data of selected macroscale observables. The MD computational setting is the one earlier in this section. These data are then exploited in a Bayesian setting to recover the "true" set of driving parameters. Attention is focused on inferring the same three parameters, ε, σ, and d, for which we built the PC surrogate above.

The analysis can be regarded as a three-stage process: first, three values of the parameters of interest $\widehat{\varepsilon} = 0.17$ (kcal/mol), $\widehat{\sigma} = 3.15$ (Å), and $\widehat{d} = 0.14$ (Å) are chosen and regarded as the "true" parameters (the "hat" will be used to denote the "true" values); secondly, these "true" values are used to run $N = 10$ replica MD simulations and obtain N realizations of density (ρ), self-diffusion (ϕ), and enthalpy (H); finally, these observations are used within a Bayesian inference framework to recover the original (or "true") subset of driving parameters. Our goal is to investigate the performance of the Bayesian approach in terms of the accuracy with which we recover the "true" parameters and characterize the main factors affecting the inference.

Formulation for a deterministic surrogate model. When using a *deterministic* PC expansion for each observable, the formulation yields the following posterior [118]

$$p\left(\xi, \widetilde{\sigma}^2 \middle| \{D_l\}_{l=1}^3\right) \propto \prod_{k=1}^{3} \prod_{i=1}^{N} \frac{\exp\left(-\dfrac{\left[d_k^i - F_k(\xi)\right]^2}{2\widetilde{\sigma}_k^2}\right)}{\sqrt{2\pi\widetilde{\sigma}_k^2}} \widetilde{p}\left(\widetilde{\sigma}_k^2\right) p(\xi), \tag{6.65}$$

where D_l, $l = 1, 2, 3$, represents that data for the l-th observable, $\widetilde{p}\left(\widetilde{\sigma}_k^2\right)$ is the prior of the noise variance, $\widetilde{\sigma}_k^2$, and $p(\xi)$ represents the probability in the ξ-space corresponding to the prior on the parameter vector $\boldsymbol{\vartheta} = \{\varepsilon, \sigma, \boldsymbol{d}\}$. We have used a tilde to distinguish between the variance, $\widetilde{\sigma}_k^2$, associated with the k-th observable, and the force field parameter, σ.

Formulation for a nondeterministic surrogate model. When using *nondeterministic* PC expansions for all three observables, the formulation is more complex. In this case, each PC coefficients vector $c^{(k)} = \left\{c_l^{(k)}\right\}_{l=0}^{P}$, $k = 1, 2, 3$, is a random vector defined by a $(P + 1)$-dimensional joint probability density. We can define a suitable likelihood function for this case as follows.

For a given sample $\xi^{(j)} = \left\{\xi_1^{(j)}, \xi_2^{(j)}, \xi_3^{(j)}\right\}$, we can construct the following constant column vector

$$a = \left\{\Psi_0\left(\xi^{(j)}\right), \ldots, \Psi_P\left(\xi^{(j)}\right)\right\}^T, \tag{6.66}$$

i.e., by substituting $\xi^{(j)}$ into the truncated PC basis. Hence, we can interpret each *nondeterministic* PC representation, $F_k(\xi)$, as a *linear combination* of the random vector $c^{(k)}$, according to

$$F_k = a^T c^{(k)}, \quad k = 1, 2, 3. \tag{6.67}$$

For this chapter, as shown in Ref. [118], the probability density describing the uncertain PC expansion of each observable closely resembles a Gaussian. We thus approximate the $(P+1)$-dimensional distribution describing the random vector $c^{(k)} = \left\{c_0^{(k)}, \ldots, c_P^{(k)}\right\}^T$, $k = 1, 2, 3$, with a $(P+1)$-variate Gaussian with mean $\mu^{(k)}$ and covariance matrix $Z^{(k)}$, $k = 1, 2, 3$. Consequently, the linear combination

$$a^T c^{(k)} = \Psi_0\left(\xi^{(j)}\right)c_0^{(k)} + \ldots + \Psi_P\left(\xi^{(j)}\right)c_P^{(k)}, \quad k = 1, 2, 3, \tag{6.68}$$

is distributed according to a *univariate* Gaussian with mean $\left(a^T \mu^{(k)}\right)$ and variance $\left(a^T Z^{(k)} a\right)$, namely as

$$a^T c^{(k)} \sim \mathcal{N}\left(a^T \mu^{(k)}, a^T Z^{(k)} a\right), \quad k = 1, 2, 3. \tag{6.69}$$

Note that the uncertainty in the PC coefficients appears only through the mean vector $\mu^{(k)}$ and the covariance $Z^{(k)}$, because the constant vector $\boldsymbol{a}$ is only ξ-dependent. Assuming an independent additive error model, the discrepancy between each observation, d_k^i, $k = 1, 2, 3$, $i = 1, \ldots, N$, and the *nondeterministic* surrogate model prediction, $F_k(\xi) = \boldsymbol{a}^T \boldsymbol{c}^{(k)}$, $k = 1, 2, 3$, can be expressed as

$$\begin{aligned} d_k^i &= F_k(\xi) + \gamma_k^i \\ &= \boldsymbol{a}^T \boldsymbol{c}^{(k)} + \gamma_k^i, \\ k &= 1, 2, 3, i = 1, \ldots, N, \end{aligned} \tag{6.70}$$

where each set $\left\{\gamma_k^i\right\}_{i=1}^N$, $k = 1, 2, 3$, comprises *i.i.d.* random variables with density p_{γ_k}, $k = 1, 2, 3$. Assuming $\gamma_k^i \sim \mathcal{N}\left(0, \widetilde{\sigma}_k^2\right)$, $i = 1, \ldots, N$, $k = 1, 2, 3$, and considering, by construction, N-independent realizations for each observable, we obtain the following likelihood function

$$p\left(\{D_l\}_{l=1}^3 | \xi\right) = \prod_{k=1}^3 \prod_{i=1}^N \frac{1}{\sqrt{2\pi\left(a^T Z^{(k)} a + \widetilde{\sigma}_k^2\right)}} \exp\left(-\frac{\left[d_k^i - \left(a^T \mu^{(k)}\right)\right]^2}{2\left(a^T Z^{(k)} a + \widetilde{\sigma}_k^2\right)}\right), \tag{6.71}$$

where the index k enumerates the observables, i enumerates the observations, $\mu^{(k)}$ and $Z^{(k)}$, respectively, denote the mean and covariance matrix of the $(P+1)$-variate Gaussian representing the $(P+1)$-dimensional distribution of the *nondeterministic* PC coefficients featuring in the expansion of the k-th observable, and the constant vector $\boldsymbol{a} = \{\Psi_0(\xi), \ldots, \Psi_P(\xi)\}^T$ is computed by evaluating the PC basis for a given ξ. We treat the variances $\widetilde{\sigma}^2 = \left\{\widetilde{\sigma}_k^2\right\}_{k=1}^3$ as hyperparameters. We remark that this likelihood function combines both data noise and surrogate uncertainty in a self-consistent manner.

Results. Fig. 6.9 shows the contour plots corresponding to 30%, 60%, and 90% of the maximum probability of the joint posteriors $p(\varepsilon, \sigma|\boldsymbol{D})$ (a), $p(\varepsilon, d|\boldsymbol{D})$ (b), and $p(\sigma, d|\boldsymbol{D})$ (c). The plots reveal that the posteriors obtained from a nondeterministic surrogate are centered on the true values, whereas those obtained with a deterministic surrogate do not capture the true values with the same accuracy. The blue and black contours plotted in the left column reveal, in fact, a similar orientation and a comparable spread. The results allow us to conclude that, for the present problem, the inference based on nondeterministic surrogates provides a more robust framework to perform the inverse problem.

6.4.2.2 Gaussian processes and efficient global Optimization strategies

We consider now an example of adaptive learning by kriging metamodels, as exposed in Section 6.3.2.2.

Calibration of a TIP4P water force field. EGO has been used in the context of the Bayesian calibration of a water force field by Cailliez et al. [77]. The TIP4P force field, as described above, is chosen to model water molecules. In addition to the three parameters σ, ε, and d optimized by Rizzi et al. [84,118], the partial charge q_H borne by each hydrogen atom was optimized (note that the partial charge of the oxygen atom is constrained by the neutrality of the molecule: $q_O + 2q_H = 0$). The target of the calibration is the experimental liquid density of water at five temperatures between 253 and 350K under a pressure of 1 bar. On the molecular simulation side, those quantities are computed with MD simulations, the length of which prevents from a direct exploration of the parameter space. Predictions of MD simulations are assigned a constant uncertainty at all temperatures and for any parameter set. Due to the very small experimental uncertainties of the calibration data, those are ignored ($u_{d_i} \ll u_{F_i}$ in Eq. 6.4).

The score function G to be minimized by EGO is $-\log(p(\boldsymbol{\vartheta}|\boldsymbol{D}, \boldsymbol{X}, \boldsymbol{M}))$. The metamodel for G is built from five GP processes, each aiming at reproducing the liquid density obtained from an MD simulation with the TIP4P model at a given temperature over the parameter space. As G is computed from noisy data, a variant of the EI has been used, adapted from Huang et al. [91]:

$$EI^*(\boldsymbol{\vartheta}) = \mathbb{E}\left[\max\left(\widetilde{G}(\boldsymbol{\vartheta}^{**}) - \widetilde{G}(\boldsymbol{\vartheta})\right), 0\right], \tag{6.72}$$

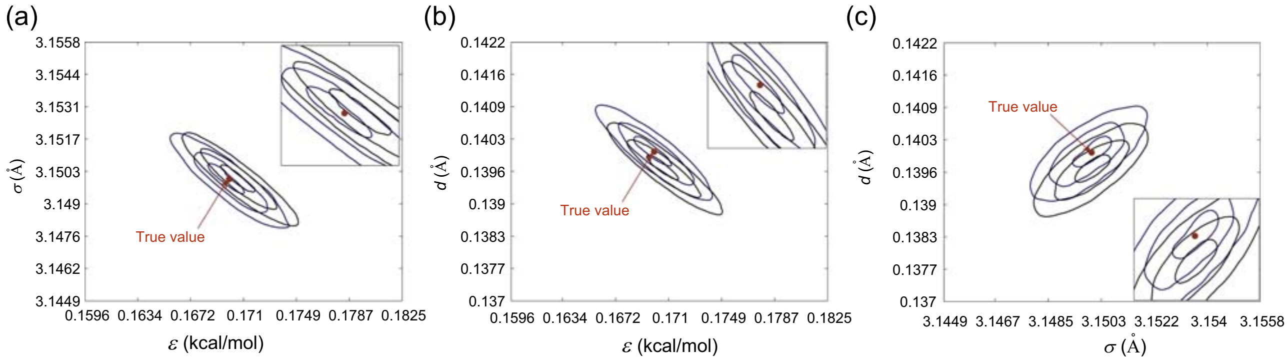

Figure 6.9 Contour plots corresponding to 30%, 60%, and 90% of the maximum probability of the marginalized joint posteriors $p(\varepsilon, \sigma|\boldsymbol{D})$ (a), $p(\varepsilon, d|\boldsymbol{D})$ (b), and $p(\sigma, d|\boldsymbol{D})$ (c). The *black line* represents the results obtained using a third-order *deterministic* surrogate model, whereas the *blue line* represents the results computed using a third-order *nondeterministic* PC surrogate with $\boldsymbol{\lambda} = 0$. The results are based on considering all three observables (ρ, D, H), with a total of 30 data points.

This figure is reproduced from F. Rizzi, H.N. Najm, B.J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, O.M. Knio, Uncertainty quantification in MD simulations. Part II: bayesian inference of force-field parameters, Multiscale Model. Simul. 10 (2012) 1460−1492. https://doi.org/10.1137/110853170 with permission.

where $\widetilde{G}(\vartheta^{**})$ is the value of the metamodel at the point ϑ^{**} of the sampling design that minimizes $\left(\widetilde{G}(\boldsymbol{\vartheta}) - u_{\widetilde{G}}(\boldsymbol{\vartheta})\right)$, and $u_{\widetilde{G}}(\boldsymbol{\vartheta})$ is an estimate of the standard deviation of $\widetilde{G}(\boldsymbol{\vartheta})$.

The calibration procedure converged within five steps, which is due to the rather large initial sampling design (84 points, which corresponds to 21 points per dimension of the parameter space), leading to best parameter values consistent with the literature (see green point in Fig. 6.10). The results of the calibration reveal that there exists a

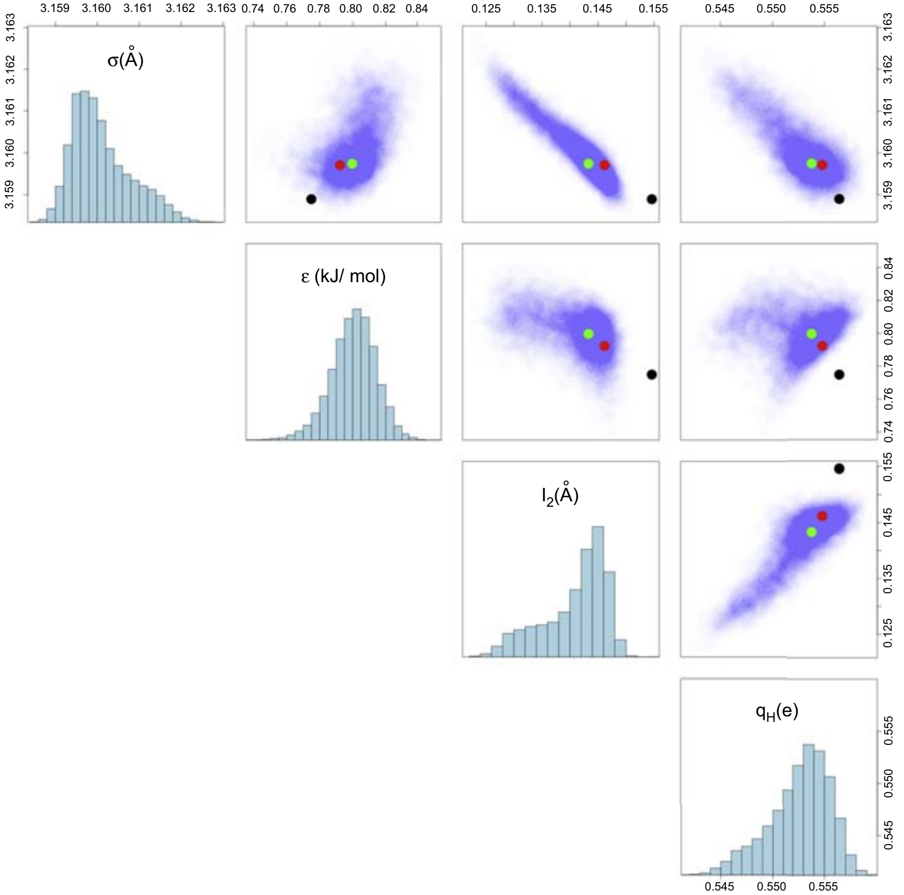

Figure 6.10 Markov chain over the PDF obtained after calibration of the TIP4P force field. Diagonal plots: marginal distribution of the parameters. Upper plots: 2D projections of the Markov chain (*blue points*). The *black point* corresponds to the TIP4P-2005 water model. The best values obtained after EGO calibration are displayed as *green and red points*. On this figure, variable d controlling the position of the oxygen point charge is named l_2.
This figure is reproduced from F. Cailliez, A. Bourasseau, P. Pernot, Calibration of forcefields for molecular simulation: sequential design of computer experiments for building cost-efficient kriging metamodels, J. Comput. Chem. 35 (2013) 130−149. https://doi.org/10.1002/jcc.23475 with permission.

unique optimal region in parameter space for the TIP4P model that allows to reproduce the evolution of the water liquid density as a function of temperature, as shown in Fig. 6.10.

The advantage of this calibration strategy with respect to "standard" methods is that one also has access to an estimation of the PDF of the parameters around the MAP. This allows to perform parametric UP for the force field parameters. As already observed in the example of LJ fluids, the contribution of parametric uncertainties to the total uncertainty was found to be greater than that of the numerical uncertainties.

This application illustrates how metamodeling combined with efficient optimization strategies can be used in the context of statistical force field calibration. Before closing this topic, it is worth commenting on possible pitfalls of this kind of procedure:

- Metamodels are built on molecular simulation data, which may not always have reached convergence, due to sampling time smaller than the relaxation time of the system. This could be the case for the whole parameter space or in some regions of the parameter space with a global limited computational budget for molecular simulations. In Ref. [77], such a situation arose, due to some parameter sets leading to "glassy water" at low temperature. Removing those "incorrect" data (around a third of the initial design) in the process of metamodel building led to similar results (see the red point in Fig. 6.10) as when using the full design. This illustrates the stability of this calibration strategy.
- In order to reduce the computational burden, it is important to minimize the number of molecular simulations to be run. Cailliez and coworkers [77] estimated that reducing the size of the initial design to 32 points (8 points per dimension of the parameter space) should lead to a viable calibration. For smaller initial designs, the EI^* utility function used in Ref. [77] may not be successful, and the use of other variants of EGO would be required.

6.4.3 Model selection and model inadequacy

A number of recent works have addressed the issue of model inadequacy in capturing the properties of molecular systems by the classical LJ potentials [17,30,35,42,52,59].

Most studies have dealt with simple monoatomic gases, for which the simple LJ force field is expected to be adequate. The LJ potential (also noted as LJ 6−12) has been the main focus, with recent developments on an LJ 6-p potential. The shift to the 6-p model was motivated by the inability of the LJ potential to predict observables in different phases [35]. However, the introduction of variability on the p exponent of the repulsive term is not sufficient to compensate for all modeling adequacy issues of the LJ potential.

HM and SEm methods both aim at designing a PDF of the force field parameters which enables some form of compatibility between the model predictions and the calibration dataset.

HM attempts to reconcile heterogeneous observables and/or physical conditions: an overall distribution is designed to contain the different parameter sets best adapted to each subset [52,59]. As mentioned above, prediction of new data for a new observable or for new physical conditions should use the overall distribution. This typically leads to large prediction uncertainties, much larger than prediction uncertainties for new data of observables contained in the calibration set (see Fig. 13 in Ref. [59] and the toy models in Ref. [30]).

As seen in Section 6.4.1.2, model inadequacy with LJ parameters might also be problematic when one considers a single observable.

Pernot and Cailliez [30] published a critical review of the available methods to manage this problem for the calibration of LJ parameters on viscosity data for Kr, among which additive correction by a GP, HM, or SEm was considered. We summarize the main results of that study here.

In the Kr example in Ref. [30] and the Ar case in Ref. [42], both SEm approaches described above (Sections 6.2.3.3–6.2.3.4) were evaluated, which lead to mitigated results. In both cases, it was possible to design a PDF for the LJ parameters which enabled prediction uncertainties to be large enough to compensate for model errors. Fig. 6.11 presents the residuals and 95% confidence and prediction bands for the LJ model calibrated on a dataset of five series of measurements of Ar viscosity [42]. The trend in the residuals is clearly visible at lower temperatures. As mentioned earlier,

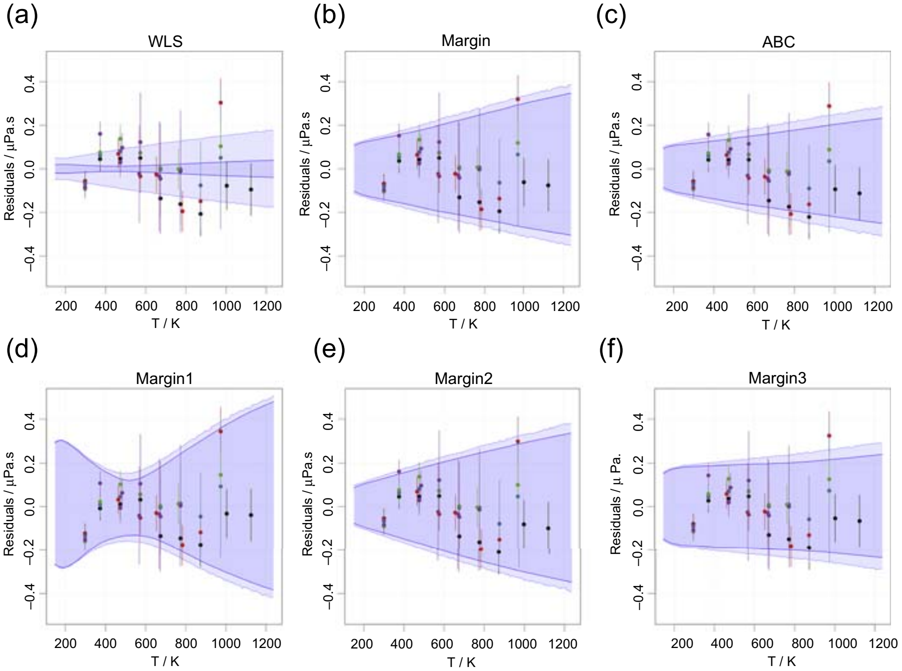

Figure 6.11 Calibration of LJ parameters for Ar over T-dependent viscosity data. Residuals and prediction uncertainty for the WLS (a), Margin (b), and ABC (c) methods and for the three degenerate solutions of the Margin method (d–f). The *dark blue bands* represent the model confidence 95% intervals, and the *light blue bands* the data prediction 95% intervals, corrected from the mean prediction.
This figure is reproduced in part from Ref. P. Pernot, The parameter uncertainty inflation fallacy, J. Chem. Phys. 147 (2017) 104102. https://doi.org/10.1063/1.4994654 with permission.

the basic calibration procedure using only data uncertainties (labeled WLS in Fig. 6.11) results in much too small prediction uncertainties, whereas both SEm approaches (labeled Margin and ABC) enable to successfully design prediction bands in agreement with the residuals. Note that the residuals are basically unchanged when compared to WLS: the fit of the data has not been improved.

Beyond this apparent success, the SEm strategies present a set of limitations, which might prevent its general applicability [30,42]:

- due to the geometry of the problem in data space [42,124], enlarging the uncertainty patch on the model manifold around the optimal parameters does not improve the statistical validity of an inadequate model (no improvement of the residuals);
- being constrained by the law of UP, the shape of the prediction uncertainty bands over the control variable(s) space does not necessarily conform to the shape of the model inadequacy errors [42];
- the elements of the covariance matrix of the stochastic parameters might have multimodal posterior distributions. This has been observed for the LJ potential calibration problem for both implementations of the SEm [30,42]. Samples of the posterior PDF for these methods reveal three modes, each one corresponding to the minimum value of one parameter of the covariance matrix. This diagnostic matches the observation that the estimates of covariance matrices in hierarchical models tend to be degenerate [125], i.e., with zero variance for some parameters or a perfect correlation among them. A major inconvenient of this degeneracy is that each mode corresponds to a different prediction uncertainty profile (see Fig. 6.11, bottom row). The posterior predictive uncertainty profile, being a weighted average of these modes, might consequently be very sensitive to the calibration dataset through mode flipping.

Considering these limitations, and notably the nonrobust shape of the prediction bands, it is still not clear whether the posterior PDFs estimated by SEm represent an improvement for the prediction of QoIs not in the calibration set. Further research is necessary to tackle the statistical treatment of model inadequacy.

In a recent work [36], HM was shown to be efficient in distinguishing between different models of coarse-grained molecular dynamics (CGMD) potentials. Such CGMD simulations are often calibrated on different experimental conditions, resulting in a plethora of models with little transferability. HM provides guidance on the model accuracy as well as on its trade-off with computational accuracy. In Fig. 6.12(a), the speed-up gained by using a CG model is plotted against its model evidence. Each model is characterized by the name of the model (1S, 2S, 2SF, 3S*, 3SF*) and the coarse-graining resolution (1, 3, 4, 5 ,6), see Ref. [36] for a detailed description of the models. In Fig. 6.12(b), the evidences of three of the models of Fig. 6.12(a) are estimated at different temperatures. In order to combine all the evidences into one number, the HM approach is adopted.

6.5 Conclusion and perspectives

The increasing use of MD and MC in academia as well as in industry calls for a rigorous management of uncertainty in molecular simulations [126]. Among the

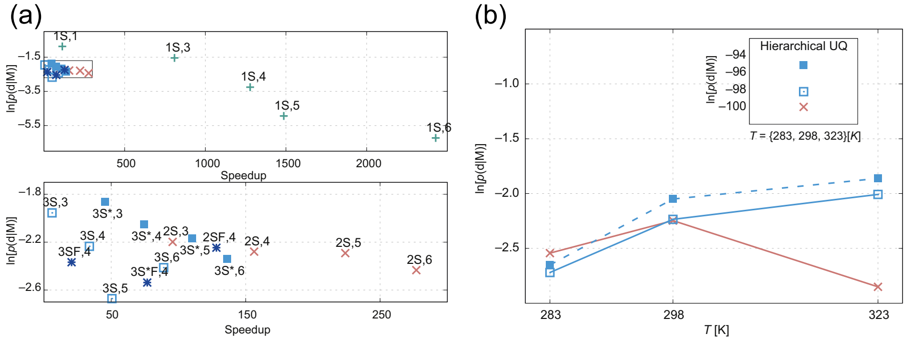

Figure 6.12 (a) Model evidences with respect to the speedup of the examined water models marked with the name and the mapping. The boxed section in the top plot is enlarged in the bottom plot. (b) Logarithm of the model evidences at different temperatures *T* for models *2SF* (×), *3S* (⊡), and *3S** (■). The inset shows the model evidences of the hierarchical UQ approach, where all three temperatures are considered concurrently.
This figure is reproduced from Ref. J. Zavadlav, G. Arampatzis, P. Koumoutsakos, Bayesian selection for coarse-grained models of liquid water, Sci. Rep. 9 (2019) 99. https://doi.org/10.1038/s41598-018-37471-0 with permission.

various sources of uncertainties in molecular simulations, this chapter focused on those originating from the values of the force field parameters. Bayesian approaches provide an operative and complete framework to deal with the determination of the parameter uncertainties and their propagation through molecular simulations. The computational cost of such methods has limited their use in the context of molecular simulations until the last decade. However, the increase in computational resources as well as the development of efficient numerical strategies currently enables the investigation of force fields parametric uncertainties.

In this chapter, we have reviewed recent approaches to address the issue of force field parameter calibration and outlined significant lessons that have been learned over the last 10 years. We have shown that uncertainty estimation for molecular simulation predictions can be more complex than the simulations themselves.

Section 6.4.1 has been designed to give a step-by-step pedagogical example of the application of Bayesian “standard” strategy to the calibration of an LJ force field for rare gases. Apart from giving an introduction to Bayesian methods, this application sheds light on the major issues arising during force field calibration. A large part of this chapter (Section 6.4.2) has been devoted to computational aspects, especially the use of metamodeling-based strategies, that are currently the only way to overcome the bottleneck of performing thousands, or more, of molecular simulations. However, the major challenge that emerged from the early studies on Bayesian calibration of force field parameters is that of model inadequacy. The physics contained in typical force fields is often too crude to enable a statistically valid calibration without employing advanced calibration schemes. Most of the time, the use of a more complex force field is computationally prohibitive, so one has to deal with inadequacy of the current force fields. Section 6.4.3 has been centered on the *pros* and *cons* of various strategies that have been used recently in the context of force field calibration with inadequate models.

Although significant advances have been made in recent years on the application of Bayesian methods to force field calibrations, the fully consistent characterization of the prediction uncertainty in molecular simulations is still an exciting research topic.

The need for reliable, reproducible, and portable molecular simulations is broadly recognized and has led, among other efforts, to the recent development of OpenKIM (openkim.org), a community-driven online framework. We suggest that frameworks such as OpenKIM could greatly benefit from adopting a systematic Bayesian inference approach to link experimental data with the results of MD simulations. Beyond being a formidable interdisciplinary scientific discovery framework, we believe that by proper integration of experimental data and simulations through Bayesian inference molecular simulation will become an effective virtual measurement tool.

Abbreviations and symbols

ABC	Approximate Bayesian computation
CGMD	Coarse-grained molecular dynamics
EGO	Efficient global optimization
GP	Gaussian process
LJ	Lennard-Jones
LUP	Linear uncertainty propagation
$\boldsymbol{M}$	Full model, comprising the physical model and the statistical model
MAP	Maximum a posteriori
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MD	Molecular dynamics
PCE	Polynomial chaos expansion
PDF	Probability density function
PES	Potential energy surface
QoI	Quantity of interest
SEm	Stochastic embedding
TMCMC	Transitional MCMC
UP	Uncertainty propagation
$\delta \boldsymbol{F}$	Discrepancy function for model F
$\boldsymbol{\varepsilon}$	Interaction energy of the Lennard-Jones potential
$\boldsymbol{\kappa}$	Hyperparameters of hierarchical model
$\boldsymbol{\mu}_{\vartheta}$	Mean value of stochastic parameters
$\Psi_{\boldsymbol{I}j}$	Univariate orthogonal polynomial used in PCE
$\Psi_{\boldsymbol{I}}$	Multivariate orthogonal polynomial used in PCE
$\boldsymbol{\sigma}$	Radius parameter of the Lennard-Jones potential
σ_i^2	Variance of the errors at point x_i
$\boldsymbol{\Sigma}_{\boldsymbol{D}}$	Covariance matrix of the reference data $\boldsymbol{D}$
$\boldsymbol{\Sigma}_{\boldsymbol{F}}$	Covariance matrix of the physical model errors
$\boldsymbol{\Sigma}_{\boldsymbol{R}}$	Covariance matrix of the residuals $\boldsymbol{R}(\vartheta)$
$\boldsymbol{\Sigma}_{\vartheta}$	Covariance matrix of stochastic parameters
τ_i	Parameter of the uncertainty model

$\boldsymbol{\vartheta}$	Parameter set of the physical model F
$\boldsymbol{\vartheta_{\delta F}}$	Parameters of the discrepancy function δF
$\boldsymbol{\varepsilon_i}$	Noise variable
$\boldsymbol{\vartheta_i}$	A parameter of the physical model F
$\boldsymbol{xi_i}$	An auxiliary random variable for PCE
$\boldsymbol{D}$	Set of reference/calibration data
$\mathscr{L}(\vartheta)$	Logarithm of the posterior PDF
$\boldsymbol{R}(\vartheta)$	Vector of residuals
$\boldsymbol{X}$	Set of physical conditions for the reference data
$\boldsymbol{d_i}$	A reference datum
$\boldsymbol{D_{KL}}$	Kullback—Leibler divergence
$\boldsymbol{F}(\boldsymbol{x};\vartheta)$	Computational or physical model
$\boldsymbol{F_i}(\vartheta)$	Value of the computational model at point x_i with parameters ϑ
$\boldsymbol{N_\tau}$	Number of parameters for the uncertainty model
$\boldsymbol{N_\vartheta}$	Number of parameters of the physical model F
$\boldsymbol{N_D}$	Size of the reference dataset $\boldsymbol{D}$
$\boldsymbol{p}(\boldsymbol{X}\vert\boldsymbol{Y})$	Conditional PDF of X knowing Y
$\boldsymbol{r_B}$	Birge ratio
$\boldsymbol{u_{d_i}^2}$	Variance of noise for datum d_i
$\boldsymbol{u_{F_i}^2}$	Computational variance for model value $F_i(\vartheta)$
$\boldsymbol{x_i}$	Physical condition(s) for a reference datum
$\boldsymbol{\bar{x}}$	Mean value of parameters x
$\boldsymbol{\widehat{x}}$	Optimal value of parameter x
$\boldsymbol{\widetilde{x}}$	Value of x out of the calibration dataset

References

[1] E. Maginn, From discovery to data: what must happen for molecular simulation to become a mainstream chemical engineering tool, AIChE J. 55 (2009) 1304—1310. https://doi.org/10.1002/aic.11932.

[2] C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J.-C. de Hemptinne, P. Ungerer, B. Rousseau, C. Adamo, A general guidebook for the theoretical prediction of physicochemical properties of chemicals for regulatory purposes, Chem. Rev. 115 (2015) 13093—13164. https://doi.org/10.1021/acs.chemrev.5b00215.

[3] K. Irikura, R. Johnson, R. Kacker, Uncertainty associated with virtual measurements from computational quantum chemistry models, Metrologia 41 (2004) 369—375. https://doi.org/10.1088/0026-1394/41/6/003.

[4] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement (GUM), Tech. Rep. 100:2008, Joint Committee for Guides in Metrology, JCGM, 2008.

[5] P.N. Patrone, A. Dienstfrey, Uncertainty Quantification for Molecular Dynamics, John Wiley & Sons, Ltd, 2018, pp. 115—169. https://doi.org/10.1002/9781119518068.ch3. Ch. 3.

[6] J. Proppe, M. Reiher, Reliable estimation of prediction uncertainty for physicochemical property models, J. Chem. Theory Comput. 13 (2017) 3297—3317. https://doi.org/10.1021/acs.jctc.7b00235.

[7] A. Chernatynskiy, S.R. Phillpot, R. LeSar, Uncertainty quantification in multiscale simulation of materials: a prospective, Annu. Rev. Mater. Res. 43 (2013) 157−182. https://doi.org/10.1146/annurev-matsci-071312-121708.

[8] M. Salloum, K. Sargsyan, R. Jones, H.N. Najm, B. Debusschere, Quantifying sampling noise and parametric uncertainty in atomistic-to-continuum simulations using surrogate models, Multiscale Model. Simul. 13 (2015) 953−976. https://doi.org/10.1137/140989601.

[9] X. Zhou, S.M. Foiles, Uncertainty quantification and reduction of molecular dynamics models, in: J.P. Hessling (Ed.), Uncertainty Quantification and Model Calibration, InTech, Rijeka, 2017, pp. 1−25. https://doi.org/10.5772/intechopen.68507. Ch. 05.

[10] Z. Li, J.R. Kermode, A. De Vita, Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces, Phys. Rev. Lett. 114 (2015) 096405. https://doi.org/10.1103/PhysRevLett.114.096405.

[11] J. Behler, Perspective: machine learning potentials for atomistic simulations, J. Chem. Phys. 145 (2016) 170901. https://doi.org/10.1063/1.4966192.

[12] Y. Li, H. Li, F.C. Pickard, B. Narayanan, F.G. Sen, M.K.Y. Chan, S.K.R.S. Sankaranarayanan, B.R. Brooks, B. Roux, Machine learning force field parameters from ab initio data, J. Chem. Theory Comput. 13 (2017) 4492−4503. https://doi.org/10.1021/acs.jctc.7b00521.

[13] P. Ungerer, C. Beauvais, J. Delhommelle, A. Boutin, B. Rousseau, A.H. Fuchs, Optimisation of the anisotropic united atoms intermolecular potential for n-alkanes, J. Chem. Phys. 112 (2000) 5499−5510. https://doi.org/10.1063/1.481116.

[14] E. Bourasseau, M. Haboudou, A. Boutin, A.H. Fuchs, New optimization method for intermolecular potentials: optimization of a new anisotropic united atoms potential for olefins: prediction of equilibrium properties, J. Chem. Phys. 118 (2003) 3020−3034. https://doi.org/10.1063/1.1537245.

[15] A. García-Sánchez, C.O. Ania, J.B. Parra, D. Dubbeldam, T.J.H. Vlugt, R. Krishna, S. Calero, Transferable force field for carbon dioxide adsorption in zeolites, J. Phys. Chem. C 113 (20) (2009) 8814−8820. https://doi.org/10.1021/jp810871f.

[16] D. Horinek, S. Mamatkulov, R. Netz, Rational design of ion force fields based on thermodynamic solvation properties, J. Chem. Phys. 130 (2009) 124507. https://doi.org/10.1063/1.3081142.

[17] F. Cailliez, P. Pernot, Statistical approaches to forcefield calibration and prediction uncertainty of molecular simulations, J. Chem. Phys. 134 (2011) 054124. https://doi.org/10.1063/1.3545069.

[18] C. Vega, J.L.F. Abascal, M.M. Conde, J.L. Aragones, What ice can teach us about water interactions: a critical comparison of the performance of different water models, Faraday Discuss. 141 (2009) 246−251. https://doi.org/10.1039/B805531A.

[19] S.B. Zhu, C.F. Wong, Sensitivity analysis of distribution-functions of liquid water, J. Chem. Phys. 99 (1993) 9047−9053. https://doi.org/10.1063/1.465572.

[20] S.B. Zhu, C.F. Wong, Sensitivity analysis of water thermodynamics, J. Chem. Phys. 98 (11) (1993) 8892−8899. https://doi.org/10.1063/1.464447.

[21] A.P. Moore, C. Deo, M.I. Baskes, M.A. Okuniewski, D.L. McDowell, Understanding the uncertainty of interatomic potentials − parameters and formalism, Comput. Mater. Sci. 126 (2017) 308−320. https://doi.org/10.1016/j.commatsci.2016.09.041.

[22] L. Sun, W.-Q. Deng, Recent developments of first-principles force fields, WIREs Comput. Mol. Sci. 7 (2017) e1282. https://doi.org/10.1002/wcms.1282.

[23] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, Evaluation of Measurement Data — Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement" — Propagation of Distributions Using a Monte Carlo Method, Tech. Rep. 101:2008, Joint Committee for Guides in Metrology, JCGM, 2008.

[24] B. Cooke, S. Schmidler, Statistical prediction and molecular dynamics simulation, Biophys. J. 95 (2008) 4497—4511. https://doi.org/10.1529/biophysj.108.131623.

[25] P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge University Press, 2005.

[26] D.S. Sivia, in: Data Analysis: A Bayesian Tutorial, second ed., Oxford University Press, New York, 2006.

[27] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, in: Bayesian Data Analysis, third ed., Chapman and Hall/CRC, 2013.

[28] P.R. Bevington, D.K. Robinson, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1992. http://shop.mheducation.com/highered/product.M0072472278.html.

[29] I. Lira, Combining inconsistent data from interlaboratory comparisons, Metrologia 44 (2007) 415—421. https://doi.org/10.1088/0026-1394/44/5/019.

[30] P. Pernot, F. Cailliez, A critical review of statistical calibration/prediction models handling data inconsistency and model inadequacy, AIChE J. 63 (2017) 4642—4665. https://doi.org/10.1002/aic.15781.

[31] T.A. Oliver, G. Terejanu, C.S. Simmons, R.D. Moser, Validating predictions of unobserved quantities, Comput. Methods Appl. Mech. Eng. 283 (2015) 1310—1335. https://doi.org/10.1016/j.cma.2014.08.023.

[32] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework, J. Chem. Phys. 137 (2012) 144103. https://doi.org/10.1063/1.4757266.

[33] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Data-driven, predictive molecular dynamics for nanoscale flow simulations under uncertainty, J. Phys. Chem. B 117 (2013) 14808—14816. https://doi.org/10.1021/jp4084713.

[34] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, X-TMCMC: Adaptive kriging for Bayesian inverse modeling, Comput. Methods Appl. Mech. Eng. 289 (2015) 409—428. https://doi.org/10.1016/j.cma.2015.01.015.

[35] L. Kulakova, G. Arampatzis, P. Angelikopoulos, P. Hadjidoukas, C. Papadimitriou, P. Koumoutsakos, Data driven inference for the repulsive exponent of the Lennard-Jones potential in molecular dynamics simulations, Sci. Rep. 7 (2017) 16576. https://doi.org/10.1038/s41598-017-16314-4.

[36] J. Zavadlav, G. Arampatzis, P. Koumoutsakos, Bayesian selection for coarse-grained models of liquid water, Sci. Rep. 9 (2019) 99. https://doi.org/10.1038/s41598-018-37471-0.

[37] J. Ching, Y.-C. Chen, Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, J. Eng. Mech. 133 (2007) 816—832. https://doi.org/10.1061/(ASCE)0733-9399(2007)133:7(816).

[38] R.N. Kacker, A. Forbes, R. Kessel, K.-D. Sommer, Classical and Bayesian interpretation of the Birge test of consistency and its generalized version for correlated results from interlaboratory evaluations, Metrologia 45 (2008) 257—264. https://doi.org/10.1088/0026-1394/45/3/001.

[39] O. Bodnar, C. Elster, On the adjustment of inconsistent data using the Birge ratio, Metrologia 51 (2014) 516—521. https://doi.org/10.1088/0026-1394/51/5/516.

[40] NIST Chemistry WebBook, in: Thermophysical Properties of Fluid Systems, srd 69 Edition, 2017. https://webbook.nist.gov/chemistry/fluid/.

[41] C. Tegeler, R. Span, W. Wagner, A new equation of state for Argon covering the fluid region for temperatures from the melting line to 700 K at pressures up to 1000 MPa, J. Phys. Chem. Ref. Data 28 (1999) 779−850. https://doi.org/10.1063/1.556037.

[42] P. Pernot, The parameter uncertainty inflation fallacy, J. Chem. Phys. 147 (2017) 104102. https://doi.org/10.1063/1.4994654.

[43] L. Zarkova, An isotropic intermolecular potential with temperature-dependent effective parameters for heavy globular gases, Mol. Phys. 88 (1996) 489−495. https://doi.org/10.1080/00268979650026488.

[44] L. Zarkova, Viscosity, second pVT-virial coefficient, and diffusion of pure and mixed small alkanes CH4, C2H6, C3H8, n-C4H10, i-C4H10, n-C5H12, i-C5H12, and C(CH3)4 calculated by means of an isotropic temperature-dependent potential. I. Pure alkanes, J. Phys. Chem. Ref. Data 35 (2006) 1331. https://doi.org/10.1063/1.2201308.

[45] L. Zarkova, U. Hohm, Effective (n-6) Lennard-Jones potentials with temperature-dependent parameters introduced for accurate calculation of equilibrium and transport properties of Ethene, Propene, Butene, and Cyclopropane, J. Chem. Eng. Data 54 (2009) 1648−1655. https://doi.org/10.1021/je800733b.

[46] J. Brynjarsdóttir, A. O'Hagan, Learning about physical parameters: the importance of model discrepancy, Inverse Probl. 30 (2014) 114007. https://doi.org/10.1088/0266-5611/30/11/114007.

[47] K. Sargsyan, X. Huan, H.N. Najm, Embedded Model Error Representation for Bayesian Model Calibration, 2018 arXiv:1801.06768, http://arxiv.org/abs/1801.06768.

[48] S.G. Walker, Bayesian inference with misspecified models, J. Stat. Plan. Inference 143 (2013) 1621−1633. https://doi.org/10.1016/j.jspi.2013.05.013.

[49] S.G. Walker, Reply to the discussion: bayesian inference with misspecified models, J. Stat. Plan. Inference 143 (2013) 1649−1652. https://doi.org/10.1016/j.jspi.2013.05.017.

[50] A. O'Hagan, Bayesian inference with misspecified models: inference about what? J. Stat. Plan. Inference 143 (2013) 1643−1648. https://doi.org/10.1016/j.jspi.2013.05.016.

[51] K. Sargsyan, H.N. Najm, R. Ghanem, On the statistical calibration of physical models, Int. J. Chem. Kinet. 47 (2015) 246−276. https://doi.org/10.1002/kin.20906.

[52] S. Wu, P. Angelikopoulos, C. Papadimitriou, R. Moser, P. Koumoutsakos, A hierarchical Bayesian framework for force field selection in molecular dynamics simulations, Phil. Trans. R. Soc. A 374 (2015) 20150032. https://doi.org/10.1098/rsta.2015.0032.

[53] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. B 63 (2001) 425−464. https://doi.org/10.1111/1467-9868.00294.

[54] K. Campbell, Statistical calibration of computer simulations, Reliab. Eng. Syst. Saf. 91 (2006) 1358−1363.

[55] P.D. Arendt, D.W. Apley, W. Chen, Quantification of model uncertainty: calibration, model discrepancy, and identifiability, J. Mech. Des. 134 (2012) 100908. https://doi.org/10.1115/1.4007390.

[56] P.S. Craig, M. Goldstein, J.C. Rougier, A.H. Seheult, Bayesian forecasting for complex systems using computer simulators, J. Am. Stat. Assoc. 96 (2001) 717−729. https://doi.org/10.1198/016214501753168370.

[57] A. Gelman, J. Hill, Data analysis using regression and multilevel/hierarchical models, in: Analytical Methods for Social Research, Cambridge University Press, 2007. https://doi.org/10.1017/CBO9780511790942.

[58] R. McElreath, Statistical Rethinking, Texts in Statistical Science, CRC Press, 2015.

[59] S. Wu, P. Angelikopoulos, G. Tauriello, C. Papadimitriou, P. Koumoutsakos, Fusing heterogeneous data for the calibration of molecular dynamics force fields using hierarchical Bayesian models, J. Chem. Phys. 145 (2016) 244112. https://doi.org/10.1063/1.4967956.

[60] G.C. Goodwin, M.E. Salgado, A stochastic embedding approach for quantifying uncertainty in the estimation of restricted complexity models, Int. J. Adapt. Control Signal Process. 3 (1989) 333−356. https://doi.org/10.1002/acs.4480030405.

[61] L. Ljung, G.C. Goodwin, J.C. Aguero, T. Chen, Model error modeling and stochastic embedding, IFAC-PapersOnLine 48 (2015) 75−79. https://doi.org/10.1016/j.ifacol.2015.12.103.

[62] J.K. Pritchard, M.T. Seielstad, A. Pérez-Lezaun, M.W. Feldman, Population growth of human Y chromosomes: a study of Y chromosome microsatellites, Mol. Biol. Evol. 16 (1999) 1791−1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091.

[63] M.A. Beaumont, W. Zhang, D.J. Balding, Approximate Bayesian computation in population genetics, Genetics 162 (2002) 2025−2035.

[64] M.A. Beaumont, Approximate Bayesian computation in evolution and ecology, Annu. Rev. Ecol. Systemat. 41 (2010) 379−406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

[65] B.M. Turner, T. Van Zandt, Hierarchical approximate Bayesian computation, Psychometrika 79 (2014) 185−209. https://doi.org/10.1007/s11336-013-9381-x.

[66] P. Marjoram, J. Molitor, V. Plagnol, S. Tavaré, Markov chain Monte Carlo without likelihoods, Proc. Natl. Acad. Sci. USA 100 (2003) 15324−15328. https://doi.org/10.1073/pnas.0306899100.

[67] M. Chiachio, J. Beck, J. Chiachio, G. Rus, Approximate Bayesian computation by subset simulation, SIAM J. Sci. Comput. 36 (2014) A1339−A1358. https://doi.org/10.1137/130932831.

[68] V. Elske, P. Dennis, S.M. Richard, Taking error into account when fitting models using Approximate Bayesian Computation, Ecol. Appl. 28 (2018) 267−274. https://doi.org/10.1002/eap.1656.

[69] L. Kulakova, P. Angelikopoulos, P.E. Hadjidoukas, C. Papadimitriou, P. Koumoutsakos, Approximate Bayesian computation for granular and molecular dynamics simulations, in: Proceedings of the Platform for Advanced Scientific Computing Conference, Vol. 4 of PASC'16, ACM, New York, NY, USA, 2016, pp. 1−12. https://doi.org/10.1145/2929908.2929918.

[70] R. Dutta, Z. Faidon Brotzakis, A. Mira, Bayesian calibration of force-fields from experimental data: TIP4P water, J. Chem. Phys. 149 (2018) 154110. https://doi.org/10.1063/1.5030950.

[71] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087−1092. https://doi.org/10.1063/1.1699114.

[72] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97−109. https://doi.org/10.2307/2334940.

[73] W. Gilks, S. Richardson, D. Spiegelhalter, Markov Chain Monte Carlo in Practice, Chapman & Hall/CRC Interdisciplinary Statistics, Taylor & Francis, 1995.

[74] B. Berg, Markov Chain Monte Carlo Simulations and Their Statistical Analysis, World Scientific, 2004.

[75] D. Gamerman, H. Lopes, in: Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, second ed., Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis, 2006.

[76] P. Hadjidoukas, P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Π4U: a high performance computing framework for Bayesian uncertainty quantification of complex models, J. Comput. Phys. 284 (2015) 1−21. https://doi.org/10.1016/j.jcp.2014.12.006.

[77] F. Cailliez, A. Bourasseau, P. Pernot, Calibration of forcefields for molecular simulation: sequential design of computer experiments for building cost-efficient kriging meta-models, J. Comput. Chem. 35 (2013) 130−149. https://doi.org/10.1002/jcc.23475.

[78] R. R Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[79] Stan Development Team, RStan: The R Interface to Stan, R Package Version 2.14.1, 2016. http://mc-stan.org/.

[80] R.A. Messerly, S.M. Razavi, M.R. Shirts, Configuration-sampling-based surrogate models for rapid parameterization of non-bonded interactions, J. Chem. Theory Comput. 14 (2018) 3144−3162. https://doi.org/10.1021/acs.jctc.8b00223.

[81] T. van Westen, T.J.H. Vlugt, J. Gross, Determining force field parameters using a physically based equation of state, J. Phys. Chem. B 115 (2011) 7872−7880. https://doi.org/10.1021/jp2026219.

[82] H. Hoang, S. Delage-Santacreu, G. Galliero, Simultaneous description of equilibrium, interfacial, and transport properties of fluids using a Mie chain coarse-grained force field, Ind. Eng. Chem. Res. 56 (2017) 9213−9226. https://doi.org/10.1021/acs.iecr.7b01397.

[83] S. Werth, K. Stöbener, M. Horsch, H. Hasse, Simultaneous description of bulk and interfacial properties of fluids by the Mie potential, Mol. Phys. 115 (2017) 1017−1030. https://doi.org/10.1080/00268976.2016.1206218.

[84] F. Rizzi, H.N. Najm, B.J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, O.M. Knio, Uncertainty quantification in MD simulations. Part II: bayesian inference of force-field parameters, Multiscale Model. Simul. 10 (2012) 1460−1492. https://doi.org/10.1137/110853170.

[85] J. Sacks, W. Welch, T. Mitchell, H. Wynn, Design and analysis of computer experiments, Stat. Sci. 4 (1989) 409−423. https://www.jstor.org/stable/2245858.

[86] T. Santner, B. Williams, W. Notz, The Design and Analysis of Computer Experiments, Springer-Verlag, 2003. https://doi.org/10.1007/978-1-4757-3799-8.

[87] O. Roustant, D. Ginsbourger, Y. Deville, DiceKriging: Kriging Methods for Computer Experiments, R Package Version 1.1, 2010. http://CRAN.R-project.org/package=DiceKriging.

[88] J.L. Loeppky, J. Sacks, W.J. Welch, Choosing the sample size of a computer experiment: a practical guide, Technometrics 51 (2009) 366−376. https://doi.org/10.1198/TECH.2009.08040.

[89] D.R. Jones, A taxonomy of global optimization methods based on response surfaces, J. Glob. Optim. 21 (2001) 345−383. https://doi.org/10.1023/A:1012771025575.

[90] D. Jones, M. Schonlau, W. Welch, Efficient global optimization of expensive black-box functions, J. Glob. Optim. 13 (1998) 455−492. https://doi.org/10.1023/A:1008306431147.

[91] D. Huang, T.T. Allen, W.I. Notz, N. Zeng, Global optimization of stochastic black-box systems via sequential kriging meta-models, J. Glob. Optim. 34 (2006) 441−466. https://doi.org/10.1007/s10898-005-2454-3.

[92] V. Picheny, D. Ginsbourger, Noisy kriging-based optimization methods: a unified implementation within the DiceOptim package, Comput. Stat. Data Anal. 71 (2014) 1035−1053. https://doi.org/10.1016/j.csda.2013.03.018.

[93] H. Jalali, I.V. Nieuwenhuyse, V. Picheny, Comparison of kriging-based algorithms for simulation optimization with heterogeneous noise, Eur. J. Oper. Res. 261 (2017) 279–301. https://doi.org/10.1016/j.ejor.2017.01.035.

[94] D. Zhan, J. Qian, Y. Cheng, Pseudo expected improvement criterion for parallel EGO algorithm, J. Glob. Optim. 68 (2017) 641–662. https://doi.org/10.1007/s10898-016-0484-7.

[95] E. Vazquez, J. Villemonteix, M. Sidorkiewicz, E. Walter, Global optimization based on noisy evaluations: an empirical study of two statistical approaches, J. Phys. Conf. Ser. 135 (2008) 012100. http://stacks.iop.org/1742-6596/135/i=1/a=012100.

[96] B. Ankenman, B.L. Nelson, J. Staum, Stochastic kriging for simulation metamodeling, Oper. Res. 58 (2010) 371–382. https://doi.org/10.1287/opre.1090.0754.

[97] N. Wiener, The homogeneous chaos, Am. J. Math. 60 (1938) 897–936. https://doi.org/10.2307/2371268.

[98] R. Ghanem, P. Spanos, Stochastic Finite Elements: A Spectral Approach, Springer Verlag, New York, 1991.

[99] O.P. Le Maître, O.M. Knio, Spectral Methods for Uncertainty Quantification, Springer, New York, 2010.

[100] D. Xiu, G.E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2002) 619–644. https://doi.org/10.1137/S1064827501387826.

[101] K. Sargsyan, B. Debusschere, H.N. Najm, Y. Marzouk, Bayesian inference of spectral expansions for predictability assessment in stochastic reaction networks, J. Comput. Theor. Nanosci. 6 (2009) 2283–2297. https://doi.org/10.1166/jctn.2009.1285.

[102] L. Fejér, On the infinite sequences arising in the theories of harmonic analysis, of interpolation, and of mechanical quadratures, Bull. Am. Math. Soc. 39 (8) (1933) 521–534. https://projecteuclid.org/euclid.bams/1183496842.

[103] G.C. Ballesteros, P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Bayesian hierarchical models for uncertainty quantification in structural dynamics, in: Vulnerability, Uncertainty, and Risk, 2014, pp. 704–714.

[104] C. Archambeau, M. Opper, Y. Shen, D. Cornford, J.S. Shawe-Taylor, Variational inference for diffusion processes, in: J.C. Platt, D. Koller, Y. Singer, S.T. Roweis (Eds.), Advances in Neural Information Processing Systems 20, Curran Associates, Inc., 2008, pp. 17–24.

[105] V. Harmandaris, E. Kalligiannaki, M. Katsoulakis, P. Plechac, Path-space variational inference for non-equilibrium coarse-grained systems, J. Comput. Phys. 314 (2016) 355–383. https://doi.org/10.1016/j.jcp.2016.03.021.

[106] PSUADE. https://computation.llnl.gov/projects/psuade-uncertainty-quantification. (Accessed 19 February 2019).

[107] VECMA. https://www.vecma.eu. (Accessed 19 February 2019).

[108] SPUX. https://www.eawag.ch/en/department/siam/projects/spux/. (Accessed 19 February 2019).

[109] P.E. Hadjidoukas, E. Lappas, V.V. Dimakopoulos, A runtime library for platform-independent task parallelism, in: 2012 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, 2012, pp. 229–236.

[110] Y. Zheng, A. Kamil, M.B. Driscoll, H. Shan, K. Yelick, UPC++: a PGAS extension for C++, in: 2014 IEEE 28th International Parallel and Distributed Processing Symposium, 2014, pp. 1105–1114. https://doi.org/10.1109/IPDPS.2014.115.

[111] K. Farrell, J.T. Oden, D. Faghihi, A bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems, J. Comput. Phys. 295 (2015) 189−208. https://doi.org/10.1016/j.jcp.2015.03.071.

[112] H. Meidani, J.B. Hooper, D. Bedrov, R.M. Kirby, Calibration and ranking of coarse-grained models in molecular simulations using Bayesian formalism, Int. J. Uncertain. Quantification 7 (2017) 99−115. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2017013407.

[113] M. Schöberl, N. Zabaras, P.-S. Koutsourelakis, Predictive coarse-graining, J. Comput. Phys. 333 (2017) 49−77. https://doi.org/10.1016/j.jcp.2016.10.073.

[114] R.A. Messerly, T.A. Knotts, W.V. Wilding, Uncertainty quantification and propagation of errors of the Lennard-Jones 12-6 parameters for n-alkanes, J. Chem. Phys. 146 (2017) 194110. https://doi.org/10.1063/1.4983406.

[115] P. Vargas, E. Muñoz, L. Rodriguez, Second virial coefficient for the Lennard-Jones potential, Phys. A 290 (2001) 92−100. https://doi.org/10.1016/s0378-4371(00)00362-9.

[116] J. Dymond, K. Marsh, R. Wilhoit, K. Wong, Virial Coefficients of Pure Gases, Vol. 21A of Landolt-Börnstein − Group IV Physical Chemistry, Springer-Verlag, 2002.

[117] G. Galliéro, C. Boned, A. Baylaucq, F. Montel, Molecular dynamics comparative study of Lennard-Jones α-6 and exponential α-6 potentials: application to real simple fluids (viscosity and pressure), Phys. Rev. E 73 (2006) 061201. https://doi.org/10.1103/PhysRevE.73.061201.

[118] F. Rizzi, H.N. Najm, B.J. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson, O.M. Knio, Uncertainty quantification in MD simulations. Part I: forward propagation, Multiscale Model. Simul. 10 (2012) 1428−1459. https://doi.org/10.1137/110853169.

[119] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, J. Chem. Phys. 79 (1983) 926−935. https://doi.org/10.1063/1.445869.

[120] H.W. Horn, W.C. Swope, J.W. Pitera, J.D. Madura, T.J. Dick, G.L. Hura, T.H. Gordon, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew, J. Chem. Phys. 120 (2004) 9665−9678. https://doi.org/10.1063/1.1683075.

[121] W.L. Jorgensen, J.D. Madura, Temperature and size dependence for Monte Carlo simulations of TIP4P water, Mol. Phys. 56 (1985) 1381−1392. https://doi.org/10.1080/00268978500103111.

[122] S.W. Rick, S.J. Stuart, B.J. Berne, Dynamical fluctuating charge force fields: application to liquid water, J. Chem. Phys. 101 (1994) 6141−6156. https://doi.org/10.1063/1.468398.

[123] M.W. Mahoney, W.L. Jorgensen, A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions, J. Chem. Phys. 112 (2000) 8910−8922. https://doi.org/10.1063/1.481505.

[124] M.K. Transtrum, B.B. Machta, J.P. Sethna, Geometry of nonlinear least squares with applications to sloppy models and optimization, Phys. Rev. E 83 (2011) 036701. https://doi.org/10.1103/physreve.83.036701.

[125] Y. Chung, A. Gelman, S. Rabe-Hesketh, J. Liu, V. Dorie, Weakly informative prior for point estimation of covariance matrices in hierarchical models, J. Educ. Behav. Stat. 40 (2015) 136−157. https://doi.org/10.3102/1076998615570945.

[126] P.S. Nerenberg, T. Head-Gordon, New developments in force fields for biomolecular simulations, Curr. Opin. Struct. Biol. 49 (2018) 129−138. https://doi.org/10.1016/j.sbi.2018.02.002.